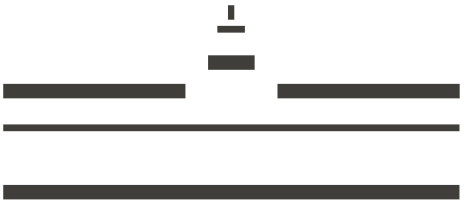


Westfälische
Wilhelms-Universität
Münster

Skript

zur Vorlesung

Mathematische Statistik

Prof. Dr. Zakhar Kabluchko

Universität Münster
Institut für Mathematische Stochastik

Inhaltsverzeichnis

Vorwort

Dies ist ein Skript zur Vorlesung "Stochastik I" bzw. "Mathematische Statistik", die an der Universität Ulm im SS 2013 bzw. an der Universität Münster im WS 2014/15 und WS 2015/16 gehalten wurde. Für die Erstellung der ersten LaTeX-Version des Skripts bedanke ich mich bei Frau Judith Olszewski. Danach wurde das Skript von mir überarbeitet, korrigiert und ergänzt. In Zukunft soll das Skript weiter ergänzt werden. Ich bedanke mich bei Hendrik Flasche, Philipp Godland und Judith Heusel für nützliche Verbesserungsvorschläge.

Bei Fragen, Wünschen und Verbesserungsvorschlägen können Sie gerne eine E-Mail an

`zakhar DOT kabluchko AT uni-muenster DOT de`

schreiben.

23. Februar 2017 Zakhar Kabluchko

Literatur

Es gibt sehr viele einführende Statistik-Lehrbücher, z. B.

- J. Lehn, H. Wegmann. *Einführung in die Statistik.*
- H. Pruscha. *Vorlesungen über Mathematische Statistik.*
- H. Pruscha. *Angewandte Methoden der Mathematischen Statistik.*
- V. Rohatgi. *Statistical Inference.*
- G. Casella, R. L. Berger. *Statistical Inference.*
- W. Pestman. *Mathematical Statistics: An Introduction.*
- K. Bosch. *Elementare Einführung in die angewandte Statistik: Mit Aufgaben und Lösungen.*

Folgende Lehrbücher behandeln sowohl Wahrscheinlichkeitstheorie als auch Statistik:

- H. Dehling und B. Haupt. *Einführung in die Wahrscheinlichkeitstheorie und Statistik.*
- U. Krengel. *Einführung in die Wahrscheinlichkeitstheorie und Statistik.*
- H.–O. Georgii. *Stochastik: Einführung in die Wahrscheinlichkeitstheorie und Statistik.*

Folgende Bücher von Lehmann sind Klassiker:

- E. L. Lehmann, G. Casella. *Theory of Point Estimation.*
- E. L. Lehmann. *Testing Statistical Hypotheses.*
- E. L. Lehmann. *Elements of Large Sample Theory.*

Sehr empfehlenswert sind diese zwei Bücher:

- L. Wasserman. *All of Statistics.*
- L. Wasserman. *All of Nonparametric Statistics.*

Drei sehr interessante Neuerscheinungen:

- L. Dümbgen. *Einführung in die Statistik.* Link
- L. Rüschendorf. *Mathematische Statistik.* Link
- L. Dümbgen. *Biometrie.* Link

Eine exzellente Referenz zur asymptotischen Statistik:

- A. W. van der Vaart. *Asymptotic Statistics.*

Statistische Entscheidungstheorie wird in folgenden Büchern behandelt:

- M. H. deGroot. *Statistical Decision Theory.*
- Th. S. Ferguson. *Mathematical Statistics: A Decision Theoretic Approach.*

Eine sehr anschauliche Darstellung bietet das folgende Buch, das leider nur auf Russisch existiert:

- M. B. Lagutin. *Nagljadnaja matematičeskaja statistika.*

Und schließlich noch einige Klassiker:

- P. J. Bickel, K. A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics.*
- S. S. Wilks. *Mathematical Statistics.*
- S. Zacks. *The Theory of Statistical Inference.*

KAPITEL 1

Stichproben und Stichprobenfunktion

1.1. Wahrscheinlichkeitstheorie und Statistik

Stochastik teilt sich in Wahrscheinlichkeitstheorie und Statistik auf, zwei Gebiete, die im gewissen Sinne entgegengesetzte Fragestellungen betrachten. Eine typische Fragestellung aus der Wahrscheinlichkeitstheorie ist diese:

> Eine Münze, die mit Wahrscheinlichkeit $p = 0.5$ „Kopf“ zeigt, wird $n = 100$ Mal geworfen. Wie groß ist die Wahrscheinlichkeit, dass diese Münze $k = 60$ Mal „Kopf“ zeigt?

In der Wahrscheinlichkeitstheorie wird also angenommen, dass eine komplette Beschreibung aller Parameter (in diesem Fall n und p) eines Zufallsexperiments vorhanden ist. Es wird dann gefragt, welche Ausgänge und mit welchen Wahrscheinlichkeiten beobachtet werden können. Eine typische Fragestellung aus der Statistik ist diese:

> Eine Münze wurde $n = 100$ Mal geworfen und hat dabei $k = 60$ Mal „Kopf“ gezeigt. Man bestimme („schätze“) die Wahrscheinlichkeit p, mit der die Münze bei einem Wurf „Kopf“ zeigt.

In der Statistik wird also angenommen, dass ein Zufallsexperiment, dessen Beschreibung nicht komplett ist, bereits durchgeführt wurde und sein Ausgang (in diesem Fall k) bekannt ist. Gefragt wird dann nach den Parametern, die dieses Zufallsexperiment beschreiben (in diesem Fall p).

Bei einer statistischen Fragestellung ist der Ausgang eines Zufallsexperiments gegeben. Diesen Ausgang nennt man eine *Stichprobe*. Ausgehend von der Stichprobe versucht man Informationen über die Parameter des Experiments zu gewinnen. Zum Beispiel kann man versuchen, die Parameter des Experiments durch gewisse Funktionen der Stichprobe zu schätzen. Im obigen Beispiel ist die Stichprobe $k = 60$ gegeben und man kann den unbekannten Parameter p durch $k/n = 0.6$ schätzen.

Es sollte aber klar sein, dass die Information darüber, dass eine Münze in 100 Würfen 60 Mal „Kopf“ gezeigt hat, nicht ausreicht, um mit absoluter Sicherheit den Wert von p zu bestimmen. Bei statistischen Entscheidungen sind also Fehler unvermeidbar. Es geht in der mathematischen Statistik darum, wie man die Wahrscheinlichkeiten oder die Größen dieser Fehler minimieren kann.

1.2. Grundbegriffe

Wir betrachten eine Serie aus n Messungen jedweder Art. In vielen Situationen muss man damit rechnen, dass die Ergebnisse der Messungen durch Zufall entstanden sind oder zumindestens durch zufällige Einflüsse beeinträchtigt werden. Man denke etwa an folgende Beispiele:

(1) mehrere zufällig ausgewählte Personen werden nach Ihrem Alter gefragt.
(2) ein verrauschtes Signal wird gemessen.
(3) die log-Returns eines Aktienpreises werden zu mehreren Zeitpunkten notiert.
(4) die Koordinaten eines Kometen am Himmel werden zu mehreren Zeitpunkten gemessen.

Wir müssen die Ergebnisse der Messungen durch Zufallsvariablen

$$X_1, \dots, X_n : \Omega \to \mathbb{R}$$

stochastisch modellieren. Dabei sei mit $(\Omega, \mathcal{A}, \mathbb{P})$ der Wahrscheinlichkeitsraum bezeichnet, auf dem die Zufallsvariablen $X_1, \dots, X_n$ definiert sind. Wenn nun die n Messungen durchgeführt werden, so heißt es, dass im Wahrscheinlichkeitsraum Ω ein Ausgang ω gemäß Verteilung $\mathbb{P}$ ausgewählt wird und wir die Resultate der Messungen erfahren:

$$x_1 := X_1(\omega), \dots, x_n := X_n(\omega).$$

Diese Resultate fassen wir in einer *Stichprobe* zusammen:

$$(x_1, \dots, x_n) \in \mathbb{R}^n.$$

Es sei bemerkt, dass $x_1, \dots, x_n$ deterministische Zahlen, wohingegen $X_1, \dots, X_n$ Zufallsvariablen sind. Man sagt, dass der deterministische Vektor $(x_1, \dots, x_n)$ eine *Realisierung* des Zufallsvektors $(X_1, \dots, X_n)$ ist. Die Anzahl der Messungen (also n) nennen wir den *Stichprobenumfang*. Die Menge aller vorstellbaren Stichproben wird der *Stichprobenraum* genannt und ist in diesem Beispiel $\mathbb{R}^n$ (oder, wenn Einschränkungen auf die Messergebnisse bestehen, eine Teilmenge von $\mathbb{R}^n$).

Wir werden sehr oft annehmen, dass die Zufallsvariablen $X_1, \dots, X_n$ *unabhängig und identisch verteilt* sind.

Man kann nun die Aufgabe der Statistik wie folgt zusammenfassen. Man betrachte einen Zufallsvektor $(X_1, \dots, X_n)$. Die Verteilung dieses Vektors sei aber nicht (oder nicht komplett) bekannt. Es werde eine Realisierung $(x_1, \dots, x_n) = (X_1(\omega), \dots, X_n(\omega))$ von $(X_1, \dots, X_n)$ beobachtet. Anhand dieser Realisierung sollen nun Rückschlüsse auf die Verteilung von $(X_1, \dots, X_n)$ gezogen werden.

Zu diesem Zweck bildet man passende Funktionen der Stichprobe.

Definition 1.2.1. Eine beliebige Borel-Funktion $\varphi : \mathbb{R}^n \to \mathbb{R}^m$ heißt *Stichprobenfunktion*, *Statistik*, oder *Schätzer*.

Wir werden sehr oft auch die zusammengesetzte Funktion betrachten:

$$\varphi(X) : \Omega \to \mathbb{R}^m,$$
$$\omega \mapsto \varphi(X_1(\omega), \ldots, X_n(\omega)).$$

Es sei bemerkt, dass $\varphi(x_1, \ldots, x_n)$ ein deterministisches Element aus $\mathbb{R}^m$ ist, wohingegen $\varphi(X)$ ein Zufallsvektor mit Werten in $\mathbb{R}^m$ ist.

Im Folgenden werden wir drei wichtige Beispiele von Stichprobenfunktionen, den empirischen Mittelwert, die empirische Varianz und die Ordnungsstatistiken, betrachten.

1.3. Empirischer Mittelwert

Es sei $(x_1, \ldots, x_n) \in \mathbb{R}^n$ eine Realisierung von unabhängigen und identisch verteilten Zufallsvariablen $X_1, \ldots, X_n$ mit Verteilungsfunktion

$$F(t) = \mathbb{P}[X_i \leq t], \quad t \in \mathbb{R},$$

die *nicht bekannt* ist. Anhand der bekannten Realisierung $(x_1, \ldots, x_n)$ sollen nun verschiedene Merkmale der Verteilungsfunktion F (z.B. der Erwartungswert $\mu = \mathbb{E}X_i$, die Varianz $\sigma^2 = \operatorname{Var} X_i$, oder sogar die komplette Funktion F) geschätzt werden. Wir werden zuerst den Erwartungswert $\mu = \mathbb{E}X_i$ schätzen.

Definition 1.3.1. Der *empirische Mittelwert* (auch das *Stichprobenmittel* oder das *arithmetische Mittel* genannt) ist definiert durch

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

Analog benutzen wir auch die Notation

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Es sei bemerkt, dass $\bar{x}_n$ eine reelle Zahl ist, wohingegen $\bar{X}_n$ eine Zufallsvariable ist. Wir fassen $\bar{x}_n$ als eine Realisierung von $\bar{X}_n$ auf: $\bar{x}_n = \bar{X}_n(\omega)$.

Satz 1.3.2. Seien $X_1, \ldots, X_n$ unabhängige Zufallsvariablen mit $\mathbb{E}X_i = \mu$ und $\operatorname{Var} X_i = \sigma^2$. Dann gilt

$$\mathbb{E}\bar{X}_n = \mu \text{ und } \operatorname{Var} \bar{X}_n = \frac{\sigma^2}{n}.$$

Beweis. Indem wir die Linearität des Erwartungswertes benutzen, erhalten wir

$$\mathbb{E}\bar{X}_n = \mathbb{E}\left[\frac{X_1 + \ldots + X_n}{n}\right] = \frac{1}{n} \cdot \mathbb{E}[X_1 + \ldots + X_n] = \frac{1}{n} \cdot n\mathbb{E}[X_1] = \mathbb{E}[X_1] = \mu.$$

Indem wir die Additivität der Varianz (bei unabhängigen Zufallsvariablen) benutzen, erhalten wir

$$\operatorname{Var} \bar{X}_n = \operatorname{Var}\left(\frac{X_1 + \ldots + X_n}{n}\right) = \frac{1}{n^2} \operatorname{Var}(X_1 + \ldots + X_n) = \frac{1}{n^2} \cdot n \operatorname{Var}(X_1) = \frac{\sigma^2}{n}.$$

□

Bemerkung 1.3.3. Der obige Satz besagt, dass beim Schätzen von $\mu = \mathbb{E}X_i$ durch $\bar{x}_n$ (oder $\bar{X}_n$) kein systematischer Fehler entsteht, in dem Sinne, dass der Erwartungswert des Schätzers $\bar{X}_n$ mit dem zu schätzenden Parameter μ übereinstimmt: $\mathbb{E}\bar{X}_n = \mu$. Man sagt, dass $\bar{X}_n$ ein *erwartungstreuer Schätzer* für μ ist.

Der nächste Satz besagt, dass bei einer immer größer werdenden Stichprobe der Schätzer $\bar{X}_n$ gegen den zu schätzenden Wert μ konvergiert.

Satz 1.3.4. Seien $X_1, X_2, \ldots$ unabhängige und identisch verteilte Zufallsvariablen mit $\mathbb{E}X_i = \mu$. Dann gilt

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

Man sagt in diesem Zusammenhang, dass $\bar{X}_n$ ein *stark konsistenter* Schätzer für μ ist.

Beweis. Das folgt direkt aus dem starken Gesetz der großen Zahlen. □

1.4. Empirische Varianz

Definition 1.4.1. Die *empirische Varianz* oder die *Stichprobenvarianz* einer Stichprobe $(x_1, \ldots, x_n)$ ist definiert durch

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

Analog benutzen wir auch die Notation

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

Die Rolle des Faktors $\frac{1}{n-1}$ (anstelle von $\frac{1}{n}$) wird in Satz 1.4.3 klar. Zuerst leiten wir eine alternative Formel für S_n^2 her.

Satz 1.4.2. Es gilt

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

Beweis. Durch ausquadrieren ergibt sich

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

Dabei haben wir die Formel $\sum_{i=1}^n X_i = n\bar{X}_n$ benutzt. □

Satz 1.4.3. Seien $X_1, \ldots, X_n$ unabhängige Zufallsvariablen mit $\mathbb{E}X_i = \mu$ und $\operatorname{Var} X_i = \sigma^2$. Dann gilt

$$\mathbb{E}[S_n^2] = \sigma^2.$$

Beweis. Mit Satz 1.4.2 ergibt sich

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

Dabei haben wir verwendet, dass

$$\begin{aligned}
\mathbb{E}[X_i^2] &= \operatorname{Var} X_i + (\mathbb{E}X_i)^2 = \sigma^2 + \mu^2 \\
\mathbb{E}[\bar{X}_n^2] &= \operatorname{Var}\bar{X}_n + (\mathbb{E}\bar{X}_n)^2 = \frac{\sigma^2}{n} + \mu^2.
\end{aligned}$$

Für die zweite Formel haben wir benutzt, dass $\mathbb{E}\bar{X}_n = \mu$ und $\operatorname{Var}\bar{X}_n = \frac{\sigma^2}{n}$, siehe Satz 1.3.2. □

Bemerkung 1.4.4. Die empirische Varianz s_n^2 (bzw. S_n^2) ist ein natürlicher Schätzer für die theoretische Varianz $\sigma^2 = \operatorname{Var} X_i$. Der obige Satz besagt, dass S_n^2 ein erwartungstreuer Schätzer für σ^2 ist im Sinne, dass der Erwartungswert des Schätzers mit dem zu schätzenden Parameter σ^2 übereinstimmt: $\mathbb{E}S_n^2 = \sigma^2$.

Bemerkung 1.4.5. Der Faktor $\frac{1}{n-1}$ in der Definition von S_n^2 wird die *Bessel-Korrektur* genannt und macht S_n^2 zu einem erwartungstreuen Schätzer. An Stelle von S_n^2 kann auch folgende Stichprobenfunktion betrachtet werden:

$$\tilde{S}_n^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Der Unterschied zwischen S_n^2 und $\tilde{S}_n^2$ ist also nur der Vorfaktor $\frac{1}{n-1}$ bzw. $\frac{1}{n}$. Allerdings ist $\tilde{S}_n^2$ *kein* erwartungstreuer Schätzer für σ^2, denn

$$\mathbb{E}[\tilde{S}_n^2] = \mathbb{E}\left[\frac{n-1}{n}S_n^2\right] = \frac{n-1}{n}\cdot\mathbb{E}[S_n^2] = \frac{n-1}{n}\cdot\sigma^2 < \sigma^2.$$

Somit wird die Varianz σ^2 „unterschätzt“. Schätzt man σ^2 durch $\tilde{S}_n^2$, so entsteht ein systematischer Fehler (Bias) von $-\frac{1}{n}\sigma^2$.

Aufgabe 1.4.6 (Verhalten von s_n^2 unter affinen Transformationen)**.** Es sei $s_n^2 = s_n^2(x_1, \ldots, x_n)$ die empirische Varianz der Stichprobe $(x_1, \ldots, x_n) \in \mathbb{R}^n$. Zeigen Sie, dass für alle $a, b \in \mathbb{R}$,

$$s_n^2(ax_1 + b, \ldots, ax_n + b) = a^2 s_n^2(x_1, \ldots, x_n).$$

Aufgabe 1.4.7 (Satz von Steiner)**.** Es sei $(x_1, \ldots, x_n) \in \mathbb{R}^n$. Zeigen Sie, dass für alle $a \in \mathbb{R}$,

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

Aufgabe 1.4.8 (Charakterisierung des empirischen Mittelwerts)**.** Sei $(x_1, \ldots, x_n) \in \mathbb{R}^n$ eine Stichprobe. Bestimmen Sie das Minimum der Funktion $f(a) := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - a)^2$, $a \in \mathbb{R}$. Zeigen Sie, dass das Minimum für $a = \bar{x}_n$ erreicht wird.

Aufgabe 1.4.9 (Charakterisierung der Erwartungswerts)**.** Sei X eine Zufallsvariable mit $\mathbb{E}[X^2] < \infty$. Bestimmen Sie das Minimum der Funktion $f(a) := \mathbb{E}[(X-a)^2]$, $a \in \mathbb{R}$. Zeigen Sie, dass das Minimum für $a = \mathbb{E}X$ erreicht wird.

Ein natürlicher Schätzer für die theoretische Standardabweichung $\sigma = \sqrt{\operatorname{Var} X_i}$ ist durch die Wurzel aus der empirischen Varianz gegeben.

Definition 1.4.10. Die *empirische Standardabweichung* ist definiert durch

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

Aufgabe 1.4.11. Ist S_n ein erwartungstreuer Schätzer für σ? Was ist größer: $\mathbb{E}S_n$ oder σ?

Bemerkung 1.4.12. Das Stichprobenmittel $\bar{x}_n$ ist ein *Lageparameter* (beschreibt die Lage der Stichprobe). Die Stichprobenvarianz s_n^2 (bzw. die empirische Standardabweichung s_n) ist ein *Streuungsparameter* (beschreibt die Ausdehnung der Stichprobe).

Aufgabe 1.4.13 (Starke Konsistenz der empirischen Varianz)**.** Seien $X_1, X_2, \ldots$ unabhängige identisch verteilte Zufallsvariablen mit $\mathbb{E}X_i = \mu$ und $\operatorname{Var} X_i = \sigma^2$. Zeigen Sie, dass

$$S_n^2 \xrightarrow[n\to\infty]{f.s.} \sigma^2, \quad \tilde{S}_n^2 \xrightarrow[n\to\infty]{f.s.} \sigma^2.$$

Das heißt, S_n^2 und $\tilde{S}_n^2$ sind stark konsistente Schätzer für σ^2.

Aufgabe 1.4.14 (Zwei Stichproben mit gleicher Varianz)**.** Seien $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ unabhängige Zufallsvariablen mit $\mathbb{E}X_i = \mu$, $\mathbb{E}Y_i = \nu$ und $\operatorname{Var} X_i = \operatorname{Var} Y_i = \sigma^2$, wobei alle drei Parameter unbekannt seien. Zeigen Sie, dass der Schätzer

$$T := \frac{1}{n+m-2} \cdot \left(\sum_{i=1}^{n} (X_i - \bar{X}_n)^2 + \sum_{j=1}^{m} (Y_i - \bar{Y}_n)^2 \right),$$

erwartungstreu für σ^2 ist, d.h. $\mathbb{E}T = \sigma^2$.

Aufgabe 1.4.15 (Varianz schätzen bei bekanntem Erwartungswert)**.** Es seien $X_1, \ldots, X_n$ unabhängige Zufallsvariablen mit *bekanntem* Erwartungswert μ und unbekannter Varianz $\sigma^2 < \infty$. Zeigen Sie, dass der Schätzer

$$T_n := \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$$

erwartungstreu für σ^2 ist.

Ist (X, Y) ein Zufallsvektor mit $\mathbb{E}[X^2] < \infty$ und $\mathbb{E}[Y^2] < \infty$, so ist seine Kovarianz durch

$$\operatorname{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

definiert. In der nächsten Aufgabe geht es darum, die Kovarianz anhand von n unabhängigen Beobachtungen von (X, Y) zu schätzen.

Aufgabe 1.4.16 (Empirische Kovarianz)**.** Seien $(X_1, Y_1), (X_2, Y_2), \ldots$ unabhängige identisch verteilte Zufallsvektoren mit $\mathbb{E}[X_i^2] < \infty$, $\mathbb{E}[Y_i^2] < \infty$ und $\rho = \operatorname{Cov}(X_i, Y_i)$. Zeigen Sie, dass die sogenannte *empirische Kovarianz*

$$R_n := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)$$

ein erwartungstreuer und stark konsistenter Schätzer für die theoretische Kovarianz ρ ist, d.h.

$$\mathbb{E}R_n = \rho \text{ und } R_n \xrightarrow[n\to\infty]{f.s.} \rho.$$

Hinweis: Zuerst kann man die folgende Formel beweisen:

$$R_n = \frac{1}{n-1}\left(\sum_{i=1}^{n}(X_iY_i - \bar{X}_n\bar{Y}_n)\right).$$

Aufgabe 1.4.17 (Formel für die empirische Varianz)**.** Zeigen Sie, dass

$$S_n^2 = \frac{1}{2\binom{n}{2}}\sum_{1\le i<j\le n}(X_i - X_j)^2.$$

Aufgabe 1.4.18 (Varianz der empirischen Varianz)**.** Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit $\mathbb{E}[X_i^4] < \infty$. Zeigen Sie, dass

$$\operatorname{Var}[S_n^2] = \frac{1}{n}\left(\kappa_4 - \frac{n-3}{n-1}\sigma^4\right),$$

wobei $\mu = \mathbb{E}X_i$, $\sigma^2 = \operatorname{Var} X_i$ und $\kappa_4 = \mathbb{E}[(X_i - \mu)^4]$ das vierte zentrale Moment von X_i ist.

1.5. Ordnungsstatistiken und Quantile

Das Stichprobenmittel ist kein *robuster* Lageparameter, d.h. es wird stark von Ausreißern beeinflusst. Das wird im folgenden Beispiel gezeigt. Betrachte zuerst die Stichprobe $(1, 2, 2, 2, 1, 1, 1, 2)$. Somit ist $\bar{x}_n = 1.5$. Ändert man nur den letzten Wert der Stichprobe in 200 um, also $(1, 2, 2, 2, 1, 1, 1, 200)$, dann gilt $\bar{x}_n = 26.25$. Wir konnten also den Wert des Stichprobenmittels stark verändern, indem wir nur ein einziges Element aus der Stichprobe verändert haben. Die Stichprobenvarianz ist ebenfalls nicht robust.

Im weiteren werden wir robuste Lage- und Streuungsparameter einführen, d.h. solche Parameter, die sich bei einer Änderung (und zwar sogar bei einer sehr starken Änderung) von nur wenigen Elementen aus der Stichprobe nicht sehr stark verändern. Zu diesem Zweck werden wir nun Ordnungsstatistiken und Quantile definieren.

Definition 1.5.1. Sei $(x_1, \ldots, x_n) \in \mathbb{R}^n$ eine Stichprobe. Wir können die Elemente der Stichprobe aufsteigend anordnen:

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}.$$

Wir nennen $x_{(i)}$ die i-te *Ordnungsstatistik* der Stichprobe.

Zum Beispiel ist $x_{(1)} = \min\limits_{i=1,\ldots,n} x_i$ das Minimum und $x_{(n)} = \max\limits_{i=1,\ldots,n} x_i$ das Maximum der Stichprobe.

Definition 1.5.2. Der *Stichprobenmedian* ist gegeben durch

$$\mathrm{Med}_n = \mathrm{Med}_n(x_1, \ldots, x_n) = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{falls } n \text{ ungerade,} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right), & \text{falls } n \text{ gerade.} \end{cases}$$

Somit befindet sich die Hälfte der Stichprobe über dem Stichprobenmedian und die andere Hälfte der Stichprobe darunter.

Beispiel 1.5.3. Der Median ist ein robuster Lageparameter. Als Beispiel dafür betrachten wir zwei Stichproben mit Stichprobenumfang $n = 8$.
Die erste Stichprobe sei

$$(x_1, \ldots, x_8) = (1, 2, 2, 2, 1, 1, 1, 2).$$

Somit sind die Ordnungsstatistiken gegeben durch

$$(x_{(1)}, \ldots, x_{(8)}) = (1, 1, 1, 1, 2, 2, 2, 2).$$

Daraus lässt sich der Median berechnen und dieser ist $\mathrm{Med}_8 = \frac{1+2}{2} = 1.5$.
Als zweite Stichprobe betrachten wir

$$(y_1, \ldots, y_8) = (1, 2, 2, 2, 1, 1, 1, 200).$$

Die Ordnungsstatistiken sind gegeben durch

$$(y_{(1)}, \ldots, y_{(n)}) = (1, 1, 1, 1, 2, 2, 2, 200),$$

und der Median ist nach wie vor $\mathrm{Med}_8 = 1.5$. Dies zeigt, dass der Median robust ist.

Bemerkung 1.5.4. Im Allgemeinen gilt $\mathrm{Med}_n \neq \bar{x}_n$.

Aufgabe 1.5.5 (Charakterisierung des empirischen Medians)**.** Sei $(x_1, \ldots, x_n) \in \mathbb{R}^n$ eine Stichprobe. Beschreiben Sie die Menge aller $a \in \mathbb{R}$, für die die Funktion $f(a) := \sum_{i=1}^n |x_i - a|$ minimal wird. *Hinweis:* Betrachten Sie die Fälle n gerade und n ungerade separat.

Ein weiterer robuster Lageparameter ist das getrimmte Mittel.

Definition 1.5.6. Das *getrimmte Mittel* einer Stichprobe $(x_1, \ldots, x_n)$ ist definiert durch

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

Die Wahl von k entscheidet, wie viele Daten nicht berücksichtigt werden. Man kann zum Beispiel $k = [0.05 \cdot n]$ wählen, dann werden 10% aller Daten nicht berücksichtigt. In diesem Fall spricht man auch vom 5%-getrimmten Mittel.

Anstatt des getrimmten Mittels betrachtet man oft das *winsorisierte Mittel*:

$$\frac{1}{n}\left(\sum_{i=k+1}^{n-k} x_{(i)} + k\, x_{(k+1)} + k\, x_{(n-k)}\right).$$

Nachdem wir nun einige robuste Lageparameter konstruiert haben, wenden wir uns den robusten Streuungsparametern zu. Dazu benötigen wir die empirischen Quantile.

Definition 1.5.7. Sei $(x_1, \ldots, x_n) \in \mathbb{R}^n$ eine Stichprobe und $\alpha \in (0,1)$. Das *empirische α-Quantil* ist definiert durch

$$q_\alpha = \begin{cases} x_{([n\alpha]+1)}, & \text{falls } n\alpha \notin \mathbb{N}, \\ \frac{1}{2}\left(x_{([n\alpha])} + x_{([n\alpha]+1)}\right), & \text{falls } n\alpha \in \mathbb{N}. \end{cases}$$

Hierbei steht $[\cdot]$ für die Gaußklammer.

Der empirische Median ist somit das empirische $\frac{1}{2}$-Quantil.

Definition 1.5.8. Die *empirischen Quartile* sind die Zahlen

$$q_{0,25}, \quad q_{0,5}, \quad q_{0,75}.$$

Die Differenz $q_{0,75} - q_{0,25}$ nennt man den *empirischen Interquartilsabstand*.

Der empirische Interquartilsabstand ist ein robuster Streuungsparameter.

Die empirischen Quantile können als Schätzer für die „theoretischen“ Quantile betrachtet werden, die wir nun einführen werden.

Definition 1.5.9. Sei X eine Zufallsvariable mit Verteilungsfunktion $F(t)$ und sei $\alpha \in (0,1)$. Das *theoretische α-Quantil* $Q(\alpha)$ von X ist definiert als die Lösung der Gleichung

$$F(Q(\alpha)) = \alpha.$$

Leider kann es passieren, dass diese Gleichung keine Lösungen hat (wenn die Funktion F den Wert α überspringt) oder dass es mehrere Lösungen gibt (wenn die Funktion F auf einem Intervall konstant und gleich α ist). Deshalb benutzt man die folgende Definition, die auch in diesen Ausnahmefällen Sinn ergibt:

$$Q(\alpha) = \inf\left\{t \in \mathbb{R} : F(t) \geq \alpha\right\}.$$

Definition 1.5.10. Der Wert $Q(\frac{1}{2})$ heißt der *(theoretische) Median*.

Aufgabe 1.5.11 (Charakterisierung des Medians)**.** Sei X eine Zufallsvariable mit einer stetigen Verteilungsfunktion F. Beschreiben Sie die Menge aller $a \in \mathbb{R}$, für die die Funktion $f(a) := \mathbb{E}|X - a|$ minimal wird. *Hinweis:* Zeigen Sie, dass $f'(a) = 2F(a) - 1$.

Beispiel 1.5.12. Weitere Lageparameter, die in der Statistik vorkommen:

(1) Das Bereichsmittel $\frac{1}{2}(x_{(n)} + x_{(1)})$ (nicht robust).

(2) Das empirische Quartilsmittel $\frac{1}{2}(q_{0,25} + q_{0,75})$ (robust).

Beispiel 1.5.13. Weitere Streuungsparameter:

(1) Die Spannweite $x_{(n)} - x_{(1)}$.

(2) Die mittlere absolute Abweichung vom Mittelwert $\frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}_n|$.

(3) Die mittlere absolute Abweichung vom Median $\frac{1}{n}\sum_{i=1}^{n} |x_i - \mathrm{Med}_n|$.

Alle drei Parameter sind nicht robust. Man kann den dritten Parameter allerdings robust machen, wenn man wieder einmal den Mittelwert durch den Median ersetzt:

$$\mathrm{Med}(|x_1 - \mathrm{Med}_n|, \ldots, |x_n - \mathrm{Med}_n|).$$

1.6. Verteilung der Ordnungsstatistiken

Satz 1.6.1. Seien $X_1, X_2, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen, die absolut stetig sind mit Dichte f und Verteilungsfunktion F. Es seien

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$$

die Ordnungsstatistiken. Dann ist die Dichte der Zufallsvariable $X_{(i)}$ gegeben durch

$$f_{X_{(i)}}(t) = \frac{n!}{(i-1)!(n-i)!} f(t) F(t)^{i-1} (1 - F(t))^{n-i}.$$

Erster Beweis. Damit $X_{(i)} = t$ ist, muss Folgendes passieren:

1. Eine der Zufallsvariablen, z.B. X_k, muss den Wert t annehmen. Es gibt n Möglichkeiten, das k auszuwählen. Die „Dichte“ des Ereignisses $X_k = t$ ist $f(t)$.

2. Unter den restlichen $n-1$ Zufallsvariablen müssen genau $i-1$ Zufallsvariablen Werte unter t annehmen. Wir haben $\binom{n-1}{i-1}$ Möglichkeiten, die $i-1$ Zufallsvariablen auszuwählen. Die Wahrscheinlichkeit, dass die ausgewählten Zufallsvariablen allesamt kleiner als t sind, ist $F(t)^{i-1}$.

3. Die verbliebenen $n-i$ Zufallsvariablen müssen allesamt größer als t sein. Die Wahrscheinlichkeit davon ist $(1 - F(t))^{n-i}$.

Indem wir nun alles ausmultiplizieren, erhalten wir das Ergebnis:

$$f_{X_{(i)}}(t) = nf(t) \cdot \binom{n-1}{i-1} F(t)^{i-1} \cdot (1 - F(t))^{n-i}.$$

Das ist genau die erwünschte Formel, denn $n\binom{n-1}{i-1} = \frac{n(n-1)!}{(i-1)!(n-i)!} = \frac{n!}{(i-1)!(n-i)!}$. □

Zweiter Beweis. Schritt 1. Die Anzahl der Elemente der Stichprobe, die unterhalb von t liegen, bezeichnen wir mit

$$N = \#\{i \in \{1, \ldots, n\} : X_i \leq t\} = \sum_{i=1}^{n} \mathbb{1}_{X_i \leq t}.$$

Dabei steht # für die Anzahl der Elemente in einer Menge. Die Zufallsvariablen $X_1, \dots, X_n$ sind unabhängig und identisch verteilt mit $\mathbb{P}[X_i \le t] = F(t)$. Somit ist die Zufallsvariable N binomialverteilt:

$$N \sim \mathrm{Bin}(n, F(t)).$$

SCHRITT 2. Es gilt $\{X_{(i)} \le t\} = \{N \ge i\}$. Daraus folgt für die Verteilungsfunktion von $X_{(i)}$, dass

$$F_{X_{(i)}}(t) = \mathbb{P}[X_{(i)} \le t] = \mathbb{P}[N \ge i] = \sum_{k=i}^{n} \binom{n}{k} F(t)^k (1 - F(t))^{n-k}.$$

SCHRITT 3. Die Dichte ist die Ableitung der Verteilungsfunktion. Somit erhalten wir

$$\begin{aligned} f_{X_{(i)}}(t) &= F'_{X_{(i)}}(t) \\ &= \sum_{k=i}^{n} \binom{n}{k} \left\{ kF(t)^{k-1} f(t)(1 - F(t))^{n-k} - (n-k)F(t)^k (1 - F(t))^{n-k-1} f(t) \right\} \\ &= \sum_{k=i}^{n} \binom{n}{k} kF(t)^{k-1} f(t)(1 - F(t))^{n-k} - \sum_{k=i}^{n} \binom{n}{k} (n-k)F(t)^k (1 - F(t))^{n-k-1} f(t). \end{aligned}$$

Wir schreiben nun den Term mit $k = i$ in der ersten Summe getrennt, und für alle anderen Terme in der ersten Summe führen wir den neuen Summationsindex $l = k - 1$ ein. Die zweite Summe lassen wir unverändert, ersetzen aber den Summationsindex k durch l:

$$\begin{aligned} f_{X_{(i)}}(t) = \binom{n}{i} iF(t)^{i-1} f(t)(1 - F(t))^{n-i} + \sum_{l=i}^{n-1} \binom{n}{l+1} (l+1)F(t)^l f(t)(1 - F(t))^{n-l-1} \\ - \sum_{l=i}^{n} \binom{n}{l} (n-l)F(t)^l f(t)(1 - F(t))^{n-l-1}. \end{aligned}$$

Der Term mit $l = n$ in der zweiten Summe ist wegen des Faktors $n - l$ gleich 0, somit können wir in der zweiten Summe bis $n - 1$ summieren. Nun sehen wir, dass die beiden Summen gleich sind, denn

$$\binom{n}{l+1} (l+1) = \frac{n!}{l!(n-l-1)} = \binom{n}{l} (n-l).$$

Die Summen kürzen sich und somit folgt

$$f_{X_{(i)}}(t) = \binom{n}{i} iF(t)^{i-1} f(t)(1 - F(t))^{n-i}. \qquad \square$$

Aufgabe 1.6.2 (Gemeinsame Dichte von $X_{(i)}$ und $X_{(j)}$)**.** Seien $X_1, \dots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit Dichte f und Verteilungsfunktion F. Man zeige, dass für alle $1 \le i < j \le n$ die gemeinsame Dichte $f_{X_{(i)}, X_{(j)}}(t, s)$ der Ordnungsstatistiken $X_{(i)}$ und $X_{(j)}$ durch die folgende Formel gegeben ist:

$$f(t)f(s) \binom{n}{2} \binom{n}{i-1, j-1-i, n-j} F(t)^{i-1} (F(s) - F(t))^{j-1-i} (1 - F(s))^{n-j}.$$

Im nächsten Satz bestimmen wir die gemeinsame Dichte *aller* Ordnungsstatistiken.

Satz 1.6.3. Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit Dichte f. Seien $X_{(1)} \leq \ldots \leq X_{(n)}$ die Ordnungsstatistiken. Dann ist die gemeinsame Dichte des Zufallsvektors $(X_{(1)}, \ldots, X_{(n)})$ gegeben durch

$$f_{X_{(1)},\ldots,X_{(n)}}(t_1, \ldots, t_n) = \begin{cases} n! \cdot f(t_1) \cdot \ldots \cdot f(t_n), & \text{falls } t_1 \leq \ldots \leq t_n, \\ 0, & \text{sonst.} \end{cases}$$

Beweis. Da die Ordnungsstatistiken per Definition aufsteigend sind, ist die Dichte gleich 0, wenn die Bedingung $t_1 \leq \ldots \leq t_n$ nicht erfüllt ist. Sei nun die Bedingung $t_1 \leq \ldots \leq t_n$ erfüllt. Damit $X_{(1)} = t_1, \ldots, X_{(n)} = t_n$ ist, muss eine der Zufallsvariablen (für deren Wahl es n Möglichkeiten gibt) gleich t_1 sein, eine andere (für deren Wahl es $n-1$ Möglichkeiten gibt) gleich t_2, usw. Wir haben also $n!$ Möglichkeiten für die Wahl der Reihenfolge der Variablen. Zum Beispiel tritt für $n = 2$ das Ereignis $\{X_{(1)} = t_1, X_{(2)} = t_2\}$ genau dann ein, wenn entweder $\{X_1 = t_1, X_2 = t_2\}$ oder $\{X_1 = t_2, X_2 = t_1\}$ eintritt, was 2 Möglichkeiten ergibt. Da alle Möglichkeiten sich nur durch Permutationen unterscheiden und somit die gleiche „Dichte“ besitzen, betrachten wir nur eine Möglichkeit und multiplizieren dann das Ergebnis mit $n!$. Die einfachste Möglichkeit ist, dass $\{X_1 = t_1, \ldots, X_n = t_n\}$ eintritt. Diesem Ereignis entspricht die „Dichte“ $f(t_1) \cdot \ldots \cdot f(t_n)$, da die Zufallsvariablen $X_1, \ldots, X_n$ unabhängig sind. Multiplizieren wir nun diese Dichte mit $n!$, so erhalten wir das gewünschte Ergebnis. □

Beispiel 1.6.4. Seien $X_1, \ldots, X_n$ unabhängig und gleichverteilt auf dem Intervall $[0, 1]$. Die Dichte von X_i ist $f(t) = \mathbb{1}_{[0,1]}(t)$. Somit gilt für die Dichte der i-ten Ordnungsstatistik

$$f_{X_{(i)}}(t) = \begin{cases} \binom{n}{i} i \cdot t^{i-1}(1-t)^{n-i}, & \text{falls } t \in [0, 1], \\ 0, & \text{sonst.} \end{cases}$$

Diese Verteilung ist ein Spezialfall der Betaverteilung, die wir nun einführen.

Definition 1.6.5. Eine Zufallsvariable Z heißt *betaverteilt* mit Parametern $\alpha, \beta > 0$, falls

$$f_Z(t) = \begin{cases} \frac{1}{\mathrm{B}(\alpha,\beta)} \cdot t^{\alpha-1}(1-t)^{\beta-1}, & \text{falls } t \in [0, 1], \\ 0, & \text{sonst.} \end{cases}$$

Bezeichnung: $Z \sim \mathrm{Beta}(\alpha, \beta)$. Hierbei ist $\mathrm{B}(\alpha, \beta)$ die Eulersche *Betafunktion*, gegeben durch

$$\mathrm{B}(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} \mathrm{d}t.$$

Indem wir nun die Dichte von $X_{(i)}$ im gleichverteilten Fall mit der Dichte der Betaverteilung vergleichen, erhalten wir, dass

$$X_{(i)} \sim \mathrm{Beta}(i, n - i + 1).$$

Dabei muss man gar nicht nachrechnen, dass $\frac{1}{\mathrm{B}(i,n-i+1)} = \binom{n}{i} i$ ist, denn in beiden Fällen handelt es sich um eine Dichte. Wären die beiden Konstanten unterschiedlich, so wäre das Integral einer der Dichten ungleich 1, was nicht möglich ist.

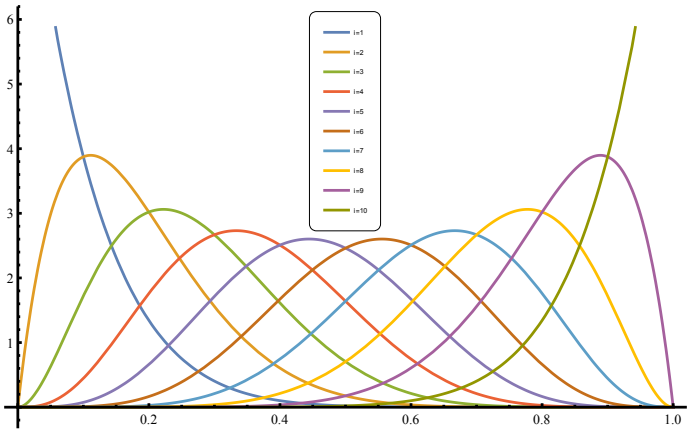


ABBILDUNG 1. Dichten der Ordnungsstatistiken $X_{(1)}, \ldots, X_{(10)}$ einer unabhängigen und auf $[0,1]$ gleichverteilten Stichprobe $X_1, \ldots, X_{10}$.

Aufgabe 1.6.6. Seien $X_1, \ldots, X_n$ unabhängig und gleichverteilt auf dem Intervall $[0,1]$. Zeigen Sie, dass

$$\mathbb{E}[X_{(i)}] = \frac{i}{n+1}.$$

Aufgabe 1.6.7. Seien $X_1, \ldots, X_n$ unabhängig und exponentialverteilt mit Parameter 1. Zeigen Sie, dass

$$\mathbb{E}[X_{(i)}] = \frac{1}{n} + \frac{1}{n-1} + \ldots + \frac{1}{n-i+1}.$$

Aufgabe 1.6.8. Es seien $X, X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit einer stetigen Verteilungsfunktion. Berechnen Sie

(a) $\mathbb{P}[X_1 < X_2 < \ldots < X_n]$.
(b) $\mathbb{P}[X_n = \max\{X_1, \ldots, X_n\}]$.

Aufgabe 1.6.9. Es seien $X, X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit einer stetigen Verteilungsfunktion. Es seien $X_{(1)} < \ldots < X_{(n)}$ die Ordnungsstatistiken von $X_1, \ldots, X_n$ (ohne Berücksichtigung von X). Berechnen Sie $\mathbb{P}[X < X_{(k)}]$ für $k = 1, \ldots, n$.

KAPITEL 2

Empirische Verteilungsfunktion

2.1. Empirische Verteilungsfunktion

Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit theoretischer Verteilungsfunktion

$$F(t) = \mathbb{P}[X_i \leq t].$$

Es sei $(x_1, \ldots, x_n)$ eine Realisierung dieser Zufallsvariablen. Wie können wir die theoretische Verteilungsfunktion F anhand der Stichprobe $(x_1, \ldots, x_n)$ schätzen? Dafür benötigen wir die empirische Verteilungsfunktion.

Definition 2.1.1. Die *empirische Verteilungsfunktion* einer Stichprobe $(x_1, \ldots, x_n) \in \mathbb{R}^n$ ist definiert durch

$$\widehat{F}_n(t) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{x_i \leq t} = \frac{1}{n} \# \{i \in \{1, \ldots, n\} : x_i \leq t\}, \quad t \in \mathbb{R}.$$

Bemerkung 2.1.2. Die oben definierte empirische Verteilungsfunktion kann wie folgt durch die Ordnungsstatistiken $x_{(1)}, \ldots, x_{(n)}$ ausgedrückt werden

$$\widehat{F}_n(t) = \begin{cases} 0, & \text{falls } t < x_{(1)}, \\ \frac{1}{n}, & \text{falls } x_{(1)} \leq t < x_{(2)}, \\ \frac{2}{n}, & \text{falls } x_{(2)} \leq t < x_{(3)}, \\ \ldots & \ldots \\ \frac{n-1}{n}, & \text{falls } x_{(n-1)} \leq t < x_{(n)}, \\ 1, & \text{falls } x_{(n)} \leq t. \end{cases}$$

Bemerkung 2.1.3. Die empirische Verteilungsfunktion $\widehat{F}_n$ hat alle Eigenschaften einer Verteilungsfunktion, denn es gilt

(1) $\lim_{t \to -\infty} \widehat{F}_n(t) = 0$ und $\lim_{t \to +\infty} \widehat{F}_n(t) = 1$.
(2) $\widehat{F}_n$ ist monoton nichtfallend.
(3) $\widehat{F}_n$ ist rechtsstetig.

Parallel werden wir auch die folgende Definition benutzen.

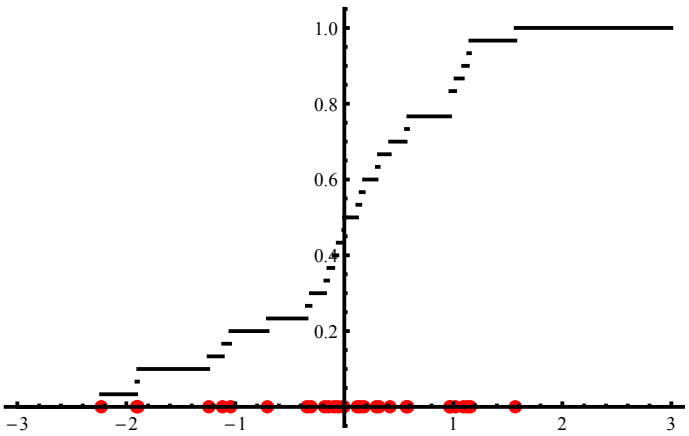


ABBILDUNG 1. Empirische Verteilungsfunktion. Die roten Punkte zeigen die Stichprobe. An jedem dieser Punkte springt die empirische Verteilungsfunktion um $1/n$.

Definition 2.1.4. Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen. Dann ist die empirische Verteilungsfunktion gegeben durch

$$\widehat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq t}, \quad t \in \mathbb{R}.$$

Es sei bemerkt, dass $\widehat{F}_n(t)$ für jedes $t \in \mathbb{R}$ eine Zufallsvariable ist. Somit ist $\widehat{F}_n$ eine zufällige Funktion. Auf die Eigenschaften von $\widehat{F}_n(t)$ gehen wir im folgenden Satz ein.

Satz 2.1.5. Seien $X_1, X_2, \ldots$ unabhängige und identisch verteilte Zufallsvariablen mit Verteilungsfunktion F. Dann gilt

(1) Die Zufallsvariable $n\widehat{F}_n(t)$ ist binomialverteilt:

$$n\widehat{F}_n(t) \sim \mathrm{Bin}(n, F(t)).$$

Das heißt:

$$\mathbb{P}\left[\widehat{F}_n(t) = \frac{k}{n}\right] = \binom{n}{k} F(t)^k (1 - F(t))^{n-k}, \quad k = 0, 1, \ldots, n.$$

(2) Für den Erwartungswert und die Varianz von $\widehat{F}_n(t)$ gilt:

$$\mathbb{E}[\widehat{F}_n(t)] = F(t), \quad \mathrm{Var}[\widehat{F}_n(t)] = \frac{F(t)(1 - F(t))}{n}.$$

Somit ist $\widehat{F}_n(t)$ ein erwartungstreuer Schätzer für $F(t)$.

(3) Für alle $t \in \mathbb{R}$ gilt

$$\widehat{F}_n(t) \xrightarrow[n\to\infty]{f.s.} F(t).$$

In diesem Zusammenhang sagt man, dass $\widehat{F}_n(t)$ ein „stark konsistenter“ Schätzer für $F(t)$ ist.
(4) Für alle $t \in \mathbb{R}$ mit $F(t) \neq 0, 1$ gilt:

$$\sqrt{n}\frac{\widehat{F}_n(t) - F(t)}{\sqrt{F(t)(1-F(t))}} \xrightarrow[n\to\infty]{d} \mathrm{N}(0,1).$$

In diesem Zusammenhang sagt man, dass $\widehat{F}_n(t)$ ein „asymptotisch normalverteilter Schätzer“ ist.

Bemerkung 2.1.6. Die Aussage von Teil 4 kann man folgendermaßen verstehen: Die Verteilung des Schätzfehlers $\widehat{F}_n(t) - F(t)$ ist für große Werte von n approximativ

$$\mathrm{N}\left(0, \frac{F(t)(1-F(t))}{n}\right).$$

Beweis von (1). Wir betrachten n Experimente. Beim i-ten Experiment überprüfen wir, ob $X_i \leq t$. Falls $X_i \leq t$, sagen wir, dass das i-te Experiment ein Erfolg ist. Die Experimente sind unabhängig voneinander, denn die Zufallsvariablen $X_1, \ldots, X_n$ sind unabhängig. Die Erfolgswahrscheinlichkeit in jedem Experiment ist $\mathbb{P}[X_i \leq t] = F(t)$. Die Anzahl der Erfolge in den n Experimenten, also die Zufallsvariable

$$n\widehat{F}_n(t) = \sum_{i=1}^{n} \mathbb{1}_{X_i \leq t}$$

muss somit binomialverteilt mit Parametern n (Anzahl der Experimente) und $F(t)$ (Erfolgswahrscheinlichkeit) sein.

Beweis von (2). Wir haben in (1) gezeigt, dass $n\widehat{F}_n(t) \sim \mathrm{Bin}(n, F(t))$. Der Erwartungswert einer binomialverteilten Zufallsvariable ist die Anzahl der Experimente multipliziert mit der Erfolgswahrscheinlichkeit. Also gilt

$$\mathbb{E}[n\widehat{F}_n(t)] = nF(t).$$

Teilen wir beide Seiten durch n, so erhalten wir $\mathbb{E}[\widehat{F}_n(t)] = F(t)$.
Die Varianz einer $\mathrm{Bin}(n,p)$-verteilten Zufallsvariable ist $np(1-p)$, also

$$\mathrm{Var}[n\widehat{F}_n(t)] = nF(t)(1-F(t)).$$

Wir können nun das n aus der Varianz herausziehen, allerdings wird daraus (nach den Eigenschaften der Varianz) n^2. Indem wir nun beide Seiten durch n^2 teilen, erhalten wir

$$\mathrm{Var}[\widehat{F}_n(t)] = \frac{F(t)(1-F(t))}{n}.$$

Beweis von (3). Wir führen die Zufallsvariablen $Y_i = \mathbb{1}_{X_i \leq t}$ ein. Diese sind unabhängig und identisch verteilt (da $X_1, X_2, \ldots,$ unabhängig und identisch verteilt sind) mit

$$\mathbb{P}[Y_i = 1] = \mathbb{P}[X_i \leq t] = F(t), \quad \mathbb{P}[Y_i = 0] = 1 - \mathbb{P}[X_i \leq t] = 1 - F(t).$$

Es gilt also $\mathbb{E}Y_i = F(t)$. Wir können nun das starke Gesetz der großen Zahlen auf die Folge $Y_1, Y_2, \ldots$ anwenden:

$$\widehat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \le t} = \frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow[n\to\infty]{f.s.} \mathbb{E}Y_1 = F(t).$$

Beweis von (4). Mit der Notation von Teil (3) gilt

$$\mathbb{E}Y_i = F(t) \quad \operatorname{Var} Y_i = F(t)(1 - F(t)).$$

Wir wenden den zentralen Grenzwertsatz auf die Folge $Y_1, Y_2, \ldots$ an:

$$\sqrt{n}\frac{\widehat{F}_n(t) - F(t)}{\sqrt{F(t)(1-F(t))}} = \sqrt{n}\frac{\frac{1}{n}\sum_{i=1}^{n} Y_i - \mathbb{E}Y_1}{\sqrt{\operatorname{Var} Y_1}} = \frac{\sum_{i=1}^{n} Y_i - n\mathbb{E}Y_1}{\sqrt{n \operatorname{Var} Y_1}} \xrightarrow[n\to\infty]{d} \mathrm{N}(0,1).$$

□

2.2. Empirische Verteilung

Mit Hilfe der empirischen Verteilungsfunktion können wir also die theoretische Verteilungsfunktion schätzen. Nun führen wir auch die empirische Verteilung ein, mit der wir die theoretische Verteilung schätzen können. Zuerst definieren wir, was die theoretische Verteilung ist.

Definition 2.2.1. Sei X eine Zufallsvariable. Die *theoretische Verteilung* von X ist ein Wahrscheinlichkeitsmaß μ auf $(\mathbb{R}, \mathcal{B})$ mit

$$\mu(A) = \mathbb{P}[X \in A] \text{ für jede Borel-Menge } A \subset \mathbb{R}.$$

Der Zusammenhang zwischen der theoretischen Verteilung μ und der theoretischen Verteilungsfunktion F einer Zufallsvariable ist dieses:

$$F(t) = \mu((-\infty, t]), \quad t \in \mathbb{R}.$$

Wie können wir die theoretische Verteilung anhand einer Stichprobe $(x_1, \ldots, x_n)$ schätzen?

Definition 2.2.2. Die *empirische Verteilung* einer Stichprobe $(x_1, \ldots, x_n) \in \mathbb{R}^n$ ist ein Wahrscheinlichkeitsmaß $\widehat{\mu}_n$ auf $(\mathbb{R}, \mathcal{B})$ mit

$$\widehat{\mu}_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x_i \in A} = \frac{1}{n}\#\{i \in \{1, \ldots, n\} : x_i \in A\}$$

für jede Borel-Menge $A \subset \mathbb{R}$.

Die theoretische Verteilung μ ordnet jeder Menge A die Wahrscheinlichkeit, dass X einen Wert in A annimmt, zu. Die empirische Verteilung $\widehat{\mu}_n$ ordnet jeder Menge A den Anteil der Stichprobe, der in A liegt, zu.

Die empirische Verteilung $\widehat{\mu}_n$ kann man sich folgendermaßen vorstellen: Sie ordnet jedem der Punkte x_i aus der Stichprobe das gleiche Gewicht $1/n$ zu. Falls ein Wert mehrmals in der Stichprobe vorkommt, wird sein Gewicht entsprechend erhöht. Dem Rest der reellen Geraden, also der Menge $\mathbb{R}\backslash\{x_1, \ldots, x_n\}$, ordnet $\widehat{\mu}_n$ Gewicht 0 zu. Am Besten kann man das mit dem Begriff des Dirac-δ-Maßes beschreiben.

Definition 2.2.3. Sei $x \in \mathbb{R}$ eine Zahl. Das *Dirac-δ-Maß* δ_x ist ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}, \mathcal{B})$ mit

$$\delta_x(A) = \begin{cases} 1, & \text{falls } x \in A, \\ 0, & \text{falls } x \notin A \end{cases} \text{ für alle Borel-Mengen } A \subset \mathbb{R}.$$

Das Dirac-δ-Maß δ_x ordnet dem Punkt x das Gewicht 1 zu. Der Menge $\mathbb{R}\backslash\{x\}$ ordnet es das Gewicht 0 zu. Die empirische Verteilung $\widehat{\mu}_n$ lässt sich nun wie folgt darstellen:

$$\widehat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}.$$

Zwischen der empirischen Verteilung $\widehat{\mu}_n$ und der empirischen Verteilungsfunktion $\widehat{F}_n$ besteht der folgende Zusammenhang:

$$\widehat{F}_n(t) = \widehat{\mu}_n((-\infty, t]).$$

Der nächste Satz fasst die wichtigsten Eigenschaften der empirischen Verteilung zusammen.

Satz 2.2.4. Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit Verteilung μ. Sei $A \subset \mathbb{R}$ eine Borel-Menge. Dann gilt

(1) Die Zufallsvariable $n\widehat{\mu}_n(A)$ ist binomialverteilt:

$$n\widehat{\mu}_n(A) \sim \text{Bin}(n, \mu(A)).$$

(2) Der Erwartungswert und die Varianz von $\widehat{\mu}_n(A)$ sind gegeben durch:

$$\mathbb{E}[\widehat{\mu}_n(A)] = \mu(A), \quad \text{Var}[\widehat{\mu}_n(A)] = \frac{\mu(A)(1-\mu(A))}{n}.$$

Insbesondere ist $\widehat{\mu}_n(A)$ ein erwartungstreuer Schätzer für $\mu(A)$.

(3) $\widehat{\mu}_n(A)$ ein stark konsistenter Schätzer für $\mu(A)$, d.h.

$$\widehat{\mu}_n(A) \xrightarrow[n\to\infty]{f.s.} \mu(A).$$

.

(4) Falls $\mu(A) \neq 0, 1$, gilt:

$$\sqrt{n}\frac{\widehat{\mu}_n(A) - \mu(A)}{\sqrt{\mu(A)(1-\mu(A))}} \xrightarrow[n\to\infty]{d} \mathrm{N}(0,1).$$

Somit ist $\widehat{\mu}_n(A)$ ein asymptotisch normalverteilter Schätzer für $\mu(A)$.

Aufgabe 2.2.5. Beweisen Sie Satz 2.2.4. Leiten Sie Satz 2.1.5 aus Satz 2.2.4 her.

Aufgabe 2.2.6. Seien $A, B \subset \mathbb{R}$ zwei Borel-Mengen. Bestimmen Sie $\mathrm{Cov}(\widehat{\mu}_n(A), \widehat{\mu}_n(B))$.

Aufgabe 2.2.7. Seien $A_1, \ldots, A_k \subset \mathbb{R}$ Borel-Mengen. Zeigen Sie, dass der Zufallsvektor

$$\big(\sqrt{n}(\widehat{\mu}_n(A_1) - \mu(A_1)), \ldots, \sqrt{n}(\widehat{\mu}_n(A_k) - \mu(A_k))\big)$$

für $n \to \infty$ in Verteilung gegen eine k-variate Normalverteilung konvergiert.

2.3. Plug-in-Schätzer

Seien $X_1, \ldots, X_n$ unabhängige identisch verteilte Zufallsvariablen mit einer Verteilung μ. Gegeben sei deren Realisierung $(x_1, \ldots, x_n)$. Angenommen, wir wollen ein bestimmtes Merkmal von μ schätzen, z.B. dessen Erwartungswert oder Varianz. Wir wollen also $\Psi(\mu)$ schätzen, wobei $\Psi : \mathcal{M} \to \mathbb{R}$ ein Funktional auf einer Menge $\mathcal{M}$ ist, die aus Wahrscheinlichkeitsmaßen auf $\mathbb{R}$ besteht. Da wir die theoretische Verteilung μ durch die empirische Verteilung $\widehat{\mu}_n$ schätzen, ist es ganz natürlich, $\Psi(\mu)$ durch $\Psi(\widehat{\mu}_n)$ zu schätzen.

Definition 2.3.1. $\Psi(\widehat{\mu}_n)$ heißt der *Plug-in-Schätzer* von $\Psi(\mu)$.

Beispiel 2.3.2 (Empirische Momente)**.** Sei $m_k := \Psi(\mu) = \int_{\mathbb{R}} x^k \mu(\mathrm{d}x) = \mathbb{E}[X_i^k]$ das k-te Moment von μ. Der Plug-in-Schätzer für m_k ist gegeben durch

$$\widehat{m}_k = \int_{\mathbb{R}} x^k \widehat{\mu}_n(\mathrm{d}x) = \frac{x_1^k + \ldots + x_n^k}{n}.$$

Man nennt $\widehat{m}_k$ das k-te *empirische Moment* der Stichprobe $(x_1, \ldots, x_n)$. Insbesondere ist $\bar{x}_n = \hat{m}_1$ der Plug-in-Schätzer für den Erwartungswert $\mu = \mathbb{E}X_1$.

Beispiel 2.3.3 (Plug-in-Schätzer für die Varianz)**.** Sei $\Psi(\mu) = \mathrm{Var}\, X_i = \int_{\mathbb{R}} x^2 \mu(\mathrm{d}x) - \big(\int_{\mathbb{R}} x\mu(\mathrm{d}x)\big)^2$ die Varianz von μ. Der Plug-in-Schätzer ist gegeben durch

$$\hat{\sigma}^2_{\text{Plug-in}} = \int_{\mathbb{R}} x^2 \widehat{\mu}_n(\mathrm{d}x) - \left(\int_{\mathbb{R}} x\widehat{\mu}_n(\mathrm{d}x)\right)^2 = \frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}_n^2 = \frac{n-1}{n} s_n^2,$$

wobei s_n^2 die empirische Varianz ist.

Beispiel 2.3.4 (Schiefe und deren Plug-in-Schätzer)**.** Es seien die ersten drei Momente einer Zufallsvariable X endlich: $\mathbb{E}X = \mu$, $\mathrm{Var}\, X = \sigma^2 > 0$, $\mathbb{E}[X^3] < \infty$. Die *Schiefe* von X ist

definiert als

$$\gamma(X) := \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right].$$

Die Schiefe bleibt invariant unter affinen Transformationen, d.h. für alle $a > 0$, $b \in \mathbb{R}$ gilt

$$\gamma(aX + b) = \gamma(X).$$

Für Verteilungen, die symmetrisch um ihren Erwartungswert sind, verschwindet die Schiefe. Dis Schiefe ist somit weder ein Lage- noch Skalenparameter, sondern ein Maß dafür, wie unsymmetrisch (um den Erwartungswert) die Verteilung einer Zufallsvariable ist. Seien nun $X_1, \ldots, X_n$ unabhängige identisch verteilte Zufallsvariablen mit der gleichen Verteilung wie X. Es sei eine Realisierung $(x_1, \ldots, x_n)$ von $(X_1, \ldots, X_n)$ gegeben. Der Plug-in-Schätzer für die Schiefe von X ist gegeben durch

$$\hat{\gamma}_{\text{Plug-in}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}_n}{\hat{\sigma}^2_{\text{Plug-in}}}\right)^3.$$

Aufgabe 2.3.5 (Kumulanten)**.** Es sei X eine Zufallsvariable für die die Funktion $\psi(t) := \log \mathbb{E}e^{tX}$ in einer Umgebung von 0 endlich ist. Die Ableitungen von ψ an der Stelle 0 heißen die *Kumulanten* von X:

$$\kappa_k(X) = \psi^{(k)}(0), \quad k \in \mathbb{N}.$$

Die Funktion ψ heißt die *kumulantenerzeugende Funktion* von X.

(1) Zeigen Sie, dass $\kappa_1 = \mathbb{E}X$, $\kappa_2 = \operatorname{Var} X$ und $\kappa_3 = \mathbb{E}[(X - \mathbb{E}X)^3]$.
(2) Zeigen Sie, dass κ_k als ein Polynom von $\mathbb{E}X, \mathbb{E}X^2, \ldots, \mathbb{E}X^k$ dargestellt werden kann.
(3) Seien $X_1, \ldots, X_n$ unabhängige Zufallsvariablen. Zeigen Sie, dass für alle $k \in \mathbb{N}$,

$$\kappa_k(X_1 + \ldots + X_n) = \kappa_k(X_1) + \ldots + \kappa_k(X_n).$$

Aufgabe 2.3.6. Zeigen Sie, dass für eine normalverteilte Zufallsvariable alle Kumulanten angefangen mit κ_3 verschwinden.

Aufgabe 2.3.7. Zeigen Sie, dass für eine Poi(λ)-verteilte Zufallsvariable alle Kumulanten gleich λ sind.

Aufgabe 2.3.8. Zeigen Sie, dass wenn $X_1, \ldots, X_n$ unabhängig und identisch verteilte Zufallsvariablen mit endlichem dritten Moment sind, dann gilt

$$\gamma(X_1 + \ldots + X_n) = \frac{\gamma(X)}{\sqrt{n}}.$$

Für $n \to \infty$ geht die Schiefe von $X_1 + \ldots + X_n$ also gegen 0, was dadurch erklärt werden kann, dass die Zufallsvariable $X_1 + \ldots + X_n$ nach dem zentralen Grenzwertsatz approximativ normalverteilt ist. Die Schiefe einer Normalverteilung mit beliebigen Parametern ist aber 0.

Beispiel 2.3.9 (Empirische charakteristische Funktion)**.** Es sei $(x_1, \ldots, x_n)$ eine Realisierung von unabhängigen und identisch verteilten Zufallsvariablen $X_1, \ldots, X_n$ mit Verteilung

μ. Die charakteristische Funktion von $X_1, \ldots, X_n$ sei mit

$$\varphi(t) = \mathbb{E}\mathrm{e}^{itX_1} = \int_{\mathbb{R}} \mathrm{e}^{itx}\mu(\mathrm{d}x), \quad t \in \mathbb{R},$$

bezeichnet. Der Plug-in-Schätzer für $\varphi(t)$ ist

$$\hat{\varphi}_n(t) = \int_{\mathbb{R}} \mathrm{e}^{itx}\hat{\mu}_n(\mathrm{d}x) = \frac{1}{n}\sum_{k=1}^{n} \mathrm{e}^{itX_k}.$$

Aufgabe 2.3.10 (Eigenschaften der empirischen charakteristischen Funktion)**.** Zeigen Sie, dass $\hat{\varphi}_n(t)$ ein erwartungstreuer und stark konsistenter Schätzer für $\varphi(t)$ ist. D.h. für jedes $t \in \mathbb{R}$ gilt

$$\mathbb{E}\hat{\varphi}_n(t) = \varphi(t), \quad \hat{\varphi}_n(t) \xrightarrow[n\to\infty]{f.s.} \varphi(t).$$

2.4. Satz von Gliwenko-Cantelli

Wir haben in Teil 3 von Satz 2.1.5 gezeigt, dass für jedes $t \in \mathbb{R}$ die Zufallsvariable $\widehat{F}_n(t)$ gegen die Konstante $F(t)$ fast sicher konvergiert. Man kann auch sagen, dass die empirische Verteilungsfunktion $\widehat{F}_n$ punktweise fast sicher gegen die theoretische Verteilungsfunktion $F(t)$ konvergiert. Im nächsten Satz beweisen wir eine viel stärkere Aussage. Wir zeigen nämlich, dass die Konvergenz mit Wahrscheinlichkeit 1 sogar *gleichmäßig* ist.

Definition 2.4.1. Der *Kolmogorov-Abstand* zwischen der empirischen Verteilungsfunktion $\widehat{F}_n$ und der theoretischen Verteilungsfunktion F wird folgendermaßen definiert:

$$D_n := \sup_{t\in\mathbb{R}} |\widehat{F}_n(t) - F(t)|.$$

Satz 2.4.2 (Gliwenko-Cantelli)**.** Für den Kolmogorov-Abstand D_n gilt

$$D_n \xrightarrow[n\to\infty]{f.s.} 0.$$

Mit anderen Worten, es gilt

$$\mathbb{P}\left[\lim_{n\to\infty} D_n = 0\right] = 1.$$

Beispiel 2.4.3. Da aus der fast sicheren Konvergenz die Konvergenz in Wahrscheinlichkeit folgt, gilt auch

$$D_n \xrightarrow[n\to\infty]{P} 0.$$

Somit gilt für alle $\varepsilon > 0$:

$$\lim_{n\to\infty} \mathbb{P}\left[\sup_{t\in\mathbb{R}} |\widehat{F}_n(t) - F(t)| > \varepsilon\right] = 0.$$

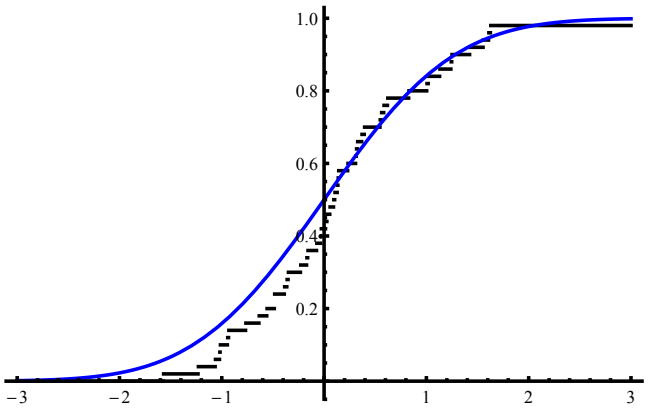


ABBILDUNG 2. Die schwarz dargestellte Funktion ist die empirische Verteilungsfunktion einer Stichprobe vom Umfang $n = 50$ aus der Standardnormalverteilung. Die blaue Kurve ist die Verteilungsfunktion der Normalverteilung. Der Satz von Gliwenko-Cantelli besagt, dass bei steigendem Stichprobenumfang n die schwarze Kurve mit Wahrscheinlichkeit 1 gegen die blaue Kurve gleichmäßig konvergiert.

Also geht die Wahrscheinlichkeit, dass bei der Schätzung von F durch $\widehat{F}_n$ an irgendeiner Stelle ein Fehler von mehr als ε entsteht, für $n \to \infty$ gegen 0.

Bemerkung 2.4.4. Für jedes $t \in \mathbb{R}$ gilt offenbar

$$0 \leq |\widehat{F}_n(t) - F(t)| \leq D_n.$$

Aus dem Satz von Gliwenko-Cantelli und dem Sandwich-Lemma folgt nun, dass für alle $t \in \mathbb{R}$

$$|\widehat{F}_n(t) - F(t)| \xrightarrow[n\to\infty]{f.s.} 0,$$

was exakt der Aussage von Satz 2.1.5, Teil 3 entspricht. Somit ist der Satz von Gliwenko-Cantelli stärker als Satz 2.1.5, Teil 3.

Beweis von Satz 2.4.2. Wir werden den Beweis nur unter der vereinfachenden Annahme führen, dass die Verteilungsfunktion F stetig ist. Sei also F stetig. Sei $m \in \mathbb{N}$ beliebig.

SCHRITT 1. Da F stetig ist und von 0 bis 1 monoton ansteigt, können wir Zahlen

$$z_1 < z_2 < \ldots < z_{m-1}$$

mit der Eigenschaft

$$F(z_1) = \frac{1}{m}, \ldots, F(z_k) = \frac{k}{m}, \ldots, F(z_{m-1}) = \frac{m-1}{m}$$

finden. Um die Notation zu vereinheitlichen, definiern wir noch $z_0 = -\infty$ und $z_m = +\infty$, so dass $F(z_0) = 0$ und $F(z_m) = 1$.

SCHRITT 2. Wir werden nun die Differenz zwischen $\widehat{F}_n(z)$ und $F(z)$ an einer beliebigen Stelle z durch die Differenzen an den Stellen z_k abschätzen. Für jedes $z \in \mathbb{R}$ können wir ein k mit $z \in [z_k, z_{k+1})$ finden. Dann gilt wegen der Monotonie von $\widehat{F}_n$ und F:

$$\widehat{F}_n(z) - F(z) \leq \widehat{F}_n(z_{k+1}) - F(z_k) = \widehat{F}_n(z_{k+1}) - F(z_{k+1}) + \frac{1}{m}.$$

Auf der anderen Seite gilt auch

$$\widehat{F}_n(z) - F(z) \geq \widehat{F}_n(z_k) - F(z_{k+1}) = \widehat{F}_n(z_k) - F(z_k) - \frac{1}{m}.$$

SCHRITT 3. Definiere für $m \in \mathbb{N}$ und $k = 0, 1, \ldots, m$ das Ereignis

$$A_{m,k} := \left\{ \omega \in \Omega : \lim_{n\to\infty} \widehat{F}_n(z_k; \omega) = F(z_k) \right\}.$$

Dabei sei bemerkt, dass $\widehat{F}_n(z_k)$ eine Zufallsvariable ist, weshalb sie auch als Funktion des Ausgangs $\omega \in \Omega$ betrachtet werden kann. Aus Satz 2.1.5, Teil 3 folgt, dass

$$\mathbb{P}[A_{m,k}] = 1 \text{ für alle } m \in \mathbb{N}, k = 0, \ldots, m.$$

SCHRITT 4. Definiere das Ereignis $A_m := \cap_{k=0}^{m} A_{m,k}$. Da ein Schnitt von endlich vielen fast sicheren Ereignis wiederum fast sicher ist, folgt, dass

$$\mathbb{P}[A_m] = 1 \text{ für alle } m \in \mathbb{N}.$$

Da nun auch ein Schnitt von abzählbar vielen fast sicheren Ereignissen wiederum fast sicher ist, gilt auch für das Ereignis $A := \cap_{m=1}^{\infty} A_m$, dass $\mathbb{P}[A] = 1$.

SCHRITT 5. Betrachte nun einen beliebigen Ausgang $\omega \in A_m$. Dann gibt es wegen der Definition von $A_{m,k}$ ein $n(\omega, m) \in \mathbb{N}$ mit der Eigenschaft

$$|\widehat{F}_n(z_k; \omega) - F(z_k)| < \frac{1}{m} \text{ für alle } n > n(\omega, m) \text{ und } k = 0, \ldots, m.$$

Aus Schritt 2 folgt, dass

$$D_n(\omega) = \sup_{z\in\mathbb{R}} |\widehat{F}_n(z; \omega) - F(z)| \leq \frac{2}{m} \text{ für alle } \omega \in A_m \text{ und } n > n(\omega, m).$$

Betrachte nun einen beliebigen Ausgang $\omega \in A$. Somit liegt ω im Ereignis A_m, und das für alle $m \in \mathbb{N}$. Wir können nun das, was oben gezeigt wurde, auch so schreiben: Für alle $m \in \mathbb{N}$ existiert ein $n(\omega, m) \in \mathbb{N}$ so dass für alle $n > n(\omega, m)$ die Ungleichung $0 \leq D_n(\omega) < \frac{2}{m}$ gilt. Das bedeutet aber, dass

$$\lim_{n\to\infty} D_n(\omega) = 0 \text{ für alle } \omega \in A.$$

Da nun die Wahrscheinlichkeit des Ereignisses A laut Schritt 4 gleich 1 ist, erhalten wir

$$\mathbb{P}\left[\left\{ \omega \in \Omega : \lim_{n\to\infty} D_n(\omega) = 0 \right\} \right] \geq \mathbb{P}[A] = 1.$$

Somit gilt $D_n \xrightarrow[n\to\infty]{f.s.} 0$. □

KAPITEL 3

Methoden zur Konstruktion von Schätzern

3.1. Aufgabe der parametrischen Statistik

In den nachfolgenden zwei Abschnitten werden wir das allgemeine Problem der parametrischen Statistik formulieren.

3.2. Zwei Beispiele

Beispiel 3.2.1. Wir betrachten ein Experiment, bei dem eine physikalische Größe (etwa eine Naturkonstante wie z.B. die Lichtgeschwindigkeit) bestimmt werden soll. Da das Ergebnis einer Messung fehlerbehaftet ist, werden n unabhängige Messungen der unbekannten Größe durchgeführt. Es ergeben sich die n Werte $(x_1, \ldots, x_n)$. Üblicherweise nimmt man an, dass diese Werte eine Realisierung von n unabhängigen und identisch verteilten Zufallsvariablen $(X_1, \ldots, X_n)$ mit einer Normalverteilung sind:

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

Dabei ist μ der wahre Wert der zu bestimmenden Größe und σ^2 die quadratische Streuung der Messung. Beide Parameter sind unbekannt. Somit stellt sich das Problem, die Parameter μ und σ^2 anhand der gegebenen Daten $(x_1, \ldots, x_n)$ zu schätzen, siehe Abbildung 1.

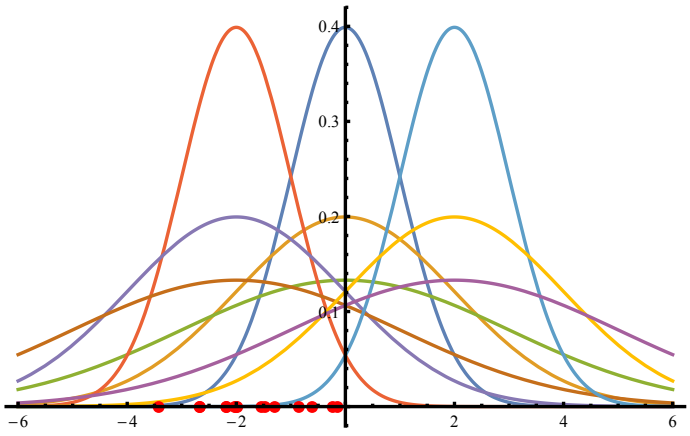


ABBILDUNG 1. Das Bild zeigt die Dichten der Normalverteilungen, die zu verschiedenen Werten der Parameter μ und σ^2 gehören. Die roten Kreise auf der x-Achse zeigen die Stichprobe. Die Aufgabe der parametrischen Statistik ist es, zu entscheiden, zu welchen Parameterwerten die Stichprobe gehört.

Als „Schätzer“ für μ und σ^2 können wir z.B. den empirischen Mittelwert und die empirische Varianz verwenden:

$$\hat{\mu}(x_1,\dots,x_n)=\frac{x_1+\ldots+x_n}{n}=\bar{x}_n,\quad \hat{\sigma}^2(x_1,\dots,x_n)=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2=s_n^2.$$

Die Menge aller möglichen Stichproben wird mit $\mathfrak{X}$ bezeichnet und der *Stichprobenraum* genannt. In diesem Beispiel ist $\mathfrak{X}=\mathbb{R}^n$. Die Menge aller möglichen Parameterwerte wird der *Parameterraum* genannt und mit Θ bezeichnet. In unserem Fall ist

$$\Theta=\{(\mu,\sigma^2)\colon \mu\in\mathbb{R},\sigma^2>0\}=\mathbb{R}\times(0,\infty).$$

Jedem Parameterwert $(\mu,\sigma^2)\in\Theta$ entspricht ein Wahrscheinlichkeitsmaß $\mathbb{P}_{\mu,\sigma^2}$ auf dem Stichprobenraum, das die Verteilung der Stichprobe $(X_1,\dots,X_n)$ unter diesem Parameterwert beschreibt. In unserem Fall ist die Lebesgue-Dichte von $\mathbb{P}_{\mu,\sigma^2}$ gegeben durch

$$L(x_1,\dots,x_n;\theta):=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x_i-\mu)^2}{2\sigma^2}}.$$

Beispiel 3.2.2. Wir betrachten ein Portfolio aus n Versicherungsverträgen. Es sei $x_i\in\{0,1,\dots\}$ die Anzahl der Schäden, die der Vertrag i in einem bestimmten Zeitraum erzeugt hat:

Vertrag	1	2	3	…	n
Schäden	x_1	x_2	x_3	…	x_n

In der Versicherungsmathematik nimmt man oft an, dass die konkrete Stichprobe $(x_1,\dots,x_n)$ eine Realisierung von n unabhängigen und identisch verteilten Zufallsvariablen $(X_1,\dots,X_n)$ ist, die eine Poissonverteilung mit einem unbekannten Parameter $\theta\geq 0$ haben. Es stellt sich

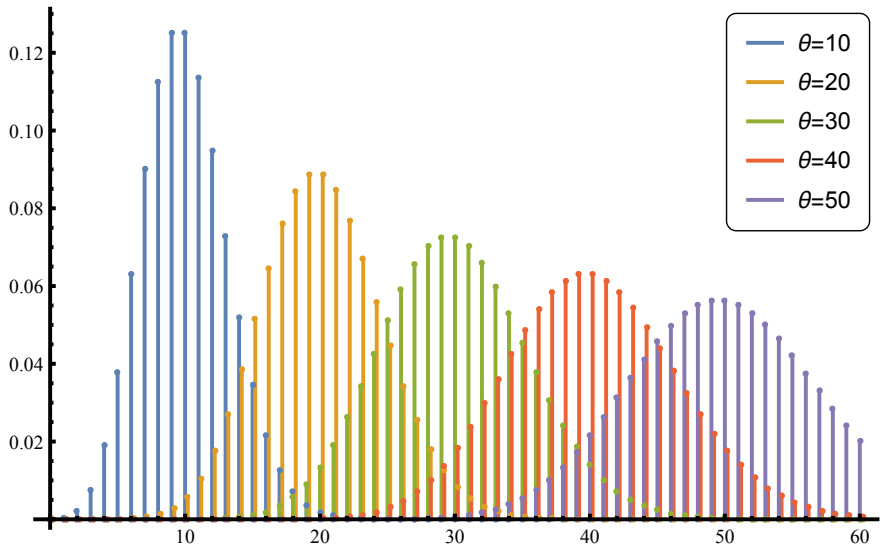


Abbildung 2. Zähldichten der Poissonverteilungen, die zu verschiedenen Werten des Parameters θ gehören.

die Aufgabe, den Parameter θ anhand der Stichprobe $(x_1,\dots,x_n)$ zu schätzen. Da θ der

Erwartungswert der Zufallsvariable X_i ist, können wir als einen natürlichen Schätzer für θ den empirischen Mittelwert $\bar{x}_n$ betrachten.

Der Stichprobenraum ist hier $\mathfrak{X} = \{0, 1, 2, \ldots\}^n = \mathbb{N}_0^n$. Der Parameterraum ist $\Theta = (0, \infty)$. Unsere Verteilungsannahme lautet

$$\mathbb{P}_\theta[X_1 = x_1, \ldots, X_n = x_n] = \prod_{i=1}^{n} \mathrm{e}^{-\theta} \frac{\theta^{x_i}}{x_i!} = \mathrm{e}^{-n\theta} \frac{\theta^{x_1+\ldots+x_n}}{x_1! \ldots x_n!}, \quad (x_1, \ldots, x_n) \in \mathbb{N}_0^n.$$

Das Wahrscheinlichkeitsmaß $\mathbb{P}_\theta$ ist gegeben durch

$$\mathbb{P}_\theta[A] = \mathrm{e}^{-n\theta} \sum_{(x_1,\ldots,x_n)\in A} \frac{\theta^{x_1+\ldots+x_n}}{x_1! \ldots x_n!}, \quad A \subset \mathbb{N}_0^n.$$

3.3. Statistische Modelle: Definition

Nun werden wir die obigen Beispiele verallgemeinern.

Definition 3.3.1. Ein *statistisches Modell* ist ein Tripel $(\mathfrak{X}, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta})$, wobei die Komponenten des Tripels die folgende Bedeutung haben.

- $\mathfrak{X}$ (genannt der *Stichprobenraum*) ist die Menge aller möglichen Stichproben.
- $\mathcal{A} \subset 2^{\mathfrak{X}}$ ist eine σ-Algebra auf $\mathfrak{X}$. Teilmengen $A \subset X$ mit $A \in \mathcal{A}$ heißen *Ereignisse* (oder messbare Teilmengen).
- Θ (genannt der *Parameterraum*) ist die Menge aller möglichen Werte des Parameters θ.
- Für jedes $\theta \in \Theta$ ist $\mathbb{P}_\theta$ ein Wahrscheinlichkeitsmaß auf $(\mathfrak{X}, \mathcal{A})$.

Die Aufgabe der Statistik kann man sich folgendermaßen vorstellen:

- Die Natur such sich einen Parameterwert $\theta \in \Theta$ aus, der uns aber vorenthalten bleibt.
- Aus dem Stichprobenraum $\mathfrak{X}$ wird zufällig und gemäß der Wahrscheinlichkeitsverteilung $\mathbb{P}_\theta$ eine Stichprobe x gezogen, die wir beobachten können.
- Anhand dieser Stichprobe sollen wir θ schätzen.

Die vernünftige Wahl eines statistischen Modells ist eine Aufgabe des Statistikers und hängt vom konkreten Problem ab. Im Weiteren werden wir mehrere Beispiele von statistischen Modellen sehen.

Wir fassen die Stichprobe x als eine Realisierung eines Zufallselements X mit Werten im Stichprobenraum $\mathfrak{X}$ auf. In vielen Beispielen ist $\mathfrak{X} = \mathbb{R}^n$. Dann ist X ein n-dimensionaler Zufallsvektor, dessen Komponenten wir mit $X_1, \ldots, X_n$ bezeichnen.

Notation: Wir bezeichnen mit X ein Zufallselement mit Werten in $\mathfrak{X}$, dessen Verteilung durch $\mathbb{P}_\theta$ gegeben ist. Das heißt, die Wahrscheinlichkeit, dass X einen Wert in der Menge $A \in \mathcal{A}$ annimmt, ist $\mathbb{P}_\theta[A]$.

Per Definition muss $X : \Omega \to \mathfrak{X}$ eine Funktion auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{B}, W_\theta)$ sein (wobei das Wahrscheinlichkeitsmaß W_θ von θ abhängt). Im Folgenden wird es egal sein, wie man diesen Wahrscheinlichkeitsraum wählt, die einfachste Wahl aber ist diese: $\Omega = \mathfrak{X}$, $\mathcal{B} = \mathcal{A}$, $W_\theta = \mathbb{P}_\theta$. Dann ist $X : \mathfrak{X} \to \mathfrak{X}$ die identische Abbildung: $X(x) = x$.

Bevor man von der Wahrscheinlichkeit eines Ereignisses $A \subset \mathfrak{X}$ spricht, muss man sagen, welchen Wert der Parameter θ annehmen soll: $\mathbb{P}_{\theta_1}[A]$ und $\mathbb{P}_{\theta_2}[A]$ können sich durchaus unterscheiden. Genauso verhält es sich mit dem Erwartungswert und der Varianz:

Mit $\mathbb{E}_\theta Z$ und $\mathrm{Var}_\theta\, Z$ bezeichnen wir den Erwartungswert und die Varianz einer Zufallsvariable $Z : \mathfrak{X} \to \mathbb{R}$, die mit Hilfe des Wahrscheinlichkeitsmaßes $\mathbb{P}_\theta$ berechnet wurden.

Nun werden wir ein einfaches statistisches Modell konstruieren.

Beispiel 3.3.2. Wir betrachten eine (im Allgemeinen unfaire) Münze. Die Wahrscheinlichkeit θ, dass die Münze bei einem Wurf „Kopf“ zeigt, sei unbekannt. Um diesen Parameter zu schätzen, werfen wir die Münze n Mal. Man kann für dieses Experiment zwei Modelle vorschlagen.

Erstes Modell. Den Ausgang des Experiments, bei dem die Münze n Mal geworfen wird, fassen wir in einer Stichprobe $(x_1, \dots, x_n) \in \{0,1\}^n$ zusammen, wobei $x_i = 1$, wenn die Münze bei Wurf i „Kopf“ gezeigt hat, und $x_i = 0$, wenn die Münze bei Wurf i „Zahl“ gezeigt hat. Der Stichprobenraum ist somit $\mathfrak{X} = \{0,1\}^n$. Wir fassen alle Teilmengen von $\mathfrak{X}$ als messbar auf, d.h. $\mathcal{A} = 2^{\mathfrak{X}}$. Wir modellieren die Stichprobe $(x_1, \dots, x_n)$ als eine Realisierung von unabhängigen und identisch verteilten Zufallsvariablen $X_1, \dots, X_n$, die Bernoulli-verteilt mit Parameter $\theta \in [0,1]$ sind, d.h.

$$\mathbb{P}_\theta[X_i = x_i] = \theta^{x_i}(1-\theta)^{1-x_i} = \begin{cases} \theta, & \text{falls } x_i = 1, \\ 1-\theta, & \text{falls } x_i = 0, \end{cases} \quad x_i \in \{0,1\}.$$

Das Wahrscheinlichkeitsmaß $\mathbb{P}_\theta$ auf $\{0,1\}^n$ sieht folgendermaßen aus:

$$\mathbb{P}_\theta(A) = \mathbb{P}_\theta[(X_1, \dots, X_n) \in A] = \sum_{(x_1,\dots,x_n)\in A} \theta^{x_1+\dots+x_n}(1-\theta)^{n-(x_1+\dots+x_n)}, \quad A \subset \{0,1\}^n.$$

Zweites Modell. In diesem Modell betrachten wir die Anzahl s der Würfe, bei denen die Münze „Kopf“ gezeigt hat, als eine Realisierung einer Zufallsvariable S, die binomialverteilt mit Parametern n und θ ist. Der Stichprobenumfang ist hier 1, der Stichprobenraum ist $\mathfrak{X} = \{0, 1, \dots, n\}$, mit der σ-Algebra $\mathcal{A} = 2^{\mathfrak{X}}$. Die Wahrscheinlichkeitsmaße $\mathbb{P}_\theta$, $\theta \in [0,1]$, sehen wie folgt aus:

$$\mathbb{P}_\theta[A] = \mathbb{P}_\theta[S \in A] = \sum_{s\in A} \binom{n}{s} \theta^s (1-\theta)^{n-s}, \quad A \subset \{0, 1, \dots, n\}.$$

Abbildung 3 zeigt die Zähldichten der Wahrscheinlichkeitsmaße P_θ für verschiedene Werte von θ.

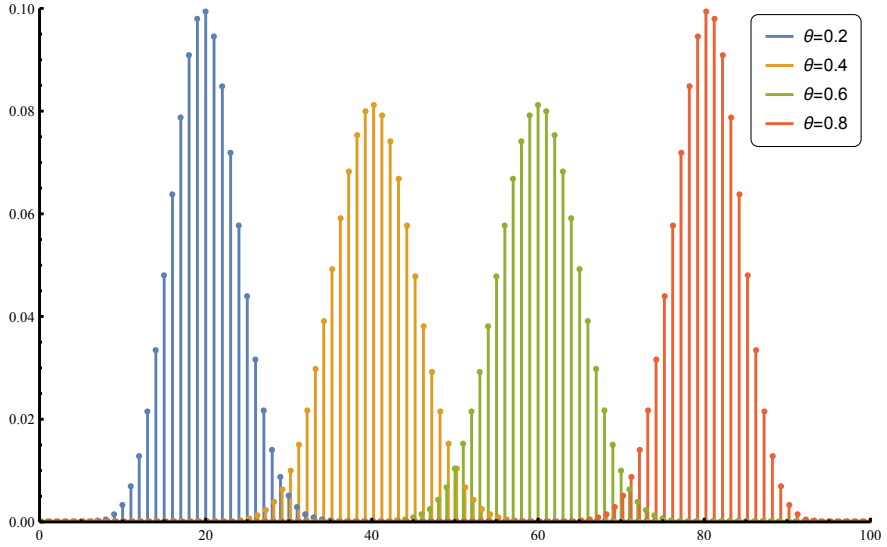


ABBILDUNG 3. Zu Beispiel 3.3.2, zweites Modell: Zähldichten der Binomialverteilungen $\mathrm{Bin}(n, \theta)$, mit $n = 100$ und $\theta = 0.2, 0.4, 0.6, 0.8$.

In diesem Skript werden wir meistens Stichproben der Form $X = (X_1, \ldots, X_n)$ betrachten, wobei $X_1, \ldots, X_n$ *unabhängige und identisch verteilte* Zufallsvariablen sind. Desweiteren werden wir annehmen, dass einer der folgenden beiden Fällen eintritt:

Der diskrete Fall: Die Zufallsvariablen $X_1, \ldots, X_n$ sind diskret (d.h. sie nehmen nur endlich viele oder abzählbar viele Werte an) und haben eine Zähldichte $h_\theta(x)$, die von einem unbekannten Parameter θ abhängt.

Der absolut stetige Fall: Die Zufallsvariablen $X_1, \ldots, X_n$ sind absolut stetig und haben eine Dichte $h_\theta(x)$, die von einem unbekannten Parameter θ abhängt.

Beispiel 3.3.3. Zum diskreten Fall gehören die folgenden Familien von Zähldichten:

(1) Bernoulli-Verteilungen $\{\mathrm{Bern}(\theta) \colon \theta \in [0, 1]\}$,
(2) Binomialverteilungen $\{\mathrm{Bin}(m, p) \colon m \in \mathbb{N}, p \in [0, 1]\}$,
(3) Poisson-Verteilungen $\{\mathrm{Poi}(\theta) \colon \theta > 0\}$, usw.

Zum absolut stetigen Fall gehören die folgenden Verteilungsfamilien:

(1) Normalverteilungen $\{\mathrm{N}(\mu, \sigma^2) \colon \mu \in \mathbb{R}, \sigma^2 > 0\}$,
(2) Betaverteilungen $\{\mathrm{Beta}(\alpha, \beta) \colon \alpha > 0, \beta > 0\}$,
(3) Gammaverteilungen $\{\mathrm{Gamma}(\alpha, \lambda) \colon \alpha > 0, \lambda > 0\}$, usw.

In den obigen Beispielen ist θ ein Vektor mit endlich vielen Komponenten, d.h. $\Theta \subset \mathbb{R}^r$. Dann spricht man von *parametrischer Statistik*. Man kann aber auch eine u.i.v. Stichprobe $(X_1, \ldots, X_n)$ betrachten, in der gar keine Bedingung an die Verteilungsfunktion von X_i gestellt wird. In diesem Fall ist Θ die Menge aller Verteilungsfunktionen. Oder man nimmt an, dass die Zufallsvariablen X_i eine Dichte besitzen, die zu einem bestimmten unendlich dimensionalen Funktionenraum gehört, der dann mit Θ identifiziert wird. Dann spricht man von *nichtparametrischer* Statistik. In diesem Kapitel befassen wir uns nur mit dem parametrischen Fall. Für nichtparametrische Fragestellungen siehe z.B. Kapitel 2 (empirische Verteilungsfunktion) und Kapitel 7 (Kerndichteschätzer).

Unsere Aufgabe besteht nun darin, den Parameter θ anhand der Stichprobe $x \in \mathfrak{X}$ zu schätzen. Zu diesem Zweck konstruiert man Schätzer.

Definition 3.3.4. Sei $\Theta \subset \mathbb{R}^r$. Ein *Schätzer* ist eine messbare Abbildung

$$\hat{\theta} : \mathfrak{X} \to \Theta, \quad x \mapsto \hat{\theta}(x).$$

Manchmal werden wir auch Funktionen $\hat{\theta} : \mathfrak{X} \to \mathbb{R}^r$, die Werte außerhalb von Θ annehmen dürfen, als Schätzer bezeichnen.

Man muss versuchen, den Schätzer so zu konstruieren, dass $\hat{\theta}(x)$ den wahren Wert des Parameters θ möglichst gut approximiert. In den nächsten drei Abschnitten werden wir die drei wichtigsten Methoden zur Konstruktion von Schätzern betrachten: die *Momentenmethode*, die *Maximum-Likelihood-Methode* und die *Bayes-Methode*. Anschließend werden wir auf eine weitere (ziemlich exotische) Methode, die *Maximum-Spacing-Methode*, eingehen.

3.4. Momentenmethode

Wir nehmen an, dass die Stichprobe $(x_1, \ldots, x_n) \in \mathbb{R}^n$ eine Realisierung von unabhängigen und identisch verteilten Zufallsvariablen $(X_1, \ldots, X_n)$ mit Verteilungsfunktion F_θ ist. Wir werden den unbekannten Parameter $\theta = (\theta_1, \ldots, \theta_r) \in \mathbb{R}^r$ schätzen. Für die Momentenmethode benötigen wir die folgenden Begriffe.

Definition 3.4.1. Das k-te *theoretische Moment* (mit $k \in \mathbb{N}$) der Zufallsvariable X_i ist definiert durch

$$m_k(\theta) = \mathbb{E}_\theta[X_i^k].$$

Zum Beispiel ist $m_1(\theta)$ der Erwartungswert von X_i. Die theoretischen Momente sind Funktionen des Parameters θ.

Definition 3.4.2. Das k-te *empirische Moment* (mit $k \in \mathbb{N}$) der Stichprobe $(x_1, \ldots, x_n)$ ist definiert durch

$$\hat{m}_k = \frac{x_1^k + \ldots + x_n^k}{n}.$$

Zum Beispiel ist $\hat{m}_1$ der empirische Mittelwert $\bar{x}_n$ der Stichprobe.

Die Idee der Momentenmethode besteht darin, die empirischen Momente den theoretischen gleichzusetzen. Dabei sind die empirischen Momente bekannt, denn sie hängen nur von der Stichprobe $(x_1, \ldots, x_n)$ ab. Die theoretischen Momente sind hingegen Funktionen des unbekannten Parameters θ, bzw. Funktionen seiner Komponenten $\theta_1, \ldots, \theta_r$. Um r unbekannte Parameter zu finden, brauchen wir normalerweise r Gleichungen. Wir betrachten also ein System aus r Gleichungen mit r Unbekannten:

$$m_1(\theta_1, \ldots, \theta_r) = \hat{m}_1, \quad \ldots, \quad m_r(\theta_1, \ldots, \theta_r) = \hat{m}_r.$$

Die Lösung dieses Gleichungssystems (falls sie existiert und eindeutig ist) nennt man den *Momentenschätzer* und bezeichnet mit $\hat{\theta}_{\mathrm{ME}}$. Dabei steht „ME“ für „Moment Estimator“.

Beispiel 3.4.3. Momentenmethode für den Parameter der Bernoulli-Verteilung Bern(θ).
In diesem Beispiel betrachten wir eine unfaire Münze. Die Wahrscheinlichkeit θ, dass die Münze bei einem Wurf „Kopf“ zeigt, sei unbekannt. Um diesen Parameter zu schätzen, werfen wir die Münze $n = 100$ Mal. Nehmen wir an, dass die Münze dabei $s = 60$ Mal „Kopf“ gezeigt hat. Das Problem besteht nun darin, θ zu schätzen.
Wir betrachten für dieses Problem das folgende mathematische Modell. Zeigt die Münze bei Wurf i Kopf, so setzen wir $x_i = 1$, ansonsten sei $x_i = 0$. Auf diese Weise erhalten wir eine Stichprobe $(x_1, \ldots, x_n) \in \{0, 1\}^n$ mit $x_1 + \ldots + x_n = s = 60$. Wir nehmen an, dass $(x_1, \ldots, x_n)$ eine Realisierung von n unabhängigen Zufallsvariablen $X_1, \ldots, X_n$ mit einer Bernoulli-Verteilung mit Parameter $\theta \in [0, 1]$ ist, d.h.

$$\mathbb{P}_\theta[X_i = 1] = \theta, \quad \mathbb{P}_\theta[X_i = 0] = 1 - \theta.$$

Da wir nur einen unbekannten Parameter haben, brauchen wir nur das erste Moment zu betrachten. Das erste theoretische Moment von X_i ist gegeben durch

$$m_1(\theta) = \mathbb{E}_\theta X_i = 1 \cdot \mathbb{P}_\theta[X_i = 1] + 0 \cdot \mathbb{P}_\theta[X_i = 0] = \theta.$$

Das erste empirische Moment ist gegeben durch

$$\hat{m}_1 = \frac{x_1 + \ldots + x_n}{n} = \frac{s}{n} = \frac{60}{100} = 0.6.$$

Setzen wir beide Momente gleich, so erhalten wir den Momentenschätzer

$$\hat{\theta}_{\mathrm{ME}} = \frac{s}{n} = 0.6.$$

Das Ergebnis ist natürlich nicht überraschend.

Beispiel 3.4.4. Momentenmethode für die Parameter der Normalverteilung $\mathrm{N}(\mu, \sigma^2)$.
Sei $(x_1, \ldots, x_n)$ eine Realisierung von unabhängigen und identisch verteilten Zufallsvariablen $X_1, \ldots, X_n$, die eine Normalverteilung mit unbekannten Parametern (μ, σ^2) haben. Als Motivation kann etwa Beispiel 3.2.1 dienen. Wir schätzen μ und σ^2 mit der Momentenmethode.

Da wir zwei Parameter haben, brauchen wir zwei Gleichungen (also Momente der Ordnungen 1 und 2), um diese zu finden. Zuerst berechnen wir die theoretischen Momente. Der Erwartungswert und die Varianz einer $\mathrm{N}(\mu, \sigma^2)$-Verteilung sind gegeben durch

$$\mathbb{E}_{\mu,\sigma^2} X_i = \mu, \quad \mathrm{Var}_{\mu,\sigma^2} X_i = \sigma^2.$$

Daraus ergeben sich die ersten zwei theoretischen Momente:

$$\begin{aligned} m_1(\mu, \sigma^2) &= \mathbb{E}_{\mu,\sigma^2}[X_i] = \mu, \\ m_2(\mu, \sigma^2) &= \mathbb{E}_{\mu,\sigma^2}[X_i^2] = \mathrm{Var}_{\mu,\sigma^2} X_i + (\mathbb{E}_{\mu,\sigma^2}[X_i])^2 = \sigma^2 + \mu^2. \end{aligned}$$

Setzt man die theoretischen und die empirischen Momente gleich, so erhält man das Gleichungssystem

$$\begin{aligned} \frac{x_1 + \ldots + x_n}{n} &= \mu, \\ \frac{x_1^2 + \ldots + x_n^2}{n} &= \sigma^2 + \mu^2. \end{aligned}$$

Dieses Gleichungssystem lässt sich wie folgt nach μ und σ^2 auflösen:

$$\begin{aligned} \mu &= \bar{x}_n, \\ \sigma^2 &= \frac{1}{n}\sum_{i=1}^n x_i^2 - \left(\frac{1}{n}\sum_{i=1}^n x_i\right)^2 = \frac{1}{n}\left(\sum_{i=1}^n x_i^2 - n\bar{x}_n^2\right) = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}_n)^2 = \frac{n-1}{n} s_n^2. \end{aligned}$$

Dabei haben wir die Identität $\sum_{i=1}^n x_i^2 - n\bar{x}_n^2 = \sum_{i=1}^n (x_i - \bar{x}_n)^2$ benutzt (Übung). Somit sind die Momentenschätzer gegeben durch

$$\hat{\mu}_{\mathrm{ME}} = \bar{x}_n, \quad \hat{\sigma}^2_{\mathrm{ME}} = \frac{n-1}{n} s_n^2.$$

Beispiel 3.4.5. Momentenmethode für den Parameter der Poisson-Verteilung $\mathrm{Poi}(\theta)$. Sei $(x_1, \ldots, x_n)$ eine Realisierung von unabhängigen und mit Parameter $\theta > 0$ verteilten Zufallsvariablen $X_1, \ldots, X_n$. Wir schätzen θ mit der Momentenmethode. Da der Erwartungswert einer $\mathrm{Poi}(\theta)$-Verteilung gleich θ ist, gilt

$$m_1(\theta) = \mathbb{E}_\theta X_i = \theta.$$

Das erste empirische Moment ist gegeben durch

$$\hat{m}_1(\theta) = \frac{x_1 + \ldots + x_n}{n} = \bar{x}_n.$$

Nun setzen wir die beiden Momente gleich und erhalten den Momentenschätzer

$$\hat{\theta}_{\mathrm{ME}} = \bar{x}_n.$$

Aufgabe 3.4.6. Es seien $X_1, \ldots, X_n$ unabhängig und gleichverteilt auf dem Intervall

(1) $[0, \theta]$, wobei $\theta > 0$ unbekannt sei.
(2) $[\theta_1, \theta_2]$, wobei $\theta_1, \theta_2 \in \mathbb{R}$ mit $\theta_1 < \theta_2$ unbekannt seien.
(3) $[\theta, \theta + 1]$, wobei $\theta \in \mathbb{R}$ unbekannt sei.
(4) $[-\theta, \theta]$, wobei $\theta > 0$ unbekannt sei.

Bestimmen Sie die Momentenschätzer für die jeweiligen Parameter. *Hinweis:* Im vierten Fall reicht eine Gleichung nicht aus.

Aufgabe 3.4.7. Es seien $X_1, \dots, X_n$ unabhängig und identisch verteilt mit

(1) $X_i \sim \text{Gamma}(\alpha, \lambda)$, wobei $\alpha > 0$ und $\lambda > 0$ unbekannt seien.
(2) $X_i \sim \text{Beta}(\alpha, \beta)$, wobei $\alpha > 0, \beta > 0$ unbekannt seien.
(3) $X_i \sim \text{Bin}(m, p)$, wobei $m \in \mathbb{N}$ bekannt und $p \in [0, 1]$ unbekannt sei.
(4) $X_i \sim \text{Bin}(m, p)$, wobei $m \in \mathbb{N}$ und $p \in [0, 1]$ beide unbekannt seien.

Bestimmen Sie den jeweiligen Momentenschätzer. Im vierten Fall darf der Schätzer für m auch Werte außerhalb von $\mathbb{N}$ annehmen.

Aufgabe 3.4.8. Es seien $X_1, \dots, X_n$ unabhängig und identisch verteilt. Die Verteilung der X_i sei eine Mischung aus zwei Poisson-Verteilungen:

$$\mathbb{P}[X_i = k] = \varepsilon \mathrm{e}^{-\lambda_1} \frac{\lambda_1^k}{k!} + (1 - \varepsilon)\mathrm{e}^{-\lambda_2} \frac{\lambda_2^k}{k!}, \quad k \in \mathbb{N}_0,$$

wobei $\varepsilon \in (0, 1)$, $\lambda_1 > 0$, $\lambda_2 > 0$ unbekannt seien. Benutzen Sie die Momentenmethode, um die Gleichungen für die Momentenschätzer $\hat{\varepsilon}_{\text{ME}}, \hat{\lambda}_{1,\text{ME}}, \hat{\lambda}_{2,\text{ME}}$ aufzustellen.

Aufgabe 3.4.9. Eine Bahnschranke steht für θ Minuten offen, dann für θ Minuten geschlossen, dann wieder für θ Minuten offen, usw. Dabei sei $\theta > 0$ unbekannt. Man hat n Personen gefragt, wie lange sie an der Bahnschranke warten mussten und die Werte $x_1, \dots, x_n$ bekommen. (Einige dieser Werte können 0 sein). Schätzen Sie θ mit der Momentenmethode.

Aufgabe 3.4.10. In einem Hochhaus gibt es n Stockwerke gleicher Bauart, im i-ten Stockwerk wohnen x_i Familien mit Haustieren. Schätzen Sie die Anzahl der Familien ohne Haustiere, die in diesem Haus wohnen.

3.5. Maximum-Likelihood-Methode

Die Maximum-Likelihood-Methode wurde von vielen Mathematikern unabhängig entdeckt, darunter Daniel Bernoulli, Joseph Louis Lagrange und Carl Friedrich Gauß. Bekannt wurde diese Methode vor allem durch die Arbeiten von Ronald Fisher um 1920. Die Maximum-Likelihood-Methode ist (wie auch die Momentenmethode) ein Verfahren, um Schätzer für den unbekannten Parameter θ zu gewinnen.

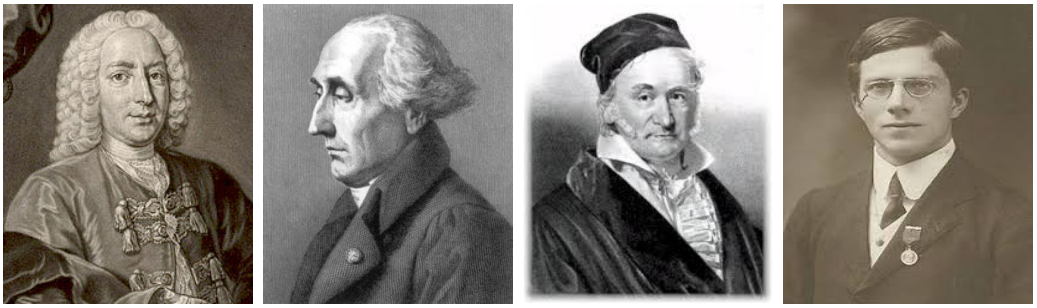


Abbildung 4. Erfinder der Maximum-Likelihood-Methode: Daniel Bernoulli, Joseph-Louis Lagrange, Carl-Friedrich Gauß, Ronald Fisher.

Der diskrete Fall. Sei $(\mathfrak{X}, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ ein statistisches Modell mit einem endlichen oder abzählbaren Stichprobenraum $\mathfrak{X}$. Dann ist die *Likelihood-Funktion* gegeben durch

$$L(x;\theta) = \mathbb{P}_\theta[X = x].$$

Die Likelihood-Funktion hängt sowohl von der Stichprobe x, als auch vom Parameterwert θ ab, wir werden sie aber hauptsächlich als Funktion von θ auffassen und deshalb auch $L(\theta)$ schreiben.

Wichtigster Spezialfall. Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit Werten in einer höchstens abzählbaren Menge $E \subset \mathbb{R}$ und Zähldichte $h_\theta(t) = \mathbb{P}_\theta[X_i = t]$, die von einem Parameter θ abhängt. Wegen Unabhängigkeit gilt für die Likelihood-Funktion

$$L(x_1, \ldots, x_n; \theta) = \mathbb{P}_\theta[X_1 = x_1, \ldots, X_n = x_n] = h_\theta(x_1) \cdot \ldots \cdot h_\theta(x_n).$$

Die Idee der Maximum-Likelihood-Methode besteht darin, einen Wert von θ zu finden, der die Likelihood-Funktion maximiert:

$$L(\theta) \to \max.$$

Definition 3.5.1. Der *Maximum-Likelihood-Schätzer* (oder der *ML-Schätzer*) ist definiert durch

$$\hat{\theta}_{ML} = \arg\max_{\theta\in\Theta} L(\theta).$$

Es kann passieren, dass dieses Maximierungsproblem mehrere Lösungen hat. In diesem Fall muss man eine dieser Lösungen als Schätzer auswählen.

Beispiel 3.5.2. Maximum-Likelihood-Schätzer für den Parameter der Bernoulli-Verteilung $\mathrm{Bern}(\theta)$.
Wir betrachten wieder eine unfaire Münze, wobei die mit θ bezeichnete Wahrscheinlichkeit von „Kopf“ wiederum unbekannt sei. Nach $n = 100$ Würfen habe die Münze $s = 60$ Mal „Kopf“ gezeigt. Wir werden nun θ mit der Maximum-Likelihood-Methode schätzen. Das kann man mit zwei verschiedenen Ansätzen machen, die aber (wie wir sehen werden) zum gleichen Ergebnis führen.

Erstes Modell. Wir modellieren die Stichprobe $(x_1, \ldots, x_n)$ als eine Realisierung von unabhängigen und identisch verteilten Zufallsvariablen $X_1, \ldots, X_n$, die Bernoulli-verteilt mit Parameter $\theta \in [0, 1]$ sind. Es handelt sich um diskrete Zufallsvariablen und die Zähldichte ist gegeben durch

$$h_\theta(x_i) = \mathbb{P}_\theta[X_i = x_i] = \begin{cases} \theta, & \text{falls } x_i = 1, \\ 1-\theta, & \text{falls } x_i = 0, \\ 0, & \text{sonst.} \end{cases}$$

Somit gilt für die Likelihood-Funktion, dass:

$$L(x_1, \ldots, x_n; \theta) = \mathbb{P}_\theta[X_1 = x_1, \ldots, X_n = x_n] = h_\theta(x_1) \cdot \ldots \cdot h_\theta(x_n) = \theta^s(1-\theta)^{n-s},$$

wobei $s = x_1 + \ldots + x_n = 60$ ist. Wir maximieren nun $L(\theta)$; siehe Abbildung 5.

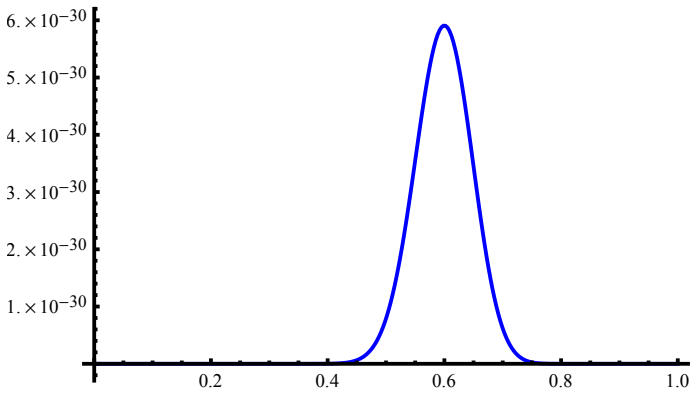


ABBILDUNG 5. Die Likelihood-Funktion $L(\theta) = \theta^{60}(1-\theta)^{40}$, $\theta \in [0,1]$, aus Beispiel 3.5.2, erstes Modell. Das Maximum wird an der Stelle $\theta = 0.6$ erreicht.

Wir benötigen eine Fallunterscheidung.

FALL 1. Sei $s = 0$. Dann ist $L(\theta) = (1-\theta)^n$ und somit gilt $\arg\max L(\theta) = 0$.

FALL 2. Sei $s = n$. Dann ist $L(\theta) = \theta^n$ und somit gilt $\arg\max L(\theta) = 1$.

FALL 3. Sei nun $s \notin \{0, n\}$. Wir leiten die Likelihood-Funktion nach θ ab:

$$\frac{d}{d\theta}L(\theta) = s\theta^{s-1}(1-\theta)^{n-s} - (n-s)\theta^s(1-\theta)^{n-s-1} = \left(\frac{s}{\theta} - \frac{n-s}{1-\theta}\right)\theta^s(1-\theta)^{n-s}.$$

Die Ableitung ist gleich 0 an der Stelle $\theta = \frac{s}{n}$. (Das würde für $s = 0$ und $s = n$ nicht stimmen). Außerdem ist L nichtnegativ und es gilt

$$L(0) = L(1) = 0.$$

Daraus folgt, dass die Stelle $\theta = \frac{s}{n}$ das globale Maximum der Funktion $L(\theta)$ ist.

Die Ergebnisse der drei Fälle können wir nun wie folgt zusammenfassen: Der Maximum-Likelihood-Schätzer ist gegeben durch

$$\hat{\theta}_{ML} = \frac{s}{n} \quad \text{für } s = 0, 1, \ldots, n.$$

Somit ist in unserem Beispiel $\hat{\theta}_{ML} = \frac{60}{100} = 0.6$.

ZWEITES MODELL. In diesem Modell betrachten wir $s = 60$ als eine Realisierung einer binomialverteilten Zufallsvariable S mit Parametern $n = 100$ (bekannt) und $\theta \in [0,1]$ (unbekannt). Somit ist die Likelihood-Funktion

$$L(s;\theta) = \mathbb{P}_\theta[S = s] = \binom{n}{s}\theta^s(1-\theta)^{n-s}.$$

Maximierung dieser Funktion führt genauso wie im ersten Modell zu dem Maximum-Likelihood-Schätzer

$$\hat{\theta}_{ML} = \frac{s}{n}.$$

Beispiel 3.5.3. Maximum-Likelihood-Schätzer für den Parameter der Poisson-Verteilung $\mathrm{Poi}(\theta)$.

Sei $(x_1, \ldots, x_n) \in \mathbb{N}_0^n$ eine Realisierung der unabhängigen und mit Parameter θ Poisson-verteilten Zufallsvariablen $X_1, \ldots, X_n$. Wir schätzen θ mit der Maximum-Likelihood-Methode.

Die Zähldichte der Poissonverteilung $\mathrm{Poi}(\theta)$ ist gegeben durch

$$h_\theta(x) = \mathrm{e}^{-\theta}\frac{\theta^x}{x!}, \quad x = 0, 1, \ldots.$$

Dies führt zu folgender Likelihood-Funktion

$$L(x_1, \ldots, x_n; \theta) = \mathrm{e}^{-\theta}\frac{\theta^{x_1}}{x_1!} \cdot \ldots \cdot \mathrm{e}^{-\theta}\frac{\theta^{x_n}}{x_n!} = \mathrm{e}^{-\theta n}\frac{\theta^{x_1+\ldots+x_n}}{x_1! \cdot \ldots \cdot x_n!}.$$

An Stelle der Likelihood-Funktion ist es in diesem Falle einfacher, die sogenannte *log-Likelihood-Funktion* zu betrachten:

$$\log L(\theta) = -\theta n + (x_1 + \ldots + x_n)\log\theta - \log(x_1! \ldots x_n!).$$

Nun wollen wir einen Wert von θ finden, der diese Funktion maximiert. Für $x_1 = \ldots = x_n = 0$ ist dieser Wert offenbar $\theta = 0$. Seien nun nicht alle x_i gleich 0. Die Ableitung von $\log L(\theta)$ ist gegeben durch

$$\frac{d}{d\theta}\log L(\theta) = -n + \frac{x_1 + \ldots + x_n}{\theta}.$$

Die Ableitung ist gleich 0 an der Stelle $\theta = \bar{x}_n$. (Das ist im Falle, wenn alle x_i gleich 0 sind, falsch, denn dann wäre die Ableitung an der Stelle 0 gleich $-n$). Um zu sehen, dass $\theta = \bar{x}_n$ tatsächlich das globale Maximum der Funktion $\log L(\theta)$ ist, kann man wie folgt vorgehen. Es gilt offenbar $\frac{d}{d\theta}\log L(\theta) > 0$ für $0 \leq \theta < \bar{x}_n$ und $\frac{d}{d\theta}\log L(\theta) < 0$ für $\theta > \bar{x}_n$. Somit ist die Funktion $\log L(\theta)$ strikt steigend auf $[0, \bar{x}_n)$ und strikt fallend auf $(\bar{x}_n, \infty)$. Die Stelle $\bar{x}_n$ ist also tatsächlich das globale Maximum. Der Maximum-Likelihood-Schätzer ist somit

$$\hat{\theta}_{ML} = \bar{x}_n = \frac{x_1 + \ldots + x_n}{n}.$$

Aufgabe 3.5.4. Seien $X_1, \ldots, X_n \sim \mathrm{Bin}(m, p)$ unabhängig, wobei $m \in \mathbb{N}$ bekannt und $p \in [0, 1]$ unbekannt sei. Schätzen Sie p mit der Maximum-Likelihood-Methode.

Aufgabe 3.5.5. Seien $X_1 \sim \mathrm{Poi}(a_1\theta), \ldots, X_n \sim \mathrm{Poi}(a_n\theta)$ unabhängig, wobei $a_1, \ldots, a_n > 0$ bekannt seien und $\theta > 0$ der unbekannte Parameter sei. Bestimmen Sie den Maximum-Likelihood-Schätzer für θ.

Aufgabe 3.5.6. Jeder Mensch trägt einen der drei Genotypen AA, Aa oder aa. Gemäß dem Hardy-Weinberg-Gesetz treten diese Genotypen mit den Wahrscheinlichkeiten $(1-p)^2$, $2p(1-p)$ und p^2 auf, wobei $0 < p < 1$ unbekannt ist. Eine Untersuchung von n Personen ergab folgende Resultate:

- Genotyp AA: x Personen,
- Genotyp Aa: y Personen,
- Genotyp aa: z Personen.

Beschreiben Sie das dazugehörige statistische Modell und bestimmen Sie den Maximum-Likelihood-Schätzer für p.

Der allgemeine Fall. Sei $(\mathfrak{X}, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ ein statistisches Modell, so dass alle Wahrscheinlichkeitsmaße $\mathbb{P}_\theta$ absolut stetig bezüglich eines σ-endlichen Maßes λ auf $(\mathfrak{X}, \mathcal{A})$ sind.

Ein solches statistisches Modell heißt *dominiert* (von λ). Die Dichte von $\mathbb{P}_\theta$ bezüglich λ wird mit

$$L(x;\theta) = \frac{d\mathbb{P}_\theta}{d\lambda}(x)$$

bezeichnet und die Likelihood-Funktion genannt.

Wichtiger Spezialfall. Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit Lebesgue-Dichte h_θ, die von einem Parameter θ abhängt. Als Stichprobenraum können wir dann $\mathfrak{X} = \mathbb{R}^n$ mit der Borel-σ-Algebra nehmen. Sei nun λ das Lebesgue-Maß auf $\mathbb{R}^n$. Dann stimmt die Likelihood-Funktion mit der gemeinsamen Dichte von $X_1, \ldots, X_n$ überein, die wegen Unabhängigkeit in Produktform dargestellt werden kann:

$$L(x_1, \ldots, x_n; \theta) = h_\theta(x_1) \cdot \ldots \cdot h_\theta(x_n).$$

In diesem Fall ist $\mathbb{P}_\theta[X_1 = x_1, \ldots, X_n = x_n] = 0$, deshalb wird die Likelihood-Funktion als Dichte und nicht als Wahrscheinlichkeit definiert!

Der Maximum-Likelihood-Schätzer ist definiert als der Wert von θ, der die Likelihood-Funktion maximiert:

$$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} L(\theta).$$

Beispiel 3.5.7. ML-Schätzer für den Endpunkt der Gleichverteilung $\mathrm{U}[0, \theta]$.
Stellen wir uns vor, dass jemand in einem Intervall $[0, \theta]$ zufällig, gleichverteilt und unabhängig voneinander n Punkte $x_1, \ldots, x_n$ ausgewählt und markiert hat. Uns werden nun die Positionen der n Punkte gezeigt, nicht aber die Position des Endpunktes θ; siehe Abbildung 6. Wir sollen θ anhand der Stichprobe $(x_1, \ldots, x_n)$ rekonstruieren.

ABBILDUNG 6. Rote Kreise zeigen eine Stichprobe vom Umfang $n = 7$, die gleichverteilt auf einem Intervall $[0, \theta]$ ist. Schwarze Kreise zeigen die Endpunkte des Intervalls. Die Position des rechten Endpunktes soll anhand der Stichprobe geschätzt werden.

Der Parameterraum ist hier $\Theta = \{\theta > 0\} = (0, \infty)$. Wir modellieren $(x_1, \ldots, x_n)$ als Realisierungen von unabhängigen und identisch verteilten Zufallsvariablen $X_1, \ldots, X_n$, die gleichverteilt auf einem Intervall $[0, \theta]$ sind. Die Zufallsvariable X_i ist somit absolut stetig und ihre Dichte ist gegeben durch

$$h_\theta(x_i) = \begin{cases} \frac{1}{\theta}, & \text{falls } x_i \in [0, \theta], \\ 0, & \text{falls } x_i \notin [0, \theta]. \end{cases}$$

Das führt zu folgender Likelihood-Funktion

$$L(x_1, \ldots, x_n; \theta) = h_\theta(x_1) \cdot \ldots \cdot h_\theta(x_n) = \frac{1}{\theta^n} \mathbb{1}_{x_1 \in [0,\theta]} \cdot \ldots \cdot \mathbb{1}_{x_n \in [0,\theta]} = \frac{1}{\theta^n} \mathbb{1}_{x_{(n)} \leq \theta}.$$

Dabei ist $x_{(n)} = \max\{x_1, \ldots, x_n\}$ der maximale Wert in dieser Stichprobe. Der Graph der Likelihood-Funktion ist auf Abbildung 7 zu sehen.

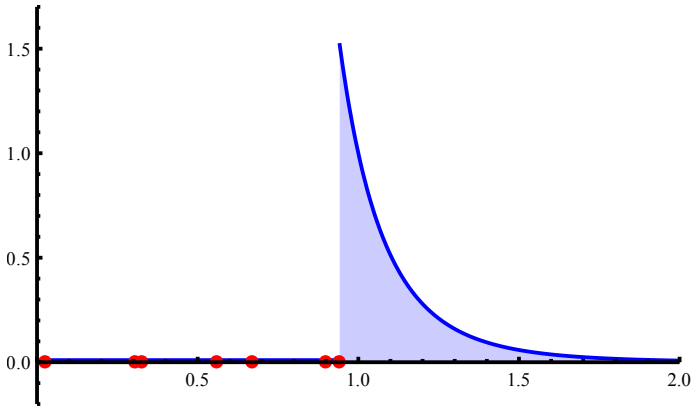


ABBILDUNG 7. Maximum-Likelihood-Schätzung des Endpunktes der Gleichverteilung. Die roten Punkte zeigen die Stichprobe. Die blaue Kurve ist die Likelihood-Funktion $L(\theta)$.

Die Funktion $L(\theta)$ ist 0 solange $\theta < x_{(n)}$, und monoton fallend für $\theta > x_{(n)}$. Somit erhalten wir den Maximum-Likelihood-Schätzer

$$\hat{\theta}_{ML} = \arg\max_{\theta>0} L(\theta) = x_{(n)}.$$

Der Maximum-Likelihood-Schätzer in diesem Beispiel ist also das Maximum der Stichprobe. Es sei bemerkt, dass dieser Schätzer den wahren Wert θ immer unterschätzt, denn die maximale Beobachtung $x_{(n)}$ ist immer kleiner als der wahre Wert des Parameters θ.

Aufgabe 3.5.8. Bestimmen Sie den Momentenschätzer im obigen Beispiel und zeigen Sie, dass er nicht mit dem Maximum-Likelihood-Schätzer übereinstimmt.

Aufgabe 3.5.9. Es seien $X_1, \ldots, X_n$ unabhängig und gleichverteilt auf dem Intervall

(a) $[\theta_1, \theta_2]$, wobei $\theta_1, \theta_2 \in \mathbb{R}$ mit $\theta_1 < \theta_2$ unbekannt seien.
(b) $[\theta, \theta + 1]$, wobei $\theta \in \mathbb{R}$ unbekannt sei.
(c) $[-\theta, \theta]$, wobei $\theta > 0$ unbekannt sei.

Bestimmen Sie den jeweiligen Maximum-Likelihood-Schätzer bzw. beschreiben Sie die Menge aller Parameterwerte, die die Likelihood-Funktion maximieren.

Beispiel 3.5.10. Maximum-Likelihood-Schätzer für die Parameter der Normalverteilung $\mathrm{N}(\mu, \sigma^2)$.
Es sei $(x_1, \ldots, x_n)$ eine Realisierung von unabhängigen und mit Parametern μ, σ^2 normalverteilten Zufallsvariablen $X_1, \ldots, X_n$. Wir schätzen μ und σ^2 mit der Maximum-Likelihood-Methode. Die Dichte von X_i ist gegeben durch

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

Dies führt zu folgender Likelihood-Funktion:

$$L(\mu, \sigma^2) = L(x_1, \ldots, x_n; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}\right).$$

Die log-Likelihood-Funktion sieht folgendermaßen aus:

$$\log L(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Wir bestimmen das Maximum dieser Funktion. Sei zunächst σ^2 fest. Wir betrachten die Funktion $\log L(\mu, \sigma^2)$ als Funktion von μ und bestimmen das Maximum dieser Funktion. Wir leiten nach μ ab:

$$\frac{\partial \log L(\mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu).$$

Die Ableitung ist gleich 0 an der Stelle $\mu = \bar{x}_n$. Für $\mu < \bar{x}_n$ ist die Ableitung positiv (und somit die Funktion steigend), für $\mu > \bar{x}_n$ ist die Ableitung negativ (und somit die Funktion fallend). Also wird bei festem σ^2 an der Stelle $\mu = \bar{x}_n$ das globale Maximum erreicht. Nun machen wir auch $s := \sigma^2$ variabel. Wir betrachten die Funktion

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

Falls alle x_i gleich sind, wird das Maximum an der Stelle $s = 0$ erreicht. Es seien nun nicht alle x_i gleich. Wir leiten nach s ab:

$$\frac{\partial \log L(\bar{x}_n, s)}{\partial s} = -\frac{n}{2s} + \frac{1}{2s^2}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2.$$

Die Ableitung ist gleich 0 an der Stelle

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

(Würden alle x_i gleich sein, so würde das nicht stimmen, denn an der Stelle 0 existiert die Ableitung nicht). Für $s < \tilde{s}_n^2$ ist die Ableitung positiv (und die Funktion somit steigend), für $s > \tilde{s}_n^2$ ist die Ableitung negativ (und die Funktion somit fallend). Somit wird an der Stelle $s = \tilde{s}_n^2$ das globale Maximum der Funktion erreicht. Wir erhalten somit die folgenden Maximum-Likelihood-Schätzer:

$$\hat{\mu}_{\mathrm{ML}} = \bar{x}_n, \quad \hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2.$$

Aufgabe 3.5.11. Seien $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ unabhängig, wobei

(a) $\sigma^2 > 0$ bekannt und $\mu \in \mathbb{R}$ unbekannt sei.

(b) $\mu \in \mathbb{R}$ bekannt und $\sigma^2 > 0$ unbekannt sei.

Schätzen Sie den jeweiligen unbekannten Parameter mit der Maximum-Likelihood-Methode.

Aufgabe 3.5.12. Seien $X_1, \ldots, X_n \sim \mathrm{N}(\theta, \theta)$ unabhängig, wobei $\theta > 0$ unbekannt sei. Schätzen Sie θ mit der Maximum-Likelihood-Methode.

Aufgabe 3.5.13. Seien $X_1, \ldots, X_n \sim \mathrm{N}(\theta, 1)$ unabhängig, wobei θ unbekannt und die Einschränkung $\theta \geq 0$ gegeben sei. Schätzen Sie θ mit der Maximum-Likelihood-Methode.

Aufgabe 3.5.14 (Mehrere Stichproben mit gleichen Varianzen)**.** Seien $X_{i,j}$, $i = 1, \ldots, n$, $j = 1, \ldots, m$, unabhängige Zufallsvariablen mit $X_{i,j} \sim \mathrm{N}(\mu_i^2, \sigma^2)$. Dabei seien $\mu_1 \in \mathbb{R}, \ldots, \mu_n \in \mathbb{R}, \sigma^2 > 0$ unbekannt (alle Varianzen sind gleich). Schätzen Sie diese Parameter mit der Maximum-Likelihood-Methode.

Im nächsten Beispiel betrachten wir die sogenannte *Rückfangmethode* (Englisch: *capture-recapture method*) zur Bestimmung der Größe einer Population.

Beispiel 3.5.15. In einem Teich befinden sich n Fische, wobei n (die Populationsgröße) unbekannt sei. Um die Populationsgröße n zu schätzen, kann man wie folgt vorgehen. Im ersten Schritt („capture“) werden aus dem Teich n_1 (eine bekannte Zahl) Fische gefangen und markiert. Danach werden die n_1 Fische wieder in den Teich zurückgeworfen. Im zweiten Schritt („recapture“) werden k Fische ohne Zurücklegen gefangen. Unter diesen k Fischen seien k_1 markiert und $k - k_1$ nicht markiert.

Anhand dieser Daten kann man n wie folgt schätzen. Man setzt den Anteil der markierten Fische unter den gefangenen Fischen dem Anteil der markierten Fische unter allen Fischen gleich:

$$\frac{k_1}{k} = \frac{n_1}{n}.$$

Aus dieser Gleichung ergibt sich der folgende Schätzer für die Populationsgröße:

$$\hat{n} = \frac{n_1 k}{k_1}.$$

Nun werden wir die Maximum-Likelihood-Methode anwenden und schauen, ob sie den gleichen Schätzer liefert. Die Anzahl k_1 der markierten Fische unter den k gefangenen Fischen betrachten wir als eine Realisierung der Zufallsvariable X mit einer hypergeometrischen Verteilung. Die Likelihood-Funktion ist somit gegeben durch

$$L(k_1; n) = \mathbb{P}[X = k_1] = \frac{\binom{n_1}{k_1} \cdot \binom{n-n_1}{k-k_1}}{\binom{n}{k}}.$$

Die Frage ist nun, für welches n diese Funktion maximal ist. Dabei darf n nur Werte $\{0, 1, 2, \ldots\}$ annehmen. Um dies herauszufinden, betrachten wir die folgende Funktion:

$$R(n) = \frac{L(k_1; n)}{L(k_1; n-1)} = \frac{\binom{n_1}{k_1} \cdot \binom{n-n_1}{k-k_1} \cdot \binom{n-1}{k}}{\binom{n}{k} \cdot \binom{n_1}{k_1} \cdot \binom{n-1-n_1}{k-k_1}} = \frac{(n-k) \cdot (n-n_1)}{n \cdot (n-n_1-k+k_1)}.$$

Eine elementare Rechnung zeigt:

(1) für $n < \hat{n}$ ist $R(n) < 1$;
(2) für $n > \hat{n}$ ist $R(n) > 1$;
(3) für $n = \hat{n}$ ist $R(n) = 1$.

Dabei benutzen wir die Notation $\hat{n} = \frac{n_1 k}{k_1}$. Daraus folgt, dass die Likelihood-Funktion $L(n)$ für $n < \hat{n}$ steigt und für $n > \hat{n}$ fällt. Ist nun $\hat{n}$ keine ganze Zahl, so wird das Maximum von $L(n)$ an der Stelle $n = [\hat{n}]$ erreicht. Ist aber $\hat{n}$ eine ganze Zahl, so gibt es zwei Maxima an den Stellen $n = \hat{n}$ und $n = \hat{n} - 1$. Dabei sind die Werte von $L(n)$ an diesen Stellen gleich,

denn $R(\hat{n}) = 1$. Dies führt zum folgenden Maximum-Likelihood-Schätzer:

$$\hat{n}_{\mathrm{ML}} = \begin{cases} \left[\frac{n_1 k}{k_1}\right], & \text{falls } \frac{n_1 k}{k_1} \notin \mathbb{Z}, \\ \frac{n_1 k}{k_1} \text{ oder } \frac{n_1 k}{k_1} - 1, & \text{falls } \frac{n_1 k}{k_1} \in \mathbb{Z}. \end{cases}$$

Im zweiten Fall ist der Maximum-Likelihood-Schätzer nicht eindeutig definiert. Der Maximum-Likelihood-Schätzer $\hat{n}_{\mathrm{ML}}$ unterscheidet sich also nur unwesentlich vom Schätzer $\hat{n}$.

Aufgabe 3.5.16. Eine Münze, die bei jedem Wurf mit Wahrscheinlichkeit $p \in (0,1)$ „Kopf“ zeigt, wurde n Mal geworfen und hat k mal „Kopf“ gezeigt. Es seien k und p bekannt. Schätzen Sie n mit der Maximum-Likelihood-Methode.

Noch einige Beispiele, in denen eine explizite Bestimmung des Maximum-Likelihood-Schätzers möglich ist, finden sich in den nachfolgenden Aufgaben.

Aufgabe 3.5.17. Seien $X_1, \ldots, X_n$ unabhängig und identisch verteilt mit Dichte

(a) $h_\theta(x) = \theta \mathrm{e}^{-\theta x}$, $x \geq 0$ (Exponentialverteilung).
(b) $h_\theta(x) = \theta c x^{c-1} \mathrm{e}^{-\theta x^c}$, $x \geq 0$, wobei $c > 0$ eine gegebene Konstante sei (Weibull-Verteilung).
(c) $h_\theta(x) = \theta x^{-\theta-1}$, $x \geq 1$ (Pareto-Verteilung).
(d) $h_\theta(x) = \theta x^{\theta-1}$, $0 < x < 1$ (Betaverteilung mit Parametern $(\theta, 1)$).
(e) $h_\theta(x) = (x/\theta^2) \mathrm{e}^{-x^2/\theta^2}$, $x \geq 0$ (Rayleigh-Verteilung).

Dabei ist $\theta > 0$ der unbekannte Parameter. Bestimmen Sie den jeweiligen Maximum-Likelihood-Schätzer für θ.

Aufgabe 3.5.18. Die Zufallsvariablen $X_1, \ldots, X_n$ seien unabhängig und identisch gemäß der Dichte

$$h_\mu(x) = \frac{1}{2} \mathrm{e}^{-|x-\mu|}, \qquad x \in \mathbb{R},$$

$\mu \in \mathbb{R}$, verteilt (verschobene zweiseitige Exponentialverteilung). Bestimmen Sie einen Maximum-Likelihood-Schätzer für μ bzw. beschreiben Sie die Menge aller Parameterwerte, die die Likelihood-Funktion maximieren. *Hinweis:* Aufgabe 1.5.5.

Aufgabe 3.5.19. Die Zufallsvariablen $X_1, \ldots, X_n$ seien unabhängig und identisch gemäß der Dichte

(a) $h_{m,\theta}(x) = \theta \mathrm{e}^{-\theta(x-m)} \mathbb{1}_{[m,\infty)}(x)$ (verschobene Exponentialverteilung).
(b) $h_{m,\theta}(x) = \theta m^\theta x^{-\theta-1} \mathbb{1}_{[m,\infty)}(x)$ (verschobene Pareto-Verteilung).

Dabei seien $m \in \mathbb{R}$ und $\theta > 0$ unbekannte Parameter. Bestimmen Sie den jeweiligen Maximum-Likelihood-Schätzer für (m, θ).

Aufgabe 3.5.20. Seien $X_1, \ldots, X_n$ unabhängig identisch verteilt mit Dichte

$$h_\theta(x) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left\{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right\}, \quad x > 0.$$

wobei $\mu > 0$, $\lambda > 0$ unbekannte Parameter seien (inverse Gauß-Verteilung). Bestimmen Sie den Maximum-Likelihood-Schätzer für (μ, λ).

Zum Schluss betrachten wir ein Beispiel, in dem die Maximum-Likelihood-Methode versagt.

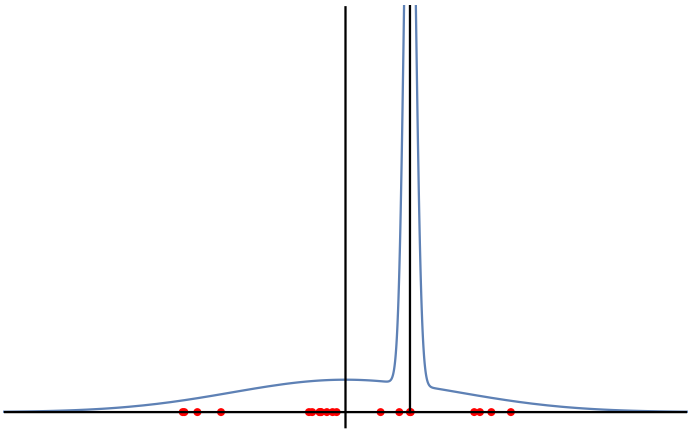


ABBILDUNG 8. Zu Beispiel 3.5.21: Diese Dichte hat eine sehr große Likelihood.

Beispiel 3.5.21 (Mischung aus zwei Normalverteilungen)**.** Betrachte folgendes Zufallsexperiment, in dem eine Zufallsvariable X erzeugt wird. Man wirft eine Münze, die mit Wahrscheinlichkeit $\varepsilon \in (0,1)$ „Kopf“ zeigt. Zeigt die Münze „Kopf“, so erzeugt man eine normalverteilte Zufallsvariable mit Parametern (μ_1, σ_1^2). Zeigt die Münze „Zahl“, so erzeugt man eine normalverteilte Zufallsvariable mit Parametern (μ_2, σ_2^2). Die auf diese Weise erzeugte Zufallsvariable hat (nach der Formel der totalen Wahrscheinlichkeit) die folgende Dichte:

$$h_{\varepsilon,\mu_1,\sigma_1^2,\mu_2,\sigma_2^2}(t) = \frac{1-\varepsilon}{\sigma_1} p\left(\frac{t-\mu_1}{\sigma_1}\right) + \frac{\varepsilon}{\sigma_2} p\left(\frac{t-\mu_2}{\sigma_2}\right), \quad t \in \mathbb{R},$$

wobei $p(t) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-t^2/2}$ die Dichte der Standardnormalverteilung ist. Sei nun $(x_1, \ldots, x_n)$ eine Realisierung von unabhängigen identisch verteilten Zufallsvariablen $X_1, \ldots, X_n$ mit Dichte $h_{\varepsilon,\mu_1,\sigma_1^2,\mu_2,\sigma_2^2}(t)$, wobei

$$\theta := (\varepsilon, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2) \in (0,1) \times \mathbb{R} \times (0,\infty) \times \mathbb{R} \times (0,\infty)$$

der unbekannte Parameter sei. Nun versuchen wir, θ mit der Maximum-Likelihood-Methode zu schätzen. Wir werden das Problem sogar etwas vereinfachen, indem wir die Parameter $\varepsilon, \mu_2, \sigma_2^2$ als gegeben betrachten und nur die restlichen Parameter μ_1, σ_1^2 schätzen. Die Likelihood-Funktion ist

$$L(x_1, \ldots, x_n; \theta) = h_\theta(x_1) \ldots h_\theta(x_n).$$

Wir versuchen, die Likelihood-Funktion zu maximieren. Sei $i \in \{1, \ldots, n\}$ beliebig. Wir werden zeigen, dass die Likelihood-Funktion gegen $+\infty$ geht, wenn $\mu_1 \to x_i$ und $\sigma_1 \to 0$; siehe Abbildung 8. Für jedes $j \in \{1, \ldots, n\} \backslash \{i\}$ gilt

$$\lim_{\substack{\mu_1 \to x_i \\ \sigma_1 \to 0}} h_\theta(x_i) = +\infty, \quad \lim_{\substack{\mu_1 \to x_i \\ \sigma_1 \to 0}} h_\theta(x_j) = \frac{\varepsilon}{\sigma_2} p\left(\frac{x_j - \mu_2}{\sigma_2}\right) > 0.$$

Folglich gilt auch (bei festen $\varepsilon, \mu_2, \sigma_2^2$)

$$\lim_{\substack{\mu_1 \to x_i \\ \sigma_1 \to 0}} L(x_1, \ldots, x_n; \theta) = +\infty.$$

Die Likelihood-Funktion ist unbeschränkt. Die Maximum-Likelihood-Methode hat in diesem Beispiel versagt.

Aufgabe 3.5.22 (ML-Methode bei sehr schweren Tails)**.** Seien $X_1, \ldots, X_n$ unabhängig und identisch gemäß der Dichte

$$h_{\mu,\sigma}(x) = \sigma^{-1} p\left(\frac{x-\mu}{\sigma}\right)$$

verteilt, wobei $p(t) > 0$ eine bekannte Dichtefunktion sei, für die $\lim_{|t|\to\infty} |t|^{1+\varepsilon} p(t) = \infty$ für jedes $\varepsilon > 0$ gelte. Zeigen Sie, dass dann

$$\lim_{\sigma \downarrow 0} L(x_1, \ldots, x_n; x_{(1)} - \sigma, \sigma) = +\infty$$

und somit der Maximum-Likelihood-Schätzer nicht wohldefiniert ist.

Aufgabe 3.5.23. Zeigen Sie, dass eine Dichtefunktion p, die der Bedingung $\lim_{|t|\to\infty} |t|^{1+\varepsilon} p(t) = \infty$ für jedes $\varepsilon > 0$ genügt, existiert.

3.6. Bayes-Methode

Für die Einführung des Bayes-Schätzers muss das parametrische Modell etwas modifiziert werden. Um die Bayes-Methode anwenden zu können, werden wir zusätzlich annehmen, dass der Parameter θ selber eine Zufallsvariable mit einer gewissen (und bekannten) Verteilung ist. Wir betrachten zuerst ein Beispiel.

Beispiel 3.6.1. Eine Versicherung teile die bei ihr versicherten Autofahrer in zwei Kategorien: Typ 1 und Typ 2 (z.B. nach dem Typ des versicherten Fahrzeugs) ein. Die Wahrscheinlichkeit, dass ein Autofahrer vom Typ 1 (bzw. Typ 2) pro Jahr einen Schaden meldet, sei $\theta_1 = 0.4$ (bzw. $\theta_2 = 0.1$). Nun betrachten wir einen Autofahrer von einem unbekannten Typ, der in $n = 10$ Jahren $s = 2$ Schäden hatte. Können wir den Typ dieses Autofahrers raten (schätzen)?

Der Parameterraum ist in diesem Fall $\Theta = \{\theta_1, \theta_2\}$. Es sei S die Zufallsvariable, die die Anzahl der Schäden modelliert, die ein Autofahrer in $n = 10$ Jahren meldet. Unter $\theta = \theta_1$ (also für Autofahrer vom Typ 1) gilt $S \sim \mathrm{Bin}(n, \theta_1)$. Unter $\theta = \theta_2$ (also für Autofahrer vom Typ 2) ist $S \sim \mathrm{Bin}(n, \theta_2)$. Dies führt zur folgenden Likelihood-Funktion:

$$L(s;\theta_1) = \mathbb{P}_{\theta_1}[S = s] = \binom{n}{s}\theta_1^s(1-\theta_1)^{n-s} = \binom{10}{2} \cdot 0.4^2 \cdot 0.6^8 = 0.1209,$$
$$L(s;\theta_2) = \mathbb{P}_{\theta_2}[S = s] = \binom{n}{s}\theta_2^s(1-\theta_2)^{n-s} = \binom{10}{2} \cdot 0.1^2 \cdot 0.9^8 = 0.1937.$$

Wir können nun die Maximum-Likelihood-Methode anwenden, indem wir $L(\theta_1)$ mit $L(\theta_2)$ vergleichen. Es gilt $L(\theta_2) > L(\theta_1)$ und somit handelt es sich vermutlich um einen Autofahrer vom Typ 2.

Sei nun zusätzlich bekannt, dass 90% aller Autofahrer vom Typ 1 und somit nur 10% vom Typ 2 seien. Mit dieser zusätzlichen Vorinformation ist es natürlich, den Parameter θ als eine Zufallsvariable zu modellieren. Die Zufallsvariable θ nimmt zwei Werte θ_1 und θ_2 an und die Wahrscheinlichkeiten dieser Werte sind

$$q(\theta_1) := \mathbb{P}[\theta = \theta_1] = 0.9 \text{ und } q(\theta_2) := \mathbb{P}[\theta = \theta_2] = 0.1.$$

Die Verteilung von θ nennt man auch die *a-priori-Verteilung*. Wie ist nun die Anzahl der Schäden S verteilt, die ein Autofahrer von einem unbekannten Typ in n Jahren meldet? Die Antwort erhält man mit der Formel der totalen Wahrscheinlichkeit:

$$\begin{aligned}\mathbb{P}[S=s] &= \mathbb{P}[\theta=\theta_1]\cdot\mathbb{P}[S=s|\theta=\theta_1] + \mathbb{P}[\theta=\theta_2]\cdot\mathbb{P}[S=s|\theta=\theta_2] \\ &= q(\theta_1)\binom{n}{s}\theta_1^s(1-\theta_1)^{n-s} + q(\theta_2)\binom{n}{s}\theta_2^s(1-\theta_2)^{n-s}.\end{aligned}$$

Es sei bemerkt, dass die Zufallsvariable S *nicht* binomialverteilt ist. Vielmehr ist die Verteilung von S eine Mischung aus zwei verschiedenen Binomialverteilungen. Man sagt auch das S *bedingt* binomialverteilt ist:

$$S|\{\theta=\theta_1\} \sim \mathrm{Bin}(n,\theta_1) \text{ und } S|\{\theta=\theta_2\} \sim \mathrm{Bin}(n,\theta_2).$$

Nun betrachten wir einen Autofahrer von einem unbekannten Typ, der $s = 2$ Schäden gemeldet hat. Die Wahrscheinlichkeit, dass 2 Schäden gemeldet werden, können wir mit der obigen Formel bestimmen:

$$\mathbb{P}[S=2] = 0.9\cdot 0.1209 + 0.1\cdot 0.1937 = 0.1282.$$

Die *a-posteriori-Verteilung* von θ ist die Verteilung von θ gegeben die Information, dass $S = 2$. Zum Beispiel ist die a-posteriori-Wahrscheinlichkeit von $\theta = \theta_1$ definiert als die bedingte Wahrscheinlichkeit, dass $\theta = \theta_1$, gegeben, dass $S = 2$. Um die a-posteriori-Verteilung zu berechnen, benutzen wir die Bayes-Formel:

$$q(\theta_1|s) := \mathbb{P}[\theta=\theta_1|S=s] = \frac{\mathbb{P}[\theta=\theta_1\cap S=s]}{\mathbb{P}[S=s]} = \frac{\mathbb{P}[\theta=\theta_1]\cdot\mathbb{P}[S=s|\theta=\theta_1]}{\mathbb{P}[S=s]}.$$

Mit den oben berechneten Werten erhalten wir, dass

$$q(\theta_1|2) = \frac{0.9\cdot 0.1209}{0.1282} = 0.8486.$$

Die a-posteriori-Wahrscheinlichkeit von $\theta = \theta_2$ kann analog berechnet werden. Es geht aber auch einfacher:

$$q(\theta_2|2) = 1 - q(\theta_1|2) = 0.1513.$$

Nun können wir die a-posteriori-Wahrscheinlichkeiten vergleichen. Da $q(\theta_1|2) > q(\theta_2|2)$, handelt es sich vermutlich um einen Autofahrer vom Typ 1.

Bemerkung 3.6.2. „A priori“ steht für „vor der Erfahrung“. „A posteriori“ steht für „nach der Erfahrung“.

Nun beschreiben wir die allgemeine Form der Bayes-Methode.

Bayes-Methode im diskreten Fall. Zuerst betrachten wir den Fall, dass θ eine diskrete Zufallsvariable ist. Die möglichen Werte für θ seien $\theta_1, \theta_2, \ldots$. Die Verteilung von θ (die auch die *a-priori-Verteilung* genannt wird) sei bekannt:

$$q(\theta_i) := \mathbb{P}[\theta=\theta_i], \quad i = 1, 2, \ldots.$$

Seien $(X_1, \ldots, X_n)$ Zufallsvariablen mit der folgenden Eigenschaft: Gegeben, dass $\theta = \theta_i$, sind die Zufallsvariablen $X_1, \ldots, X_n$ unabhängig und identisch verteilt mit Zähldichte/Dichte $h_{\theta_i}(x)$. Es sei bemerkt, dass die Zufallsvariablen $X_1, \ldots, X_n$ nicht unabhängig, sondern lediglich bedingt unabhängig sind. Es werde nun eine Realisierung $(x_1, \ldots, x_n)$ von $(X_1, \ldots, X_n)$ beobachtet.

Definition 3.6.3. Die *a-posteriori-Verteilung* von θ ist die bedingte Verteilung von θ gegeben die Information, dass $X_1 = x_1, \ldots, X_n = x_n$, d.h.

$$q(\theta_i|x_1, \ldots, x_n) := \mathbb{P}[\theta = \theta_i | X_1 = x_1, \ldots, X_n = x_n], \quad i = 1, 2, \ldots.$$

Hier nehmen wir der Einfachheit halber an, dass die Zufallsvariablen $X_1, \ldots, X_n$ diskret sind. Diese Wahrscheinlichkeit berechnet man mit der Bayes-Formel:

$$\begin{aligned} q(\theta_i|x_1, \ldots, x_n) &= \mathbb{P}[\theta = \theta_i | X_1 = x_1, \ldots, X_n = x_n] \\ &= \frac{\mathbb{P}[X_1 = x_1, \ldots, X_n = x_n, \theta = \theta_i]}{\mathbb{P}[X_1 = x_1, \ldots, X_n = x_n]} \\ &= \frac{\mathbb{P}[\theta = \theta_i] \cdot \mathbb{P}[X_1 = x_1, \ldots, X_n = x_n | \theta = \theta_i]}{\mathbb{P}[X_1 = x_1, \ldots, X_n = x_n]} \\ &= \frac{q(\theta_i) h_{\theta_i}(x_1) \ldots h_{\theta_i}(x_n)}{\sum\limits_j q(\theta_j) h_{\theta_j}(x_1) \ldots h_{\theta_j}(x_n)}. \end{aligned}$$

Wir haben dabei angenommen, dass X_i diskret sind, die Endformel hat aber auch für absolut stetige Variablen X_i Sinn.
In der Bayes-Statistik schreibt man oft $A(t) \propto B(t)$, wenn es eine Konstante C (die von t nicht abhängt) mit $A(t) = C \cdot B(t)$ gibt. Das Zeichen $\propto$ steht also für die Proportionalität von Funktionen. Die Formel für die a-posteriori-Zähldichte von θ kann man dann auch wie folgt schreiben:

$$q(\theta_i|x_1, \ldots, x_n) \propto q(\theta_i) h_{\theta_i}(x_1) \ldots h_{\theta_i}(x_n).$$

Die a-posteriori-Zähldichte $q(\theta_i|x_1, \ldots, x_n)$ ist somit proportional zur a-priori-Zähldichte $q(\theta_i)$ und zur Likelihood-Funktion $L(x_1, \ldots, x_n; \theta_i) = h_{\theta_i}(x_1) \ldots h_{\theta_i}(x_n)$.

Nach der Anwendung der Bayes-Methode erhalten wir als Endergebnis die a-posteriori-Verteilung des Parameters θ. Oft möchte man allerdings das Endergebnis in Form einer Zahl haben. In diesem Fall kann man z. B. folgendermaßen vorgehen.

Definition 3.6.4. Der *Bayes-Schätzer* wird definiert als der Erwartungswert der a-posteriori-Verteilung:

$$\hat{\theta}_{\text{Bayes}} = \sum_i \theta_i q(\theta_i|x_1, \ldots, x_n).$$

Alternativ kann man den Bayes-Schätzer auch als den Median der a-posteriori-Verteilung definieren.

Bayes-Methode im absolut stetigen Fall. Sei nun θ eine absolut stetige Zufallsvariable (bzw. Zufallsvektor) mit Werten in $\mathbb{R}^r$ und einer Dichte $q(\tau)$. Dabei bezeichnen wir mit $\tau \in \mathbb{R}^r$ mögliche Werte von θ. Die Dichte $q(\tau)$ wird auch die a-priori-Dichte genannt. Seien $(X_1, \ldots, X_n)$ Zufallsvariablen mit der folgenden Eigenschaft: Gegeben, dass $\theta = \tau$, sind die Zufallsvariablen $X_1, \ldots, X_n$ unabhängig und identisch verteilt mit Zähldichte/Dichte $h_\tau(x)$. Sei $(x_1, \ldots, x_n)$ eine Realisierung von $(X_1, \ldots, X_n)$. Die a-posteriori-Verteilung von θ ist die

bedingte Verteilung von θ gegeben die Information, dass $X_1 = x_1, \ldots, X_n = x_n$. Indem wir in der Formel aus dem diskreten Fall die Zähldichte von θ durch die Dichte von θ ersetzen, erhalten wir die folgende Formel für die a-posteriori-Dichte von θ:

$$q(\tau|x_1, \ldots, x_n) = \frac{q(\tau)h_\tau(x_1)\ldots h_\tau(x_n)}{\int_{\mathbb{R}^r} q(t)h_t(x_1)\ldots h_t(x_n)\mathrm{d}t}.$$

Das können wir auch wie folgt schreiben:

$$q(\tau|x_1, \ldots, x_n) \propto q(\tau)h_\tau(x_1)\ldots h_\tau(x_n).$$

Die a-posteriori-Dichte $q(\tau|x_1, \ldots, x_n)$ ist somit proportional zur a-priori-Dichte $q(\tau)$ und zur Likelihood-Funktion $L(x_1, \ldots, x_n; \tau) = h_\tau(x_1)\ldots h_\tau(x_n)$.
Genauso wie im diskreten Fall ist der Bayes-Schätzer definiert als der Erwartungswert der a-posteriori-Verteilung, also

$$\hat{\theta}_{\text{Bayes}} = \int_{\mathbb{R}^r} \tau q(\tau|x_1, \ldots, x_n)\mathrm{d}\tau.$$

Aufgabe 3.6.5. Zeigen Sie, dass im diskreten Fall (bzw. im stetigen Fall) $q(\tau|x_1, \ldots, x_n)$ als Funktion von τ tatsächlich eine Zähldichte (bzw. eine Dichte) ist.

Beispiel 3.6.6. Ein Unternehmen möchte ein neues Produkt auf den Markt bringen. Die a-priori-Information sei, dass der Marktanteil θ bei ähnlichen Produkten in der Vergangenheit immer zwischen 0.1 und 0.3 lag. Da keine weiteren Informationen über die Verteilung von θ vorliegen, kann man z.B. die Gleichverteilung auf $[0.1, 0.3]$ als die a-priori-Verteilung von θ ansetzen. Die a-priori-Dichte für den Marktanteil θ ist somit

$$q(\tau) = \begin{cases} 5, & \text{falls } \tau \in [0.1, 0.3], \\ 0, & \text{sonst.} \end{cases}$$

Man kann nun den a-priori-Schätzer für den Marktanteil z.B. als den Erwartungswert dieser Verteilung berechnen:

$$\hat{\theta}_{\text{apr}} = \mathbb{E}\theta = \int_{\mathbb{R}} \tau q(\tau)\mathrm{d}\tau = 0.2.$$

Außerdem seien n Kunden befragt worden, ob sie das neue Produkt kaufen würden. Sei $x_i = 1$, falls der i-te Kunde die Frage bejaht und 0, sonst. Es sei $s = x_1 + \ldots + x_n$ die Anzahl der Kunden in dieser Umfrage, die das neue Produkt kaufen würden. Wir könnten nun den Marktanteil des neuen Produkts z.B. mit der Momentenmethode (Beispiel 3.4.3) oder mit der Maximum-Likelihood-Methode (Beispiel 3.5.2) schätzen:

$$\hat{\theta}_{\text{ME}} = \hat{\theta}_{\text{ML}} = \frac{s}{n}.$$

Dieser Schätzer ignoriert allerdings die a-priori-Information. Mit der Bayes-Methode können wir einen Schätzer konstruieren, der sowohl die a-priori Information, als auch die Befragung berücksichtigt. Wir betrachten $(x_1, \ldots, x_n)$ als eine Realisierung der Zufallsvariablen $(X_1, \ldots, X_n)$. Wir nehmen an, dass bei einem gegebenen θ die Zufallsvariablen $X_1, \ldots, X_n$ unabhängig und mit Parameter θ Bernoulli-verteilt sind:

$$q_\theta(0) := \mathbb{P}_\theta[X_i = 0] = 1 - \theta, \quad q_\theta(1) := \mathbb{P}_\theta[X_i = 1] = \theta.$$

Die Likelihood-Funktion ist

$$L(x_1, \ldots, x_n; \tau) = h_\tau(x_1) \ldots h_\tau(x_n) = \tau^s (1-\tau)^{n-s},$$

wobei $s = x_1 + \ldots + x_n$. Die a-posteriori-Dichte von θ ist proportional zu $q(\tau)$ und $L(x_1, \ldots, x_n; \tau)$ und ist somit gegeben durch

$$q(\tau | x_1, \ldots, x_n) = \begin{cases} \frac{5\tau^s(1-\tau)^{n-s}}{\int_{0.1}^{0.3} 5t^s(1-t)^{n-s}\mathrm{d}t}, & \text{für } \tau \in [0.1, 0.3], \\ 0, & \text{sonst.} \end{cases}$$

Es sei bemerkt, dass die a-posteriori-Dichte (genauso wie die a-priori-Dichte) außerhalb des Intervalls $[0.1, 0.3]$ verschwindet. Wir können nun den Bayes-Schätzer für den Marktanteil θ bestimmen:

$$\hat{\theta}_{\text{Bayes}} = \int_{0.1}^{0.3} \tau q(\tau | x_1, \ldots, x_n) \mathrm{d}\tau = \frac{\int_{0.1}^{0.3} \tau^{s+1}(1-\tau)^{n-s}\mathrm{d}\tau}{\int_{0.1}^{0.3} t^s(1-t)^{n-s}\mathrm{d}t}.$$

Der Bayes-Schätzer liegt im Intervall $[0.1, 0.3]$ (denn außerhalb dieses Intervalls verschwindet die a-posteriori-Dichte) und widerspricht somit der a-priori Information nicht.

Nehmen wir nun an, wir möchten ein Bayes-Modell konstruieren, in dem wir z.B. Bernoulli-verteilte Zufallsvariablen mit einem Parameter θ betrachten, der selber eine Zufallsvariable ist. Wie sollen wir die a-priori-Verteilung von θ wählen? Es wäre schön, wenn die a-posteriori Verteilung eine ähnliche Form haben würde, wie die a-priori-Verteilung. Wie man das erreicht, sehen wir im nächsten Beispiel.

Beispiel 3.6.7 (Bernoulli-Beta-Modell)**.** Bei einem gegebenen $\theta \in [0, 1]$ seien $X_1, \ldots, X_n$ unabhängige Zufallsvariablen, die Bernoulli-verteilt mit Parameter θ sind. Somit gilt

$$h_\theta(0) = 1 - \theta, \quad h_\theta(1) = \theta.$$

Die a-priori-Verteilung von θ sei die Betaverteilung $\text{Beta}(\alpha, \beta)$. Somit ist die a-priori-Dichte von θ gegeben durch

$$q(\tau) = \frac{1}{B(\alpha, \beta)} \tau^{\alpha-1}(1-\tau)^{\beta-1} \propto \tau^{\alpha-1}(1-\tau)^{\beta-1}, \quad \tau \in [0, 1].$$

Es werde nun eine Realisierung $(x_1, \ldots, x_n)$ von $(X_1, \ldots, X_n)$ beobachtet. Die Likelihood-Funktion ist

$$L(x_1, \ldots, x_n; \tau) = h_\tau(x_1) \ldots h_\tau(x_n) = \tau^s (1-\tau)^{n-s}, \quad \tau \in [0, 1],$$

wobei $s = x_1 + \ldots + x_n$. Für die a-posteriori-Dichte von θ gilt somit

$$q(\tau | x_1, \ldots, x_n) \propto q(\tau) L(x_1, \ldots, x_n; \tau) \propto \tau^{\alpha+s-1}(1-\tau)^{\beta+n-s-1}, \quad \tau \in [0, 1].$$

In dieser Formel haben wir die multiplikative Konstante nicht berechnet. Diese muss aber so sein, dass die a-posteriori-Dichte tatsächlich eine Dichte ist, also

$$q(\tau | x_1, \ldots, x_n) = \frac{1}{\mathrm{B}(\alpha + s, \beta + n - s)} \tau^{\alpha+s-1}(1-\tau)^{\beta+n-s-1}, \quad \tau \in [0, 1].$$

Somit ist die a-posteriori-Verteilung von θ eine Betaverteilung:

$$\text{Beta}(\alpha + s, \beta + n - s).$$

Die a-posteriori-Verteilung stammt also aus derselben Betafamilie, wie die a-priori-Verteilung, bloß die Parameter sind anders. Eine a-priori-Verteilung mit dieser Eigenschaft heißt *konjugierte* a-priori-Verteilung. Der Bayes-Schätzer für θ ist der Erwartungswert der a-posteriori-Betaverteilung:

$$\hat{\theta}_{\text{Bayes}} = \frac{\alpha + s}{\alpha + \beta + n}.$$

Weitere Beispiele von Bayes-Modellen, in denen die a-posteriori-Verteilung zur selben Verteilungsfamilie gehört, wie die a-priori-Verteilung, finden sich in folgenden Aufgaben.

Aufgabe 3.6.8 (Poisson-Gamma-Modell)**.** Bei einem gegebenen Wert des Parameters $\lambda > 0$ seien die Zufallsvariablen $X_1, \ldots, X_n$ unabhängig und Poisson-verteilt mit Parameter λ. Dabei wird für λ eine a-priori-Gammaverteilung mit (deterministischen und bekannten) Parametern $b > 0$, $\alpha > 0$ angenommen, d.h.

$$q(\lambda) = \frac{b^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} \mathrm{e}^{-b\lambda} \text{ für } \lambda > 0.$$

Man beobachtet nun eine Realisierung $(x_1, \ldots, x_n)$ von $(X_1, \ldots, X_n)$. Bestimmen Sie die a-posteriori-Verteilung von λ und den Bayes-Schätzer $\hat{\lambda}_{\text{Bayes}}$.

Aufgabe 3.6.9 (Exp-Gamma-Modell)**.** Bei einem gegebenen Wert des Parameters $\lambda > 0$ seien die Zufallsvariablen $X_1, \ldots, X_n$ unabhängig und Exponential-verteilt mit Parameter λ. Dabei wird für λ eine a-priori-Gammaverteilung mit (deterministischen und bekannten) Parametern $b > 0$, $\alpha > 0$ angenommen, d.h.

$$q(\lambda) = \frac{b^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} \mathrm{e}^{-b\lambda} \text{ für } \lambda > 0.$$

Man beobachtet nun eine Realisierung $(x_1, \ldots, x_n)$ von $(X_1, \ldots, X_n)$. Bestimmen Sie die a-posteriori-Verteilung von λ und den Bayes-Schätzer $\hat{\lambda}_{\text{Bayes}}$.

Aufgabe 3.6.10 (Geo-Beta-Modell)**.** Bei einem gegebenen Wert des Parameters $p \in (0, 1)$ seien die Zufallsvariablen $X_1, \ldots, X_n$ unabhängig und geometrisch verteilt mit Parameter p. Dabei wird für p eine a-priori-Betaverteilung mit (deterministischen und bekannten) Parametern $\alpha > 0$, $\beta > 0$ angenommen. Man beobachtet eine Realisierung $(x_1, \ldots, x_n)$ von $(X_1, \ldots, X_n)$. Bestimmen Sie die a-posteriori-Verteilung von p und den Bayes-Schätzer $\hat{p}_{\text{Bayes}}$.

Aufgabe 3.6.11 (Uniform-Pareto-Modell)**.** Bei einem gegebenen Wert des Parameters $\theta > 0$ seien die Zufallsvariablen $X_1, \ldots, X_n$ unabhängig und gleichverteilt auf $[0, \theta]$. Dabei wird für θ eine a–priori–Paretoverteilung mit Parameter $\beta > 0$ angenommen, d.h. die a–priori–Dichte von θ sei

$$q(\tau) = \beta \tau^{-\beta-1} \mathbb{1}_{\tau>1}.$$

Dabei sei $\beta > 0$ ein deterministischer und bekannter Parameter. Man beobachtet eine Realisierung $(x_1, \ldots, x_n) \in (0, \infty)^n$ von $(X_1, \ldots, X_n)$. Bestimmen Sie die a–posteriori–Dichte von θ.

Die Familie der Normalverteilungen mit einem bekannten σ^2 ist selbstkonjugiert:

Aufgabe 3.6.12 (A-priori-Verteilung für den Erwartungswert einer Normalverteilung bei bekannter Varianz)**.** Bei einem gegebenen Wert des Parameters $\mu \in \mathbb{R}$ seien die Zufallsvariablen $X_1, \ldots, X_n$ unabhängig und normalverteilt mit Parametern (μ, σ^2), wobei σ^2 bekannt sei. Dabei wird für μ eine a-priori-Normalverteilung mit (deterministischen und bekannten) Parametern $\mu_0 \in \mathbb{R}$, $\sigma_0^2 > 0$ angenommen. Man beobachtet eine Realisierung $(x_1, \ldots, x_n)$ von $(X_1, \ldots, X_n)$. Bestimmen Sie die a-posteriori-Verteilung von μ und den Bayes-Schätzer $\hat{\mu}_{\text{Bayes}}$.

Aufgabe 3.6.13 (A-priori-Verteilung für die Varianz einer Normalverteilung bei bekanntem Erwartungswert)**.** Bei einem gegebenen Wert des Parameters $\tau \in \mathbb{R}$ seien die Zufallsvariablen $X_1, \ldots, X_n$ unabhängig und normalverteilt mit Parametern (μ, σ^2), wobei μ bekannt sei. Dabei wird für σ^2 eine a-priori inverse Gammaverteilung mit (deterministischen und bekannten) Parametern $b > 0$, $\alpha > 0$ angenommen. Das heißt, es wird angenommen, dass $\tau := 1/\sigma^2$ Gammaverteilt mit Parametern b und α ist. Man beobachtet eine Realisierung $(x_1, \ldots, x_n)$ von $(X_1, \ldots, X_n)$. Bestimmen Sie die a-posteriori-Verteilung von $\tau = 1/\sigma^2$.

3.7. Maximum-Spacing-Methode

Zum Schluss dieses Kapitels betrachten eine weitere Methode zur Konstruktion von Schätzern, die eine Modifikation der Maximum-Likelihood-Methode ist. Seien $X_1, \ldots, X_n$ unabhängige identisch verteilte Zufallsvariablen mit Verteilungsfunktion F_θ, wobei $\theta \in \Theta$ der zu schätzende Parameter sei. Gegeben sei eine Realisierung $(x_1, \ldots, x_n)$ von $(X_1, \ldots, X_n)$. Die Maximal-Spacings-Methode basiert auf den folgenden zwei Beobachtungen.

Lemma 3.7.1. Die Verteilungsfunktion F_θ sei stetig und strikt monoton steigend. Unter $\mathbb{P}_\theta$ sind die Zufallsvariablen $F_\theta(X_1), \ldots, F_\theta(X_n)$ unabhängig und gleichverteilt auf dem Intervall $(0, 1)$.

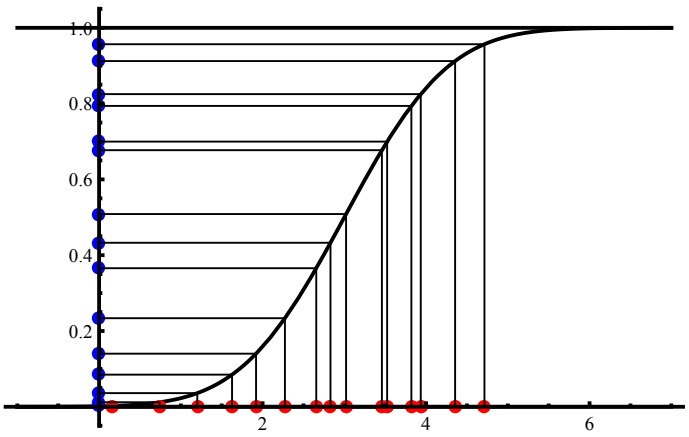


ABBILDUNG 9. Lemma 3.7.1: Eine Stichprobe wird gleichverteilt auf $[0, 1]$, wenn man auf Sie ihre eigene Verteilungsfunktion anwendet.

Beweis. Die Unabhängigkeit folgt aus der Unabhängigkeit von $X_1, \ldots, X_n$. Wir zeigen, dass $F_\theta(X_i)$ gleichverteilt auf $(0,1)$ ist. Es ist klar, dass $F_\theta(X_i)$ nur Werte im Intervall $(0,1)$ annehmen kann. Sei $x \in (0,1)$. Dann gilt

$$\mathbb{P}_\theta[F_\theta(X_i) \leq x] = \mathbb{P}_\theta[X_i \leq F_\theta^{-1}(x)] = F_\theta(F_\theta^{-1}(x)) = x.$$

Somit ist die Zufallsvariable $F_\theta(X_i)$ gleichverteilt auf $(0,1)$. □

Lemma 3.7.2. Es seien $z_1, \ldots, z_k \in [0,1]$ Zahlen mit der Nebenbedingung $z_1 + \ldots + z_k = 1$. Dann gilt

$$z_1 \ldots z_k \leq \frac{1}{k^k}$$

und die Gleichheit tritt genau dann ein, wenn alle Zahlen gleich $\frac{1}{k}$ sind.

Beweis. Dies folgt aus der Cauchy-Ungleichung zwischen dem geometrischen und dem arithmetischen Mittel. □

Nun betrachten wir die Ordnungsstatistiken $x_{(1)} \leq \ldots \leq x_{(n)}$ der Stichprobe $(x_1, \ldots, x_n)$ und definieren zusätzlich $x_{(0)} = -\infty$, $x_{(n+1)} = +\infty$. Der gesuchte Wert des Parameters θ ist dadurch charakterisiert, dass die so-genannten *spacings*

$$D_i(\theta) = F_\theta(x_{(i)}) - F_\theta(x_{(i-1)}), \quad i = 1, \ldots, n+1,$$

eine Realisierung der Spacings der n unabhängigen und auf $(0,1)$ gleichverteilten Zufallsvariablen sind. Somit sind die spacings $D_1(\theta), \ldots, D_{n+1}(\theta)$ ungefähr gleich. Nach Lemma 3.7.2 sollte das Produkt der Spacings nah am maximalen möglichen Wert $(n+1)^{-(n+1)}$ liegen. Dies motiviert die folgende Definition:

Definition 3.7.3. Der *Maximum-Spacing-Schätzer* ist definiert durch

$$\hat{\theta}_{MS} = \arg\max_{\theta \in \Theta} \prod_{i=1}^{n+1} (F_\theta(x_{(i)}) - F_\theta(x_{(i-1)})).$$

Beispiel 3.7.4. Sei $(x_1, \ldots, x_n) \in (0, \infty)^n$ eine Realisierung von unabhängigen, auf dem Intervall $[0, \theta]$ gleichverteilten Zufallsvariablen $X_1, \ldots, X_n$. Die Verteilungsfunktion F_θ ist gegeben durch

$$F_\theta(x) = \begin{cases} 0, & t \leq 0, \\ \frac{t}{\theta}, & t \in [0, \theta], \\ 1, & t \geq \theta. \end{cases}$$

Wir berechnen das Produkt der Spacings als Funktion von θ. Für $\theta \leq x_{(n)}$ gilt $F_\theta(x_{(n+1)}) - F_\theta(x_{(n)}) = 1 - 1 = 0$, denn $F_\theta(+\infty) = 1$. Somit ist das Produkt der Spacings gleich 0. Sei also $\theta > x_{(n)}$. In diesem Fall ergibt sich, dass

$$\prod_{i=1}^{n+1} (F_\theta(x_{(i)}) - F_\theta(x_{(i-1)})) = \frac{1}{\theta^{n+1}} x_{(1)}(x_{(2)} - x_{(1)}) \ldots (x_{(n)} - x_{(n-1)})(\theta - x_{(n)}).$$

Der Maximum-Spacing-Schätzer ist also gegeben durch

$$\hat{\theta}_{MS} = \arg\max_{\theta > x_{(n)}} \frac{\theta - x_{(n)}}{\theta^{n+1}} = \frac{n+1}{n} x_{(n)}.$$

Später werden wir sehen, dass dieser Schätzer erwartungstreu ist und unter allen erwartungstreuen Schätzern die kleinste Varianz besitzt.

Aufgabe 3.7.5. Sei $(x_1, \ldots, x_n) \in \mathbb{R}^n$ eine Realisierung von unabhängigen, auf einem Intervall $[\theta_1, \theta_2]$ gleichverteilten Zufallsvariablen $(X_1, \ldots, X_n)$. Dabei seien $\theta_1, \theta_2 \in \mathbb{R}$ mit $\theta_1 < \theta_2$ die unbekannten Parameter. Schätzen Sie θ_1 und θ_2 mit der Maximum-Spacing-Methode.

KAPITEL 4

Erwartungstreue Schätzer

Ein statistisches Modell ist ein Tripel $(\mathfrak{X}, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta})$, wobei $\mathfrak{X}$ (genannt der Stichprobenraum) die Menge aller möglichen Stichproben, $\mathcal{A}$ eine σ-Algebra auf $\mathfrak{X}$, und $(\mathbb{P}_\theta)_{\theta\in\Theta}$ eine Familie von Wahrscheinlichkeitsmaßen auf $(\mathfrak{X}, \mathcal{A})$ ist. In diesem Kapitel sei der Parameterraum Θ eine Teilmenge von $\mathbb{R}^r$. Eine Stichprobe X wird gemäß einem Wahrscheinlichkeitsmaß $\mathbb{P}_\theta$ zufällig aus $\mathfrak{X}$ gezogen, wobei $\theta \in \Theta$ unbekannt ist. Unsere Aufgabe besteht darin, θ anhand von X zu schätzen.

Definition 4.0.1. Ein *Schätzer* ist eine beliebige (Borel-messbare) Funktion

$$\hat{\theta} : \mathfrak{X} \to \Theta, \quad x \mapsto \hat{\theta}(x).$$

Bemerkung 4.0.2. Manchmal werden wir erlauben, dass ein Schätzer auch Werte außerhalb von Θ annimmt.

Man möchte nun Schätzer konstruieren, für die $\hat{\theta}(X)$ möglichst „nah“ an θ liegt. Dabei ist es ganz natürlich zu fordern, dass der Erwartungswert von $\hat{\theta}(X)$ mit dem zu schätzenden Wert θ übereinstimmen soll. Solche Schätzer heißen erwartungstreu. In diesem Kapitel werden wir versuchen, unter allen erwartungstreuen Schätzern in einem gewissen Sinne „den besten“ zu finden.

4.1. Erwartungstreue, Bias, mittlerer quadratischer Fehler

Definition 4.1.1. Ein Schätzer $\hat{\theta}$ heißt *erwartungstreu* (oder *unverzerrt*), falls

$$\mathbb{E}_\theta[\hat{\theta}(X)] = \theta \text{ für alle } \theta \in \Theta.$$

Der *Bias* (die *Verzerrung*) eines Schätzers $\hat{\theta}$ ist

$$\mathrm{Bias}_\theta(\hat{\theta}) = \mathbb{E}_\theta[\hat{\theta}(X)] - \theta.$$

Wir betrachten $\mathrm{Bias}_\theta(\hat{\theta})$ als eine Funktion von $\theta \in \Theta$.

Bemerkung 4.1.2. Ein Schätzer $\hat{\theta}$ ist genau dann erwartungstreu, wenn $\mathrm{Bias}_\theta(\hat{\theta}) = 0$ für alle $\theta \in \Theta$.

Aufgabe 4.1.3. Zeigen Sie, dass die Menge aller erwartungstreuen Schätzer ein affiner Unterraum des Vektorraumes aller Schätzer ist. D.h. sind $\hat{\theta}_1$ und $\hat{\theta}_2$ erwartungstreu, so ist auch $t\hat{\theta}_1 + (1-t)\hat{\theta}_2$ für alle $t \in \mathbb{R}$ erwartungstreu.

Manchmal möchte man nicht den Parameter θ, sondern eine Funktion $g(\theta)$ schätzen.

Definition 4.1.4. Ein Schätzer φ heißt *erwartungstreu für* $g(\theta)$, falls
$$\mathbb{E}_\theta[\varphi(X)] = g(\theta) \text{ für alle } \theta \in \Theta.$$

Aufgabe 4.1.5. Seien $X_1, \ldots, X_n$ unabhängig und mit Parameter $\theta \in [0,1]$ Bernoulliverteilt. Zeigen Sie, dass es keinen erwartungstreuen Schätzer für $\frac{1}{\theta}$ gibt. Es gibt also statistische Modelle ohne erwartungstreue Schätzer.

Beispiel 4.1.6. In diesem Beispiel werden wir verschiedene Schätzer für den Endpunkt der Gleichverteilung konstruieren. Es seien $X_1, \ldots, X_n \sim \mathrm{U}[0,\theta]$ unabhängige und auf dem Intervall $[0,\theta]$ gleichverteilte Zufallsvariablen, wobei $\theta > 0$ der zu schätzende Parameter sei. Es seien $X_{(1)} < \ldots < X_{(n)}$ die Ordnungsstatistiken von $X_1, \ldots, X_n$.

Erster Schätzer. Zuerst betrachten wir den Maximum-Likelihood-Schätzer
$$\hat{\theta}_1(X_1, \ldots, X_n) = X_{(n)} = \max\{X_1, \ldots, X_n\}.$$
Es ist offensichtlich, dass $\hat{\theta}_1 < \theta$. Somit hat $\hat{\theta}_1$ einen negativen Bias.

Zweiter Schätzer. Wir versuchen nun den Schätzer $\hat{\theta}_1$ zu verbessern, indem wir ihn etwas vergrößern. Wir würden ihn gerne um $\theta - X_{(n)}$ vergrößern, allerdings ist θ unbekannt. Deshalb machen wir den folgenden Ansatz. Wir gehen davon aus, dass die beiden Intervalle $(0, X_{(1)})$ und $(X_{(n)}, \theta)$ ungefähr gleich lang sind, d.h.
$$X_{(1)} \overset{!}{=} \theta - X_{(n)}.$$
Lösen wir diese Gleichung bzgl. θ, so erhalten wir den Schätzer
$$\hat{\theta}_2(X_1, \ldots, X_n) = X_{(n)} + X_{(1)}.$$

Dritter Schätzer. Es gibt aber auch einen anderen natürlichen Ansatz. Wir können davon ausgehen, dass die Intervalle
$$(0, X_{(1)}), (X_{(1)}, X_{(2)}), \ldots, (X_{(n)}, \theta)$$
ungefähr gleich lang sind. Dann kann man die Länge des letzten Intervalls durch das arithmetische Mittel der Längen aller vorherigen Intervalle schätzen, was zu folgender Gleichung führt:
$$\theta - X_{(n)} \overset{!}{=} \frac{1}{n}(X_{(1)} + (X_{(2)} - X_{(1)}) + (X_{(3)} - X_{(2)}) + \ldots + (X_{(n)} - X_{(n-1)})).$$
Da auf der rechten Seite eine Teleskop-Summe steht, erhalten wir die Gleichung
$$\theta - X_{(n)} \overset{!}{=} \frac{1}{n} X_{(n)}.$$
Auf diese Weise ergibt sich der Schätzer
$$\hat{\theta}_3(X_1, \ldots, X_n) = \frac{n+1}{n} X_{(n)}.$$

Vierter Schätzer. Wir können auch den Momentenschätzer betrachten. Setzen wir den Erwartungswert von X_i dem empirischen Mittelwert gleich, so erhalten wir

$$\mathbb{E}_\theta[X_i] = \frac{\theta}{2} \stackrel{!}{=} \bar{X}_n.$$

Dies führt zum Schätzer

$$\hat{\theta}_4(X_1, \ldots, X_n) = 2\bar{X}_n.$$

Aufgabe 4.1.7. Zeigen Sie, dass $\hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4$ erwartungstreu sind, $\hat{\theta}_1$ jedoch nicht.

Man sieht an diesem Beispiel, dass es für ein parametrisches Problem mehrere natürliche (und sogar mehrere erwartungstreue) Schätzer geben kann. Die Frage ist nun, welcher Schätzer der beste ist.

Definition 4.1.8. Sei $\Theta = (a, b) \subset \mathbb{R}$ ein Intervall. Der *mittlere quadratische Fehler* (*mean square error*, MSE) eines Schätzers $\hat{\theta} : \mathfrak{X} \to \mathbb{R}$ ist definiert durch

$$\mathrm{MSE}_\theta(\hat{\theta}) = \mathbb{E}_\theta[(\hat{\theta}(X) - \theta)^2].$$

In dieser Definition wird stillschweigend vorausgesetzt, dass $\hat{\theta}$ *quadratisch integrierbar* ist, d.h.

$$\mathbb{E}_\theta[f(X)^2] < \infty \text{ für alle } \theta \in \Theta.$$

Wir bezeichnen mit L^2 die Menge aller quadratisch integrierbaren Schätzer.

Wir fassen $\mathrm{MSE}_\theta(\hat{\theta})$ als eine Funktion von $\theta \in (a, b)$ auf.

Lemma 4.1.9. Es gilt der folgende Zusammenhang zwischen dem mittleren quadratischen Fehler und dem Bias:

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

Beweis. Um die Notation zu vereinfachen, benutzen wir $\hat{\theta}$ als eine Abkürzung für die Zufallsvariable $\hat{\theta}(X)$. Wir benutzen die Definition des mittleren quadratischen Fehlers, erweitern mit $\mathbb{E}_\theta[\hat{\theta}]$ und quadrieren:

$$\begin{aligned}
\mathrm{MSE}_\theta(\hat{\theta}) &= \mathbb{E}_\theta[(\hat{\theta} - \theta)^2] \\
&= \mathbb{E}_\theta[(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}] + \mathbb{E}_\theta[\hat{\theta}] - \theta)^2] \\
&= \mathbb{E}_\theta[(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}])^2] + 2\mathbb{E}_\theta[(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}]) \cdot (\mathbb{E}_\theta[\hat{\theta}] - \theta)] + \mathbb{E}_\theta[(\mathbb{E}_\theta[\hat{\theta}] - \theta)^2] \\
&= \mathrm{Var}_\theta(\hat{\theta}) + 2(\mathbb{E}_\theta[\hat{\theta}] - \theta) \cdot \mathbb{E}_\theta[\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}]] + (\mathrm{Bias}_\theta(\hat{\theta}))^2.
\end{aligned}$$

Dabei haben wir benutzt, dass $\mathbb{E}_\theta[\hat{\theta}] - \theta$ nicht zufällig ist. Der mittlere Term auf der rechten Seite verschwindet, denn $\mathbb{E}_\theta[\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}]] = \mathbb{E}_\theta[\hat{\theta}] - \mathbb{E}_\theta[\hat{\theta}] = 0$. Daraus ergibt sich die gewünschte Identität. □

Bemerkung 4.1.10. Ist $\hat{\theta}$ erwartungstreu, so gilt $\mathrm{Bias}_\theta(\hat{\theta}) = 0$ für alle $\theta \in \Theta$ und somit vereinfacht sich Lemma 4.1.9 zu

$$\mathrm{MSE}_\theta(\hat{\theta}) = \mathrm{Var}_\theta(\hat{\theta}).$$

Bemerkung 4.1.11. Der Bias ist der „systematische Fehler“ eines Schätzers, die Standardabweichung $\sqrt{\mathrm{Var}_\theta\,\hat{\theta}}$ kann als der „zufällige Fehler“ eines Schätzers angesehen werden. Der mittlere quadratische Fehler MSE berücksichtigt beide Arten von Fehlern.

Mit dem Begriff des mittleren quadratischen Fehlers können wir nun Schätzer vergleichen: je kleiner der Fehler, umso besser der Schätzer.

Definition 4.1.12. Seien $\hat{\theta}_1$ und $\hat{\theta}_2$ zwei Schätzer. Wir sagen, dass $\hat{\theta}_1$ *gleichmäßig besser* als $\hat{\theta}_2$ ist, falls

$$\mathrm{MSE}_\theta(\hat{\theta}_1) \leq \mathrm{MSE}_\theta(\hat{\theta}_2) \text{ für alle } \theta \in \Theta.$$

Bemerkung 4.1.13. Falls $\hat{\theta}_1$ und $\hat{\theta}_2$ erwartungstreu sind, dann ist $\hat{\theta}_1$ gleichmäßig besser als $\hat{\theta}_2$, wenn

$$\mathrm{Var}_\theta(\hat{\theta}_1) \leq \mathrm{Var}_\theta(\hat{\theta}_2) \text{ für alle } \theta \in \Theta.$$

Aufgabe 4.1.14. Es sei X eine Zufallsvariable mit

$$\mathbb{P}_\theta[X = n] = \frac{\mathrm{e}^{-\theta}}{1 - \mathrm{e}^{-\theta}} \frac{\theta^n}{n!}, \quad n \in \mathbb{N}, \quad \theta > 0.$$

Bestimmen Sie alle erwartungstreuen Schätzer für $g(\theta) = \mathrm{e}^{-\theta}$ und den mittleren quadratischen Fehler für jeden solchen Schätzer.

4.2. Bester erwartungstreuer Schätzer

In einem statistischen Modell kann es mehrere erwartungstreue Schätzer geben. Wir versuchen nun, unter diesen Schätzern denjenigen mit der kleinsten Varianz zu finden.

Definition 4.2.1. Ein Schätzer $\hat{\theta}$ heißt *bester erwartungstreuer Schätzer* (für θ), falls er erwartungstreu ist und für jeden anderen erwartungstreuen Schätzer $\tilde{\theta}$ gilt, dass

$$\mathrm{Var}_\theta\,\hat{\theta} \leq \mathrm{Var}_\theta\,\tilde{\theta} \text{ für alle } \theta \in \Theta.$$

Bemerkung 4.2.2. Der entsprechende englische Begriff lautet „UMVU estimator“ (uniformly minimal variance unbiased).

Im nächsten Satz zeigen wir, dass es höchstens einen besten erwartungstreuen Schätzer geben kann.

Satz 4.2.3. Seien $\hat{\theta}_1, \hat{\theta}_2 : \mathfrak{X} \to \Theta$ zwei beste erwartungstreue Schätzer, dann gilt

$$\hat{\theta}_1 = \hat{\theta}_2 \text{ fast sicher unter } \mathbb{P}_\theta \text{ für alle } \theta \in \Theta.$$

Beweis. Schritt 1. Da beide Schätzer beste erwartungstreue Schätzer sind, stimmen die Varianzen dieser beiden Schätzer überein, d.h.

$$\operatorname{Var}_\theta \hat{\theta}_1 = \operatorname{Var}_\theta \hat{\theta}_2 \text{ für alle } \theta \in \Theta.$$

Ist nun $\operatorname{Var}_\theta \hat{\theta}_1 = \operatorname{Var}_\theta \hat{\theta}_2 = 0$ für ein $\theta \in \Theta$, so sind $\hat{\theta}_1$ und $\hat{\theta}_2$ fast sicher konstant unter $\mathbb{P}_\theta$. Da beide Schätzer erwartungstreu sind, muss diese Konstante gleich θ sein und somit muss $\hat{\theta}_1 = \hat{\theta}_2$ fast sicher unter $\mathbb{P}_\theta$ gelten. Die Behauptung des Satzes wäre somit gezeigt. Wir können also im Folgenden annehmen, dass die beiden Varianzen $\operatorname{Var}_\theta \hat{\theta}_1 = \operatorname{Var}_\theta \hat{\theta}_2$ strikt positiv sind.

Schritt 2. Da beide Schätzer erwartungstreu sind, ist auch $\theta^* = \frac{\hat{\theta}_1 + \hat{\theta}_2}{2}$ erwartungstreu und für die Varianz von θ^* gilt

$$\operatorname{Var}_\theta \theta^* = \frac{1}{4}\operatorname{Var}_\theta \hat{\theta}_1 + \frac{1}{4}\operatorname{Var}_\theta \hat{\theta}_2 + \frac{1}{2}\operatorname{Cov}_\theta(\hat{\theta}_1, \hat{\theta}_2) \leq \frac{1}{2}\operatorname{Var}_\theta \hat{\theta}_1 + \frac{1}{2}\sqrt{\operatorname{Var}_\theta \hat{\theta}_1}\sqrt{\operatorname{Var}_\theta \hat{\theta}_2} = \operatorname{Var}_\theta \hat{\theta}_1.$$

Dabei wurde die Cauchy-Schwarzsche Ungleichung angewendet. Somit folgt, dass $\operatorname{Var}_\theta \theta^* \leq \operatorname{Var}_\theta \hat{\theta}_1$. Allerdings ist $\hat{\theta}_1$ der beste erwartungstreue Schätzer, also muss $\operatorname{Var}_\theta \theta^* = \operatorname{Var}_\theta \hat{\theta}_1$ gelten. Daraus folgt, dass die Cauchy-Schwarz-Ungleichung in Wirklichkeit eine Gleichheit gewesen sein muss, also

$$\operatorname{Cov}_\theta(\hat{\theta}_1, \hat{\theta}_2) = \operatorname{Var}_\theta \hat{\theta}_1 = \operatorname{Var}_\theta \hat{\theta}_2.$$

Schritt 3. Der Korrelationskoeffizient von $\hat{\theta}_1$ und $\hat{\theta}_2$ ist also gleich 1. Somit besteht ein linearer Zusammenhang zwischen $\hat{\theta}_1$ und $\hat{\theta}_2$, d.h. es gibt $a = a(\theta)$, $b = b(\theta)$ mit

$$\hat{\theta}_2 = a(\theta) \cdot \hat{\theta}_1 + b(\theta) \text{ fast sicher unter } \mathbb{P}_\theta \text{ für alle } \theta \in \Theta.$$

Setzen wir diesen Zusammenhang bei der Betrachtung der Kovarianz ein und berücksichtigen zusätzlich, dass wie oben gezeigt $\operatorname{Var}_\theta \hat{\theta}_1 = \operatorname{Cov}_\theta(\hat{\theta}_1, \hat{\theta}_2)$, so erhalten wir, dass

$$\operatorname{Var}_\theta \hat{\theta}_1 = \operatorname{Cov}_\theta(\hat{\theta}_1, \hat{\theta}_2) = \operatorname{Cov}_\theta(\hat{\theta}_1, a(\theta) \cdot \hat{\theta}_1 + b(\theta)) = a(\theta) \cdot \operatorname{Var}_\theta \hat{\theta}_1.$$

Also ist $a(\theta) = 1$, denn $\operatorname{Var}_\theta \hat{\theta}_1 \neq 0$ gemäß Schritt 1.

Schritt 4. Somit gilt $\hat{\theta}_2 = \hat{\theta}_1 + b(\theta)$. Auf Grund der Erwartungstreue der Schätzer ist $b(\theta) = 0$, denn

$$\theta = \mathbb{E}_\theta \hat{\theta}_2 = \mathbb{E}_\theta \hat{\theta}_1 + b(\theta) = \theta + b(\theta).$$

Somit folgt, dass $\hat{\theta}_1 = \hat{\theta}_2$ fast sicher unter $\mathbb{P}_\theta$ für alle $\theta \in \Theta$. □

Bemerkung 4.2.4. Der beste erwartungstreue Schätzer muss nicht in jedem parametrischen Modell existieren. Z.B. kann es passieren, dass es überhaupt keine erwartungstreuen Schätzer gibt (Aufgabe 4.1.5).

4.3. Bester erwartungstreuer Schätzer im Bernoulli-Modell

Im Folgenden werden wir für mehrere statistische Modelle den besten erwartungstreuen Schätzer konstruieren. Um unsere Vorgehensweise zu erklären, betrachten wir ein Beispiel, das trotz seiner Einfachheit die beiden wichtigsten Ideen, *Suffizienz* und *Vollständigkeit*, beinhaltet, die man für die allgemeine Konstruktion braucht.

Wir werden den besten erwartungstreuen Schätzer für die Erfolgswahrscheinlichkeit im n-fachen Bernoulli-Experiment bestimmen. Es seien $X_1, \dots, X_n \sim \mathrm{Bern}(\theta)$ unabhängige, mit Parameter $\theta \in [0,1]$ Bernoulli-verteilte Zufallsvariablen. Wir beobachten eine Realisierung $(x_1, \dots, x_n)$ und sollen θ schätzen.

Statistisches Modell. Der Stichprobenraum ist $\mathfrak{X} = \{0,1\}^n$. Als σ-Algebra der messbaren Ereignisse nehmen wir die Potenzmenge $\mathcal{A} = 2^{\mathfrak{X}}$. Die möglichen Verteilungen von $(X_1, \dots, X_n)$ sehen wie folgt aus. Für $\theta \in [0,1]$ ist $\mathbb{P}_\theta$ das Wahrscheinlichkeitsmaß auf $\mathfrak{X}$ mit

$$\mathbb{P}_\theta[A] = \sum_{(x_1,\dots,x_n)\in A} \theta^{x_1+\dots+x_n}(1-\theta)^{n-(x_1+\dots+x_n)}, \quad A \subset \mathfrak{X}.$$

Der folgende Satz sollte nicht überraschend sein.

Satz 4.3.1. Der Schätzer $\hat{\theta}(x_1, \dots, x_n) = \bar{x}_n$ ist der beste erwartungstreue Schätzer von θ im n-fachen Bernoulli-Experiment.

Beweis. Zuerst müssen wir anmerken, dass der Schätzer $\bar{X}_n$ erwartungstreu ist. Es bleibt zu zeigen, dass es keinen besseren erwartungstreuen Schätzer gibt. Sei $\varphi : \mathfrak{X} \to [0,1]$ ein weiterer erwartungstreuer Schätzer von θ. Wir wollen zeigen, dass

$$\mathrm{Var}_\theta\, \varphi \geq \mathrm{Var}_\theta\, \bar{X}_n \text{ für alle } \theta \in [0,1].$$

Erste Idee: Suffizienz. Intuitiv erscheint es plausibel, dass ein „guter“ Schätzer nur die Information verwenden sollte, *wieviele* Erfolge in den Bernoulli-Experimenten beobachtet wurden. Es sollte egal sein, *wann* die Erfolge eintgereten sind. So sollte z.B. ein Schätzer φ mit $\varphi(0,0,1,1,1) \neq \varphi(1,0,1,0,1)$ kein guter Schätzer sein.

Wie können wir das beweisen? Für $k = 0, 1, \dots, n$ definieren wir die Mengen

$$A_k := \{x = (x_1, \dots, x_n) \in \{0,1\}^n \colon x_1 + \dots + x_n = k\}.$$

Dann ist $A_0 \cup \dots \cup A_n = \mathfrak{X}$ eine disjunkte Zerlegung von $\mathfrak{X}$. Die Anzahl der Elemente in A_k ist $\binom{n}{k}$. Für jeden Schätzer $\varphi : \mathfrak{X} \to [0,1]$ betrachten wir nun seine *Rao-Blackwell-Verbesserung* $\varphi^* : \mathfrak{X} \to [0,1]$ mit

$$\varphi^*(x) = \frac{1}{\binom{n}{k}} \sum_{y\in A_k} \varphi(y), \text{ falls } x = (x_1, \dots, x_n) \in A_k.$$

Der Schätzer φ^* ist konstant auf jeder der Mengen $A_0, \dots, A_n$, und der Wert von φ^* auf A_k ist einfach der Mittelwert von φ über A_k; siehe Abbildung 1.

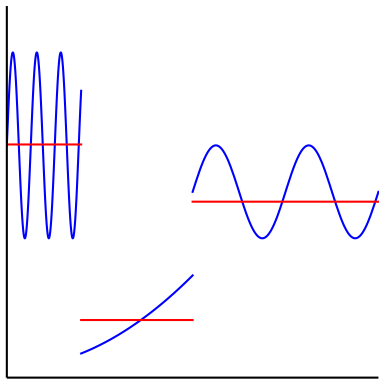


ABBILDUNG 1. Die Idee der Rao-Blackwell-Verbesserung. Eine Funktion (blau) wird durch Mittelwerte (rot) über Mengen aus einer disjunkten Zerlegung ersetzt. Der Erwartungswert bleibt unverändert, die Varianz wird kleiner.

Wir behaupten nun, dass φ^* ebenfalls erwartungstreu ist. In der Tat, wegen der Definition von φ^* gilt

$$\mathbb{E}\varphi^* = \frac{1}{2^n}\sum_{x\in\mathfrak{X}} \varphi^*(x) = \frac{1}{2^n}\sum_{k=0}^{n}\sum_{x\in A_k} \varphi^*(x) = \frac{1}{2^n}\sum_{k=0}^{n}\sum_{y\in A_k} \varphi(y) = \mathbb{E}\varphi = \theta.$$

Und nun zeigen wir, dass $\operatorname{Var}_\theta \varphi^* \le \operatorname{Var}_\theta \varphi$. Mit der Ungleichung $\frac{a_1^2+\ldots+a_N^2}{N} \ge \left(\frac{a_1+\ldots+a_N}{N}\right)^2$ (vom arithmetischen und quadratischen Mittel) erhalten wir

$$\sum_{y\in A_k}(\varphi(y)-\theta)^2 = \binom{n}{k}\frac{1}{\binom{n}{k}}\sum_{y\in A_k}(\varphi(y)-\theta)^2 \ge \binom{n}{k}\left(\frac{1}{\binom{n}{k}}\sum_{y\in A_k}(\varphi(y)-\theta)\right)^2 = \sum_{x\in A_k}(\varphi^*(x)-\theta)^2.$$

Da die Wahrscheinlichkeit (unter $\mathbb{P}_\theta$) von jedem Ausgang $y \in A_k$ gleich $\theta^k(1-\theta)^{n-k}$ ist, können wir schreiben:

$$\begin{aligned}\operatorname{Var}_\theta \varphi = \mathbb{E}_\theta[(\varphi-\theta)^2] &= \sum_{k=0}^{n}\theta^k(1-\theta)^{n-k}\sum_{y\in A_k}(\varphi(y)-\theta)^2 \\ &\ge \sum_{k=0}^{n}\theta^k(1-\theta)^{n-k}\sum_{x\in A_k}(\varphi^*(x)-\theta)^2 = \mathbb{E}_\theta[(\varphi^*-\theta)^2] = \operatorname{Var}_\theta \varphi^*.\end{aligned}$$

Zweite Idee: Vollständigkeit. Im Rest des Beweises können wir also annehmen, dass φ konstant auf jeder der Mengen $A_0, \ldots, A_n$ bleibt (denn andernfalls können wir φ durch φ^* ersetzen, was den Schätzer verbessert). Außerdem muss φ erwartungstreu sein. Gibt es viele solche Schätzer? Zum Glück gibt es nur einen, nämlich $\bar{X}_n$! Um das zu zeigen, bezeichnen wir den Wert von φ auf A_k mit a_k. Dann lautet die Bedingung der Erwartungstreue wie folgt:

$$\mathbb{E}_\theta\varphi = \sum_{k=0}^{n} a_k\binom{n}{k}\theta^k(1-\theta)^{n-k} \stackrel{!}{=} \theta \text{ für alle } \theta\in[0,1].$$

Es ist eine (nicht ganz triviale) Übung zu zeigen, dass das nur für $a_k = k/n$ möglich ist. Für die Lösung verweisen wir auf Abschnitt 4.7. □

Im Rest dieses Kapitels werden wir den obigen Beweis auf eine viel größere Klasse von statistischen Modellen erweitern.

4.4. Definition der Suffizienz im diskreten Fall

Beispiel 4.4.1. Betrachten wir eine unfaire Münze, wobei die Wahrscheinlichkeit θ, dass die Münze Kopf zeigt, geschätzt werden soll. Dafür werde die Münze n mal geworfen. Falls die Münze beim i-ten Wurf Kopf zeigt, definieren wir $x_i = 1$, sonst sei $x_i = 0$. Die komplette Information über unser Zufallsexperiment ist somit in der Stichprobe $(x_1, \ldots, x_n)$ enthalten. Es erscheint aber intuitiv klar, dass für die Beantwortung der statistischen Fragen über θ nur die Information darüber, *wie oft* die Münze Kopf gezeigt hat (also die Zahl $x_1 + \ldots + x_n$) relevant ist. Hingegen ist die Information, *bei welchen Würfen* die Münze Kopf gezeigt hat, nicht nützlich. Deshalb nennt man in diesem Beispiel die Stichprobenfunktion $T(x_1, \ldots, x_n) = x_1 + \ldots + x_n$ eine suffiziente (d.h. ausreichende) Statistik. Anstatt das Experiment durch die ganze Stichprobe $(x_1, \ldots, x_n)$ zu beschreiben, können wir es lediglich durch den Wert von $x_1 + \ldots + x_n$ beschreiben, ohne dass dabei nützliche statistische Information verloren geht.
Ein guter Schätzer für θ muss eine Funktion von $x_1 + \ldots + x_n$ sein. Das garantiert nämlich, dass der Schätzer nur nützliche statistische Information verwendet und nicht durch die Verwendung von unnützlichem Zufallsrauschen die Varianz des Schätzers gesteigert wird.

Nun werden wir eine allgemeine Definition der Suffizienz geben. Sei $(\mathfrak{X}, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ ein statistisches Modell. In diesem Abschnitt betrachten wir nur den Fall eines endlichen oder abzählbar unendlichen Stichprobenraumes $\mathfrak{X}$.

Definition 4.4.2. Eine Funktion $T : \mathfrak{X} \to \mathbb{R}^r$ heißt eine *suffiziente Statistik*, wenn für alle $x \in \mathfrak{X}$ und für alle $t \in \mathbb{R}^r$ die Funktion

$$\theta \mapsto \mathbb{P}_\theta[X = x | T(X) = t]$$

konstant ist. D.h. es soll gelten, dass

$$\mathbb{P}_{\theta_1}[X = x | T(X) = t] = \mathbb{P}_{\theta_2}[X = x | T(X) = t]$$

für alle $t \in \mathbb{R}^r$ und alle $\theta_1, \theta_2 \in \Theta$ mit $\mathbb{P}_{\theta_1}[T(X) = t] \neq 0$, $\mathbb{P}_{\theta_2}[T(X) = t] \neq 0$.

Man kann die obige Definition auch wie folgt formulieren: T ist suffizient, wenn die bedingte Verteilung von X gegeben, dass $T(X) = t$ nicht von θ abhängt.

Beispiel 4.4.3 (Fortsetzung von Beispiel 4.4.1)**.** Seien $X_1, \ldots, X_n \sim \text{Bern}(\theta)$ unabhängige Zufallsvariablen, wobei $\theta \in (0, 1)$ zu schätzen sei. Wir behaupten, dass $T(x_1, \ldots, x_n) = x_1 + \ldots + x_n$ eine suffiziente Statistik ist.

Beweis. Sei $t \in \{0, \ldots, n\}$, denn für alle anderen Werte von t ist das Ereignis $T(X) = t$ unmöglich. Betrachte für $(x_1, \ldots, x_n) \in \{0,1\}^n$ den Ausdruck

$$\begin{aligned} P(\theta) &:= \mathbb{P}_\theta[X_1 = x_1, \ldots, X_n = x_n | X_1 + \ldots + X_n = t] \\ &= \frac{\mathbb{P}_\theta[X_1 = x_1, \ldots, X_n = x_n, X_1 + \ldots + X_n = t]}{\mathbb{P}_\theta[X_1 + \ldots + X_n = t]}. \end{aligned}$$

Fall 1. Ist $x_1 + \ldots + x_n \neq t$, so gilt $P(\theta) = 0$. In diesem Fall hängt $P(\theta)$ von θ nicht ab.

Fall 2. Sei nun $x_1 + \ldots + x_n = t$. Dann gilt

$$P(\theta) = \frac{\mathbb{P}_\theta[X_1 = x_1, \ldots, X_n = x_n, X_1 + \ldots + X_n = t]}{\mathbb{P}_\theta[X_1 + \ldots + X_n = t]} = \frac{\mathbb{P}_\theta[X_1 = x_1, \ldots, X_n = x_n]}{\mathbb{P}_\theta[X_1 + \ldots + X_n = t]}.$$

Indem wir nun benutzen, dass $X_1, \ldots, X_n$ unabhängig sind und $X_1 + \ldots + X_n \sim \mathrm{Bin}(n, \theta)$ ist, erhalten wir, dass

$$P(\theta) = \frac{\theta^{x_1}(1-\theta)^{1-x_1} \cdot \ldots \cdot \theta^{x_n}(1-\theta)^{1-x_n}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{1}{\binom{n}{t}}.$$

Dieser Ausdruck ist ebenfalls von θ unabhängig. □

Bemerkung 4.4.4. Wir haben gezeigt, dass für alle $t \in \{0, \ldots, n\}$

$$\mathbb{P}_\theta[X_1 = x_1, \ldots, X_n = x_n | X_1 + \ldots + X_n = t] = \frac{\mathbb{1}_{x_1 + \ldots + x_n = t}}{\binom{n}{t}}, \quad (x_1, \ldots, x_n) \in \{0,1\}^n.$$

Somit ist die bedingte Verteilung von $(X_1, \ldots, X_n)$ gegeben, dass $X_1 + \ldots + X_n = t$, eine Gleichverteilung auf der Menge

$$\{(x_1, \ldots, x_n) \in \{0,1\}^n : x_1 + \ldots + x_n = t\}.$$

Diese Menge besteht aus $\binom{n}{t}$ Elementen. Die bedingte Verteilung hängt nicht von θ ab (Suffizienz).

Aufgabe 4.4.5. Seien $X_1, \ldots, X_n$ unabhängige und

(a) mit Parameter $\theta > 0$ Poisson-verteilte Zufallsvariablen;
(b) mit Parameter $\theta \in (0,1)$ geometrisch-verteilte Zufallsvariablen.

Zeigen Sie, dass $T(X_1, \ldots, X_n) = X_1 + \ldots + X_n$ eine suffiziente Statistik ist. Bestimmen Sie für $t \in \mathbb{N}_0$ die bedingte Verteilung von $(X_1, \ldots, X_n)$ gegeben, dass $X_1 + \ldots + X_n = t$.

Aufgabe 4.4.6. Seien $X_1, \ldots, X_n$ unabhängige, auf der endlichen Menge $\{1, \ldots, \theta\}$ gleichverteilte Zufallsvariablen, wobei $\theta \in \mathbb{N}$ ein Parameter sei. Zeigen Sie, dass $T(X_1, \ldots, X_n) = \max\{X_1, \ldots, X_n\}$ eine suffiziente Statistik ist und bestimmen Sie für $t \in \mathbb{N}$ die bedingte Verteilung von $(X_1, \ldots, X_n)$ gegeben, dass $\max\{X_1, \ldots, X_n\} = t$.

Im obigen Beispiel haben wir gezeigt, dass $X_1 + \ldots + X_n$ eine suffiziente Statistik ist. Ist dann z.B. auch $\bar{X}_n = \frac{X_1 + \ldots + X_n}{n}$ eine suffiziente Statistik? Im folgenden Lemma zeigen wir, dass die Antwort positiv ist.

Lemma 4.4.7. Sei $T : \mathfrak{X} \to \mathbb{R}^r$ eine suffiziente Statistik und sei $g : \operatorname{Im} T \to \mathbb{R}^k$ eine *injektive* Funktion. Dann ist auch

$$g \circ T : \mathfrak{X} \to \mathbb{R}^k, \quad x \mapsto g(T(x))$$

eine suffiziente Statistik.

Beweis. Seien $t \in \mathbb{R}^k$ und $\theta_1, \theta_2 \in \Theta$ mit $\mathbb{P}_{\theta_i}[g(T(X)) = t] \neq 0$, $i = 1, 2$. Wegen der Suffizienz von T ist

$$P(\theta_i) := \mathbb{P}_{\theta_i}[X = x | g(T(X)) = t] = \mathbb{P}_{\theta_i}[X = x | T(X) = g^{-1}(t)]$$

unabhängig von der Wahl von $i = 1, 2$. Dabei ist das Urbild $g^{-1}(t)$ wohldefiniert, da g injektiv ist. □

4.5. Faktorisierungssatz von Neyman-Fisher

In diesem Abschnitt beweisen wir den Faktorisierungssatz von Neyman-Fisher. Dieser Satz bietet eine einfache Methode zur Überprüfung der Suffizienz. Sei $(\mathfrak{X}, \mathcal{A}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ ein statistisches Modell mit einer höchstens abzählbaren Stichprobenmenge $\mathfrak{X}$. Sei $T : \mathfrak{X} \to \mathbb{R}^r$ eine Statistik. Im nächsten Lemma benutzen wir die folgende Notation:

- $L(x; \theta) = \mathbb{P}_\theta[X = x]$ ist die Likelihood-Funktion.
- $q(t; \theta) = \mathbb{P}_\theta[T(X) = t]$, wobei $t \in \mathbb{R}^r$, ist die Zähldichte von $T(X)$ unter $\mathbb{P}_\theta$.

Lemma 4.5.1. Eine Funktion $T : \mathfrak{X} \to \mathbb{R}^r$ ist genau dann eine suffiziente Statistik, wenn für alle $x \in \mathfrak{X}$ die Funktion

$$\theta \mapsto \frac{L(x; \theta)}{q(T(x); \theta)} \tag{4.5.1}$$

konstant ist. (Der Definitionsbereich besteht aus allen $\theta \in \Theta$ mit $q(T(x); \theta) \neq 0$).

Beweis. Betrachte den Ausdruck

$$P(\theta) := \mathbb{P}_\theta[X = x | T(X) = t].$$

Im Falle $t \neq T(x)$ ist $P(\theta) = 0$, was unabhängig von θ ist. Sei deshalb $t = T(x)$. Dann gilt

$$P(\theta) = \frac{\mathbb{P}_\theta[X = x, T(X) = t]}{\mathbb{P}_\theta[T(X) = T(x)]} = \frac{\mathbb{P}_\theta[X = x]}{\mathbb{P}_\theta[T(X) = T(x)]} = \frac{L(x; \theta)}{q(T(x); \theta)}.$$

Somit ist T eine suffiziente Statistik genau dann, wenn (4.5.1) nicht von θ abhängt.

Satz 4.5.2 (Faktorisierungssatz von Neyman-Fisher, diskreter Fall)**.** Eine Funktion $T : \mathfrak{X} \to \mathbb{R}^r$ ist eine suffiziente Statistik genau dann, wenn es Funktionen $g : \mathbb{R}^r \times \Theta \to \mathbb{R}$ und $h : \mathfrak{X} \to \mathbb{R}$ gibt, so dass die folgende Faktorisierung gilt:

$$L(x; \theta) = g(T(x); \theta) \cdot h(x) \text{ für alle } x \in \mathfrak{X}, \theta \in \Theta. \tag{4.5.2}$$

Beweis von „$\Longrightarrow$“. Sei T eine suffiziente Statistik. Definiere die Funktion

$$h(x) := \frac{L(x;\theta)}{q(T(x);\theta)}, \quad x \in \mathfrak{X}.$$

Dabei können wir auf der rechten Seite ein beliebiges θ mit $q(T(x);\theta) \neq 0$ einsetzen, denn nach Lemma 4.5.1 ist der Term unabhängig von θ. Gibt es kein θ mit $q(T(x);\theta) \neq 0$, so definieren wir einfach $h(x) = 0$.
Mit diesem h und $g(t;\theta) = q(t;\theta)$ gilt die Faktorisierung (4.5.2). □
Beweis von „$\Longleftarrow$“. Es gelte die Faktorisierung (4.5.2). Sei $x \in \mathfrak{X}$ fest. Es bezeichne

$$A := \{y \in \mathfrak{X} \colon T(y) = T(x)\}$$

die Niveaumenge von T, die x enthält. Dann gilt für alle $\theta \in \Theta$ mit $q(T(x);\theta) \neq 0$, dass

$$\frac{L(x;\theta)}{q(T(x);\theta)} = \frac{g(T(x);\theta)h(x)}{\sum_{y\in A} L(y;\theta)} = \frac{g(T(x);\theta)h(x)}{\sum_{y\in A} g(T(y);\theta)h(y)} = \frac{h(x)}{\sum_{y\in A} h(y)}.$$

Dieser Ausdruck hängt nicht von θ ab. Nach Lemma 4.5.1 ist T suffizient. □

Beispiel 4.5.3. Seien $X_1, \ldots, X_n \sim \mathrm{Bern}(\theta)$ unabhängig, wobei $\theta \in (0,1)$. Für die Likelihood-Funktion gilt

$$\begin{aligned} L(x_1,\ldots,x_n;\theta) &= \mathbb{P}_\theta[X_1 = x_1, \ldots, X_n = x_n] \\ &= \theta^{x_1}(1-\theta)^{1-x_1}\mathbb{1}_{x_1\in\{0,1\}} \cdot \ldots \cdot \theta^{x_n}(1-\theta)^{1-x_n}\mathbb{1}_{x_n\in\{0,1\}} \\ &= \theta^{x_1+\ldots+x_n}(1-\theta)^{n-(x_1+\ldots+x_n)}\mathbb{1}_{x_1,\ldots,x_n\in\{0,1\}}. \end{aligned}$$

Daraus ist ersichtlich, dass die Neyman-Fisher-Faktorisierung (4.5.2) mit

$$T(x_1,\ldots,x_n) = x_1 + \ldots + x_n, \quad g(t;\theta) = \theta^t(1-\theta)^{n-t}, \quad h(x_1,\ldots,x_n) = \mathbb{1}_{x_1,\ldots,x_n\in\{0,1\}}$$

gilt. Nach dem Faktorisierungssatz von Neyman-Fisher ist T suffizient.

4.6. Definition der Suffizienz im absolut stetigen Fall

Bisher haben wir nur den Fall eines höchstens abzählbaren Stichprobenraums $\mathfrak{X}$ betrachtet. Sei nun $(\mathfrak{X}, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ ein statistisches Modell mit einem beliebigen Stichprobenraum $\mathfrak{X}$. Die Funktion $T : \mathfrak{X} \to \mathbb{R}^r$ sei Borel-messbar. Im diskreten Fall haben wir T suffizient genannt, wenn die bedingte Verteilung von X gegeben, dass $T(X) = t$ nicht von θ abhängt. Im Allgemeinen kann die Wahrscheinlichkeit des Ereignisses $T(X) = t$ gleich 0 sein und es ist nicht klar, wie man die bedingte Verteilung definiert. Dieses Problem hat eine Lösung, die im Abschnitt 4.11 besprochen wird.

Definition 4.6.1. Sei $(\mathfrak{X}, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ ein statistisches Modell. Eine Statistik $T : \mathfrak{X} \to \mathbb{R}^r$ heißt *suffizient*, falls es eine (von θ unabhängige!) Familie von Wahrscheinlichkeitsmaßen $\{\pi_t \colon t \in \mathbb{R}^r\}$ gibt mit

(1) Für jedes $t \in \mathbb{R}^r$ ist π_t ist ein Wahrscheinlichkeitsmaß auf der Niveaumenge

$$T^{-1}(t) = \{x \in \mathfrak{X} \colon T(x) = t\}.$$

(2) Für alle $\theta \in \Theta$ gilt „die Formel der totalen Wahrscheinlichkeit“

$$\mathbb{P}_\theta[A] = \int_{\mathbb{R}^r} \pi_t(A \cap T^{-1}(t))\, \mu_\theta(\mathrm{d}t), \quad A \in \mathcal{A},$$

wobei $\mu_\theta(B) = \mathbb{P}_\theta[T \in B]$, $B \subset \mathbb{R}^r$ Borel, die Verteilung von T unter $\mathbb{P}_\theta$ ist.

Bemerkung 4.6.2. Man kann sich π_t als die „bedingte Verteilung“ von X gegeben, dass $T(X) = t$, vorstellen. Entscheidend für die Suffizienz ist die Forderung, dass π_t keine Funktion von θ sein darf. Stellen wir uns vor, es wurde eine Stichprobe X gemäß $\mathbb{P}_\theta$ gezogen, uns wurde allerdings lediglich der Wert $T(X)$ mitgeteilt. Wir wissen also, dass X irgendwo in der Niveaumenge $T^{-1}(t)$ liegt. Die bedingte Verteilung von X ist π_t. Da aber diese Verteilung nicht mehr von θ abhängt, würde uns die Information über die genaue Position von X auf der Niveaumenge nichts nützen. Wir könnten aus dieser Information keine Rückschlüsse auf θ ziehen. Deshalb heißt T suffizient.

Beispiel 4.6.3. Im n-fachen Bernoulli-Model mit der suffizienten Statistik $T(x_1, \ldots, x_n) = x_1 + \ldots + x_n$ ist π_t die Gleichverteilung auf der Menge $\{(x_1, \ldots, x_n) \in \{0,1\}^n \colon x_1 + \ldots + x_n = t\}$, für alle $t \in \{0, \ldots, n\}$. Es ist egal, wie man π_t für andere Werte von t definiert, da diese eine Nullmenge bzgl. μ_θ bilden.

Bemerkung 4.6.4. Es lässt sich zeigen, dass auch die folgende „Formel der totalen Erwartung“ gilt:

$$\mathbb{E}_\theta[f(X)] = \int_{\mathbb{R}^r} \left(\int_{T^{-1}(t)} f d\pi_t \right) \mu_\theta(\mathrm{d}t),$$

für alle Funktionen $f : \mathfrak{X} \to \mathbb{R}$ mit $\mathbb{E}_\theta |f(X)| < \infty$.

Leider lassen die Bedingungen der obigen Definition nicht so einfach überprüfen. Zum Glück gibt es eine allgemeine Version des Satzes von Neyman-Fisher, die als eine alternative Definition der Suffizienz benutzt werden kann. Wir nehmen an, dass das statistische Modell $(\mathfrak{X}, \mathcal{A}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ dominiert ist, d.h. es gibt ein σ-endliches Maß λ auf $\mathfrak{X}$ (typischerweise das Lebesgue- oder das Zählmaß), so dass jedes $\mathbb{P}_\theta$ eine Dichte bezüglich λ besitzt. Diese Dichte wird mit $L(x; \theta)$ bezeichnet und heißt die Likelihood-Funktion.

Satz 4.6.5 (Faktorisierungssatz von Neyman-Fisher, allgemeiner Fall)**.** Eine Statistik $T : \mathfrak{X} \to \mathbb{R}^r$ ist suffizient genau dann, wenn es messbare Funktionen $g : \mathbb{R}^r \times \Theta \to \mathbb{R}$ und $h : \mathfrak{X} \to \mathbb{R}$ gibt mit

$$L(x; \theta) = g(T(x); \theta) \cdot h(x), \text{ für alle } x \in \mathfrak{X}, \theta \in \Theta.$$

Beispiel 4.6.6. Seien $X_1, \ldots, X_n$ unabhängige und auf dem Intervall $[0, \theta]$ gleichverteilte Zufallsvariablen, wobei $\theta > 0$ der unbekannte Parameter sei. Wir zeigen, dass $T(X_1, \ldots, X_n) = \max\{X_1, \ldots, X_n\}$ eine suffiziente Statistik ist.

Die Dichte von X_i gegeben durch

$$h_\theta(x_i) = \frac{1}{\theta} \mathbb{1}_{x_i \in [0,\theta]}.$$

Für die Likelihood-Funktion (also die Dichte von $(X_1, \ldots, X_n)$ bzgl. des Lebesgue-Maßes auf $\mathbb{R}^n$) gilt

$$\begin{aligned} L(x_1, \ldots, x_n; \theta) &= h_\theta(x_1) \ldots h_\theta(x_n) \\ &= \frac{1}{\theta^n} \mathbb{1}_{x_1 \in [0,\theta]} \cdot \ldots \cdot \mathbb{1}_{x_n \in [0,\theta]} \\ &= \frac{1}{\theta^n} \mathbb{1}_{\max(x_1,\ldots,x_n) \leq \theta} \cdot \mathbb{1}_{x_1,\ldots,x_n \geq 0} \\ &= g(T(x_1, \ldots, x_n); \theta) \cdot h(x_1, \ldots, x_n), \end{aligned}$$

wobei

$$g(t; \theta) = \frac{1}{\theta^n} \mathbb{1}_{t \leq \theta}, \quad h(x_1, \ldots, x_n) = \mathbb{1}_{x_1,\ldots,x_n \geq 0}.$$

Somit ist $T(X_1, \ldots, X_n) = \max\{X_1, \ldots, X_n\}$ eine suffiziente Statistik. Ein guter Schätzer für θ muss also eine Funktion von $\max\{X_1, \ldots, X_n\}$ sein. Insbesondere ist der Schätzer $2\bar{X}_n$ in diesem Sinne nicht gut, denn er benutzt überflüssige Information. Diese überflüssige Information steigert die Varianz des Schätzers. Das ist der Grund dafür, dass der Schätzer $\frac{n+1}{n} \max\{X_1, \ldots, X_n\}$ (der suffizient und erwartungstreu ist) eine kleinere Varianz als der Schätzer $2\bar{X}_n$ (der nur erwartungstreu ist) hat.

Beispiel 4.6.7. Seien $X_1, \ldots, X_n$ unabhängige und mit Parameter $\theta > 0$ exponentialverteilte Zufallsvariablen. Somit ist die Dichte von X_i gegeben durch

$$h_\theta(x_i) = \theta \exp(-\theta x_i) \mathbb{1}_{x_i \geq 0}.$$

Wir zeigen, dass $T(x_1, \ldots, x_n) = x_1 + \ldots + x_n$ eine suffiziente Statistik ist. Für die Likelihood-Funktion gilt

$$\begin{aligned} L(x_1, \ldots, x_n; \theta) &= h_\theta(x_1) \ldots h_\theta(x_n) \\ &= \theta^n \exp(-\theta(x_1 + \ldots + x_n)) \mathbb{1}_{x_1,\ldots,x_n \geq 0} \\ &= g(T(x_1, \ldots, x_n); \theta) \cdot h(x_1, \ldots, x_n), \end{aligned}$$

wobei

$$g(t; \theta) = \theta^n \exp(-\theta t), \quad h(x_1, \ldots, x_n) = \mathbb{1}_{x_1,\ldots,x_n \geq 0}.$$

Ein guter Schätzer für θ muss also eine Funktion von $X_1 + \ldots + X_n$ sein.

Beispiel 4.6.8. Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit $X_i \sim N(\mu, \sigma^2)$. Der unbekannte Parameter ist $\theta = (\mu, \sigma^2)$, wobei $\mu \in \mathbb{R}$ und $\sigma^2 > 0$. Die Aufgabe besteht nun darin, eine suffiziente Statistik zu finden. Da wir normalverteilte Zufallsvariablen betrachten, gilt für die Dichte

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

Somit ist die Likelihood-Funktion gegeben durch

$$\begin{aligned} L(x_1,\dots,x_n;\mu,\sigma^2) &= h_{\mu,\sigma^2}(x_1)\dots h_{\mu,\sigma^2}(x_n) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\right) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^n x_i^2 - 2\mu\sum_{i=1}^n x_i + n\mu^2\right]\right). \end{aligned}$$

Nun betrachten wir die Statistik $T:\mathbb{R}^n\to\mathbb{R}^2$ mit

$$(x_1,\dots,x_n)\mapsto\left(\sum_{i=1}^n x_i^2,\sum_{i=1}^n x_i\right)=(T_1,T_2).$$

Diese Statistik T ist suffizient, denn wir haben die Neyman-Fisher-Faktorisierung

$$L(x_1,\dots,x_n;\mu,\sigma^2)=g(T_1,T_2;\mu,\sigma^2)h(x_1,\dots,x_n)$$

mit $h(x_1,\dots,x_n)=1$ und

$$g(T_1,T_2;\mu,\sigma^2)=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n\exp\left(-\frac{T_1-2\mu T_2+n\mu^2}{2\sigma^2}\right).$$

Allerdings ist T_1 oder T_2 allein betrachtet nicht suffizient.

Bemerkung 4.6.9. Im obigen Beispiel ist die Statistik $(\bar{x}_n, s_n^2)$ mit

$$\bar{x}_n=\frac{x_1+\ldots+x_n}{n}\text{ und }s_n^2=\frac{1}{n-1}\left(\sum_{i=1}^n x_i^2-n\bar{x}_n^2\right)$$

ebenfalls suffizient, denn

$$T_1=n\bar{x}_n\text{ und }T_2=(n-1)s_n^2+n\bar{x}_n^2.$$

Wir können also $g(T_1,T_2;\mu,\sigma^2)$ auch als eine Funktion von $\bar{x}_n, s_n^2$ und μ,σ^2 schreiben.

Beispiel 4.6.10. Es seien $X_1,\dots,X_n$ unabhängige identisch verteilte Zufallsvariablen mit Dichte h_θ, wobei θ unbekannt sei. Dann ist die Statistik

$$T:(x_1,\dots,x_n)\mapsto(x_{(1)},\dots,x_{(n)})$$

suffizient. Das heißt, die Angabe der Werte der Stichprobe ohne die Angabe der Reihenfolge, in der diese Werte beobachtet wurden, ist suffizient. In der Tat, für die Likelihood-Funktion gilt

$$L(x_1,\dots,x_n;\theta)=h_\theta(x_1)\dots h_\theta(x_n).$$

Diese Funktion ändert sich bei Permutationen von $x_1,\dots,x_n$ nicht und kann somit als eine Funktion von $(x_{(1)},\dots,x_{(n)})$ und θ dargestellt werden. Somit haben wir eine Neyman-Fisher-Faktorisierung angegeben.

Aufgabe 4.6.11. Seien $X_1,\dots,X_n$ unabhängige Zufallsvariablen, die auf einem Intervall (θ_1,θ_2) gleichverteilt sind. Dabei seien $\theta_1\in\mathbb{R}$ und $\theta_2\in\mathbb{R}$ die zu schätzenden Parameter mit $\theta_1<\theta_2$. Zeigen Sie die Suffizienz der Statistik $T(x_1,\dots,x_n)=(x_{(1)},x_{(n)})$.

Aufgabe 4.6.12. Seien $X_1, \ldots, X_n$ unabhängig und identisch verschoben exponentialverteilt, d.h. verteilt gemäß der Dichte

$$h_{m,\theta}(x) = \frac{1}{\theta} \mathrm{e}^{-\theta(x-\mu)} \mathbb{1}_{[m,\infty)}(x),$$

wobei $m \in \mathbb{R}$, $\theta > 0$ Parameter seien. Zeigen Sie die Suffizienz der Statistiken

$$T(x_1, \ldots, x_n) = \left(x_{(1)}, \sum_{i=1}^{n} x_i \right) \quad \text{und} \quad T'(x_1, \ldots, x_n) = \left(x_{(1)}, \frac{1}{n} \sum_{i=1}^{n} (x_i - x_{(1)}) \right).$$

Aufgabe 4.6.13. Es seien $X_1 \sim \mathrm{N}(\mu, \sigma_1^2), \ldots, X_n \sim \mathrm{N}(\mu, \sigma_n^2)$ unabhängige Zufallsvariablen, wobei $\sigma_1^2, \ldots, \sigma_n^2 > 0$ bekannt seien und μ der unbekannte Parameter sei. Zeigen Sie, dass die Statistik $T(x_1, \ldots, x_n) := x_1/\sigma_1^2 + \ldots + x_n/\sigma_n^2$ suffizient ist.

4.7. Vollständigkeit

Sei $(\mathbb{P}_\theta)_{\theta \in \Theta}$ eine Familie von Wahrscheinlichkeitsmaßen auf dem Stichprobenraum $(\mathfrak{X}, \mathcal{A})$.

Definition 4.7.1. Eine Stichprobenfunktion $\varphi : \mathfrak{X} \to \mathbb{R}$ heißt *erwartungstreuer Schätzer von* 0, falls

$$\mathbb{E}_\theta \varphi = 0 \text{ für alle } \theta \in \Theta.$$

Beispiel 4.7.2. Sine $\hat{\theta}_1$ und $\hat{\theta}_2$ beide erwartungstreue Schätzer von θ, so ist ihre Differenz $\hat{\theta}_1 - \hat{\theta}_2$ erwartungstreuer Schätzer von 0.

Definition 4.7.3. Eine Statistik $T : \mathfrak{X} \to \mathbb{R}^r$ heißt *vollständig*, falls für alle Borel-Funktionen $g : \mathbb{R}^r \to \mathbb{R}$ aus der Gültigkeit von

$$\mathbb{E}_\theta g(T) = 0 \text{ für alle } \theta \in \Theta$$

folgt, dass $g(T) = 0$ fast sicher bezüglich $\mathbb{P}_\theta$ für alle $\theta \in \Theta$.

Mit anderen Worten: Es gibt keinen nichttrivialen erwartungstreuen Schätzer von 0, der nur auf dem Wert der Statistik T basiert.

Beispiel 4.7.4. Seien $X_1, \ldots, X_n$ unabhängige und mit Parameter $\theta \in (0, 1)$ Bernoulliverteilte Zufallsvariablen. In diesem Fall ist die Statistik

$$T : (X_1, \ldots, X_n) \to (X_1, \ldots, X_n)$$

nicht vollständig für $n \geq 2$. Um die Unvollständigkeit zu zeigen, betrachten wir die Funktion $g(X_1, \ldots, X_n) = X_2 - X_1$. Dann gilt für den Erwartungswert

$$\mathbb{E}_\theta g(T(X_1, \ldots, X_n)) = \mathbb{E}_\theta g(X_1, \ldots, X_n) = \mathbb{E}_\theta [X_2 - X_1] = 0,$$

denn X_2 hat die gleiche Verteilung wie X_1. Dabei ist $X_2 - X_1 \neq 0$ fast sicher, also ist die Bedingung aus der Definition der Vollständigkeit nicht erfüllt.

Beispiel 4.7.5. Seien $X_1, \ldots, X_n$ unabhängige und mit Parameter $\theta \in (0,1)$ Bernoulli-verteilte Zufallsvariablen. Dann ist die Statistik

$$T(X_1, \ldots, X_n) = X_1 + \ldots + X_n$$

vollständig.

Beweis. Sei $g : \mathbb{R} \to \mathbb{R}$ eine Funktion mit $\mathbb{E}_\theta g(X_1 + \ldots + X_n) = 0$ für alle $\theta \in (0,1)$. Somit gilt

$$0 = \sum_{i=0}^{n} g(i) \binom{n}{i} \theta^i (1-\theta)^{n-i} = (1-\theta)^n \sum_{i=0}^{n} g(i) \binom{n}{i} \left(\frac{\theta}{1-\theta}\right)^i.$$

Betrachte die Variable $z := \frac{\theta}{1-\theta}$. Nimmt θ alle möglichen Werte im Intervall $(0,1)$ an, so nimmt z alle möglichen Werte im Intervall $(0,\infty)$ an. Es folgt, dass

$$\sum_{i=0}^{n} g(i) \binom{n}{i} z^i = 0 \text{ für alle } z > 0.$$

Also gilt für alle $i = 0, \ldots, n$, dass $g(i)\binom{n}{i} = 0$ und somit auch $g(i) = 0$. Hieraus folgt, dass $g = 0$ und die Vollständigkeit ist bewiesen. □

Aufgabe 4.7.6. Seien $X_1, \ldots, X_n \sim \text{Poi}(\theta)$, $\theta > 0$, unabhängig. Zeigen Sie die Vollständigkeit der Statistik $X_1 + \ldots + X_n$.

Aufgabe 4.7.7. Seien $X_1, \ldots, X_n \sim \text{Bin}(\theta)$ unabhängige Zufallsvariablen, wobei über den Parameter θ zusätzlich bekannt sei, dass $\theta \in \Theta \subset (0,1)$ mit $|\Theta| \geq n+1$. Zeigen Sie, dass die Statistik $X_1 + \ldots + X_n$ vollständig ist.

Beispiel 4.7.8. Seien $X_1, \ldots, X_n$ unabhängige und auf $[0,\theta]$ gleichverteilte Zufallsvariablen, wobei $\theta > 0$. Dann ist die Statistik $T(X_1, \ldots, X_n) = \max\{X_1, \ldots, X_n\}$ vollständig.

Beweis. Die Verteilungsfunktion von T unter $\mathbb{P}_\theta$ ist gegeben durch

$$\mathbb{P}_\theta[T \leq x] = \begin{cases} 0, & x \leq 0, \\ (\frac{x}{\theta})^n, & 0 \leq x \leq \theta, \\ 1, & x \geq \theta. \end{cases}$$

Die Dichte von T unter $\mathbb{P}_\theta$ erhält man indem man die Verteilungsfunktion ableitet:

$$q(x;\theta) = n x^{n-1} \theta^{-n} \mathbb{1}_{0 \leq x \leq \theta}.$$

Sei nun $g : \mathbb{R} \to \mathbb{R}$ eine Borel-Funktion mit $\mathbb{E}_\theta g(T(X_1, \ldots, X_n)) = 0$ für alle $\theta > 0$. Das heißt, es gilt

$$\theta^{-n} \int_0^\theta n x^{n-1} g(x) \mathrm{d}x = 0 \text{ für alle } \theta > 0.$$

Wir können durch θ^{-n} teilen:

$$\int_0^\theta n x^{n-1} g(x) \mathrm{d}x = 0 \text{ für alle } \theta > 0.$$

Nun können wir nach θ ableiten: $n\theta^{n-1} g(\theta) = 0$ und somit $g(\theta) = 0$ für Lebesgue-fast alle $\theta > 0$. Somit ist $g(x) = 0$ fast sicher bzgl. der Gleichverteilung auf $[0,\theta]$ für alle $\theta > 0$. Es sei

bemerkt, dass g auf der negativen Halbachse durchaus ungleich 0 sein kann, allerdings hat die negative Halbachse Wahrscheinlichkeit 0 bzgl. der Gleichverteilung auf $[0, \theta]$. □

Aufgabe 4.7.9. Seien $X_1, \ldots, X_n \sim \mathrm{U}[\theta_1, \theta_2]$ unabhängig, wobei $\theta_1 < \theta_2$ unbekannt seien. Zeigen Sie, dass die Statistik $(X_{(1)}, X_{(n)})$ vollständig ist.

Aufgabe 4.7.10. Seien $X_1, \ldots, X_n$ unabhängige, auf dem Intervall $[\theta, 2\theta]$ gleichverteilte Zufallsvariablen. Dabei sei $\theta > 0$ ein unbekannter Parameter. Zeigen Sie, dass die Statistik $(X_{(1)}, X_{(n)})$ suffizient aber *nicht* vollständig ist.

4.8. Eine Charakterisierung des besten erwartungstreuen Schätzers

Sei $(\mathbb{P}_\theta)_{\theta\in\Theta}$ eine Familie von Wahrscheinlichkeitsmaßen auf dem Stichprobenraum $(\mathfrak{X}, \mathcal{A})$, wobei $\Theta \subset \mathbb{R}$. [1]

Satz 4.8.1. Sei $\hat\theta : \mathfrak{X} \to \mathbb{R}$ ein erwartungstreuer Schätzer für θ. Dann sind die folgenden Bedingungen äquivalent:

(1) $\hat\theta$ ist der beste erwartungstreue Schätzer für θ.
(2) Für jeden (quadratisch integrierbaren) erwartungstreuen Schätzer φ für 0 gilt, dass $\mathrm{Cov}_\theta(\hat\theta, \varphi) = 0$ für alle $\theta \in \Theta$.

Also ist ein erwartungstreuer Schätzer genau dann der beste erwartungstreue Schätzer, wenn er zu jedem erwartungstreuen Schätzer für 0 orthogonal ist.

Beweis von „$\Longrightarrow$“. Sei $\hat\theta$ der beste erwartungstreue Schätzer für θ und sei $\varphi : \mathbb{R}^n \to \Theta$ eine Stichprobenfunktion mit $\mathbb{E}_\theta\varphi = 0$ für alle $\theta \in \Theta$. Somit müssen wir zeigen, dass

$$\mathrm{Cov}_\theta(\hat\theta, \varphi) = 0 \text{ für alle } \theta \in \Theta.$$

Definieren wir uns hierfür $\tilde\theta = \hat\theta + a\varphi$, $a \in \mathbb{R}$. Dann ist $\tilde\theta$ ebenfalls ein erwartungstreuer Schätzer für θ, denn

$$\mathbb{E}_\theta\tilde\theta = \mathbb{E}_\theta\hat\theta + a \cdot \mathbb{E}_\theta\varphi = \theta.$$

Es gilt für die Varianz von $\tilde\theta$, dass

$$\mathrm{Var}_\theta\,\tilde\theta = \mathrm{Var}_\theta\,\hat\theta + a^2\,\mathrm{Var}_\theta\,\varphi + 2a\,\mathrm{Cov}_\theta(\hat\theta, \varphi) = \mathrm{Var}_\theta\,\hat\theta + g(a),$$

wobei $g(a) = a^2\,\mathrm{Var}_\theta\,\varphi + 2a\,\mathrm{Cov}_\theta(\hat\theta, \varphi)$. Wäre nun $\mathrm{Cov}_\theta(\hat\theta, \varphi) \neq 0$, dann hätte die quadratische Funktion g zwei verschiedene Nullstellen bei 0 und $-2\,\mathrm{Cov}_\theta(\hat\theta, \varphi)/\,\mathrm{Var}_\theta\,\varphi$. (Wir dürfen hier annehmen, dass $\mathrm{Var}_\theta\,\varphi \neq 0$, denn andernfalls wäre φ fast sicher konstant unter $\mathbb{P}_\theta$ und dann würde $\mathrm{Cov}_\theta(\hat\theta, \varphi) = 0$ trivialerweise gelten). Zwischen diesen Nullstellen gäbe es ein $a \in \mathbb{R}$ mit $g(a) < 0$ und es würde folgen, dass $\mathrm{Var}_\theta\,\tilde\theta < \mathrm{Var}_\theta\,\hat\theta$. Das widerspricht aber der Annahme, dass $\hat\theta$ der beste erwartungstreue Schätzer für θ ist. Somit muss $\mathrm{Cov}_\theta(\hat\theta, \varphi) = 0$ gelten. □

Beweis von „$\Longleftarrow$“. Sei $\hat\theta$ ein erwartungstreuer Schätzer für θ. Sei außerdem $\mathrm{Cov}_\theta(\varphi, \hat\theta) = 0$ für alle erwartungstreuen Schätzer φ für 0. Jetzt werden wir zeigen, dass $\hat\theta$ der beste

[1] Die Ergebnisse dieses Abschnitts werden im Folgenden nicht verwendet.

erwartungstreue Schätzer ist. Mit $\tilde{\theta}$ bezeichnen wir einen anderen erwartungstreuen Schätzer für θ. Somit genügt es zu zeigen, dass

$$\operatorname{Var}_\theta \hat{\theta} \leq \operatorname{Var}_\theta \tilde{\theta}.$$

Um das zu zeigen, schreiben wir $\tilde{\theta} = \hat{\theta} + (\tilde{\theta} - \hat{\theta}) =: \hat{\theta} + \varphi$. Da $\hat{\theta}$ und $\tilde{\theta}$ beide erwartungstreue Schätzer für θ sind, ist $\varphi := (\tilde{\theta} - \hat{\theta})$ ein erwartungstreuer Schätzer für 0. Für die Varianzen von $\tilde{\theta}$ und $\hat{\theta}$ gilt:

$$\operatorname{Var}_\theta \tilde{\theta} = \operatorname{Var}_\theta \hat{\theta} + \operatorname{Var}_\theta \varphi + 2 \operatorname{Cov}_\theta(\hat{\theta}, \varphi) = \operatorname{Var}_\theta \hat{\theta} + \operatorname{Var}_\theta \varphi \geq \operatorname{Var}_\theta \hat{\theta}.$$

Die letzte Ungleichung gilt, da $\operatorname{Var}_\theta \varphi \geq 0$. Somit ist $\hat{\theta}$ der beste erwartungstreue Schätzer. □

Aufgabe 4.8.2. Seien $\hat{\nu}_1, \ldots, \hat{\nu}_k$ beste erwartungstreue Schätzer für die Funktionen $\nu_1(\theta), \ldots, \nu_k(\theta)$. Zeigen Sie, dass für beliebige Konstanten $c_1, \ldots, c_k \in \mathbb{R}$ der beste erwartungstreue Schätzer für $c_1\nu_1(\theta) + \ldots + c_k\nu_k(\theta)$ durch $c_1\hat{\nu}_1 + \ldots + c_k\hat{\nu}_k$ gegeben ist.

4.9. Exponentialfamilien

In diesem Abschnitt führen wir den Begriff der Exponentialfamilie ein. Dieser Begriff ist aus mindestens zwei Gründen sehr nützlich. Auf der einen Seite, lässt sich für eine Exponentialfamilie sehr schnell eine suffiziente und vollständige Statistik (und somit, wie wir später sehen werden, der beste erwartungstreue Schätzer) konstruieren. Auf der anderen Seite, sind praktisch alle Verteilungsfamilien, die wir bisher betrachtet haben, Exponentialfamilien.
Sei $\{h_\theta(x) : \theta \in \Theta\}$ eine Familie von Dichten bzw. Zähldichten.

Definition 4.9.1. Die Familie $\{h_\theta(x) : \theta \in \Theta\}$ heißt *Exponentialfamilie*, falls es Funktionen $a(\theta)$, $b(x)$, $c(\theta)$, $d(x)$ gibt mit

$$h_\theta(x) = a(\theta)b(x)\mathrm{e}^{c(\theta)d(x)}.$$

Beispiel 4.9.2. Betrachten wir die Familie der Binomialverteilungen mit Parametern n (bekannt) und $\theta \in (0, 1)$ (unbekannt). Für $x \in \{0, \ldots, n\}$ ist die Zähldichte gegeben durch

$$h_\theta(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x} = (1-\theta)^n\binom{n}{x}\left(\frac{\theta}{1-\theta}\right)^x = (1-\theta)^n\binom{n}{x}\exp\left(\log\left(\frac{\theta}{1-\theta}\right)x\right).$$

Somit haben wir die Darstellung $h_\theta(x) = a(\theta)b(x)\mathrm{e}^{c(\theta)d(x)}$ mit

$$a(\theta) = (1-\theta)^n, \quad b(x) = \binom{n}{x}, \quad c(\theta) = \log\left(\frac{\theta}{1-\theta}\right), \quad d(x) = x.$$

Beispiel 4.9.3. Für die Normalverteilung mit Parametern $\mu \in \mathbb{R}$ und $\sigma^2 > 0$ ist die Dichte gegeben durch:

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

Unbekanntes μ, bekanntes σ^2. Betrachten wir den Parameter μ als unbekannt und σ^2 als gegeben und konstant, so gilt die Darstellung $h_{\mu,\sigma^2}(x) = a(\mu)b(x)\mathrm{e}^{c(\mu)d(x)}$ mit

$$a(\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right), \quad b(x) = \exp\left(-\frac{x^2}{2\sigma^2}\right), \quad c(\mu) = \frac{\mu}{\sigma^2}, \quad d(x) = x.$$

Bekanntes μ, unbekanntes σ^2. Betrachten wir μ als gegeben und konstant und σ^2 als unbekannt, so gilt die Darstellung $h_{\mu,\sigma^2}(x) = a(\sigma^2)b(x)\mathrm{e}^{c(\sigma^2)d(x)}$ mit

$$a(\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right), \quad b(x) = 1, \quad c(\sigma^2) = \frac{1}{2\sigma^2}, \quad d(x) = 2x\mu - x^2.$$

Aufgabe 4.9.4. Zeigen Sie, dass folgende Familien von Verteilungen Exponentialfamilien sind:

(1) $\{\mathrm{Exp}(\theta)\colon \theta > 0\}$.
(2) $\{\mathrm{Poi}(\theta)\colon \theta > 0\}$.

Kein Beispiel hingegen ist die Familie der Gleichverteilungen auf $[0, \theta]$. Das liegt daran, dass der Träger der Gleichverteilung von θ abhängt.

Leider bildet die Familie der Normalverteilungen, wenn man sowohl μ als auch σ^2 als unbekannt betrachtet, keine Exponentialfamilie im Sinne der obigen Definition. Deshalb werden wir die obige Definition etwas erweitern.

Definition 4.9.5. Eine Familie $\{h_\theta : \theta \in \Theta\}$ von Dichten oder Zähldichten heißt eine *m-parametrige Exponentialfamilie*, falls es eine Darstellung der Form

$$h_\theta(x) = a(\theta)b(x)\mathrm{e}^{c_1(\theta)d_1(x)+\ldots+c_m(\theta)d_m(x)}$$

gibt.

Beispiel 4.9.6. Die Familie der Normalverteilungen mit Parametern $\mu \in \mathbb{R}$ und $\sigma^2 > 0$ (die beide als unbekannt betrachtet werden) ist eine 2-parametrige Exponentialfamilie, denn

$$h_{\mu,\sigma^2}(x) = a(\mu, \sigma^2)b(x)\mathrm{e}^{c_1(\mu,\sigma^2)d_1(x)+c_2(\mu,\sigma^2)d_2(x)}$$

mit

$$a(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\mu^2}{\sigma^2}\right), \quad b(x) = 1,$$

$$c_1(\mu, \sigma^2) = -\frac{1}{2\sigma^2}, \quad d_1(x) = x^2, \quad c_2(\mu, \sigma^2) = \frac{\mu}{\sigma^2}, \quad d_2(x) = x.$$

Weitere Beispiele zwei-parametriger Exponentialfamilien sind die Familie der Gammaverteilungen und die Familie der Betaverteilungen, die später eingeführt werden.

4.10. Vollständige und suffiziente Statistik für Exponentialfamilien

Für eine Exponentialfamilie lässt sich sehr leicht eine suffiziente und vollständige Statistik angeben. Nämlich ist die Statistik $(T_1, \dots, T_m)$ mit

$$T_1(X_1, \dots, X_n) = \sum_{j=1}^{n} d_1(X_j), \quad \dots, \quad T_m(X_1, \dots, X_n) = \sum_{j=1}^{n} d_m(X_j)$$

suffizient. Um dies zu zeigen, schreiben wir die Likelihood-Funktion wie folgt um:

$$\begin{aligned} L(x_1, \dots, x_n; \theta) &= h_\theta(x_1) \dots h_\theta(x_n) \\ &= (a(\theta))^n b(x_1) \dots b(x_n) \exp\left(\sum_{i=1}^{m} c_i(\theta) d_i(x_1)\right) \dots \exp\left(\sum_{i=1}^{m} c_i(\theta) d_i(x_n)\right) \\ &= (a(\theta))^n b(x_1) \dots b(x_n) \exp\left(T_1 c_1(\theta) + \dots + T_m c_m(\theta)\right). \end{aligned}$$

Die Suffizienz von $(T_1, \dots, T_m)$ folgt aus dem Faktorisierungssatz von Neyman-Fisher mit

$$h(x_1, \dots, x_n) = b(x_1) \dots b(x_n), \quad g(T_1, \dots, T_m; \theta) = (a(\theta))^n \exp\left(T_1 c_1(\theta) + \dots + T_m c_m(\theta)\right).$$

Man kann zeigen, dass diese Statistik auch vollständig ist, wenn die Menge

$$\{(c_1(\theta), \dots, c_m(\theta)) \colon \theta \in \Theta\} \subset \mathbb{R}^m$$

mindestens einen m-dimensionalen Ball enthält (ohne Beweis).

Beispiel 4.10.1. Betrachten wir die Familie der Normalverteilungen mit Parametern $\mu \in \mathbb{R}$ und $\sigma^2 > 0$, wobei beide Parameter als unbekannt betrachtet werden. Wir haben bereits gesehen, dass diese Familie eine zweiparametrige Exponentialfamilie mit $d_1(x) = x^2$ und $d_2(x) = x$ ist. Somit st die Statistik (T_1, T_2) mit

$$\begin{aligned} T_1(X_1, \dots, X_n) &= \sum_{j=1}^{n} d_1(X_j) = \sum_{j=1}^{n} X_j^2, \\ T_2(X_1, \dots, X_n) &= \sum_{j=1}^{n} d_2(X_j) = \sum_{j=1}^{n} X_j \end{aligned}$$

suffizient und vollständig.

4.11. Bedingter Erwartungswert und bedingte Wahrscheinlichkeiten

In diesem Abschnitt werden wir eine allgemeine Definition der bedingten Wahrscheinlichkeiten und Erwartungswerte vorstellen.

Elementare bedingte Wahrscheinlichkeiten und Erwartungswerte. Zuerst erinnern wir uns an die Definition, die bereits mehrmals benutzt wurde.

Definition 4.11.1. Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Seien $A \in \mathcal{A}$ und $B \in \mathcal{A}$ zwei Ereignisse mit $\mathbb{P}[B] \neq 0$. Dann ist die *bedingte Wahrscheinlichkeit* von A gegeben

B folgendermaßen definiert:

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}.$$

Diese Definition ergibt nur dann Sinn, wenn $\mathbb{P}[B] \neq 0$. Analog kann man den bedingten Erwartungswert definieren.

Definition 4.11.2. Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $X : \Omega \to \mathbb{R}$ eine Zufallsvariable mit $\mathbb{E}|X| < \infty$. Sei $B \in \mathcal{A}$ ein Ereignis mit $\mathbb{P}[B] \neq 0$. Dann ist der *bedingte Erwartungswert* von X gegeben B folgendermaßen definiert:

$$\mathbb{E}[X|B] = \frac{\mathbb{E}[X\mathbb{1}_B]}{\mathbb{P}[B]}.$$

Auch diese Definition ergibt nur dann Sinn, wenn $\mathbb{P}[B] \neq 0$. Bedingte Wahrscheinlichkeiten können als Spezialfall des bedingten Erwartungswerts angesehen werden, denn

$$\mathbb{P}[A|B] = \mathbb{E}[\mathbb{1}_A|B].$$

Wir haben gesehen (z.B. bei der Definition der Suffizienz im absolut stetigen Fall), dass man bedingte Wahrscheinlichkeiten oder Erwartungswerte oft auch im Falle $\mathbb{P}[B] = 0$ betrachten muss. In diesem Abschnitt werden wir eine allgemeine Definition des bedingten Erwartungswerts geben, die das (zumindest in einigen Fällen) möglich macht.

Bedingter Erwartungswert gegeben eine σ-Algebra. Sei $X : \Omega \to \mathbb{R}$ eine auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ definierte Zufallsvariable. Wir nehmen an, dass X integrierbar ist, d.h. $\mathbb{E}|X| < \infty$. Sei $\mathcal{B} \subset \mathcal{A}$ eine Teil-σ-Algebra von $\mathcal{A}$, d.h. für jede Menge $B \in \mathcal{B}$ gelte auch $B \in \mathcal{A}$. In diesem Abschnitt werden wir den bedingten Erwartungswert von X gegeben die σ-Algebra $\mathcal{B}$ definieren.

Sei zunächst $X \geq 0$ fast sicher.

Schritt 1. Sei Q ein Maß auf dem Messraum $(\Omega, \mathcal{B})$ mit

$$Q(B) = \mathbb{E}[X\mathbb{1}_B] \text{ für alle } B \in \mathcal{B}.$$

Das Maß Q ist endlich, denn $Q(\Omega) = \mathbb{E}X < \infty$ nach Voraussetzung. Es sei bemerkt, dass das Maß Q auf $(\Omega, \mathcal{B})$ und nicht auf $(\Omega, \mathcal{A})$ definiert wurde. Das Wahrscheinlichkeitsmaß $\mathbb{P}$ hingegen ist auf $(\Omega, \mathcal{A})$ definiert, wir können es aber auch auf die kleinere σ-Algebra $\mathcal{B}$ einschränken und als ein Wahrscheinlichkeitsmaß auf $(\Omega, \mathcal{B})$ betrachten.

Schritt 2. Ist nun $B \in \mathcal{B}$ eine Menge mit $\mathbb{P}[B] = 0$, so folgt, dass $Q(B) = \mathbb{E}[X\mathbb{1}_B] = 0$, denn die Zufallsvariable $X\mathbb{1}_B$ ist $\mathbb{P}$-fast sicher gleich 0. Somit ist Q absolut stetig bezüglich $\mathbb{P}$ auf $(\Omega, \mathcal{B})$. Nach dem Satz von Radon-Nikodym gibt es eine Funktion Z, die messbar bezüglich $\mathcal{B}$ ist, mit

$$\mathbb{E}[Z\mathbb{1}_B] = \mathbb{E}[X\mathbb{1}_B] \text{ für alle } B \in \mathcal{B}.$$

Es sei bemerkt, dass X $\mathcal{A}$-messbar ist, wohingegen Z lediglich $\mathcal{B}$-messbar ist. Wir nennen die Zufallsvariable Z den bedingten Erwartungswert von X gegeben $\mathcal{B}$ und schreiben

$$\mathbb{E}[X|\mathcal{B}] = Z.$$

SCHRITT 3. Sei nun X eine beliebige (nicht unbedingt positive) Zufallsvariable auf $(\Omega, \mathcal{A}, \mathbb{P})$ mit $\mathbb{E}|X| < \infty$. Sei $\mathcal{B} \subset \mathcal{A}$ nach wie vor eine Teil-σ-Algebra. Wir haben die Darstellung $X = X^+ - X^-$ mit $X^+ \geq 0$ und $X^- \geq 0$. Die bedingte Erwartung von X gegeben $\mathcal{B}$ ist definiert durch

$$\mathbb{E}[X|\mathcal{B}] = \mathbb{E}[X^+|\mathcal{B}] - \mathbb{E}[X^-|\mathcal{B}].$$

Wir können nun die obigen Überlegungen zu folgender Definition zusammenfassen.

Definition 4.11.3. Sei X eine Zufallsvariable mit $\mathbb{E}|X| < \infty$, definiert auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$. (Somit ist X $\mathcal{A}$-messbar). Sei $\mathcal{B} \subset \mathcal{A}$ eine Teil-σ-Algebra. Eine Funktion $Z : \Omega \to \mathbb{R}$ heißt *bedingter Erwartungswert von X gegeben $\mathcal{B}$*, falls

(1) Z ist $\mathcal{B}$-messbar.
(2) $\mathbb{E}[Z\mathbb{1}_B] = \mathbb{E}[X\mathbb{1}_B]$ für alle $B \in \mathcal{B}$.

Wir schreiben dann $\mathbb{E}[X|\mathcal{B}] = Z$.

Bemerkung 4.11.4. Die bedingte Erwartung $\mathbb{E}[X|\mathcal{B}]$ ist eine Zufallsvariable, keine Konstante! Die Existenz von $\mathbb{E}[X|\mathcal{B}]$ wurde bereits oben mit dem Satz von Radon-Nikodym bewiesen. Der bedingte Erwartungswert ist bis auf $\mathbb{P}$-Nullmengen eindeutig definiert. Das folgt aus der entsprechenden Eigenschaft der Dichte im Satz von Radon-Nikodym.

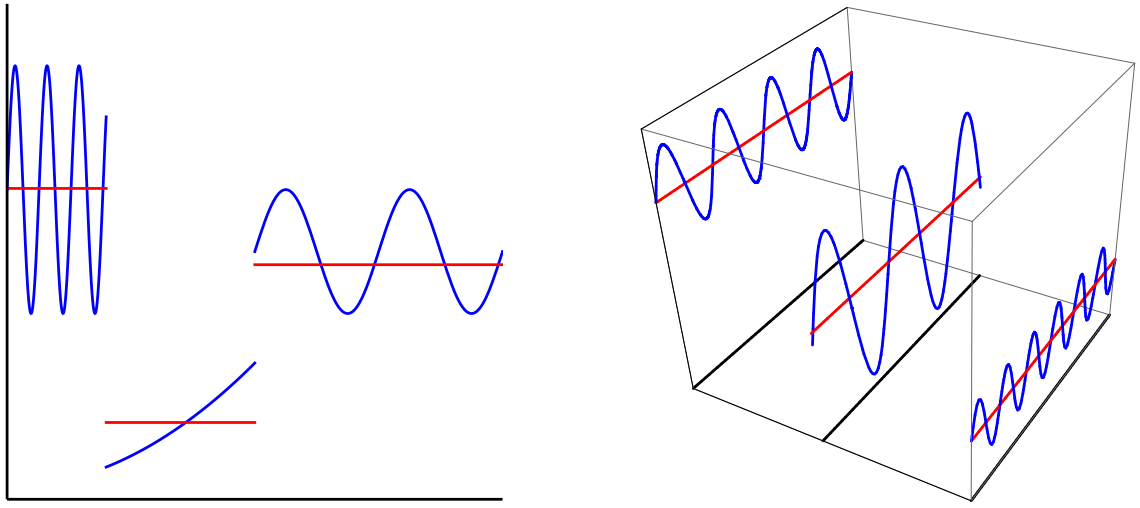


ABBILDUNG 2. Bedingter Erwartungswert in Beispiel 4.11.5 (links) und Beispiel 4.11.6 (rechts). Blau: Der Graph von X. Rot: Der bedingte Erwartungswert.

Beispiel 4.11.5. Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Betrachte eine disjunkte abzählbare Zerlegung $\Omega = \cup_{n\in\mathbb{N}}\Omega_n$, wobei $\Omega_n \in \mathcal{A}$ und $\mathbb{P}[\Omega_n] \neq 0$. (Wir betrachten hier eine unendliche Zerlegung, der Fall einer endlichen Zerlegung ist völlig analog). Sei $\mathcal{B}$ die σ-Algebra, die von

der Familie $\{\Omega_1, \Omega_2, \ldots\}$ erzeugt wird. Somit ist

$$\mathcal{B} = \left\{ \bigcup_{n\in\mathbb{N}} \Omega_n^{\varepsilon_n} : \varepsilon_1, \varepsilon_2, \ldots \in \{0,1\} \right\},$$

wobei $\Omega_n^1 = \Omega_n$ und $\Omega_n^0 = \varnothing$. Sei X eine beliebige ($\mathcal{A}$-messbare) Zufallsvariable auf Ω mit $\mathbb{E}|X| < \infty$. Für den bedingten Erwartungswert von X geegeben $\mathcal{B}$ gilt:

$$\mathbb{E}[X|\mathcal{B}](\omega) = \frac{\mathbb{E}[X\mathbb{1}_{\Omega_n}]}{\mathbb{P}[\Omega_n]}, \quad \text{falls } \omega \in \Omega_n.$$

Beweis. Beachte, dass $Z := \mathbb{E}[X|\mathcal{B}]$ $\mathcal{B}$-messbar sein muss. Also ist Z konstant auf jeder Menge Ω_n. Sei also $Z(\omega) = c_n$ für $\omega \in \Omega_n$. Es muss außerdem gelten, dass

$$\mathbb{E}[X\mathbb{1}_{\Omega_n}] = \mathbb{E}[Z\mathbb{1}_{\Omega_n}] = c_n\mathbb{P}[\Omega_n].$$

Daraus folgt, dass $c_n = \mathbb{E}[X\mathbb{1}_{\Omega_n}]/\mathbb{P}[\Omega_n]$ sein muss. □

Beispiel 4.11.6. Sei $\Omega = [0,1]^2$. Sei $\mathcal{A}$ die Borel-σ-Algebra auf $[0,1]^2$ und $\mathbb{P}$ das Lebesgue-Maß. Sei $X : [0,1]^2 \to \mathbb{R}$ eine ($\mathcal{A}$-messbare) Zufallsvariable mit $\mathbb{E}|X| < \infty$. Sei $\mathcal{B} \subset \mathcal{A}$ eine Teil-σ-Algebra von $\mathcal{A}$ mit

$$\mathcal{B} = \{C \times [0,1] : C \subset [0,1] \text{ ist Borel}\}.$$

Dann ist der bedingte Erwartungswert von X gegeben $\mathcal{B}$ gegeben durch:

$$Z(s,t) := \mathbb{E}[X|\mathcal{B}](s,t) = \int_0^1 X(s,u)\mathrm{d}u, \quad (s,t) \in [0,1]^2.$$

Beweis. Wir zeigen, dass die soeben definierte Funktion Z die beiden Bedingungen aus der Definition der bedingten Erwartung erfüllt. Zunächst ist $Z(s,t)$ eine Funktion, die nur von s abhängt. Somit ist Z messbar bzgl. $\mathcal{B}$. Außerdem gilt für jede $\mathcal{B}$-messbare Menge $B = C \times [0,1]$, dass

$$\mathbb{E}[Z\mathbb{1}_{C\times[0,1]}] = \int_{C\times[0,1]} Z(s,t)\mathrm{d}s\mathrm{d}t = \int_C \left(\int_0^1 Z(s,t)\mathrm{d}t \right) \mathrm{d}s = \mathbb{E}[X\mathbb{1}_{C\times[0,1]}].$$

Somit ist auch die zweite Bedingung erfüllt.

Beispiel 4.11.7. Sei $X : \Omega \to \mathbb{R}$ eine Zufallsvariable mit $\mathbb{E}|X| < \infty$, definiert auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$. Dann gilt

(1) $\mathbb{E}[X|\{\Omega, \varnothing\}] = \mathbb{E}X$.
(2) $\mathbb{E}[X|\mathcal{A}] = X$.

Beweis. Übung. □

Satz 4.11.8. Seien $X, Y : \Omega \to \mathbb{R}$ Zufallsvariablen (beide $\mathcal{A}$-messbar) mit $\mathbb{E}|X| < \infty$, $\mathbb{E}|Y| < \infty$, definiert auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$. Sei $\mathcal{B} \subset \mathcal{A}$ eine Teil-σ-Algebra von $\mathcal{A}$.

(1) Es gilt die Formel der totalen Erwartung: $\mathbb{E}[\mathbb{E}[X|\mathcal{B}]] = \mathbb{E}X$.
(2) Aus $X \le Y$ fast sicher folgt, dass $\mathbb{E}[X|\mathcal{B}] \le \mathbb{E}[Y|\mathcal{B}]$ fast sicher.
(3) Für alle $a, b \in \mathbb{R}$ gilt $\mathbb{E}[aX + bY|\mathcal{B}] = a\,\mathbb{E}[X|\mathcal{B}] + b\,\mathbb{E}[Y|\mathcal{B}]$ fast sicher.

(4) Falls Y sogar $\mathcal{B}$-messbar ist und $\mathbb{E}|XY| < \infty$, dann gilt

$$\mathbb{E}[XY|\mathcal{B}] = Y\mathbb{E}[X|\mathcal{B}] \text{ fast sicher.}$$

Beweis. Übung. □

Bedingter Erwartungswert gegeben eine Zufallsvariable. Besonders oft wird die Definition der bedingten Erwartung im folgenden Spezialfall benutzt.

Definition 4.11.9. Sei Y eine Zufallsvariable auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$. Die *von Y erzeugte σ-Algebra* ist definiert durch

$$\sigma(Y) = \{Y^{-1}(C) : C \subset \mathbb{R} \text{ Borel}\}.$$

Man kann sich die σ-Algebra $\sigma(Y)$ folgendermaßen vorstellen. Zunächst einmal liegen alle Niveaumengen der Form $Y^{-1}(t) := \{\omega \in \Omega \colon Y(\omega) = t\}$ in $\sigma(Y)$, für alle $t \in \mathbb{R}$. Dabei ist Ω eine disjunkte Vereinigung dieser Niveaumengen: $\Omega = \cup_{t\in\mathbb{R}} Y^{-1}(t)$. Außerdem beinhaltet $\sigma(Y)$ alle Vereinigungen der Niveaumengen der Form $\cup_{t\in C} Y^{-1}(t)$, wobei $C \subset \mathbb{R}$ eine beliebige Borel-Menge ist.

Definition 4.11.10. Seien X und Y zwei $\mathcal{A}$-messbare Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$. Sei X integrierbar. Der bedingte Erwartungswert von X gegeben Y ist definiert durch

$$\mathbb{E}[X|Y] = \mathbb{E}[X|\sigma(Y)].$$

Bemerkung 4.11.11. $\mathbb{E}[X|Y]$ ist eine Zufallsvariable! Aus der Messbarkeit von $\mathbb{E}[X|Y] = \mathbb{E}[X|\sigma(Y)]$ bzgl. $\sigma(Y)$ (s. Definition 4.11.3) kann man herleiten, dass $\mathbb{E}[X|Y]$ eine Borel-Funktion von Y sein muss (s. das Faktorisierungslemma 4.11.16 im nächsten Abschnitt). Es gibt also eine Borel-Funktion $\varphi : \mathbb{R} \to \mathbb{R}$ mit

$$\mathbb{E}[X|Y] = \varphi(Y).$$

D.h. der bedingte Erwartungswert $\mathbb{E}[X|Y]$ bleibt konstant auf jeder Niveaumenge $\Omega_t = \{\omega \in \Omega \colon Y(\omega) = t\}$, für alle $t \in \mathbb{R}$:

$$\mathbb{E}[X|Y](\omega) = \varphi(t) \text{ falls } Y(\omega) = t.$$

Dabei ist Ω eine disjunkte Vereinigung dieser Niveaumengen: $\Omega = \cup_{t\in\mathbb{R}} \Omega_t$. Den Wert von $\mathbb{E}[X|Y]$ auf einer Niveaumenge Ω_t kann man sich als den „Mittelwert“ von X über Ω_t vorstellen, vgl. Beispiele 4.11.5 und 4.11.6.

Definition 4.11.12. Wir definieren den bedingten Erwartungswert von X gegeben, dass $Y = t$ ist, durch

$$\mathbb{E}[X|Y = t] = \mathbb{E}[X|Y](\omega) = \varphi(t), \quad t \in \mathbb{R},$$

wobei $\omega \in \Omega$ ein beliebiges Element mit $Y(\omega) = t$ sei.

Dabei darf $\mathbb{P}[Y = t]$ auch 0 sein! Wir müssen hier allerdings die Frage der Eindeutigkeit klären. Die Zufallsvariable $\mathbb{E}[X|Y]$ ist nur bis auf $\mathbb{P}$-Nullmengen eindeutig definiert. Inwieweit ist die Funktion φ eindeutig? Sei μ_Y die Verteilung von Y, d.h. μ_Y ist ein Wahrscheinlichkeitsmaß auf $\mathbb{R}$ mit $\mu_Y(A) = \mathbb{P}[Y \in A]$. Verändern wir nun φ auf einer μ_Y-Nullmenge, so verändert sich $\varphi(Y)$ auf einer $\mathbb{P}$-Nullmenge. Somit ist die Funktion φ nur bis auf μ_Y-Nullmengen eindeutig definiert. Es folgt, dass auch die Borel-Funktion $t \mapsto \mathbb{E}[X|Y = t]$ bis auf Nullmengen von μ_Y eindeutig definiert ist. Für einen vorgegebenen Wert $t \in \mathbb{R}$ kann man also in der Regel leider nicht sagen, was $\mathbb{E}[X|Y = t]$ ist! Man muss die Funktion $t \mapsto \mathbb{E}[X|Y = t]$ immer als Ganzes betrachten.

Nun können wir auch bedingte Wahrscheinlichkeuten als Spezialfall des bedingten Erwartungswerts definieren.

Definition 4.11.13. Sei $Y : \Omega \to \mathbb{R}$ eine Zufallsvariable auf $(\Omega, \mathcal{A}, \mathbb{P})$. Für ein Ereignis $A \in \mathcal{A}$ definieren wir die die bedingte Wahrscheinlichkeit von A gegeben, dass $Y = t$, durch

$$\mathbb{P}[A|Y = t] := \mathbb{E}[\mathbb{1}_A|Y](\omega), \quad t \in \mathbb{R},$$

wobei $\omega \in \Omega$ beliebig mit $Y(\omega) = t$ ist.

Beispiel 4.11.14. Sei Y eine diskrete Zufallsvariable auf $(\Omega, \mathcal{A}, \mathbb{P})$. Das Bild von Y, also die Menge

$$\operatorname{Im} Y = \{t \in \mathbb{R} \colon \mathbb{P}[Y = t] > 0\}$$

ist somit höchstens abzählbar. Die von Y erzeugte σ-Algebra $\sigma(Y)$ wird von den Mengen $\Omega_t = \{Y = t\}$, $t \in \operatorname{Im} Y$, erzeugt und hat somit die gleiche Gestalt wie in Beispiel 4.11.5. Für der bedingten Erwartungswert einer integrierbaren Zufallsvariable $X : \Omega \to \mathbb{R}$ gilt somit

$$\mathbb{E}[X|Y = t] = \mathbb{E}[X|Y](\omega) = \frac{\mathbb{E}[X\,\mathbb{1}_{\Omega_t}]}{\mathbb{P}[\Omega_t]},$$

wobei $\omega \in \Omega_t$ beliebig ist.
Seien nun X und Y beide diskret mit gemeinsamer Zähldichte $f_{X,Y}(s,t) = \mathbb{P}[X = s, Y = t]$ und die Zähldichte von Y sei $f_Y(t) = \mathbb{P}[Y = t]$. Dann gilt für den bedingten Erwartungswert

$$\mathbb{E}[X|Y = t] = \frac{\sum_{s \in \operatorname{Im} X} \mathbb{P}[X = s \cap Y = t]s}{\mathbb{P}[Y = t]} = \frac{\sum_{s \in \operatorname{Im} X} f_{X,Y}(s,t)s}{f_Y(t)}.$$

Beispiel 4.11.15. Seien X, Y Zufallsvariablen mit gemeinsamer Dichte $f_{X,Y}(s,t)$ und die Dichte von Y sei $f_Y(t)$. Man kann zeigen, dass dann für den bedingten Erwartungswert eine ähnliche Formel gilt:

$$\mathbb{E}[X|Y](\omega) = \mathbb{E}[X|Y = t] = \frac{\int_{\mathbb{R}} f_{X,Y}(s,t)s\mathrm{d}s}{f_Y(t)},$$

wobei $\omega \in \Omega$ beliebig mit $Y(\omega) = t$ ist. Diese Formel ergibt Sinn, wenn $f_Y(t) \neq 0$.

Faktorisierungslemma. Hier beweisen wir die im vorherigen Abschnitt angekündigte Aussage. Es ist klar, dass eine Zufallsvariable der Form $\varphi(Y)$, wobei $\varphi : \mathbb{R} \to \mathbb{R}$ eine Borel-Funktion ist, $\sigma(Y)$-messbar ist. Wir beweisen nun, dass jede $\sigma(Y)$-messbare Zufallsvariable von dieser Form ist.

Lemma 4.11.16 (Faktorisierungslemma)**.** Sei $Y : \Omega \to \mathbb{R}$ eine Zufallsvariable auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$. Ist $Z : \Omega \to \mathbb{R}$ eine $\sigma(Y)$-messbare Zufallsvariable, so gibt es eine Borel-Funktion $\varphi : \mathbb{R} \to \mathbb{R}$ mit $Z = \varphi(Y)$.

Beweis. Schritt 1. Sei zuerst $Z \geq 0$. Wir zeigen, dass es Mengen $A_1, A_2, \ldots \in \sigma(Y)$ und Konstanten $\alpha_1, \alpha_2, \ldots \geq 0$ gibt mit

$$Z = \sum_{n=1}^{\infty} \alpha_n \mathbb{1}_{A_n}.$$

Definiere $Z_n := 2^{-n}\lfloor 2^n Z \rfloor \wedge n$, wobei $\wedge$ für das Minimum steht. Dann gilt $Z_n \uparrow Z$ punktweise und Z_n nimmt Werte in der Menge $\{0, \frac{1}{2^n}, \frac{2}{2^n}, \ldots, \frac{n2^n}{2^n}\}$ an. Definiere $\sigma(Y)$-messbare Mengen

$$B_{n,i} = \{\omega \in \Omega \colon Z_n(\omega) - Z_{n-1}(\omega) = i2^{-n}\}, \quad n \in \mathbb{N}, \;\; i = 1, 2, \ldots, 2^n.$$

Es gilt $\cup_{i=1}^{2^n} B_{n,i} = \Omega$ und somit $Z_n - Z_{n-1} = \sum_{i=1}^{2^n} \frac{i}{2^n} \mathbb{1}_{B_{n,i}}$. Wir können nun Z als eine Teleskop-Summe schreiben:

$$Z = \sum_{n=1}^{\infty} (Z_n - Z_{n-1}) = \sum_{n=1}^{\infty} \sum_{i=1}^{2^n} \frac{i}{2^n} \mathbb{1}_{B_{n,i}},$$

denn $Z_0 = 0$. Nach einer Umnummerierung der Mengen $B_{n,i}$ und der Konstanten $\frac{i}{2^n}$ erhalten wir die gesuchte Darstellung.

Schritt 2. Sei nach wie vor $Z \geq 0$ mit der Darstellung $Z = \sum_{n=1}^{\infty} \alpha_n \mathbb{1}_{A_n}$ wie in Schritt 1. Nach Definition der σ-Algebra $\sigma(Y)$ gibt es zu jeder Menge $A_n \in \sigma(Y)$ eine Borel-Menge $B_n \subset \mathbb{R}$ mit $A_n = Y^{-1}(B_n)$ und somit

$$\mathbb{1}_{A_n}(\omega) = \mathbb{1}_{B_n}(Y(\omega)).$$

Definiere nun die Funktion $\varphi(t) = \sum_{n=1}^{\infty} \alpha_n \mathbb{1}_{B_n}$. Dann gilt

$$Z(\omega) = \sum_{n=1}^{\infty} \alpha_n \mathbb{1}_{A_n}(\omega) = \sum_{n=1}^{\infty} \alpha_n \mathbb{1}_{B_n}(Y(\omega)) = \varphi(Y(\omega)),$$

was die gesuchte Darstellung von Z ist.

Schritt 3. Sei nun Z nicht unbedingt nicht-negativ. Dann gibt es eine Darstellung $Z = Z_+ - Z_-$, wobei $Z_+ = \max\{Z, 0\}$ und $Z_- = \max\{-Z, 0\}$ ebenfalls $\sigma(Y)$-messbar und nicht-negativ sind. Nach Schritt 2 gibt es Darstellungen $Z_+ = \varphi_+(Y)$, $Z_- = \varphi_-(Y)$ für geeignete

Borel-Funktionen φ_+ und φ_-. Es folgt, dass $Z = \varphi_+(Y) - \varphi_-(Y) = \varphi(Y)$ mit $\varphi = \varphi_+ - \varphi_-$. □

Markow-Kerne und reguläre bedingte Wahrscheinlichkeiten. Sei $Y : \Omega \to \mathbb{R}$ eine Zufallsvariable. Für jedes Ereignis $A \in \mathcal{A}$ ist die Borel-Funktion $t \mapsto \mathbb{P}[A|Y = t]$ bis auf eine μ_Y-Nullmenge eindeutig definiert. Stellen wir uns nun vor, dass wir uns für jedes Ereignis A auf irgendeine Version dieser Funktion geeinigt haben. Man kann zeigen, dass folgende Eigenschaften gelten:

(1) $\mathbb{P}[\Omega|Y = t] = 1$ für μ_Y-fast alle $t \in \mathbb{R}$.
(2) Für jede disjunkte Familie von Ereignissen $A_1, A_2, \ldots \in \mathcal{A}$ gilt

$$\mathbb{P}\left[\cup_{n=1}^{\infty} A_n | Y = t\right] = \sum_{n=1}^{\infty} \mathbb{P}[A_n|Y = t] \quad \text{für } \mu_Y\text{-fast alle } t \in \mathbb{R}.$$

Naiv könnte man nun glauben, dass für jedes $t \in \mathbb{R}$ die Zuordnung $A \mapsto \mathbb{P}[A|Y = t]$ ein Wahrscheinlichkeitsmaß auf der Niveaumenge $\{Y = t\}$ definiert (und das man sich sozusagen als „Einschränkung“ von $\mathbb{P}$ auf die Niveaumenge vorstellen könnte). Leider gibt es hier ein Problem: alle Relationen gelten lediglich für μ_Y-fast alle $t \in \mathbb{R}$. Noch schlimmer, die Ausnahmemenge derjenigen t, für die die zweite Relation nicht gilt, hängt von $A_1, A_2, \ldots$ ab. Im schlimmsten Fall kann es passieren, dass man für jedes t eine Familie $A_1, A_2, \ldots$ finden kann, für die die σ-Additivität falsch ist. Glücklicherweise kann man dieses Problem durch eine geschickte Wahl von Versionen der Funktionen $t \mapsto \mathbb{P}[A|Y = t]$ vermeiden.

Definition 4.11.17. Ein *Markow-Kern* von $(\Omega, \mathcal{A})$ nach $(\Omega', \mathcal{A}')$ ist eine Familie $\{\pi_\omega : \omega \in \Omega\}$ mit den folgenden zwei Eigenschaften:

- Für jedes $\omega \in \Omega$ ist π_ω ein Wahrscheinlichkeitsmaß auf $(\Omega', \mathcal{A}')$.
- Für alle $A' \in \mathcal{A}'$ ist die Abbildung $\omega \mapsto \pi_\omega(A')$ eine Borel-Funktion auf $(\Omega, \mathcal{A})$.

Beispiel 4.11.18. Sei E eine höchstens abzählbare Menge und $p : E \times E \to [0, 1]$ eine Übergangswahrscheinlichkeit, d.h. $p(i, j) \geq 0$ und $\sum_{j \in E} p(i, j) = 1$ für alle $i \in E$. Dann definiert

$$\pi_i(A) = \sum_{j \in A} p(i, j)$$

einen Markow-Kern von $(E, 2^E)$ nach sich selbst.

Definition 4.11.19. Sei $Y : \Omega \to \mathbb{R}$ eine Zufallsvariable auf $(\Omega, \mathcal{A}, \mathbb{P})$. Ein Markow-Kern $\{\pi_t : t \in \mathbb{R}\}$ von $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ nach $(\Omega, \mathcal{A})$ heißt *reguläre Version* von $\mathbb{P}$ gegeben Y, wenn

(a) Das Wahrscheinlichkeitsmaß π_t ist auf der Niveaumenge $\Omega_t = \{Y = t\}$ konzentriert, d.h. $\pi_t(\Omega_t) = 1$ für alle $t \in \mathbb{R}$.

(b) Für jedes Ereignis $A \in \mathcal{A}$ gilt die „Formel der totalen Wahrscheinlichkeit“

$$\mathbb{P}[A] = \int_{\mathbb{R}} \pi_t(A) \mu_Y(\mathrm{d}t).$$

Eine reguläre Version von $\mathbb{P}$ existiert unter sehr allgemeinen Voraussetzungen an den Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$.

Definition 4.11.20. Zwei Messräume $(\Omega, \mathcal{A})$ und $(\Omega', \mathcal{A}')$ heißen *Borel-isomorph*, falls es eine bijektive Abbilding $T : \Omega \to \Omega'$ gibt mit

$$A \in \mathcal{A} \Longleftrightarrow T(A) \in \mathcal{A}'.$$

Man kann also die beiden Räume aufeinander bijektiv abbilden, sodass messbare Mengen auf messbare Mengen abgebildet werden.

Definition 4.11.21. Ein Messraum $(\Omega, \mathcal{A})$ heißt *standard Borel*, falls er zu einem der folgenden Messräume isomorph ist:

- $(E, 2^E)$, wobei E eine höchstens abzählbare Menge ist.
- Das Intervall $[0, 1]$ mit der Borel-σ-Algebra.

Satz 4.11.22. Sei (M, ρ) ein vollständiger separabler metrischer Raum (z.B. ein kompakter metrischer Raum). Es sei $\mathcal{B}(M)$ die Borel-σ-Algebra auf M, also die von allen offenen Mengen erzeugte σ-Algebra. Dann ist der Raum $(M, \mathcal{B}(M))$ standard Borel.

Ohne Beweis.

Beispiel 4.11.23. Die folgenden metrischen Räume mit ihren jeweiligen Borel-σ-Algebren sind allesamt isomorph zum Intervall $[0, 1]$ mit der Borel-σ-Algebra:

(a) Der Euklidische Raum $\mathbb{R}^n$.
(b) Der unendlich-dimensionale Hilbert-Raum $L^2[0, 1]$.
(c) Der Raum der stetigen Funktionen $C[0, 1]$ mit der Supremumsmetrik.

Die meisten Meßräume, die man in der Stochastik verwendet, sind standard Borel-Räume. Nun können wir endlich den Satz über die Existenz der regulären Version der bedingten Verteilung formulieren.

Satz 4.11.24. Sei $Y : \Omega \to \mathbb{R}$ eine Zufallsvariable auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ derart, dass $(\Omega, \mathcal{A})$ standard Borel ist. Dann existiert eine reguläre Version von $\mathbb{P}$ gegeben Y, s. Definition 4.11.19.

Ohne Beweis.

4.12. Satz von Rao-Blackwell

Eine suffiziente Statistik beinhaltet alles, was man über das Ergebnis eines statistischen Experiments wissen muss. Es ist deshalb plausibel, dass ein „guter“ Schätzer eine Funktion der suffizienten Statistik sein muss. Der nächste Satz bestätigt diese Vermutung.

Satz 4.12.1 (Rao-Blackwell)**.** Sei $(\mathbb{P}_\theta)_{\theta\in\Theta}$ eine Familie von Wahrscheinlichkeitsmaßen auf dem Stichprobenraum $(\mathfrak{X}, \mathcal{A})$, wobei $\Theta \subset \mathbb{R}$ ein Intervall sei. Es seien weiterhin

- $T : \mathfrak{X} \to \mathbb{R}^m$ eine suffiziente Statistik und
- $\hat\theta : \mathfrak{X} \to \mathbb{R}$ ein erwartungstreuer Schätzer von θ mit $\mathbb{E}_\theta\hat\theta^2 < \infty$ für alle $\theta \in \Theta$.

Definiere $\tilde\theta := \mathbb{E}_\theta[\hat\theta|T]$. Dann ist $\tilde\theta$ ebenfalls ein erwartungstreuer Schätzer von θ und es gilt

$$\operatorname{Var}_\theta \tilde\theta \le \operatorname{Var}_\theta \hat\theta \text{ für alle } \theta \in \Theta.$$

Der Schätzer $\tilde\theta$ ist somit mindestens so gut wie $\hat\theta$ und heißt aus diesem Grunde die *Rao-Blackwell-Verbesserung* von $\hat\theta$. Da $\tilde\theta$ als bedingter Erwartungswert $\sigma(T)$-messbar ist, kann man nach dem Faktorisierungslemma 4.11.16 $\tilde\theta$ als eine Borel-Funktion von T darstellen. Somit basiert $\tilde\theta$ nur auf dem Wert der suffizienten Statistik T.

Beweis. Schritt 1. Zuallererst müssen wir zeigen, dass $\tilde\theta = \mathbb{E}_\theta[\hat\theta|T]$ keine Funktion von θ ist. (Das darf nämlich ein Schätzer auf keinen Fall sein!) Wir schreiben den bedingten Erwartungswert an der Stelle $x \in \mathfrak{X}$ als das Integral bzgl. der bedingten Verteilung $\mathbb{P}_\theta[\cdot|T = T(x)]$:

$$\tilde\theta(x) = \mathbb{E}_\theta[\hat\theta|T](x) = \mathbb{E}_\theta[\hat\theta|T = T(x)] = \int_{\mathfrak{X}} \hat\theta(y)\mathbb{P}_\theta[\mathrm{d}y|T = T(x)].$$

Da T suffizient ist, hängt das Wahrscheinlichkeitsmaß $A \mapsto \mathbb{P}_\theta[A|T = T(x)]$ nicht von θ ab! Somit hängt das Integral auf der rechten Seite nicht von θ ab.

Schritt 2. Nun zeigen wir, dass $\tilde\theta$ erwartungstreu ist. Mit der Turmeigenschaft des bedingten Erwartungswerts gilt

$$\mathbb{E}_\theta\tilde\theta = \mathbb{E}_\theta\mathbb{E}_\theta[\hat\theta|T] = \mathbb{E}_\theta\hat\theta = \theta,$$

denn $\hat\theta$ ist erwartungstreu.

Schritt 3. Für die Varianz von $\tilde\theta$ gilt

$$\operatorname{Var}_\theta \tilde\theta = \mathbb{E}_\theta[(\tilde\theta - \theta)^2] = \mathbb{E}_\theta\left[(\mathbb{E}_\theta[\hat\theta|T] - \theta)^2\right] = \mathbb{E}_\theta\left[(\mathbb{E}_\theta[\hat\theta - \theta|T])^2\right].$$

Die Jensen-Ungleichung für den bedingten Erwartungswert besagt, dass $\varphi(\mathbb{E}[X|\mathcal{F}]) \le \mathbb{E}[\varphi(X)|\mathcal{F}]$ f.s., falls φ konvex ist. Mit dieser Ungleichung für $\varphi(x) = x^2$ ergibt sich

$$\mathbb{E}_\theta\left[(\mathbb{E}_\theta[\hat\theta - \theta|T])^2\right] \le \mathbb{E}_\theta\mathbb{E}_\theta\left[(\hat\theta - \theta)^2|T\right] = \mathbb{E}_\theta[(\hat\theta - \theta)^2] = \operatorname{Var}_\theta \hat\theta,$$

was die behauptete Ungleichung $\operatorname{Var}_\theta \tilde\theta \le \operatorname{Var}_\theta \hat\theta$ beweist. □

Beispiel 4.12.2. Seien $X_1, \ldots, X_n$ unabhängig und Bernoulli-verteilt mit Parameter $\theta \in (0, 1)$. Die Statistik $T(x_1, \ldots, x_n) = x_1 + \ldots + x_n$ ist suffizient. Als einen sehr einfachen

erwartungstreuen Schätzer von θ betrachten wir $\hat{\theta} = X_1$. Für die Varianz dieses Schätzers gilt $\mathrm{Var}_\theta \hat{\theta} = \theta(1-\theta)$. Nun definieren wir die Rao-Blackwell-Verbesserung von $\hat{\theta}$:

$$\tilde{\theta} = \mathbb{E}_\theta[X_1 | X_1 + \ldots + X_n].$$

Diesen bedingten Erwartungswert kann man direkt mir der Definition berechnen, es gibt allerdings eine viel elegantere Methode. Wegen Symmetrie gilt

$$\mathbb{E}_\theta[X_1 | X_1 + \ldots + X_n] = \mathbb{E}_\theta[X_2 | X_1 + \ldots + X_n] = \ldots = \mathbb{E}_\theta[X_n | X_1 + \ldots + X_n].$$

(Man erwartet im ersten Wurf genauso viel, wie im zweiten, auch dann, wenn die Summe $X_1 + \ldots + X_n$ gegeben ist). Es folgt, dass

$$\begin{aligned}\tilde{\theta} = \mathbb{E}_\theta[X_1 | X_1 + \ldots + X_n] &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\theta[X_i | X_1 + \ldots + X_n] \\ &= \frac{1}{n} \mathbb{E}_\theta[X_1 + \ldots + X_n | X_1 + \ldots + X_n] = \frac{1}{n}(X_1 + \ldots + X_n) = \bar{X}_n.\end{aligned}$$

Für die Varianz von $\tilde{\theta} = \bar{X}_n$ gilt $\mathrm{Var}_\theta \bar{X}_n = \frac{\theta(1-\theta)}{n}$, eine klare Verbesserung im Vergleich zu $\hat{\theta} = X_1$ (es sei denn $n = 1$, in welchem Fall beide Schätzer übereinstimmen).

4.13. Satz von Lehmann-Scheffé

Der folgende Satz wird uns in vielen Beispielen erlauben, den besten erwartungstreuen Schätzer zu konstruieren.

Satz 4.13.1 (Lehmann-Scheffé)**.** Sei $(\mathbb{P}_\theta)_{\theta \in \Theta}$ eine Familie von Wahrscheinlichkeitsmaßen auf dem Stichprobenraum $(\mathfrak{X}, \mathcal{A})$, wobei $\Theta \subset \mathbb{R}$ ein Intervall sei. Es seien weiterhin

- $T : \mathfrak{X} \to \mathbb{R}^m$ eine suffiziente und vollständige Statistik;
- $\hat{\theta} : \mathfrak{X} \to \mathbb{R}$ ein erwartungstreuer Schätzer von θ mit $\mathbb{E}_\theta \hat{\theta}^2 < \infty$ für alle $\theta \in \Theta$.

Dann ist $\tilde{\theta} := \mathbb{E}_\theta[\hat{\theta} | T]$ der beste erwartungstreue Schätzer von θ.

Beweis. Im Satz von Rao-Blackwell wurde gezeigt, dass $\tilde{\theta}$ wohldefiniert (d.h. keine Funktion von θ) und erwartungstreu ist.

Sei H ein weiterer erwartungstreuer Schätzer von θ mit $\mathbb{E}_\theta H^2 < \infty$ für alle $\theta \in \Theta$. Zu zeigen ist, dass $\tilde{\theta}$ besser ist, als H. Wir betrachten den Schätzer $\tilde{H} := \mathbb{E}_\theta[H | T]$. Nach dem Satz von Rao-Blackwell ist $\tilde{H}$ besser als H. Wir zeigen nun, dass $\tilde{H}$ mit $\tilde{\theta}$ übereinstimmt. Beide Schätzer sind $\sigma(T)$-messbar, da sie als bedingte Erwartungswerte gegeben T definiert sind. Nach dem Faktorisierungslemma 4.11.16 gibt es zwei Borel-Funktionen f und g mit $\tilde{\theta} = f(T)$ und $\tilde{H} = g(T)$. Dann ist $\tilde{\theta} - \tilde{H} = f(T) - g(T)$ ein erwartungstreuer Schätzer von 0, der nur auf dem Wert von T basiert. Wegen der Vollständigkeit von T muss $f(T) - g(T) = 0$ $\mathbb{P}_\theta$-f.s. gelten, woraus sich ergibt, dass $\tilde{\theta} = \tilde{H}$ $\mathbb{P}_\theta$-f.s. für alle $\theta \in \Theta$. □

Korollar 4.13.2. Sei $\hat{\theta}$ ein erwartungstreuer, suffizienter und vollständiger Schätzer von θ mit $\mathbb{E}_\theta \hat{\theta}^2 < \infty$ für alle $\theta \in \Theta$. Dann ist $\hat{\theta}$ der beste erwartungstreue Schätzer von θ.

Beweis. Folgt aus dem Satz von Lehmann-Scheffé mit $T = \hat{\theta}$ und $\tilde{\theta} = \mathbb{E}_\theta[\hat{\theta}|\hat{\theta}] = \hat{\theta}$. □

Beispiel 4.13.3. Seien $X_1, \ldots, X_n$ unabhängige, mit Parameter $\theta \in [0,1]$ Bernoulli-verteilte Zufallsvariablen. Der Schätzer $\bar{X}_n$ ist erwartungstreu, suffizient und vollständig und somit nach Korollar 4.13.2 bester erwartungstreuer Schätzer für θ. Diese Argumentation greift auch für unabhängige, mit Parameter $\theta > 0$ Poisson-verteilte Zufallsvariablen. Dabei ist der Beweis der Suffizienz und Vollständigkeit eine Übung.

Aus dem Satz von Lehmann-Scheffé ergibt sich die folgende Methode zur Konstruktion des besten erwartungstreuen Schätzers:

- Finde eine vollständige, suffiziente Statistik T.
- Finde Funktion g mit der Eigenschaft, dass $g(T)$ ein erwartungstreuer Schätzer von θ ist.
- Dann ist $g(T)$ der beste erwartungstreue Schätzer von θ.

Beweis. Folgt aus dem Satz von Lehmann-Scheffé mit $\hat{\theta} = g(T)$ und $\tilde{\theta} = \mathbb{E}_\theta[g(T)|T] = g(T)$. □

Beispiel 4.13.4. Seien $X_1, \ldots, X_n$ unabhängig und gleichverteilt auf $[0, \theta]$, wobei $\theta > 0$ geschätzt werden soll. Wir haben bereits gezeigt, dass $X_{(n)} = \max\{X_1, \ldots, X_n\}$ eine suffiziente und vollständige Statistik ist. Jedoch ist der Schätzer $X_{(n)}$ nicht erwartungstreu, denn

$$\mathbb{E}_\theta X_{(n)} = \frac{n}{n+1}\theta.$$

Deshalb betrachten wir den Schätzer

$$\tilde{\theta} := \frac{n+1}{n} X_{(n)} = \frac{n+1}{n} \max\{X_1, \ldots, X_n\},$$

der erwartungstreu und eine Funktion von $X_{(n)}$ ist. Nach den obigen Überlegungen ist $\frac{n+1}{n} X_{(n)}$ der beste erwartungstreue Schätzer für θ.

In den folgenden Beispielen werden wir den Satz von Lehmann-Scheffé in einer etwas allgemeineren Form benutzen. Der Parameterraum Θ sei beliebig und sei $\nu : \Theta \to \mathbb{R}$ eine Funktion des Parameters. Ist $\hat{\nu} : \mathfrak{X} \to \mathbb{R}$ ein erwartungstreuer und quadratisch integrierbarer Schätzer von $\nu(\theta)$ und T eine suffiziente und vollständige Statistik, so ist $\tilde{\nu} = \mathbb{E}_\theta[\hat{\nu}|T]$ der beste erwartungstreue Schätzer von $\nu(\theta)$. (Der obige Beweis funktioniert mit minimalen Veränderungen).

Beispiel 4.13.5. Seien $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ unabhängig und normalverteilt, wobei beide Parameter unbekannt seien. Wir behaupten, dass

- $\bar{X}_n$ der beste erwartungstreue Schätzer von μ ist;
- $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ der beste erwartungstreue Schätzer von σ^2 ist.

Beweis. Die Statistik $(\bar{X}_n, S_n^2)$ ist vollständig und suffizient. Sowohl $\bar{X}_n$ als auch S_n^2 sind Funktionen dieser Statistik, die erwartungstreu für μ bzw. σ^2 sind. □

Beispiel 4.13.6. Es seien $X_1, \ldots, X_n \sim \mathrm{Poi}(\theta)$ unabhängig. Der beste erwartungstreue Schätzer von θ ist, wie bereits gezeigt wurde, $\bar{X}_n$. Wie sieht nun der beste erwartungstreue Schätzer für

$$\nu(\theta) := \mathrm{e}^{-\theta} = \mathbb{P}_\theta[X_i = 0]$$

aus? Ein natürlicher Schätzer von $\mathrm{e}^{-\theta}$ ist $\mathrm{e}^{-\bar{X}_n}$, dieser Schätzer ist aber nicht erwartungstreu:

Aufgabe 4.13.7. Bestimmen Sie $\mathbb{E}_\theta \mathrm{e}^{-\bar{X}_n}$.

Um den besten erwartungstreuen Schätzer für $\mathrm{e}^{-\theta}$ zu konstruieren, benutzen wir den Satz von Lehmann-Scheffé. Es ist bekannt, dass die Statistik $T = X_1 + \ldots + X_n$ vollständig und suffizient ist. Nun brauchen wir noch einen erwartungstreuen Schätzer von $\mathrm{e}^{-\theta}$. Als solchen nehmen wir z.B.

$$\hat{\nu} = \mathbb{1}_{\{X_1 = 0\}}.$$

Die Erwartungstreue von $\hat{\nu}$ folgt aus $\mathbb{E}_\theta \hat{\nu} = \mathbb{P}_\theta[X_1 = 0] = \mathrm{e}^{-\theta}$. Nach dem Satz von Lehmann-Scheffé ist $\tilde{\nu} = \mathbb{E}_\theta[\hat{\nu}|S_n]$ der beste erwartungstreue Schätzer. Wir müssen nur noch diesen bedingten Erwartungswert berechnen. Für $s \in \mathbb{N}_0$ betrachten wir

$$f(s) := \mathbb{E}_\theta[\hat{\nu}|T = s] = \mathbb{P}_\theta[X_1 = 0|T = s] = \frac{\mathbb{P}_\theta[X_1 = 0, T = s]}{\mathbb{P}_\theta[T = s]} = \frac{\mathbb{P}_\theta[X_1 = 0]\mathbb{P}_\theta[X_2 + \ldots + X_n]}{\mathbb{P}_\theta[X_1 + \ldots + X_n = s]}.$$

Nun benutzen wir die Faltungseigenschaft der Poisson-Verteilung: Unter $\mathbb{P}_\theta$ gilt

$$X_1 \sim \mathrm{Poi}(\theta), \quad X_2 + \ldots + X_n \sim \mathrm{Poi}((n-1)\theta), \quad X_1 + \ldots + X_n \sim \mathrm{Poi}(n\theta).$$

Somit erhalten wir, dass

$$f(s) = \frac{\mathrm{e}^{-\theta}\mathrm{e}^{-(n-1)\theta}((n-1)\theta)^s/s!}{\mathrm{e}^{-n\theta}(n\theta)^s/s!} = \left(1 - \frac{1}{n}\right)^s.$$

Aus dem Satz von Lehmann-Scheffé folgt nun, dass

$$\tilde{\nu} = \mathbb{E}_\theta[\hat{\nu}|T] = f(T) = \left(1 - \frac{1}{n}\right)^T$$

der beste erwartungstreue Schätzer von $\mathrm{e}^{-\theta}$ ist.

Aufgabe 4.13.8. Es seien $X_1, \ldots, X_n \sim \mathrm{Poi}(\theta)$ unabhängig. Definiere $T := X_1 + \ldots + X_n$. Zeigen Sie, dass der beste erwartungstreue Schätzer von

$$\nu_k = \mathrm{e}^{-\theta}\frac{\theta^k}{k!} = \mathbb{P}_\theta[X_1 = k], \quad k \in \mathbb{N}_0,$$

durch

$$\hat{\nu}_{k,n} := \binom{T}{k}\left(\frac{1}{n}\right)^k\left(1 - \frac{1}{n}\right)^{T-k}$$

gegeben ist. Zeigen Sie auch, dass $\hat{\nu}_{k,n}$ ein stark konsistenter Schätzer von ν_k ist.

Beispiel 4.13.9. Seien $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ unabhängig und normalverteilt mit bekannter Varianz $\sigma^2 > 0$ und unbekanntem Erwartungswert $\mu \in \mathbb{R}$. Diese Verteilungen bilden eine Exponentialfamilie und die Statistik $\bar{X}_n$ ist vollständig und suffizient. Außerdem ist $\bar{X}_n$ erwartungstreu. Der beste erwartungstreue Schätzer für μ ist somit $\bar{X}_n$.

Versuchen wir nun, μ^2 als Parameter zu betrachten und zu schätzen. Der Schätzer $\bar{X}_n^2$ ist nicht erwartungstreu, denn

$$\mathbb{E}_\mu \bar{X}_n^2 = \operatorname{Var}_\mu \bar{X}_n + (\mathbb{E}_\mu \bar{X}_n)^2 = \frac{1}{n}\sigma^2 + \mu^2.$$

Deshalb betrachten wir den Schätzer $\bar{X}_n^2 - \frac{\sigma^2}{n}$. Dieser Schätzer ist erwartungstreu, suffizient und vollständig (Übung) und somit bester erwartungstreuer Schätzer für μ^2.

Aufgabe 4.13.10. Seien $X_1, \dots, X_n$ unabhängig und Bernoulli-verteilt mit Parameter $p \in (0,1)$. Bestimmen Sie den besten erwartungstreuen Schätzer für p^2.

Aufgabe 4.13.11. Gegeben sei eine Urne mit einer unbekannten Anzahl $N \in \mathbb{N}$ Kugeln, die von 1 bis N durchnummeriert sind. Es werden n Kugeln mit Zurücklegen gezogen und die zugehörigen Nummern $X_1, \dots, X_n$ notiert. Zeigen Sie, dass $X_{(n)}$ eine suffiziente und vollständige Statistik ist und konstruieren Sie den besten erwartungstreuen Schätzer für N.

Aufgabe 4.13.12. Seien $X_1, \dots, X_n \sim \mathrm{Geo}(\theta)$ unabhängig, $\theta \in (0,1)$. Konstruieren Sie für jedes $k = 1, 2, \dots$ den besten erwartungstreuen Schätzer von $\mathbb{P}_\theta[X_1 > k] = (1-\theta)^k$.

Aufgabe 4.13.13. Seien $X_1, \dots, X_n \sim \mathrm{Exp}(\theta)$ unabhängig, $\theta > 0$. Konstruieren Sie den besten erwartungstreuen Schätzer von $\mathbb{P}_\theta[X_1 > a] = \mathrm{e}^{-a\theta}$, wobei $a > 0$ gegeben ist.

Aufgabe 4.13.14. Seien $X_1, \dots, X_n \sim \mathrm{N}(\mu, 1)$ unabhängig, wobei $\mu \in \mathbb{R}$ unbekannt sei. Konstruieren Sie den besten erwartungstreuen Schätzer von $\mathrm{e}^{a\mu}$, wobei $a \in \mathbb{R}$ gegeben ist.

Aufgabe 4.13.15. Seien $X_1, \dots, X_n \sim \mathrm{U}[\theta_1, \theta_2]$ unabhängig, wobei $\theta_1 < \theta_2$ unbekannt seien. Konstruieren Sie die besten erwartungstreuen Schätzer von θ_1 und θ_2.

Aufgabe 4.13.16. Es seien $X_1, \dots, X_n$ identisch $\mathrm{Poi}(2\lambda)$ und $Y_1, \dots, Y_m$ identisch $\mathrm{Poi}(3\lambda)$-verteilte Zufallsgrößen, die allesamt unabhängig voneinander seien. Der Parameter $\lambda > 0$ sei unbekannt.

(a) Beschreiben Sie die Situation durch ein geeignetes statistisches Experiment und bestimmen Sie den Maximum–Likelihood–Schätzer für λ.

(b) Bestimmen Sie einen besten erwartungstreuen Schätzer für λ. *Hinweis:* Die Vollständigkeit kann mit Hilfe der Definition gezeigt werden.

4.14. Satz von Basu

Sei $(\mathfrak{X}, \mathcal{A}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ ein statistisches Modell.

Definition 4.14.1. Eine Statistik $S : \mathfrak{X} \to \mathbb{R}^p$ heißt *verteilungsfrei*, falls

$$\mathbb{P}_{\theta_1}[S \in A] = \mathbb{P}_{\theta_2}[S \in A] \text{ für alle } \theta_1, \theta_2 \in \Theta, \quad A \subset \mathbb{R}^p (\text{Borel}).$$

D.h., die Verteilung von S unter $\mathbb{P}_\theta$ hängt nicht von θ ab.

Beispiel 4.14.2. Seien $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ unabhängig, wobei $\mu \in \mathbb{R}$ unbekannter Parameter ist und σ^2 bekannt sei. Wir behaupten, dass die Spannweite $X_{(n)} - X_{(1)}$ und die empirische Varianz $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ verteilungsfreie Statistiken sind.

Beweis. Die Idee ist, dass μ ein „Verschiebungsparameter" ist, der sich sowohl in der Spannweite, als auch in S_n^2 aufhebt. Seien $\xi_1, \ldots, \xi_n \sim \mathrm{N}(0, \sigma^2)$ unabhängig (mit Erwartungswert 0). Unter $\mathbb{P}_\mu$ hat $(X_1, \ldots, X_n)$ die gleiche Verteilung wie $(\xi_1 + \mu, \ldots, \xi_n + \mu)$. Somit hat S_n^2 unter $\mathbb{P}_\mu$ die gleiche Verteilung wie

$$\frac{1}{n-1}\sum_{i=1}^n \left(\xi_i + \mu - \frac{(\xi_1 + \mu) + \ldots + (\xi_n + \mu)}{n}\right)^2 = \frac{1}{n-1}\sum_{i=1}^n \left(\xi_i - \frac{\xi_1 + \ldots + \xi_n}{n}\right)^2.$$

Die Verteilung der rechten Seite hängt aber nicht von μ ab, denn diese Variable taucht dort gar nicht erst auf. Für die Spannweite ist der Beweis analog. □

Der folgende Satz von Basu besagt, dass jede verteilungsfreie Statistik von jeder vollständig suffizienten Statistik unabhängig ist.

Satz 4.14.3 (Basu, 1955)**.** Sei $S : \mathfrak{X} \to \mathbb{R}^p$ eine verteilungsfreie Statistik und $T : \mathfrak{X} \to \mathbb{R}^m$ eine vollständige suffiziente Statistik. Dann sind die Zufallsvektoren S und T unabhängig unter $\mathbb{P}_\theta$ für alle $\theta \in \Theta$.

Beweis. Betrachte das folgende Wahrscheinlichkeitsmaß Q auf $\mathbb{R}^p$:

$$Q(A) = \mathbb{P}_\theta[S \in A], \quad A \subset \mathbb{R}^p \text{ Borel}.$$

Es sei bemerkt, dass Q unabhängig von θ wegen der Verteilungsfreiheit von S ist. Im Folgenden halten wir die Menge A fest. Betrachte die Funktion

$$f_A(t) = \mathbb{P}_\theta[S \in A | T = t] = \mathbb{E}_\theta[\mathbb{1}_{\{S \in A\}} | T = t], \quad t \in \mathbb{R}^m.$$

Diese Funktion ist unabhängig von θ, da die Statistik T suffizient ist! Es gilt dann auch $f_A(T) = \mathbb{E}_\theta[\mathbb{1}_{\{S \in A\}} | T]$ und somit

$$\mathbb{E}_\theta[f_A(T)] = \mathbb{E}_\theta\left[\mathbb{E}_\theta[\mathbb{1}_{\{S \in A\}} | T]\right] = \mathbb{E}_\theta \mathbb{1}_{\{S \in A\}} = \mathbb{P}_\theta[S \in A] = Q(A).$$

Es folgt, dass

$$\mathbb{E}_\theta[f_A(T) - Q(A)] = 0 \text{ für alle } \theta \in \Theta.$$

Somit ist $f_A(T) - Q(A)$ ein erwartungstreuer Schätzer von 0, der auf der Statistik T basiert. Wegen der Vollständigkeit von T folgt daraus, dass

$$f_A(T) = Q(A) \;\; \mathbb{P}_\theta\text{-f.s. für alle } \theta \in \Theta.$$

Wir haben also gezeigt, dass

$$\mathbb{P}_\theta[S \in A | T = t] = \mathbb{P}_\theta[S \in A].$$

Ist nun $B \subset \mathbb{R}^m$ eine Borel-Menge und bezeichnen wir mit μ_T die Verteilung von T, so ergibt sich mit der Formel der totalen Wahrscheinlichkeit, dass

$$\mathbb{P}_\theta[S \in A, T \in B] = \int_B \mathbb{P}_\theta[S \in A | T = t]\mu_T(\mathrm{d}t) = \int_B \mathbb{P}_\theta[S \in A]\mu_T(\mathrm{d}t) = \mathbb{P}_\theta[S \in A]\mathbb{P}_\theta[T \in B],$$

was die Unabhängigkeit von S und T beweist. □

Hier ist eine typische Anwendung des Satzes von Basu. Dieses Korollar wird sich bei der Konstruktion des Student-t-Tests als sehr wichtig erweisen.

Korollar 4.14.4. Seien $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ unabhängig. Dann sind die Zufallsvariablen $\bar{X}_n$ und S_n^2 unabhängig.

Die Behauptung ist sehr überraschend, denn $\bar{X}_n$ taucht in der Definition von S_n^2 explizit auf!

Beweis. Wir fassen $\mu \in \mathbb{R}$ als unbekannten Parameter auf und halten σ^2 konstant. Die Statistik S_n^2 ist verteilungsfrei, während die Statistik $\bar{X}_n$ vollständig suffizient ist. Die Behauptung folgt nun aus dem Satz von Basu. □

Aufgabe 4.14.5. Seien $X_1, X_2, \ldots \sim \mathrm{Exp}(\lambda)$ unabhängig. Betrachten Sie die Ankunftszeiten des Poisson-Punktprozesses $S_k = X_1 + \ldots + X_k$, $k \in \mathbb{N}$. Zeigen Sie, dass

$$\left(\frac{S_1}{S_n}, \frac{S_2}{S_n}, \ldots, \frac{S_{n-1}}{S_n}\right) \text{ und } S_n$$

unabhängig sind.

Aufgabe 4.14.6. Seien $X_1, \ldots, X_n \sim \mathrm{U}[0,1]$ unabhängig und gleichverteilt auf $[0,1]$. Seien $U_{(1)} < \ldots < U_{(n)}$ die Ordnungsstatistiken. Zeigen Sie, dass

$$\left(\frac{U_{(1)}}{U_{(n)}}, \frac{U_{(2)}}{U_{(n)}}, \ldots, \frac{U_{(n-1)}}{U_{(n)}}\right) \text{ und } U_{(n)}$$

unabhängig sind.

Aufgabe 4.14.7. Seien $X, Y \sim \mathrm{N}(\mu, \sigma^2)$ unabhängig. Zeigen Sie, dass $X + Y$ und $X - Y$ ebenfalls unabhängig sind.

4.15. Einige Gegenbeispiele

Kann man vielleicht sogar unter allen quadratisch integrierbaren (nicht unbedingt erwartungstreuen) Schätzern einen finden, der gleichmäßig besser, als alle anderen ist? Die Antwort ist leider „nein“. Sei $\theta_0 \in \Theta$ beliebig. Wir können dann den konstanten Schätzer $\varphi = \theta_0$ betrachten. Es gilt

$$\mathrm{MSE}_{\theta_0}(\varphi) = 0.$$

Wäre nun ein Schätzer $\hat{\theta}$ gleichmäßig besser als φ, so müsste $\mathrm{MSE}_{\theta_0}(\hat{\theta}) = 0$ gelten. Das bedeutet aber, dass $\hat{\theta} = \theta_0$ f.s. unter $\mathbb{P}_{\theta_0}$. Wäre $\hat{\theta}$ gleichmäßig besser als alle Schätzer, so müsste $\hat{\theta} = \theta$ f.s. unter $\mathbb{P}_\theta$ für alle $\theta \in \Theta$. Das heißt, der Schätzer $\hat{\theta}$ müsste den richtigen Wert θ mit Wahrscheinlichkeit 1 exakt treffen. Solche Schätzer gibt es aber nur in trivialen Situationen.

Beispiel 4.15.1. Sei X_1 eine Zufallsvariable, die gleichverteilt auf dem Intervall $[\theta, \theta + 1]$ ist, wobei $\theta \in \mathbb{Z}$ unbekannt ist. Beobachtet man nun einen Wert x_1 zwischen 2 und 3, so weiß man ganz genau, dass $\theta = 2$ ist. In diesem Fall können wir anhand der Stichprobe den

Wert von θ richtig erraten. In allen interessanten statistischen Modellen ist aber so etwas nicht möglich.

Die folgenden Aufgaben zeigen, dass der beste erwartungstreue Schätzer einen gleichmäßig größeren quadratischen Fehler haben kann, als einige nicht-erwartungstreue Schätzer.

Aufgabe 4.15.2. Seien $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ unabhängig. Betrachten Sie den Schätzer $T = \frac{1}{c} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$ für σ^2. Zeigen Sie, dass das Minimum des mittleren quadratischen Fehlers für $c = n + 1$ erreicht wird. Dabei entspricht der Wert $c = n - 1$ dem besten erwartungstreuen Schätzer.

Aufgabe 4.15.3. Seien $X_1, \dots, X_n$ unabhängig und auf $[0, \theta]$ gleichverteilt, $\theta > 0$. Zeigen Sie, dass der nicht-erwartungstreue Schätzer $\hat{\theta}_1 := X_{(n)}$ für alle $n \geq 3$ einen gleichmäßig kleineren mittleren quadratischen Fehler als der beste erwartungstreue Schätzer $\hat{\theta}_3 := \frac{n+1}{n} X_{(n)}$ hat, nämlich

$$\mathrm{MSE}_\theta(\hat{\theta}_1) = \frac{2\theta^2}{(n+2)(n+1)}, \quad \mathrm{MSE}_\theta(\hat{\theta}_3) = \frac{\theta^2}{3n},$$

KAPITEL 5

Asymptotische Eigenschaften von Schätzern

Im Folgenden definieren wir eine Reihe von asymptotischen Eigenschaften, die ein „guter“ Schätzer besitzen sollte.

5.1. Konsistenz und asymptotische Normalverteiltheit

In der Statistik ist der Stichprobenumfang n typischerweise groß. Wir schauen uns deshalb die asymptotischen Güteeigenschaften von Schätzern an. Wir betrachten eine Folge von Schätzern

$$\hat{\theta}_1(X_1), \hat{\theta}_2(X_1, X_2), \ldots, \hat{\theta}_n(X_1, \ldots, X_n), \ldots.$$

Sei im Folgenden Θ eine Teilmenge von $\mathbb{R}^m$.

Definition 5.1.1. Eine Folge von Schätzern $\hat{\theta}_n : \mathbb{R}^n \to \Theta$ heißt *asymptotisch erwartungstreu*, falls

$$\lim_{n\to\infty} \mathbb{E}_\theta \hat{\theta}_n(X_1, \ldots, X_n) = \theta \text{ für alle } \theta \in \Theta.$$

Beispiel 5.1.2. In Beispiel 4.1.6 ist $X_{(n)}$ eine asymptotisch erwartungstreue (aber nicht erwartungstreue) Folge von Schätzern, denn (Übungsaufgabe)

$$\lim_{n\to\infty} \mathbb{E}_\theta X_{(n)} = \lim_{n\to\infty} \frac{n}{n+1}\theta = \theta.$$

Definition 5.1.3. Eine Folge von Schätzern $\hat{\theta}_n : \mathbb{R}^n \to \Theta$ heißt *schwach konsistent*, falls

$$\hat{\theta}_n(X_1, \ldots, X_n) \xrightarrow[n\to\infty]{P} \theta \text{ unter } \mathbb{P}_\theta \text{ für alle } \theta \in \Theta.$$

Mit anderen Worten, für jedes $\varepsilon > 0$ und jedes $\theta \in \Theta$ soll gelten:

$$\lim_{n\to\infty} \mathbb{P}_\theta[|\hat{\theta}_n(X_1, \ldots, X_n) - \theta| > \varepsilon] = 0.$$

Definition 5.1.4. Eine Folge von Schätzern $\hat{\theta}_n : \mathbb{R}^n \to \Theta$ heißt *stark konsistent*, falls

$$\hat{\theta}_n(X_1, \ldots, X_n) \xrightarrow[n\to\infty]{f.s.} \theta \text{ unter } \mathbb{P}_\theta \text{ für alle } \theta \in \Theta.$$

Mit anderen Worten, es soll für alle $\theta \in \Theta$ gelten:

$$\mathbb{P}_\theta \left[\lim_{n\to\infty} \hat{\theta}_n(X_1, \ldots, X_n) = \theta \right] = 1.$$

Bemerkung 5.1.5. Eine fast sicher konvergente Folge von Zufallsvariablen konvergiert auch in Wahrscheinlichkeit. Aus der starken Konsistenz folgt somit die schwache Konsistenz.

Definition 5.1.6. Eine Folge von Schätzern $\hat{\theta}_n : \mathbb{R}^n \to \Theta$ heißt L^2-*konsistent*, falls

$$\hat{\theta}_n(X_1, \ldots, X_n) \xrightarrow{L^2} \theta \text{ für alle } \theta \in \Theta.$$

Mit anderen Worten, es soll für alle $\theta \in \Theta$ gelten:

$$\lim_{n\to\infty} \mathbb{E}_\theta |\hat{\theta}_n(X_1, \ldots, X_n) - \theta|^2 = 0.$$

Bemerkung 5.1.7. Aus der L^2-Konsistenz folgt die schwache Konsistenz.

Beispiel 5.1.8. Bei vielen Familien von Verteilungen, z.B. Bern(θ), Poi(θ) oder N(θ, σ^2) stimmt der Parameter θ mit dem Erwartungswert der entsprechenden Verteilung überein. In diesem Fall ist die Folge von Schätzern $\hat{\theta}_n = \bar{X}_n$ stark konsistent, denn für jedes θ gilt

$$\bar{X}_n \xrightarrow[n\to\infty]{f.s.} \mathbb{E}_\theta X_1 = \theta \text{ unter } \mathbb{P}_\theta$$

nach dem starken Gesetz der großen Zahlen.

Definition 5.1.9. Eine Folge von Schätzern $\hat{\theta}_n : \mathbb{R}^n \to \Theta \subset \mathbb{R}$ heißt *asymptotisch normalverteilt*, wenn es zwei Folgen $a_n(\theta) \in \mathbb{R}$ und $b_n(\theta) > 0$ gibt, sodass für alle $\theta \in \Theta$

$$\frac{\hat{\theta}_n(X_1, \ldots, X_n) - a_n(\theta)}{b_n(\theta)} \xrightarrow[n\to\infty]{d} \mathrm{N}(0,1) \text{ unter } \mathbb{P}_\theta.$$

Normalerweise wählt man $a_n(\theta) = \mathbb{E}_\theta \hat{\theta}(X_1, \ldots, X_n)$ und $b_n^2(\theta) = \mathrm{Var}_\theta\, \hat{\theta}(X_1, \ldots, X_n)$, sodass die Bedingung folgendermaßen lautet: Für alle $\theta \in \Theta$

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

Beispiel 5.1.10. Seien $X_1, X_2, \ldots$ unabhängige identisch verteilte Zufallsvariablen mit $\mathbb{E}X_i = \mu$ und $\operatorname{Var} X_i = \sigma^2 > 0$. Der zentrale Grenzwertsatz besagt, dass

$$\frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} \xrightarrow[n\to\infty]{d} \mathrm{N}(0,1).$$

Das kann man auch wie folgt schreiben:

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

Somit ist der Schätzer $\bar{X}_n$ asymptotisch normalverteilt.

Bemerkung 5.1.11. Viele Schätzer, die in der Statistik benutzt werden, sind asymptotisch normalverteilt. Für den Maximum-Likelihood-Schätzer werden wir das in Satz 5.2.6 zeigen. Für die empirischen Quantile $X_{([\alpha n])}$, $\alpha \in (0,1)$, werden wir das in Satz 5.4.1 beweisen. Auch das getrimmte Mittel und das winsorisierte Mittel sind (unter allgemeinen Bedingungen) asymptotisch normalverteilt.

5.2. Güteeigenschaften des ML-Schätzers

In diesem Abschnitt sei $\Theta = (a,b)$ ein Intervall. Sei $\{h_\theta : \theta \in \Theta\}$ eine Familie von Dichten oder Zähldichten und sei $\theta_0 \in \Theta$ fest. Seien $X, X_1, X_2, \ldots$ unabhängige und identisch verteilte Zufallsvariablen mit Dichte h_{θ_0}, wobei θ_0 als der „wahre Wert des Parameters“ aufgefasst wird. Uns ist der wahre Wert allerdings unbekannt und wir schätzen ihn mit dem Maximum-Likelihood-Schätzer $\hat{\theta}_{ML} = \hat{\theta}_{ML}(X_1, \ldots, X_n)$. In diesem Abschnitt wollen wir die Güteeigenschaften des Maximum-Likelihood-Schätzers untersuchen. Um die Güteeigenschaften von $\hat{\theta}_{ML}$ zu beweisen, muss man gewisse Regularitätsbedingungen an die Familie $\{h_\theta : \theta \in \Theta\}$ stellen. Leider sind diese Bedingungen nicht besonders schön. Deshalb werden wir nur die Ideen der jeweiligen Beweise zeigen. Wir werden hier nur eine der vielen Regularitätsbedingungen formulieren: Alle Dichten (oder Zähldichten) h_θ sollen den gleichen Träger haben, d.h. die Menge

$$J := \{x \in \mathbb{R} : h_\theta(x) \neq 0\}$$

soll nicht von θ abhängen.

Konsistenz des ML-Schätzers. Zuerst fragen wir, ob der Maximum-Likelihood-Schätzer stark konsistent ist, d.h. ob

$$\hat{\theta}_{ML}(X_1, \ldots, X_n) \xrightarrow[n\to\infty]{f.s.} \theta_0.$$

Wir werden zeigen, dass das stimmt. Hierfür betrachten wir die durch n geteilte log-Likelihood-Funktion

$$L_n(X_1, \ldots, X_n; \theta) := \frac{1}{n} \log L(X_1, \ldots, X_n; \theta) = \frac{1}{n} \sum_{i=1}^{n} \log h_\theta(X_i).$$

Nach dem Gesetz der großen Zahlen gilt

$$L_n(X_1, \ldots, X_n; \theta) \xrightarrow[n\to\infty]{f.s.} \mathbb{E}_{\theta_0} \log h_\theta(X) =: L_\infty(\theta) = \int_J \log h_\theta(t) h_{\theta_0}(t) \mathrm{d}t.$$

Lemma 5.2.1. Für alle $\theta \in \Theta$ gilt:

$$L_\infty(\theta) \leq L_\infty(\theta_0).$$

Beweis. Mit der Definition von $L_\infty(\theta)$ ergibt sich

$$L_\infty(\theta) - L_\infty(\theta_0) = \mathbb{E}_{\theta_0}[\log h_\theta(X) - \log h_{\theta_0}(X)] = \mathbb{E}_{\theta_0}\left[\log \frac{h_\theta(X)}{h_{\theta_0}(X)}\right].$$

Nun wenden wir auf die rechte Seite die Ungleichung $\log t \leq t - 1$ (wobei $t > 0$) an:

$$\begin{aligned} L_\infty(\theta) - L_\infty(\theta_0) &\leq \mathbb{E}_{\theta_0}\left[\frac{h_\theta(X)}{h_{\theta_0}(X)} - 1\right] \\ &= \int_J \left(\frac{h_\theta(t)}{h_{\theta_0}(t)} - 1\right) h_{\theta_0}(t)\mathrm{d}t \\ &= \int_J h_\theta(t)dt - \int_J h_{\theta_0}(t)\mathrm{d}t \\ &= 0, \end{aligned}$$

denn $\int_J h_\theta(t)\mathrm{d}t = \int_J h_{\theta_0}(t)\mathrm{d}t = 1$. □

Satz 5.2.2. Seien $X_1, X_2, \ldots$ unabhängige und identisch verteilte Zufallsvariablen mit Dichte oder Zähldichte h_{θ_0}. Unter Regularitätsbedingungen an die Familie $\{h_\theta : \theta \in \Theta\}$ gilt, dass

$$\hat{\theta}_{ML}(X_1, \ldots, X_n) \xrightarrow[n\to\infty]{f.s.} \theta_0.$$

Beweisidee. Per Definition des Maximum-Likelihood-Schätzers ist

$$\hat{\theta}_{ML}(X_1, \ldots, X_n) = \arg\max L_n(X_1, \ldots, X_n; \theta).$$

Indem wir zum Grenzwert für $n \to \infty$ übergehen, erhalten wir

$$\begin{aligned} \lim_{n\to\infty} \hat{\theta}_{ML}(X_1, \ldots, X_n) &= \lim_{n\to\infty} \arg\max L_n(X_1, \ldots, X_n; \theta) \\ &= \arg\max \lim_{n\to\infty} L_n(X_1, \ldots, X_n; \theta) \\ &= \arg\max L_\infty(X_1, \ldots, X_n; \theta) \\ &= \theta_0, \end{aligned}$$

wobei der letzte Schritt aus Lemma 5.2.1 folgt. □
Der obige Beweis ist nicht streng. Insbesondere bedarf der Schritt $\lim_{n\to\infty} \arg\max = \arg\max \lim_{n\to\infty}$ einer Begründung.

Asymptotische Normalverteiltheit des ML-Schätzers. Wir werden zeigen, dass unter gewissen Regularitätsbedingungen der Maximum-Likelihood-Schätzer asymptotisch normalverteilt ist:

$$\sqrt{n}(\hat{\theta}_{\mathrm{ML}}(X_1, \ldots, X_n) - \theta_0) \xrightarrow[n\to\infty]{d} \mathrm{N}(0, \sigma^2_{\mathrm{ML}}) \text{ unter } \mathbb{P}_{\theta_0},$$

wobei die Varianz σ^2_{ML} später identifiziert werden soll. Wir bezeichnen mit

$$l_\theta(x) = \log h_\theta(x)$$

die log-Likelihood einer einzelnen Beobachtung x. Die Ableitung nach θ wird mit

$$D = \frac{d}{d\theta}$$

bezeichnet. Insbesondere schreiben wir

$$Dl_\theta(x) = \frac{d}{d\theta} l_\theta(x), \quad D^2 l_\theta(x) = \frac{d^2}{d\theta^2} l_\theta(x).$$

Definition 5.2.3. Sei $\{h_\theta : \theta \in \Theta\}$, wobei $\Theta = (a, b)$ ein Intervall ist, eine Familie von Dichten oder Zähldichten. Die *Fisher-Information* ist eine Funktion $I : \Theta \to \mathbb{R}$ mit

$$I(\theta) = \mathbb{E}_\theta (Dl_\theta(X))^2.$$

Lemma 5.2.4. Unter Regularitätsbedingungen an die Familie $\{h_\theta : \theta \in \Theta\}$ gilt für jedes $\theta \in (a, b)$, dass

(1) $\mathbb{E}_\theta Dl_\theta(X) = 0$.
(2) $\mathbb{E}_\theta D^2 l_\theta(X) = -I(\theta)$.

Beweisidee. Für den Beweis formen wir zuerst $Dl_\theta(x)$ und $D^2 l_\theta(x)$ wie folgt um:

$$Dl_\theta(x) = D \log h_\theta(x) = \frac{Dh_\theta(x)}{h_\theta(x)},$$
$$D^2 l_\theta(x) = \frac{(D^2 h_\theta(x)) h_\theta(x) - (Dh_\theta(x))^2}{h_\theta^2(x)} = \frac{D^2 h_\theta(x)}{h_\theta(x)} - (Dl_\theta(x))^2.$$

Außerdem gilt für alle θ, dass $\int_J h_\theta(t)\mathrm{d}t = 1$, denn h_θ ist eine Dichte. Wir können nun diese Identität nach θ ableiten:

$$D \int_J h_\theta(t)\mathrm{d}t = 0 \text{ und somit } \int_J Dh_\theta(t)\mathrm{d}t = 0,$$
$$D^2 \int_J h_\theta(t)\mathrm{d}t = 0 \text{ und somit } \int_J D^2 h_\theta(t)\mathrm{d}t = 0.$$

Dabei haben wir die Ableitung und das Integral vertauscht, was unter gewissen Regularitätsbedingungen möglich ist. Mit diesen Resultaten erhalten wir, dass

$$\mathbb{E}_\theta Dl_\theta(X) = \int_J \frac{Dh_\theta(t)}{h_\theta(t)} h_\theta(t)\mathrm{d}t = \int_J Dh_\theta(t)\mathrm{d}t = 0.$$

Somit ist die erste Behauptung des Lemmas bewiesen. Die zweite Behauptung des Lemmas kann man wie folgt zeigen:

$$\begin{aligned}\mathbb{E}_\theta D^2 l_\theta(X) &= \int_J (D^2 l_\theta(t)) h_\theta(t) \mathrm{d}t \\ &= \int_J \left(\frac{D^2 h_\theta(t)}{h_\theta(t)} - (D l_\theta(t))^2 \right) h_\theta(t) \mathrm{d}t \\ &= \int_J D^2 h_\theta(t) \mathrm{d}t - \mathbb{E}_\theta (D l_\theta(X))^2 \\ &= -\mathbb{E}_\theta (D l_\theta(X))^2 \\ &= -I(\theta),\end{aligned}$$

wobei der letzte Schritt aus der Definition der Fisher-Information folgt. □
Indem wir die Notation $L_\infty(\theta) = \mathbb{E}_{\theta_0} \log h_\theta(X)$ verwenden, können wir das obige Lemma wie folgt formulieren.

Lemma 5.2.5. Unter Regularitätsbedingungen an die Familie $\{h_\theta : \theta \in \Theta\}$ gilt, dass
(1) $\mathbb{E}_{\theta_0} D l_{\theta_0}(X) = D L_\infty(\theta_0) = 0$.
(2) $\mathbb{E}_{\theta_0} D^2 l_{\theta_0}(X) = D^2 L_\infty(\theta_0) = -I(\theta_0)$.

Beweis. Unter Regularitätsbedingungen kann man den Erwartungswert $\mathbb{E}$ und die Ableitung D vertauschen. Somit gilt

$$D L_\infty(\theta_0) = D \mathbb{E}_{\theta_0} l_\theta(X)|_{\theta=\theta_0} = \mathbb{E}_{\theta_0} D l_{\theta_0}(X) = 0$$

und

$$D^2 L_\infty(\theta_0) = D^2 \mathbb{E}_{\theta_0} l_\theta(X)|_{\theta=\theta_0} = \mathbb{E}_{\theta_0} D^2 l_{\theta_0}(X) = -I(\theta_0).$$

□

Satz 5.2.6. Sei $\{h_\theta : \theta \in \Theta\}$ mit $\Theta = (a, b)$ eine Familie von Dichten oder Zähldichten. Sei $\theta_0 \in (a, b)$ und seien $X_1, X_2, \ldots$ unabhängige Zufallsvariablen mit Dichte oder Zähldichte h_{θ_0}. Unter Regularitätsbedingungen gilt für den Maximum-Likelihood-Schätzer $\hat{\theta}_{ML}(X_1, \ldots, X_n)$, dass

$$\sqrt{n}(\hat{\theta}_{ML}(X_1, \ldots, X_n) - \theta_0) \xrightarrow[n\to\infty]{d} \mathrm{N}\left(0, \frac{1}{I(\theta_0)}\right) \text{ unter } \mathbb{P}_{\theta_0}.$$

Beweisidee. Schritt 1. Für den Maximum-Likelihood-Schätzer gilt

$$\hat{\theta}_{ML} = \arg\max_{\theta\in\Theta} \frac{1}{n} \sum_{i=1}^n \log h_\theta(X_i) = \arg\max_{\theta\in\Theta} L_n(\theta).$$

Somit gilt

$$D L_n(\hat{\theta}_{ML}) = 0.$$

Der Mittelwertsatz aus der Analysis besagt, dass wenn f eine differenzierbare Funktion auf einem Intervall $[x, y]$ ist, dann lässt sich ein c in diesem Intervall finden mit

$$f(y) = f(x) + f'(c)(y - x).$$

Wir wenden nun diesen Satz auf die Funktion $f(\theta) = DL_n(\theta)$ und auf das Intervall mit den Endpunkten θ_0 und $\hat{\theta}_{ML}$ an. Es lässt sich also ein ξ_n in diesem Intervall finden mit

$$0 = DL_n(\hat{\theta}_{ML}) = DL_n(\theta_0) + D^2L_n(\xi_n)(\hat{\theta}_{ML} - \theta_0).$$

Daraus ergibt sich, dass

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) = -\frac{\sqrt{n}DL_n(\theta_0)}{DL_n(\xi_n)}.$$

Schritt 2. Anwendung von Lemma 5.2.5 führt zu $\mathbb{E}_{\theta_0} Dl_{\theta_0}(X) = 0$. Somit gilt

$$\sqrt{n}DL_n(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Dl_{\theta_0}(X_i) - 0) = \frac{\sum_{i=1}^{n} Dl_{\theta_0}(X_i) - n\mathbb{E}_{\theta_0} Dl_{\theta_0}(X_i)}{\sqrt{n}}.$$

Indem wir nun den zentralen Grenzwertsatz anwenden, erhalten wir, dass

$$\sqrt{n}DL_n(\theta_0) \xrightarrow[n\to\infty]{d} N \sim \mathrm{N}(0, \mathrm{Var}_{\theta_0} Dl_{\theta_0}(X)) = \mathrm{N}(0, I(\theta_0)),$$

denn $\mathrm{Var}_{\theta_0} Dl_{\theta_0}(X) = I(\theta_0)$ nach Lemma 5.2.5.

Schritt 3. Da sich ξ_n zwischen $\hat{\theta}_{ML}$ und θ_0 befindet und $\lim_{n\to\infty} \hat{\theta}_{ML} = \theta_0$ wegen Satz 5.2.2 (Konsistenz) gilt, erhalten wir, dass $\lim_{n\to\infty} \xi_n = \theta_0$. Nach dem Gesetz der großen Zahlen gilt somit

$$D^2L_n(\xi_n) = \frac{1}{n} \sum_{i=1}^{n} D^2 l_{\xi_n}(X_i) \underset{n\to\infty}{\longrightarrow} \mathbb{E}_{\theta_0} D^2 l_{\theta_0}(X) = -I(\theta_0),$$

wobei wir im letzten Schritt Lemma 5.2.5 benutzt haben.

Schritt 4. Kombiniert man nun diese Eigenschaften, so führt dies zu

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) = -\frac{\sqrt{n}L_n(\theta_0)}{D^2L_n(\xi_n)} \xrightarrow[n\to\infty]{d} \frac{N}{I(\theta_0)} \sim \mathrm{N}\left(0, \frac{1}{I(\theta_0)}\right),$$

wobei $N \sim \mathrm{N}(0, I(\theta_0))$. □

Beispiel 5.2.7. In diesem Beispiel betrachten wir die Familie der Bernoulli-Verteilungen mit Parameter $\theta \in (0, 1)$. Die Zähldichte ist gegeben durch

$$h_\theta(1) = \theta, \quad h_\theta(0) = 1 - \theta.$$

Eine andere Schreibweise dafür ist diese:

$$h_\theta(x) = \theta^x (1 - \theta)^{1-x}, \quad x \in \{0, 1\}.$$

Die log-Likelihood einer einzelnen Beobachtung $x \in \{0, 1\}$ ist

$$l_\theta(x) = \log h_\theta(x) = x \log \theta + (1 - x) \log(1 - \theta).$$

Ableiten nach θ führt zu

$$Dl_\theta(x) = \frac{x}{\theta} - \frac{1 - x}{1 - \theta}, \quad D^2 l_\theta(x) = -\left(\frac{x}{\theta^2} + \frac{1 - x}{(1 - \theta)^2}\right).$$

Sei X eine mit Parameter θ Bernoulli-verteilte Zufallsvariable. Somit ist die Fisher-Information gegeben durch

$$I(\theta) = -\mathbb{E}_\theta D^2 l_\theta(X) = \mathbb{E}_\theta \left[\frac{X}{\theta^2} + \frac{1-X}{(1-\theta)^2}\right] = \frac{\theta}{\theta^2} + \frac{1-\theta}{(1-\theta)^2} = \frac{1}{\theta(1-\theta)}.$$

Seien nun $X_1, X_2, \ldots, X_n$ unabhängige, mit Parameter θ Bernoulli-verteilte Zufallsvariablen. Dann ist der Maximum-Likelihood-Schätzer für den Parameter θ gegeben durch

$$\hat{\theta}_{ML}(X_1, \ldots, X_n) = \frac{X_1 + \ldots + X_n}{n} = \bar{X}_n.$$

Mit Satz 5.2.6 erhalten wir die asymptotische Normalverteiltheit von $\hat{\theta}_{ML} = \bar{X}_n$:

$$\sqrt{n}(\bar{X}_n - \theta) \xrightarrow[n\to\infty]{d} \mathrm{N}(0, \theta(1-\theta)) \text{ unter } \mathbb{P}_\theta.$$

Diese Aussage können wir auch aus dem Zentralen Grenzwertsatz herleiten, denn

$$\sqrt{n}(\bar{X}_n - \theta) = \frac{X_1 + \ldots + X_n - n\theta}{\sqrt{n}} \xrightarrow[n\to\infty]{d} \mathrm{N}(0, \theta(1-\theta)) \text{ unter } \mathbb{P}_\theta,$$

da $\mathbb{E}X_i = \theta$ und $\operatorname{Var} X_i = \theta(1-\theta)$.

Beispiel 5.2.8. Nun betrachten wir ein Beispiel, in dem der Maximum-Likelihood-Schätzer *nicht* asymptotisch normalverteilt ist. Der Grund hierfür ist, dass eine Regularitätsbedingung verletzt ist. Im folgenden Beispiel sind nämlich die Träger der Verteilungen, die zu verschiedenen Werten des Parameters gehören, nicht gleich.
Wir betrachten die Familie der Gleichverteilungen auf den Intervallen der Form $[0, \theta]$ mit $\theta > 0$. Die Dichte ist gegeben durch

$$h_\theta(x) = \frac{1}{\theta} \mathbb{1}_{x \in [0,\theta]}.$$

Seien $X_1, X_2, \ldots$ unabhängige und auf dem Intervall $[0, \theta]$ gleichverteilte Zufallsvariablen. Der Maximum-Likelihood-Schätzer für θ ist gegeben durch

$$\hat{\theta}_{ML}(X_1, \ldots, X_n) = \max\{X_1, \ldots, X_n\} =: M_n.$$

Wir zeigen nun, dass dieser Schätzer nicht asymptotisch normalverteilt, sondern asymptotisch exponentialverteilt ist.

Satz 5.2.9. Es seien $X_1, X_2, \ldots$ unabhängige und auf dem Intervall $[0, \theta]$ gleichverteilte Zufallsvariablen. Dann gilt für $M_n = \max\{X_1, \ldots, X_n\}$, dass

$$n\left(1 - \frac{M_n}{\theta}\right) \xrightarrow[n\to\infty]{d} \mathrm{Exp}(1).$$

Beweis. Sei $x \geq 0$. Es gilt

$$\mathbb{P}\left[n\left(1 - \frac{M_n}{\theta}\right) > x\right] = \mathbb{P}\left[\frac{M_n}{\theta} < 1 - \frac{x}{n}\right] = \mathbb{P}\left[X_1 < \theta\left(1 - \frac{x}{n}\right), \ldots, X_n < \theta\left(1 - \frac{x}{n}\right)\right].$$

Für genügend großes n ist $0 \leq \theta(1 - \frac{x}{n}) \leq \theta$ und somit

$$\mathbb{P}\left[X_i < \theta\left(1 - \frac{x}{n}\right)\right] = 1 - \frac{x}{n},$$

denn $X_i \sim \mathrm{U}[0,\theta]$. Wegen der Unabhängigkeit von $X_1, \ldots, X_n$ erhalten wir, dass

$$\lim_{n\to\infty} \mathbb{P}\left[n\left(1-\frac{M_n}{\theta}\right) > x\right] = \lim_{n\to\infty}\left(1-\frac{x}{n}\right)^n = \mathrm{e}^{-x}.$$

Somit erhalten wir, dass

$$\lim_{n\to\infty} \mathbb{P}\left[n\left(1-\frac{M_n}{\theta}\right) \le x\right] = \begin{cases} \mathrm{e}^{-x}, & x \ge 0, \\ 0, & x < 0. \end{cases}$$

Für $x < 0$ ist der Grenzwert gleich 0, denn das Ereignis $M_n > \theta$ ist unmöglich. Daraus ergibt sich die zu beweisende Aussage. □

Beispiel 5.2.10. In diesem Beispiel betrachten wir die Familie der Exponentialverteilungen. Die Dichte der Exponentialverteilung mit Parameter $\theta > 0$ ist gegeben durch

$$h_\theta(x) = \theta \exp(-\theta x), \quad x > 0.$$

Die log-Likelihood einer einzelnen Beobachtung $x > 0$ ist

$$l_\theta(x) = \log h_\theta(x) = \log\theta - \theta x.$$

Zweimaliges Ableiten nach θ führt zu

$$D^2 l_\theta(x) = -\frac{1}{\theta^2}.$$

Das Ergebnis ist übrigens unabhängig von x. Sei $X \sim \mathrm{Exp}(\theta)$. Die Fisher-Information ist gegeben durch

$$I(\theta) = -\mathbb{E}_\theta D^2 l_\theta(X) = \frac{1}{\theta^2}, \quad \theta > 0.$$

Seien $X_1, X_2, \ldots$ unabhängige, mit Parameter θ exponentialverteilte Zufallsvariablen. Der Maximum-Likelihood-Schätzer (und auch der Momentenschätzer) ist in diesem Beispiel

$$\hat\theta_{ML} = \frac{1}{\bar X_n}$$

Mit Satz 5.2.6 erhalten wir die asymptotische Normalverteiltheit von $\hat\theta_{ML}$:

$$\sqrt{n}\left(\frac{1}{\bar X_n} - \theta\right) \xrightarrow[n\to\infty]{d} \mathrm{N}(0,\theta^2) \text{ unter } \mathbb{P}_\theta. \tag{5.2.1}$$

Auf der anderen Seite, ergibt sich aus dem zentralen Grenzwertsatz, dass

$$\sqrt{n}\left(\bar X_n - \frac{1}{\theta}\right) \xrightarrow[n\to\infty]{d} \mathrm{N}\left(0, \frac{1}{\theta^2}\right) \text{ unter } \mathbb{P}_\theta, \tag{5.2.2}$$

denn $\mathbb{E}X_i = \frac{1}{\theta}$ und $\mathrm{Var}\, X_i = \frac{1}{\theta^2}$.

Sind nun (5.2.1) und (5.2.2) äquivalent? Etwas allgemeiner kann man auch fragen: Wenn ein Schätzer asymptotisch normalverteilt ist, muss dann auch eine Funktion von diesem Schätzer asymptotisch normalverteilt sein? Wir werden nun zeigen, dass unter gewissen Voraussetzungen die Antwort positiv ist.

Lemma 5.2.11. Seien $Z_1, Z_2, \ldots$ Zufallsvariablen und $\mu \in \mathbb{R}$ und $\sigma^2 > 0$ Zahlen mit

$$\sqrt{n}(Z_n - \mu) \xrightarrow[n\to\infty]{d} \mathrm{N}(0, \sigma^2).$$

Außerdem sei φ eine differenzierbare Funktion mit $\varphi'(\mu) \neq 0$. Dann gilt:

$$\sqrt{n}(\varphi(Z_n) - \varphi(\mu)) \xrightarrow[n\to\infty]{d} \mathrm{N}(0, (\varphi'(\mu)\sigma)^2).$$

Beweisidee. Durch die Taylorentwicklung von φ um den Punkt μ gilt

$$\varphi(Z_n) = \varphi(\mu) + \varphi'(\mu)(Z_n - \mu) + \text{Rest}.$$

Multipliziert man nun beide Seiten mit $\sqrt{n}$, so führt dies zu

$$\sqrt{n}(\varphi(Z_n) - \varphi(\mu)) = \varphi'(\mu)\sqrt{n}(Z_n - \mu) + \text{Rest}.$$

Nach Voraussetzung gilt für den ersten Term auf der rechten Seite, dass

$$\varphi'(\mu)\sqrt{n}(Z_n - \mu) \xrightarrow[n\to\infty]{d} \mathrm{N}(0, (\varphi'(\mu)\sigma)^2).$$

Der Restterm hat eine kleinere Ordnung als dieser Term, geht also gegen 0. Daraus folgt die Behauptung. □

Beispiel 5.2.12. Als Spezialfall von Lemma 5.2.11 mit $\varphi(x) = \frac{1}{x}$ und $\varphi'(x) = -\frac{1}{x^2}$ ergibt sich die folgende Implikation:

$$\sqrt{n}(Z_n - \mu) \xrightarrow[n\to\infty]{d} \mathrm{N}(0, \sigma^2) \Longrightarrow \sqrt{n}\left(\frac{1}{Z_n} - \frac{1}{\mu}\right) \xrightarrow[n\to\infty]{d} \mathrm{N}\left(0, \frac{\sigma^2}{\mu^4}\right).$$

Daraus ergibt sich die Äquivalenz von (5.2.1) und (5.2.2).

5.3. Cramér-Rao-Schranke

Sei $\{h_\theta(x) : \theta \in \Theta\}$, wobei $\Theta = (a, b)$, eine Familie von Dichten oder Zähldichten. Wir haben bereits gesehen, dass es mehrere erwartungstreue Schätzer für den Parameter θ geben kann. Unter diesen Schätzern versucht man einen Schätzer mit einer möglichst kleinen Varianz zu finden. Kann man vielleicht sogar für jedes vorgegebene $\varepsilon > 0$ einen erwartungstreuen Schätzer konstruieren, dessen Varianz kleiner als ε ist? Der nächste Satz zeigt, dass die Antwort negativ ist. Er gibt eine untere Schranke an die Varianz eines erwartungstreuen Schätzers.

Regularitätsbedingungen:

(a) $\Theta \subset \mathbb{R}$ ist ein (möglicherweise unendliches) Intervall.
(b) Der Träger $J := \{x \in \mathbb{R} \colon h_\theta(x) > 0\}$ ist unabhängig von θ.
(c) Die Ableitungen $\frac{\partial}{\partial\theta} h_\theta(x)$ und $\frac{\partial^2}{\partial\theta^2} h_\theta(x)$ existieren.
(c) Die Ableitungen $\frac{\partial}{\partial\theta} h_\theta(x)$ existieren für alle $(x, \theta) \in J \times \Theta$.

Satz 5.3.1 (Cramér-Rao)**.** Sei $\{h_\theta(x) : \theta \in \Theta\}$, wobei $\Theta = (a, b)$, eine Familie von Dichten oder Zähldichten. Seien weiterhin $X, X_1, X_2, \ldots$ unabhängige und identisch verteilte

Zufallsvariablen mit Dichte $h_\theta(x)$. Sei $\hat{\theta}(X_1, \ldots, X_n)$ ein erwartungstreuer Schätzer für θ. Unter Regularitätsbedingungen gilt die folgende Ungleichung:

$$\operatorname{Var}_\theta \hat{\theta}(X_1, \ldots, X_n) \geq \frac{1}{nI(\theta)}.$$

Beweisidee. Da $\hat{\theta}$ ein erwartungstreuer Schätzer ist, gilt für alle $\theta \in (a, b)$, dass

$$\theta = \mathbb{E}_\theta \hat{\theta}(X_1, \ldots, X_n) = \int_{\mathbb{R}^n} \hat{\theta}(x_1, \ldots, x_n) h_\theta(x_1) \ldots h_\theta(x_n) \mathrm{d}x_1 \ldots \mathrm{d}x_n.$$

Nun leiten wir nach θ ab:

$$\begin{aligned} 1 &= D \int_{\mathbb{R}^n} \hat{\theta}(x_1, \ldots, x_n) h_\theta(x_1) \ldots h_\theta(x_n) \mathrm{d}x_1 \ldots \mathrm{d}x_n \\ &= \int_{\mathbb{R}^n} \hat{\theta}(x_1, \ldots, x_n) D[h_\theta(x_1) \cdot \ldots \cdot h_\theta(x_n)] \mathrm{d}x_1 \ldots \mathrm{d}x_n. \end{aligned}$$

Indem wir nun die Formel $D \log f(\theta) = \frac{Df(\theta)}{f(\theta)}$ mit $f(\theta) = h_\theta(x_1) \cdot \ldots \cdot h_\theta(x_n)$ benutzen, erhalten wir, dass

$$\begin{aligned} 1 &= \int_{\mathbb{R}^n} \hat{\theta}(x_1, \ldots, x_n) h_\theta(x_1) \ldots h_\theta(x_n) \left(\sum_{i=1}^n D \log h_\theta(x_i) \right) \mathrm{d}x_1 \ldots \mathrm{d}x_n \\ &= \mathbb{E}_\theta[\hat{\theta}(X_1, \ldots, X_n) U_\theta(X_1, \ldots, X_n)], \end{aligned}$$

wobei

$$U_\theta(x_1, \ldots, x_n) := \sum_{i=1}^n D \log h_\theta(x_i).$$

Es sei bemerkt, dass U_θ die Ableitung der log-Likelihood-Funktion ist. Für den Erwartungswert von U_θ gilt nach Lemma 5.2.4, dass

$$\mathbb{E}_\theta U_\theta(X_1, \ldots, X_n) = \sum_{i=1}^n \mathbb{E}_\theta D \log h_\theta(X_i) = 0.$$

Für die Varianz von U_θ erhalten wir wegen der Unabhängigkeit von $X_1, \ldots, X_n$, dass

$$\mathbb{E}_\theta[U_\theta^2(X_1, \ldots, X_n)] = \operatorname{Var}_\theta U_\theta(X_1, \ldots, X_n) = \sum_{i=1}^n \operatorname{Var}_\theta[D \log h_\theta(X_i)] = nI(\theta),$$

denn

$$\operatorname{Var}_\theta[D \log h_\theta(X_i)] = \mathbb{E}_\theta(D \log h_\theta(X_i))^2 = I(\theta),$$

da $\mathbb{E}_\theta D \log h_\theta(X_i) = 0$ nach Lemma 5.2.4.
Nun erweitern wir mit dem Erwartungswert und wenden die Cauchy-Schwarz-Ungleichung an:

$$\begin{aligned} 1 &= \mathbb{E}_\theta[(\hat{\theta}(X_1, \ldots, X_n) - \theta) \cdot U_\theta(X_1, \ldots, X_n)] \\ &\leq \sqrt{\operatorname{Var}_\theta[\hat{\theta}(X_1, \ldots, X_n)] \cdot \operatorname{Var}_\theta[U_\theta(X_1, \ldots, X_n)]} \\ &= \sqrt{\operatorname{Var}_\theta[\hat{\theta}(X_1, \ldots, X_n)] \cdot nI(\theta)}. \end{aligned}$$

Umgestellt führt dies zu $\operatorname{Var}_\theta \hat{\theta}(X_1, \ldots, X_n) \geq \frac{1}{nI(\theta)}$. □

Definition 5.3.2. Ein erwartungstreuer Schätzer $\hat{\theta} : \mathbb{R}^n \to \Theta$ heißt *Cramér-Rao-effizient*, falls für jedes $\theta \in \Theta$

$$\operatorname{Var}_\theta \hat{\theta}(X_1, \ldots, X_n) = \frac{1}{nI(\theta)}.$$

Beispiel 5.3.3. Es seien $X_1, \ldots, X_n$ unabhängig und Bernoulli-verteilt mit Parameter θ. Der Maximum-Likelihood-Schätzer und gleichzeitig der Momentenschätzer für θ ist der empirische Mittelwert:

$$\hat{\theta}(X_1, \ldots, X_n) = \bar{X}_n = \frac{X_1 + \ldots + X_n}{n}.$$

Die Varianz von $\hat{\theta}$ lässt sich wie folgt berechnen

$$\operatorname{Var}_\theta \hat{\theta}(X_1, \ldots, X_n) = \operatorname{Var}_\theta \left[\frac{X_1 + \ldots + X_n}{n} \right] = \frac{n}{n^2} \operatorname{Var}_\theta X_1 = \frac{\theta(1-\theta)}{n} = \frac{1}{nI(\theta)},$$

denn wir haben in Beispiel 5.2.7 gezeigt, dass $I(\theta) = \frac{1}{\theta(1-\theta)}$. Somit ist der Schätzer $\bar{X}_n$ Cramér-Rao-effizient. Es ist also unmöglich, einen erwartungstreuen Schätzer mit einer kleineren Varianz als die von $\bar{X}_n$ zu konstruieren. Somit ist $\bar{X}_n$ der beste erwartungstreue Schätzer für den Parameter der Bernoulli-Verteilung.

Aufgabe 5.3.4. Zeigen Sie, dass der Schätzer $\hat{\theta} = \bar{X}_n$ für die folgenden Familien von Verteilungen Cramér-Rao-effizient ist:

(1) $\{\mathrm{Poi}(\theta) : \theta > 0\}$.
(2) $\{\mathrm{N}(\theta, \sigma^2) : \theta \in \mathbb{R}\}$, wobei σ^2 bekannt ist.

Somit ist $\bar{X}_n$ in beiden Fällen der beste erwartungstreue Schätzer.

Aufgabe 5.3.5. Es seien $X_1, \ldots, X_n$ unabhängig mit $X_i \sim \mathrm{Bin}(m, \theta)$, wobei $m \in \mathbb{N}$ bekannt und $\theta \in [0, 1]$ der zu schätzende unbekannte Parameter sei. Zeigen Sie, dass der Schätzer $\hat{\theta} = \frac{1}{m}\bar{X}_n$ Cramér-Rao-effizient (und somit der beste erwartungstreue Schätzer) ist.

Aufgabe 5.3.6. Sei $n \geq 3$ und $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit $X_i \sim \mathrm{Exp}(\theta)$, wobei $\theta > 0$ unbekannt sei.

(a) Zeigen Sie, dass $\bar{X}_n$ ein erwartungstreuer Schätzer für $1/\theta$ ist.
(b) Zeigen Sie, dass $1/\bar{X}_n$ *kein* erwartungstreuer Schätzer für θ ist und bestimmen Sie eine Konstante c (in Abhängigkeit von n), so dass $c/\bar{X}_n$ ein erwartungstreuer Schätzer für θ ist.
(c) Zeigen Sie, dass $c/\bar{X}_n$ nicht Cramér-Rao-effizient ist.
(d) Zeigen Sie, dass $c/\bar{X}_n$ der beste erwartungstreue Schätzer für θ ist.

Bemerkung 5.3.7. Der Maximum-Likelihood-Schätzer muss nicht immer Cramér-Rao-effizient (und sogar nicht einmal erwartungstreu) sein. Wir wollen allerdings zeigen, dass

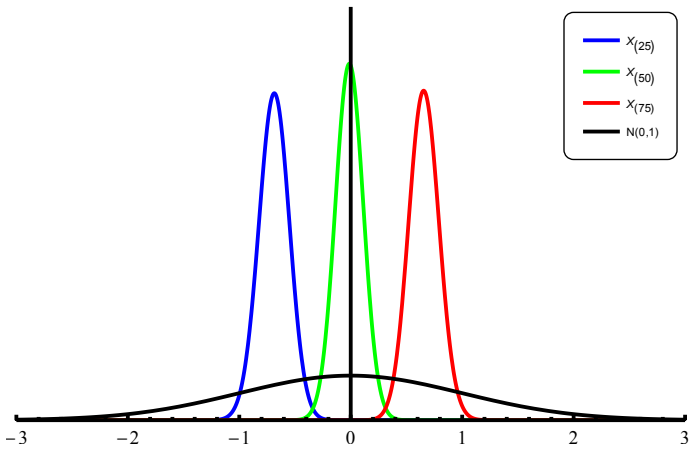


ABBILDUNG 1. Dichten der Ordnungsstatistiken $X_{(25)}, X_{(50)}, X_{(75)}$ einer unabhängigen und standardnormalverteilten Stichprobe $X_1, \ldots, X_{100}$. Die schwarze Kurve ist die Dichte der Standardnormalverteilung.

bei einem großen Stichprobenumfang n der Maximum-Likelihood-Schätzer $\hat{\theta}_{ML}$ die Cramér-Rao-Schranke asymptotisch erreicht. Nach Satz 5.2.6 gilt unter $\mathbb{P}_\theta$, dass

$$\sqrt{n}(\hat{\theta}_{ML} - \theta) \xrightarrow[n\to\infty]{d} \mathrm{N}\left(0, \frac{1}{I(\theta)}\right).$$

Die Verteilung von $\hat{\theta}_{ML}$ ist also approximativ $\mathrm{N}(\theta, \frac{1}{nI(\theta)})$ und die Varianz ist approximativ $\frac{1}{nI(\theta)}$, bei großem n. Somit nähert sich der Maximum-Likelihood-Schätzer der Cramér-Rao-Schranke asymptotisch an.

5.4. Asymptotische Normalverteiltheit der empirischen Quantile

Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit Dichte $h(x)$ und Verteilungsfunktion $F(x)$. Sei $\alpha \in (0,1)$. Das theoretische α-Quantil Q_α der Verteilungsfunktion F ist definiert als die Lösung der Gleichung $F(Q_\alpha) = \alpha$. Wir nehmen an, dass es eine eindeutige Lösung gibt. Das empirische α-Quantil der Stichprobe $X_1, \ldots, X_n$ ist $X_{([\alpha n])}$, wobei $X_{(1)} < \ldots < X_{(n)}$ die Ordnungsstatistiken von $X_1, \ldots, X_n$ seien. Wir haben hier die Definition des empirischen Quantils etwas vereinfacht, alle nachfolgenden Ergebnisse stimmen aber auch für die alte Definition. Das empirische Quantil $X_{[\alpha n]}$ ist ein Schätzer für das theoretische Quantil Q_α. Wir zeigen nun, dass dieser Schätzer asymptotisch normalverteilt ist.

Satz 5.4.1. Sei $\alpha \in (0,1)$ und seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit Dichte h, wobei h stetig in einer Umgebung von Q_α sei und $h(Q_\alpha) > 0$ gelte. Dann gilt

$$\sqrt{n}(X_{([\alpha n])} - Q_\alpha) \xrightarrow[n\to\infty]{d} \mathrm{N}\left(0, \frac{\alpha(1-\alpha)}{h^2(Q_\alpha)}\right).$$

Beweisidee. Sei $t \in \mathbb{R}$. Wir betrachten die Verteilungsfunktion

$$F_n(t) := \mathbb{P}[\sqrt{n}(X_{([\alpha n])} - Q_\alpha) \leq t] = \mathbb{P}\left[X_{([\alpha n])} \leq Q_\alpha + \frac{t}{\sqrt{n}}\right] = \mathbb{P}[K_n \geq \alpha n],$$

wobei die Zufallsvariable K_n wie folgt definiert wird:

$$K_n = \sum_{i=1}^{n} \mathbb{1}_{X_i \leq Q_\alpha + \frac{t}{\sqrt{n}}} = \#\left\{i \in \{1, \ldots, n\} : X_i \leq Q_\alpha + \frac{t}{\sqrt{n}}\right\}.$$

Somit ist $K_n \sim \mathrm{Bin}(n, F(Q_\alpha + \frac{t}{\sqrt{n}}))$. Für den Erwartungswert und die Varianz von K_n gilt somit

$$\mathbb{E}K_n = nF\left(Q_\alpha + \frac{t}{\sqrt{n}}\right), \quad \mathrm{Var}\, K_n = nF\left(Q_\alpha + \frac{t}{\sqrt{n}}\right)\left(1 - F\left(Q_\alpha + \frac{t}{\sqrt{n}}\right)\right).$$

Indem wir nun die Taylor-Entwicklung der Funktion F benutzen, erhalten wir, dass

$$\begin{aligned}\mathbb{E}K_n &= n\left(F(Q_\alpha) + F'(Q_\alpha)\frac{t}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right)\right) = \alpha n + \sqrt{n}h(Q_\alpha)t + o(\sqrt{n}),\\ \mathrm{Var}\, K_n &= n\alpha(1-\alpha) + o(n).\end{aligned}$$

Nun kann die Verteilungsfunktion $F_n(t)$ wie folgt berechnet werden

$$F_n(t) = \mathbb{P}[K_n \geq \alpha n] = \mathbb{P}\left[\frac{K_n - \mathbb{E}[K_n]}{\sqrt{\mathrm{Var}(K_n)}} \geq \frac{\alpha n - \mathbb{E}[K_n]}{\sqrt{\mathrm{Var}(K_n)}}\right].$$

Benutzen wir nun die Entwicklungen von $\mathbb{E}K_n$ und $\mathrm{Var}\, K_n$, so erhalten wir

$$F_n(t) = \mathbb{P}\left[\frac{K_n - \mathbb{E}[K_n]}{\sqrt{\mathrm{Var}(K_n)}} \geq \frac{\alpha n - \alpha n - \sqrt{n}h(Q_\alpha)t - o(\sqrt{n})}{\sqrt{n\alpha(1-\alpha) + o(n)}}\right].$$

Und nun benutzen wir den zentralen Grenzwertsatz für K_n. Mit N standardnormalverteilt, erhalten wir, dass

$$\lim_{n\to\infty} F_n(t) = \mathbb{P}\left[N \geq -\frac{h(Q_\alpha)}{\sqrt{\alpha(1-\alpha)}}t\right] = \mathbb{P}\left[\frac{N\sqrt{\alpha(1-\alpha)}}{h(Q_\alpha)} \leq t\right],$$

wobei wir im letzten Schritt die Symmetrie der Standardnormalverteilung, also die Formel $\mathbb{P}[N \geq -x] = \mathbb{P}[N \leq x]$ benutzt haben.

Die Behauptung des Satzes folgt nun aus der Tatsache, dass $\frac{N\sqrt{\alpha(1-\alpha)}}{h(Q_\alpha)} \sim \mathrm{N}\left(0, \frac{\alpha(1-\alpha)}{h^2(Q_\alpha)}\right)$. □

Bemerkung 5.4.2. Satz 5.4.1 behandelt nur „zentrale“ Ordnungsstatistiken $X_{([\alpha n])}$, $\alpha \in (0,1)$, und macht keine Aussage über $X_{(n)} = \max_{i=1,\ldots,n} X_i$ und $X_{(1)} = \min_{i=1,\ldots,n} X_i$. Die Ordnungsstatistiken $X_{(n)}$ und $X_{(1)}$ sind *nicht* asymptotisch normalverteilt. Die möglichen Grenzwertverteilungen von $X_{(n)}$ und $X_{(1)}$ heißen *Extremwertverteilungen*, für deren Beschreibung verweisen wir auf das Skript „Extremwerttheorie“.

5.5. Asymptotische relative Effizienz

In vielen statistischen Problemen lassen sich mehrere natürliche Schätzer für den unbekannten Parameter konstruieren. Wie soll man entscheiden, welcher Schätzer besser ist? Handelt es sich um zwei Folgen von asymptotisch normalverteilten Schätzern, so ist es natürlich, denjenigen Schätzer zu bevorzugen, der eine kleinere asymptotische Varianz hat.

Definition 5.5.1. Seien $\hat{\theta}_n^{(1)}$ und $\hat{\theta}_n^{(2)}$ zwei Folgen von Schätzern für einen Parameter θ, wobei n den Stichprobenumfang bezeichnet. Beide Schätzer seien asymptotisch normalverteilt mit

$$\sqrt{n}(\hat{\theta}_n^{(1)} - \theta) \xrightarrow[n\to\infty]{d} \mathrm{N}(0, \sigma_1^2(\theta)) \text{ und } \sqrt{n}(\hat{\theta}_n^{(2)} - \theta) \xrightarrow[n\to\infty]{d} \mathrm{N}(0, \sigma_2^2(\theta)) \text{ unter } \mathbb{P}_\theta.$$

Die *asymptotische relative Effizienz* der beiden Schätzer ist definiert durch

$$\mathrm{ARE}_{\hat{\theta}_n^{(1)}, \hat{\theta}_n^{(2)}}(\theta) = \frac{\sigma_2^2(\theta)}{\sigma_1^2(\theta)}.$$

Bemerkung 5.5.2. Ist z.B. $\mathrm{ARE}_{\hat{\theta}_n^{(1)}, \hat{\theta}_n^{(2)}}(\theta) > 1$ so heißt es, dass unter $\mathbb{P}_\theta$ der Schätzer $\hat{\theta}_n^{(1)}$ besser als $\hat{\theta}_n^{(2)}$ ist.

Als Beispiel werden wir den Median und den Mittelwert als Schätzer für den Lageparameter einer Dichte vergleichen. Sei $h(x)$, $x \in \mathbb{R}$, eine Dichte. Wir machen folgende Annahmen:

(1) h ist symmetrisch, d.h. $h(x) = h(-x)$.
(2) h ist stetig in einer Umgebung von 0.
(3) $h(0) > 0$.

Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit Dichte

$$h_\theta(x) = h(x - \theta),$$

wobei $\theta \in \mathbb{R}$ der unbekannte Lageparameter ist. Die Aufgabe besteht nun darin, θ zu schätzen. Dabei sei die Dichte h bekannt. Da sowohl der theoretische Median $Q(\frac{1}{2})$ als auch der Erwartungswert der Dichte h_θ wegen der Symmetrie gleich θ sind, können wir folgende natürliche Schätzer für θ betrachten:

(1) den empirischen Median $X_{([n/2])}$ (dessen Deinition wir hier etwas vereinfacht haben).
(2) den empirischen Mittelwert $\bar{X}_n$.

Welcher Schätzer ist nun besser und um wieviel? Wir beantworten dieser Frage mit Hilfe des Begriffes der asymptotischen relativen Effizienz.

Nach Satz 5.4.1 und nach dem zentralen Grenzwertsatz sind beide Schätzer asymptotisch normalverteilt mit

$$\sqrt{n}(X_{([n/2])} - \theta) \xrightarrow[n\to\infty]{d} \mathrm{N}\left(0, \frac{1}{4h^2(0)}\right) \text{ unter } \mathbb{P}_\theta, \tag{5.5.1}$$

$$\sqrt{n}(\bar{X}_n - \theta) \xrightarrow[n\to\infty]{d} \mathrm{N}(0, \mathrm{Var}_\theta X_1) \text{ unter } \mathbb{P}_\theta. \tag{5.5.2}$$

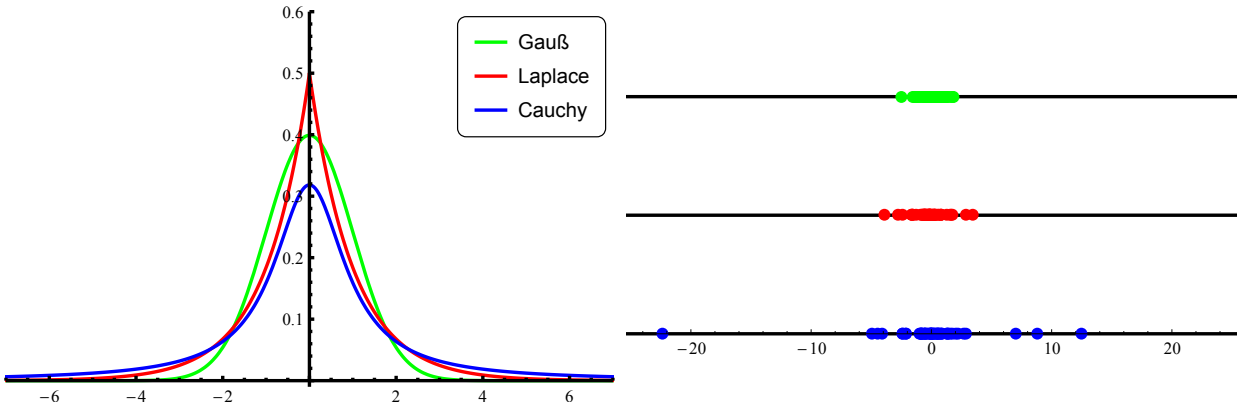


ABBILDUNG 2. Links: Die drei Dichten: Gauß-Verteilung (grün), Laplace-Verteilung (rot), Cauchy-Verteilung (blau). Rechts: Stichproben vom Umfang $n = 50$ aus den drei Verteilungen.

Die asymptotischen Varianzen sind also unabhängig von θ und gegeben durch

$$\sigma_1^2 = \frac{1}{4h^2(0)}, \quad \sigma_2^2 = \mathrm{Var}_\theta\, X_1 = \int_{\mathbb{R}} x^2 h^2(x)\mathrm{d}x.$$

Nun betrachten wir drei Beispiele.

Beispiel 5.5.3. Die Gauß-Dichte $h(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}$, $x \in \mathbb{R}$. Es gilt $h(0) = \frac{1}{\sqrt{2\pi}}$, $\mathrm{Var}_\theta\, X_1 = 1$ und somit

$$\sigma_1^2 = \frac{\pi}{2}, \quad \sigma_2^2 = 1, \quad \mathrm{ARE}_{\mathrm{Med,MW}} = \frac{2}{\pi} \approx 0.6366 < 1.$$

Der empirische Mittelwert ist also besser als der empirische Median. Das Ergebnis kann man wie folgt interpretieren: Der Median erreicht bei einer Stichprobe vom Umfang 100 in etwa die gleiche Präzision, wie der Mittelwert bei einer Stichprobe vom Umfamg 64.

Beispiel 5.5.4. Die Laplace-Dichte $h(x) = \frac{1}{2}\mathrm{e}^{-|x|}$, $x \in \mathbb{R}$. Es gilt $h(0) = \frac{1}{2}$, $\mathrm{Var}_\theta\, X_1 = 2$ und somit

$$\sigma_1^2 = 1, \quad \sigma_2^2 = 2, \quad \mathrm{ARE}_{\mathrm{Med,MW}} = 2 > 1.$$

In diesem Beispiel ist also der Median besser. Bei einer Stichprobe vom Umfang 100 erreicht der Median in etwa die gleiche Präzision, wie der Mittelwert bei einer Stichprobe vom Umfang 200.

Beispiel 5.5.5. Die Cauchy-Dichte $h(x) = \frac{1}{\pi}\frac{1}{1+x^2}$, $x \in \mathbb{R}$. Es gilt $h(0) = \frac{1}{\pi}$ und somit ist die asymptotische Varianz des Medians gegeben durch

$$\sigma_1^2 = \frac{\pi^2}{4}.$$

Der Mittelwert ist aber gar nicht asymptotisch normalverteilt, da die Cauchy-Verteilung keinen wohldefinierten Erwartungswert (und auch keine Varianz) besitzt und der zentrale Grenzwertsatz (auf dem (5.5.2) basiert) nicht anwendbar ist!

Aufgabe 5.5.6. Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit der Cauchy-Dichte

$$h_\theta(x) = \frac{1}{\pi} \frac{1}{1 + (x - \theta)^2}, \quad x \in \mathbb{R}, \theta \in \mathbb{R}.$$

Zeigen Sie, dass die Dichte des Mittelwerts $\bar{X}_n$ ebenfalls durch $h_\theta(x)$ gegeben ist. *Hinweis:* Berechnen Sie die charakteristische Funktion von $\bar{X}_n$.

Aufgabe 5.5.7. Zeigen Sie, dass $\bar{X}_n$ *kein* schwach konsistenter Schätzer für θ ist.

Aus Aufgabe 5.5.6 folgt, dass sich die Qualität des Schätzers $\bar{X}_n$ bei wachsendem Stichprobenumfang nicht verbessert: $\bar{X}_n$ hat die gleiche Verteilung wie X_1. Der statistische Fehler $\bar{X}_n - \theta$ hat Größenordnung 1 für jedes n. Dabei hat der Fehler $X_{([n/2])} - \theta$ die Größenordnung $1/\sqrt{n}$, siehe (5.5.1). In diesem Beispiel ist also der Mittelwert unendlich viel schechter als der Median! Symbolisch können wir das zusammenfassen als

$$\mathrm{ARE}_{\mathrm{Med,MW}} = +\infty.$$

Die obigen drei Beispiele zeigen, dass die Antwort auf die Frage, welcher Schätzer (der Median oder der Mittelwert) besser ist, von der zugrundeliegenden Verteilung abhängt. Als Kompromisslösung, die bei allen möglichen h's brauchbare Ergebnisse liefert, kann man den sogenannten *Hodges-Lehmann-Schätzer* für den Lageparameter betrachten:

$$\mathrm{HL} := \mathrm{Median} \left\{ \frac{X_i + X_j}{2} \right\}_{1 \le i < j \le n}.$$

Man kann zeigen, dass HL einem ähnlichen zentralen Grenzwertsatz wie der Median genügt und dass die asymptotische Varianz durch

$$\sigma^2_{\mathrm{HL}} = \frac{1}{12(\int_{\mathbb{R}} h^2(x)\mathrm{d}x)^2}$$

gegeben ist. Im Fall der Normalverteilung ergibt sich

$$\mathrm{ARE}_{\mathrm{HL,MW}} = \frac{3}{\pi} \approx 0.955.$$

Das heißt, HL ist nur unwesentlich schlechter als der Mittelwert. Außerdem kann man zeigen, dass für jede symmetrische Dichte h mit endlicher Varianz

$$e_{\mathrm{HL,MW}} \ge \frac{108}{125} = 0.864.$$

Das heißt, HL ist niemals wesentlich schlechter als der Mittelwert. Dabei gibt es Verteilungen (wie die Cauchy-Verteilung), bei denen HL unendlich viel besser als der Mittelwert ist.

Aufgabe 5.5.8. Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit $X_i \sim \mathrm{Poi}(\lambda)$, wobei $\lambda > 0$ ein unbekannter Parameter sei. Betrachten Sie die folgenden beiden Schätzer für $\theta := \mathrm{e}^{-\lambda} = \mathbb{P}_\theta[X_i = 0]$:

$$\hat{\theta}^{(1)}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i = 0\}}, \quad \hat{\theta}^{(2)}_n = \mathrm{e}^{-\overline{X}_n}.$$

Sind diese Schätzer erwartungstreu/asymptotisch erwartungstreu? Zeigen Sie, dass beide Schätzer asymptotisch normal verteilt sind und bestimmen Sie die asymptotische relative Effizienz. Welcher Schätzer ist im Sinne der asymptotischen relativen Effizienz besser?

Aufgabe 5.5.9. Seien $X_1, X_2, \ldots$ unabhängige Zufallsvariablen mit der Dichte $h(x-\theta)$, wobei h die sogenannte *logistische Dichte*

$$h(x) = \frac{1}{(\mathrm{e}^{-x/2} + \mathrm{e}^{+x/2})^2}, \quad x \in \mathbb{R},$$

bezeichne und $\theta \in \mathbb{R}$ der unbekannte Lageparameter sei. Zeigen Sie, dass der Hodges-Lehmann-Schätzer (genauso wie der Maximum-Likelihood-Schätzer) asymptotisch die Cramér-Rao-Schranke erreicht.

Aufgabe 5.5.10. Seien $X_1, \ldots, X_n \sim \mathrm{N}(0, \sigma^2)$ unabhängig mit einem unbekannten σ^2. Zeigen Sie, dass die folgenden beiden Schätzer für σ

$$\hat{\sigma}_n^{(1)} := \sqrt{\frac{\pi}{2}} \frac{1}{n} \sum_{i=1}^{n} |X_i|, \quad \hat{\sigma}_n^{(2)} := \sqrt{\frac{1}{n} \sum_{i=1}^{n} X_i^2},$$

stark konsistent und asymptotisch normalverteilt sind und bestimmen Sie die asymptotische relative Effizienz. Welcher Schätzer ist asymptotisch besser?

KAPITEL 6

Statistische Entscheidungstheorie

6.1. Verlustfunktion, Risiko, Minimax-Schätzer

Es sei $(\mathfrak{X}, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ ein statistisches Modell. Das heißt, $\mathfrak{X}$ ist die Menge aller möglichen Stichproben, $\mathcal{A}$ ist eine σ-Algebra auf $\mathfrak{X}$ und $(\mathbb{P}_\theta)_{\theta\in\Theta}$ ist eine Familie von Wahrscheinlichkeitsmaßen auf $(\mathfrak{X}, \mathcal{A})$. Nun wird eine Stichprobe X gemäß einem Wahrscheinlichkeitsmaß $\mathbb{P}_\theta$ zufällig aus $\mathfrak{X}$ gezogen, wobei $\theta \in \Theta$ unbekannt bleibt. Wir konstruieren einen Schätzer $\hat\theta : \mathfrak{X} \to \Theta$ und versuchen, den Parameter $\theta \in \Theta$ durch $\hat\theta(X)$ zu schätzen. Stellen wir uns nun vor, dass wir für den Fehler, den wir dabei typischerweise machen, eine Strafe der Größe $D(\theta, \hat\theta(X))$ auferlegt bekommen, wobei $D : \Theta \times \Theta \to [0, \infty)$ eine vorgegebene Funktion ist, die *Verlustfunktion* genannt wird. Natürliche Beispiele von Verlustfunktionen sind:

(1) quadratische Verlustfunktion: $D(\theta) = \|\hat\theta - \theta\|^2$;
(2) absoluter Fehler: $D(\theta) = \|\hat\theta - \theta\|$;
(3) L^p-Verlust: $D(\theta) = \|\hat\theta - \theta\|^p$;
(4) Null-Eins-Verlust: $D(\theta) = \mathbb{1}_{\{\hat\theta\neq\theta\}}$,

wobei wir in den ersten drei Fällen voraussetzen, dass $\Theta \subset \mathbb{R}^m$ ist, und $\|\hat\theta - \theta\|$ den Euklidischen Abstand zwischen $\hat\theta$ und θ bezeichnet.

Definition 6.1.1. Das *Risiko* eines Schätzers $\hat\theta$ ist

$$R(\theta; \hat\theta) = \mathbb{E}_\theta D(\theta, \hat\theta(X)).$$

Im Folgenden werden wir sehr oft $\hat\theta$ als eine Abkürzung für die Zufallsvariable $\hat\theta(X)$ benutzen.

Beispiel 6.1.2. Für die quadratische Verlustfunktion $D(\theta, \hat\theta) = (\hat\theta - \theta)^2$ (wobei $\Theta \subset \mathbb{R}$ vorausgesetzt wird) stimmt das Risiko mit dem mittleren quadratischen Fehler überein:

$$R(\theta; \hat\theta) = \mathbb{E}_\theta(\hat\theta - \theta)^2 = \mathrm{MSE}_\theta(\hat\theta) = \mathrm{Var}_\theta\,\hat\theta + (\mathrm{Bias}_\theta\,\hat\theta)^2.$$

Man kann das Risiko benutzen, um verschiedene Schätzer miteinander zu vergleichen: je kleiner das Risiko, umso besser der Schätzer.

Definition 6.1.3. Wir sagen, dass ein Schätzer $\hat\theta_1$ *gleichmäßig besser* als ein anderer Schätzer $\hat\theta_2$ ist, wenn

$$R(\theta; \hat\theta_1) \le R(\theta; \hat\theta_2) \text{ für alle } \theta \in \Theta.$$

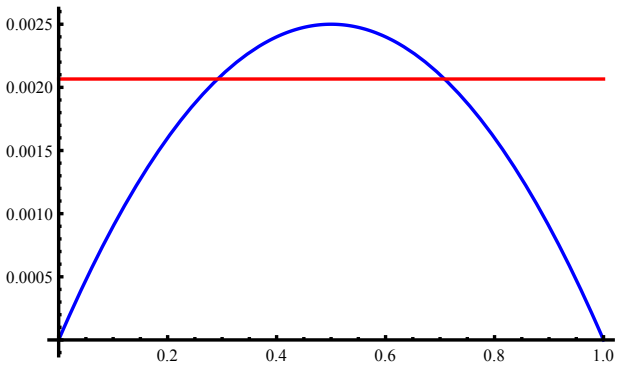


ABBILDUNG 1. Die Risikofunktionen $R(\theta;\hat{\theta}_1)$ (blau) und $R(\theta;\hat{\theta}_2)$ (rot) aus Beispiel 6.1.6. Der Stichprobenumfang ist $n = 100$.

Es kann aber durchaus passieren, dass $R(\theta';\hat{\theta}_1) < R(\theta';\hat{\theta}_2)$ für ein gewisses $\theta' \in \Theta$ und $R(\theta'';\hat{\theta}_1) > R(\theta'';\hat{\theta}_2)$ für ein anderes $\theta'' \in \Theta$. In diesem Fall ist kein Schätzer gleichmäßig besser als der andere. Es wäre deshalb schön, die Qualität eines Schätzers durch eine Zahl (und nicht durch eine Funktion von θ) charakterisieren zu können. Dazu gibt es zwei natürliche Ansätze. Zuerst betrachten wir den **Minimax-Ansatz**.

Definition 6.1.4. Das *maximale Risiko* eines Schätzers $\hat{\theta}$ ist

$$M(\hat{\theta}) = \sup_{\theta\in\Theta} R(\theta;\hat{\theta}).$$

Definition 6.1.5. Ein Schätzer $\hat{\theta}$ heißt *Minimax-Schätzer*, wenn das maximale Risiko von $\hat{\theta}$ nicht größer ist, als das maximale Risiko jedes anderen Schätzers $\tilde{\theta}$, d.h. wenn

$$\sup_{\theta\in\Theta} R(\theta;\hat{\theta}) = \inf_{\tilde{\theta}} \sup_{\theta\in\Theta} R(\theta;\tilde{\theta}).$$

Beispiel 6.1.6. Seien $X_1,\dots,X_n$ unabhängige, mit Parameter $\theta \in (0,1)$ Bernoulli-verteilte Zufallsvariablen. Wir benutzen die quadratische Verlustfunktion $D(\theta,\hat{\theta}) = (\hat{\theta}-\theta)^2$. Wir werden nun zeigen, dass der Mittelwert $\hat{\theta}_1 := \bar{X}_n$ erstaunlicherweise *kein* Minimax-Schätzer für θ ist. Die Risikofunktion von $\hat{\theta}_1$ ist gegeben durch

$$R(\theta;\hat{\theta}_1) = \mathrm{Var}_\theta\,\hat{\theta}_1 + (\mathrm{Bias}_\theta\,\hat{\theta}_1)^2 = \mathrm{Var}_\theta\,\hat{\theta}_1 = \frac{\theta(1-\theta)}{n}.$$

Für das maximale Risiko von $\hat{\theta}_1$ erhalten wir somit

$$M(\hat{\theta}) = \sup_{\theta\in(0,1)} \frac{\theta(1-\theta)}{n} = \frac{1}{4n}.$$

Betrachte nun den Schätzer

$$\hat{\theta}_2 = \frac{S_n + \alpha}{\alpha + \beta + n},$$

wobei $S_n = X_1 + \ldots + X_n$ und die Konstanten $\alpha, \beta > 0$ noch zu wählen sind. Dieser Schätzer ist der Bayes-Schätzer für θ wenn die a-priori-Verteilung von θ eine Beta(α, β)-Verteilung ist, vgl. Beispiel 3.6.7. Das Risiko von $\hat{\theta}_2$ ist

$$R(\theta; \hat{\theta}_2) = \operatorname{Var}_\theta \hat{\theta}_1 + (\operatorname{Bias}_\theta \hat{\theta}_1)^2 = \frac{n\theta(1-\theta)}{(\alpha+\beta+n)^2} + \left(\frac{n\theta + \alpha}{\alpha+\beta+n} - \theta\right)^2.$$

Wir wollen nun α und β so wählen, dass die Funktion $R(\theta; \hat{\theta}_2)$ nicht von θ abhängt. (Der Grund dafür wird später ersichtlich sein, siehe Satz 6.3.3). Nach einer einfachen Rechnung kann man sich überzeugen, dass dies für $\alpha = \beta = \frac{1}{2}\sqrt{n}$ der Fall ist. Der Schätzer $\hat{\theta}_2$ und sein Risiko sehen in diesem Fall wie folgt aus:

$$\hat{\theta}_2 = \frac{S_n + \frac{1}{2}\sqrt{n}}{n + \sqrt{n}}, \quad R(\theta; \hat{\theta}_2) = \frac{n}{4(n+\sqrt{n})^2} < \frac{1}{4n}.$$

Somit ist $\hat{\theta}_1$ kein Minimax-Schätzer. Später werden wir zeigen, dass $\hat{\theta}_2$ der Minimax-Schätzer ist. Es sei aber bemerkt, dass die Menge aller θ, für die $\hat{\theta}_2$ besser als $\hat{\theta}_2$ ist, ein (bei großem n) sehr kleines Intervall um $\frac{1}{2}$ ist und dass der Unterschied zwischen den Risiken von θ_1 und θ auf diesem Intervall sehr gering ist. Für alle anderen θ's ist der konventionelle Schätzer $\hat{\theta}_1$ besser.

Aufgabe 6.1.7. Welche Länge hat das Intervall $\{\theta \in (0,1) \colon R(\theta; \hat{\theta}_1) > R(\theta, \hat{\theta}_2)\}$?

6.2. Bayes-Schätzer

Nun betrachten wir den **Bayes-Ansatz** zur Charakterisierung des Risikos eines Schätzers. Wir nehmen an, dass für den Parameter θ eine a-priori-Verteilung, also ein Wahrscheinlichkeitsmaß auf Θ, vorgegeben ist. Der Einfachheit halber beschränken wir uns auf den Fall, wenn $\Theta \subset \mathbb{R}^m$ eine messbare Menge mit positivem Lebesgue-Maß ist und die a-priori-Verteilung von θ eine Lebesgue-Dichte $q : \Theta \to \mathbb{R}_+$ besitzt.

Definition 6.2.1. Das *Bayes-Risiko* eines Schätzers $\hat{\theta}$ unter der a-priori-Verteilung q ist gegeben durch

$$B_q(\hat{\theta}) = \int_\Theta R(\theta; \hat{\theta}) q(\theta) \mathrm{d}\theta.$$

Definition 6.2.2. Ein Schätzer $\hat{\theta}$ heißt der *Bayes-Schätzer* (unter der a-priori-Verteilung q), wenn das Bayes-Risiko von $\hat{\theta}$ nicht größer ist, als das Bayes-Risiko jedes anderen

Schätzers $\tilde{\theta}$, d.h. wenn

$$B_q(\hat{\theta}) = \inf_{\tilde{\theta}} B_q(\tilde{\theta}).$$

Annahme: Es gibt ein σ-endliches Maß λ auf $(\mathfrak{X}, \mathcal{A})$, sodass alle Wahrscheinlichkeitsmaße $\mathbb{P}_\theta$, $\theta \in \Theta$, absolut stetig bzgl. λ sind. Die Dichte von $\mathbb{P}_\theta$ bzgl. λ heißt die Likelihood-Funktion und wird mit $L(x;\theta)$ bezeichnet.

Wir leiten nun eine alternative Formel für das Bayes-Risiko $B_q(\hat{\theta})$ her. Stellen wir uns vor, dass die Stichprobe $x \in \mathfrak{X}$ bekannt ist. Nach dem Bekanntwerden von x ändert sich unsere Vorstellung über die Verteilung von θ: Die a-posteriori-Dichte von θ bei bekannter Stichprobe x ist gegeben durch

$$q(\theta|x) = \frac{q(\theta)L(x;\theta)}{m(x)}, \quad \text{wobei } m(x) = \int_\Theta L(x;t)q(t)\mathrm{d}t. \tag{6.2.1}$$

Dabei ist $m(x)$ die Dichte (bzgl. λ), dass eine Stichprobe $x \in \mathfrak{X}$ bei einem zufälligen und gemäß der a-priori-Dichte q verteilten Parameter θ beobachtet wird.

Definition 6.2.3. Schätzen wir bei einer gegebenen Stichprobe $x \in \mathfrak{X}$ den Parameter θ durch einen Wert $a \in \Theta$, so ist das dadurch entstehende *a-posteriori-Risiko* gegeben durch

$$r(a|x) = \int_\Theta D(\theta, a)q(\theta|x)\mathrm{d}\theta.$$

Insbesondere definieren wir das *a-posteriori-Risiko eines Schätzers* $\hat{\theta} : \mathfrak{X} \to \Theta$ gegeben die Stichprobe $x \in \mathfrak{X}$ als

$$r(\hat{\theta}|x) = r(\hat{\theta}(x)|x) = \int_\Theta D(\theta, \hat{\theta}(x))q(\theta|x)\mathrm{d}\theta.$$

Der nächste Satz besagt, dass wir das Bayes-Risiko berechnen können, indem wir das a-posteriori-Risiko über alle möglichen Stichproben $x \in \mathfrak{X}$ integrieren, wobei der Beitrag der Stichprobe x mit der Dichte $m(x)$ gewichtet wird.

Satz 6.2.4. Für das Bayes-Risiko $B_q(\hat{\theta})$ gilt die Formel

$$B_q(\hat{\theta}) = \int_{\mathfrak{X}} r(\hat{\theta}|x)m(x)\lambda(\mathrm{d}x).$$

Beweis. Mit der Definition von $B_q(\hat{\theta})$ gilt

$$B_q(\hat{\theta}) = \int_\Theta \mathbb{E}_\theta D(\theta, \hat{\theta}(X))q(\theta)\mathrm{d}\theta = \int_\Theta \int_{\mathfrak{X}} D(\theta, \hat{\theta}(x))L(x;\theta)q(\theta)\lambda(\mathrm{d}x)\mathrm{d}\theta,$$

wobei wir die Formel

$$\mathbb{E}_\theta D(\theta, \hat{\theta}(X)) = \int_{\mathfrak{X}} D(\theta, \hat{\theta}(x)) L(x;\theta) \lambda(\mathrm{d}x)$$

benutzt haben. Indem wir nun (6.2.1) und danach den Satz von Fubini benutzen, erhalten wir

$$B_q(\hat{\theta}) = \int_\Theta \int_{\mathfrak{X}} D(\theta, \hat{\theta}(x)) q(\theta|x) m(x) \lambda(\mathrm{d}x) \mathrm{d}\theta = \int_{\mathfrak{X}} \int_\Theta D(\theta, \hat{\theta}(x)) q(\theta|x) m(x) \mathrm{d}\theta \lambda(\mathrm{d}x).$$

Mit Definition 6.2.3 ergibt sich die Behauptung des Satzes. □

Nun können wir den Bayes-Schätzer sogar berechnen. Bei einer gegebenen Stichprobe $x \in \mathfrak{X}$ muss man den Wert $a \in \Theta$ finden, der das a-posteriori-Risiko $r(a|x)$ minimiert. Dieser Wert ist dann der Bayes-Schätzer.

Satz 6.2.5. Für alle $x \in \mathfrak{X}$ sei

$$\hat{\theta}(x) = \arg\min_{a \in \Theta} r(a|x) = \arg\min_{a \in \Theta} \int_\Theta D(\theta, a) q(\theta|x) \mathrm{d}\theta.$$

Dann ist $\hat{\theta}$ ein Bayes-Schätzer von θ zur a-priori-Verteilung q.

Beweis. Aus der Definition von $\hat{\theta}(x)$ folgt, dass $r(a|x) \geq r(\hat{\theta}(x)|x)$ für alle $a \in \Theta$. Sei $\tilde{\theta} : \mathfrak{X} \to \Theta$ ein Schätzer von θ. Mit Satz 6.2.4 gilt

$$B_q(\tilde{\theta}) = \int_{\mathfrak{X}} r(\tilde{\theta}(x)|x) m(x) \lambda(\mathrm{d}x) \geq \int_{\mathfrak{X}} r(\hat{\theta}(x)|x) m(x) \lambda(\mathrm{d}x) = B_q(\hat{\theta}).$$

Somit ist $\hat{\theta}$ ein Bayes-Schätzer. □

Korollar 6.2.6. Sei $\Theta \subset \mathbb{R}$ ein Intervall und betrachte die quadratische Verlustfunktion $D(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$. Dann ist der Bayes-Schätzer gegeben durch den Erwartungswert der a-posteriori-Verteilung, d.h.

$$\hat{\theta}(x) = \int_\Theta \theta q(\theta|x) \mathrm{d}\theta.$$

Fassen wir θ als Zufallselement mit Werten in Θ und Dichte q, so können wir auch schreiben

$$\hat{\theta}(x) = \mathbb{E}[\theta | X = x].$$

Beweis. Gemäß Satz 6.2.5 müssen wir die folgende Funktion minimieren:

$$f(a) = \int_\Theta (\theta - a)^2 q(\theta|x) \mathrm{d}\theta = \mathbb{E}[(Z - a)^2], \quad a \in \Theta,$$

wobei Z eine Zufallsvariable mit der Dichte $q(\theta|x)$ sei. Es ist eine Übung zu zeigen, dass das Minimum von $\mathbb{E}[(Z-a)^2]$ für $a = \mathbb{E}Z$ erreicht wird. □

Korollar 6.2.7. Sei $\Theta \subset \mathbb{R}$ ein Intervall und betrachte die Verlustfunktion $D(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$. Dann ist der Bayes-Schätzer gegeben durch den Median der a-posteriori-Verteilung, d.h. $\hat{\theta}(x)$ ist die Lösung von

$$\int_{-\infty}^{\hat{\theta}(x)} q(\theta|x)\mathrm{d}\theta = \frac{1}{2},$$

wobei wir hier die Schwierigkeiten ignorieren, die wegen Nichtexistenz oder Nichteindeutigkeit der Lösung entstehen können.

Beweis. Laut Satz 6.2.5 müssen wir die folgende Funktion minimieren:

$$f(a) = \int_\Theta |\theta - a| q(\theta|x)\mathrm{d}\theta = \mathbb{E}|Z - a|, \quad a \in \Theta,$$

wobei Z eine Zufallsvariable mit der Dichte $q(\theta|x)$ sei. Das Minimum von $\mathbb{E}|Z - a|$ wird erreicht, wenn a der Median von Z ist (Übung). □

6.3. Konstruktion des Minimax-Schätzers

Nun werden wir einen Zusammenhang zwischen den Bayes-Schätzern und dem Minimax-Schätzer herstellen.

Satz 6.3.1. Sei $\hat{\theta}$ der Bayes-Schäzer, der einer a-priori-Verteilung $q(\theta)$ entspricht. Falls

$$R(\theta; \hat{\theta}) \leq B_q(\hat{\theta}) \text{ für alle } \theta \in \Theta,$$

dann ist $\hat{\theta}$ ein Minimax-Schätzer.

Beweis. Sei $\hat{\theta}$ nicht minimax. Dann gibt es einen anderen Schätzer $\hat{\theta}_0$ mit

$$\sup_{\theta\in\Theta} R(\theta; \hat{\theta}_0) < \sup_{\theta\in\Theta} R(\theta; \hat{\theta}).$$

Nun ist aber der Erwartungswert einer Zufallsvariable immer kleiner als ihr Maximum:

$$B_q(\hat{\theta}_0) = \int_\Theta R(\theta; \hat{\theta}_0) q(\theta)\mathrm{d}\theta \leq \sup_{\theta\in\Theta} R(\theta; \hat{\theta}_0).$$

Es folgt aus den obigen Ungleichungen, dass

$$B_q(\hat{\theta}_0) < \sup_{\theta\in\Theta} R(\theta; \hat{\theta}) \leq B_q(\hat{\theta}),$$

wobei wir die Voraussetzung benutzt haben. Dies ist ein Widerspruch, denn wir haben vorausgesetzt, dass $\hat{\theta}$ ein Bayes-Schätzer ist. □

Bemerkung 6.3.2. Die Verteilung q aus Satz 6.3.1 heißt die *ungünstigste a-priori-Verteilung* für den Schätzer $\hat{\theta}$, denn für jede andere a-priori-Verteilung $q_0(\theta)$ gilt

$$B_{q_0}(\hat{\theta}) = \int_\Theta R(\theta; \hat{\theta}) q_0(\theta)\mathrm{d}\theta \leq \int_\Theta B_q(\hat{\theta}) q_0(\theta)\mathrm{d}\theta = B_q(\hat{\theta}).$$

Somit ist die Qualität des Schätzers $\hat{\theta}$ unter der a-priori-Verteilung schlechter, als unter jeder anderen a-priori-Verteilung q_0.

Satz 6.3.3. Sei $\hat{\theta}$ ein Bayes-Schätzer für die a-priori-Verteilung $q(\theta)$ mit

$$C := R(\theta; \hat{\theta}) = \text{const für alle } \theta \in \Theta.$$

Dann ist $\hat{\theta}$ minimax.

Beweis. Für das Bayes-Risiko von $\hat{\theta}$ gilt

$$B_q(\hat{\theta}) = \int_\Theta R(\theta; \hat{\theta}) q(\theta) \mathrm{d}\theta = C \int_\Theta q(\theta) \mathrm{d}\theta = C \geq R(\theta; \hat{\theta}) \text{ für alle } \theta \in \Theta.$$

Nun folgt die Behauptung aus Satz 6.3.1. □

Beispiel 6.3.4. Seien $X_1, \ldots, X_n \sim \mathrm{Bern}(\theta)$ unabhängig, wobei $\theta \in [0, 1]$. Die Verlustfunktion sei quadratisch. Wir haben bereits gesehen, dass die Risikofunktion des Schätzers

$$\hat{\theta}_2 = \frac{S_n + \frac{1}{2}\sqrt{n}}{n + \sqrt{n}}$$

nicht von θ abhängt. Außerdem ist $\hat{\theta}_2$ der Bayes-Schätzer, wenn wir $\mathrm{Beta}(\frac{1}{2}\sqrt{n}, \frac{1}{2}\sqrt{n})$ als a-priori-Verteilung wählen, s. Beispiel 3.6.7. Somit ist $\hat{\theta}_2$ ein Minimax-Schätzer. Die ungünstigste a-priori-Verteilung ist in diesem Modell die Beta-Verteilung

$$\mathrm{Beta}\left(\frac{1}{2}\sqrt{n}, \frac{1}{2}\sqrt{n}\right).$$

Der konventionelle Schätzer $\bar{X}_n$ ist erstaunlicherweise nicht minimax für die quadratische Verlustfunktion. Damit $\bar{X}_n$ minimax wird, muss man eine unkonventionelle Verlustfunktion betrachten.

Aufgabe 6.3.5. Seien $X_1, \ldots, X_n \sim \mathrm{Bern}(\theta)$ unabhängig und identisch verteilt, wobei $\theta \in (0, 1)$. Betrachten Sie die Verlustfunktion

$$D(\theta, \hat{\theta}) = \frac{(\theta - \hat{\theta})^2}{\theta(1 - \theta)}.$$

(a) Bestimmen Sie das Risiko des Schätzers $\hat{\theta} = \overline{X}_n$ als Funktion von θ.
(b) Zeigen Sie, dass $\overline{X}_n$ der Bayes-Schätzer von θ für die a-priori-Dichte $q(\theta) = \mathbb{1}_{[0,1]}(\theta)$ ist.
(c) Zeigen Sie, dass $\overline{X}_n$ der Minimax-Schätzer von θ für die angegebene Verlustfunktion D ist.

Aufgabe 6.3.6. Seien $X_1, \ldots, X_n \sim \mathrm{N}(\theta, 1)$ unabhängige und identisch verteilte Zufallsvariablen, wobei $\theta \in \mathbb{R}$. Betrachten Sie die Verlustfunktion $D(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$. Zeigen Sie, dass $\bar{X}_n$ der Minimax-Schätzer von θ ist, wie folgt:

(a) Bestimmen Sie den Bayes-Schätzer $\tilde{\theta}$ von θ für die a-priori-Verteilung $\mathrm{N}(0, c^2)$, wobei $c > 0$ sei.

(b) Bestimmen Sie das Bayes-Risiko dieses Schätzers und zeigen Sie damit, dass fur einen beliebigen Schätzer $\hat{\theta}$ die folgende Ungleichung gilt:

$$\sup_{\theta\in\mathbb{R}} R(\theta;\hat{\theta}) \geq \frac{1}{n}.$$

(c) Folgern Sie, dass $\bar{X}_n$ der Minimax-Schätzer von θ ist.

Aufgabe 6.3.7. Sei $X \sim \mathrm{N}(\theta, 1)$ (d.h. die Stichprobe besteht aus einem Element), wobei der Parameterraum $\Theta := [-m, m]$ mit $0 < m < 1$ sei. Die Verlustfunktion sei quadratisch, d.h. $D(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$.

(a) Betrachten Sie die a-priori-Verteilung (keine Dichte!) μ mit $\mu(\{-m\}) = \mu(\{+m\}) = 1/2$ und beweisen Sie für den entsprechenden Bayes-Schätzer die Formel

$$\hat{\theta}(x) = m\tanh(mx), \qquad \text{wobei } \tanh z = \frac{\mathrm{e}^z - \mathrm{e}^{-z}}{\mathrm{e}^z + \mathrm{e}^{-z}}.$$

(b) Zeigen Sie, dass für diesen Schätzer $R(\theta;\hat{\theta}) \leq B_\mu(\hat{\theta})$ gilt und dass das Risiko $R(\theta;\hat{\theta})$ nicht konstant ist.

(c) Folgern Sie, dass $\hat{\theta}$ ein Minimax-Schätzer ist.

6.4. Statistik als Zweipersonenspiel

Es gibt eine interessante Interpretation der statistischen Entscheidungstheorie als ein Zweipersonenspiel. Man betrachte ein Spiel mit zwei Spielern A und B. Spieler A habe Strategien $A_1, \ldots, A_m$ zur Verfügung, während die möglichen Strategien des Spielers B mit $B_1, \ldots, B_n$ bezeichnet seien. In jeder Runde des Spiels wählt jeder Spieler unabhängig vom anderen Spieler eine der ihm zur Verfügung stehenden Strategien. Wählt A Strategie A_i und B Strategie B_j, so sei der Gewinn von A in der entsprechenden Runde gleich u_{ij} (wobei diese Zahl auch negativ sein darf). Das Spiel wird also durch die Matrix (u_{ij}) eindeutig beschrieben, die auch die *Auszahlungsmatrix* genannt wird. Das Ziel von A ist es, seinen Gewinn zu maximieren. Im Gegenteil ist Spieler B bestrebt, den Gewinn von A zu minimieren.

Beispiel 6.4.1. Beim Spiel „Schere, Stein, Papier, Brunnen“ hat jeder der beiden Spieler die 4 genannten Strategien zur Verfügung. Die Regeln sind wie folgt: Stein schlägt Schere, Papier schlägt Stein, Brunnen schlägt Stein, Schere schlägt Papier, Brunnen schlägt Schere und Papier schlägt Brunnen. Der Gewinn ist bei verschiedenen Symbolen gleich ± 1. Bei gleichen Symbolen ist der Gewinn 0. Die Auszahlungsmatrix sieht somit folgendermaßen aus:

		Spieler B			
		Schere	Stein	Papier	Brunnen
Spieler A	Schere	0	-1	$+1$	-1
	Stein	$+1$	0	-1	-1
	Papier	-1	$+1$	0	$+1$
	Brunnen	$+1$	$+1$	-1	0

Das Spiel besteht aus unendlich vielen Runden. Am obigen Beispiel sieht man, dass es ungünstig ist, sich für eine Strategie zu entscheiden, und diese immer zu verwenden (dieses

Vorgehen nennt man *reine Strategie*). Spielt z.B. einer der Spieler immer Schere, so merkt das der andere Spieler nach einigen Runden und antwortet dann mit Stein. Günstiger ist es, eine *gemischte Strategie* zu verwenden, bei der in jeder Runde eine der vier Strategien zufällig ausgewählt wird.

Definition 6.4.2. Eine gemischte Strategie von A ist ein Vektor $x = (x_1, \ldots, x_m)$ mit $x_1 + \ldots + x_m = 1$ und $x_1, \ldots, x_m \geq 0$. Eine gemischte Strategie von B ist ein Vektor $y = (y_1, \ldots, y_n)$ mit $y_1 + \ldots + y_n = 1$ und $y_1, \ldots, y_n \geq 0$.

Interpretation: A wählt Strategie A_i mit Wahrscheinlichkeit x_i; B wählt Strategie B_j mit Wahrscheinlichkeit y_j. Da die Spieler unabhängig voneinander agieren, ist der erwartete Gewinn von A gegeben durch

$$G(x, y) = \sum_{i=1}^{m} \sum_{j=1}^{n} u_{ij} x_i y_j.$$

Aufgabe 6.4.3. Nehmen wir an, B spielt die „naive“ gemischte Strategie $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. Welche gemischte Strategie ist für A optimal?

Wir betrachten nun das Spiel aus Sicht der beiden Spieler.

A denkt: Angenommen, ich wähle eine gemischte Strategie x. Nach genügend vielen Runden wird B meine Strategie erkennen und seinerseits eine gemischte Strategie y benutzen, die $G(x, y)$ minimiert. Ich muss also das folgende Optimierungsproblem lösen:

Bestimme x, das $\min_y G(x, y)$ maximiert.

B denkt: Angenommen, ich wähle eine gemischte Strategie y. Nach genügend vielen Runden wird A meine Strategie erkennen und seinerseits eine gemischte Strategie x benutzen, die $G(x, y)$ maximiert. Ich muss also das folgende Optimierungsproblem lösen:

Bestimme y, das $\max_x G(x, y)$ minimiert.

Es stellt sich heraus, dass die Lösungen der beiden Probleme, die *Maximin* und *Minimax* heißen, übereinstimmen.

Satz 6.4.4 (von Neumann, 1928)**.** Es gilt

$$\max_{\substack{x_1 \ldots, x_m \geq 0 \\ x_1 + \ldots + x_m = 1}} \min_{\substack{y_1 \ldots, y_n \geq 0 \\ y_1 + \ldots + y_n = 1}} G(x, y) = \min_{\substack{y_1 \ldots, y_n \geq 0 \\ y_1 + \ldots + y_n = 1}} \max_{\substack{x_1 \ldots, x_m \geq 0 \\ x_1 + \ldots + x_m = 1}} G(x, y).$$

Definition 6.4.5. Ein Paar von gemischten Strategien (x^*, y^*) heißt *Nash-Gleichgewicht*, wenn keiner der beiden Spieler seine Position durch einseitiges Abweichen von seiner gemischten Strategie verbessern kann, d.h.

- Für jede gemischte Strategie x von A gilt $G(x, y^*) \leq G(x^*, y^*)$.

- Für jede gemischte Strategie y von B gilt $G(x^*, y) \geq G(x^*, y^*)$.

Aufgabe 6.4.6. Zeigen Sie, dass im Spiel „Schere, Stein, Papier, Brunnen“ das Paar $x^* = y^* = (\frac{1}{3}, 0, \frac{1}{3}, \frac{1}{3})$ ein Nash-Gleichgewicht bildet.

Aufgabe 6.4.7. Zeigen Sie, dass im Spiel „Schere, Stein, Papier“ (ohne Brunnen) das Paar $x^* = y^* = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ ein Nash-Gleichgewicht bildet.

Nun können wir eine Interpretation der statistischen Entscheidungstheorie als Zweipersonenspiel beschreiben. Sei $(\mathfrak{X}, \mathcal{A}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ ein statistisches Modell und $D : \Theta \times \Theta \to [0, \infty)$ eine Verlustfunktion. Wir betrachten zwei Spieler, die als *Natur* und *Statistiker* bezeichnet werden. Das Spiel verläuft wie folgt. Die Natur wählt ein $\theta \in \Theta$. Gleichzeitig und unabhängig von der Natur wählt der Statistiker einen Schätzer $\hat{\theta} : \mathfrak{X} \to \Theta$. Nachdem das Paar $(\theta, \hat{\theta})$ gewählt wurde, wird eine Stichprobe X gemäß $\mathbb{P}_\theta$ zufällig gezogen und der Gewinn der Natur (bzw. der Verlust des Statistikers) ist $D(\theta, \hat{\theta}(X))$. Bei einem gegebenen Paar $(\theta, \hat{\theta})$ ist der erwartete Gewinn der Natur gegeben durch das Risiko

$$R(\theta; \hat{\theta}) = \mathbb{E}_\theta D(\theta, \hat{\theta}(X)).$$

Die Funktion R ist somit ein Analogon der Auszahlungsmatrix.

Die reinen Strategien der Natur sind alle möglichen Parameter $\theta \in \Theta$. Die Natur kann aber auch eine gemischte Strategie verwenden, die durch ein Wahrscheinlichkeitsmaß (zur Vereinfachung: Dichte q) auf dem Parameterraum Θ beschrieben wird. Antwortet der Statistiker auf eine solche gemischte Strategie mit einem Schätzer $\hat{\theta}$, so ist sein erwarteter Verlust nichts anderes als das Bayes-Risiko des Schätzers $\hat{\theta}$:

$$B_q(\hat{\theta}) = \int_\Theta R(\theta, \hat{\theta}) q(\theta) \mathrm{d}\theta.$$

Die beste Antwort des Statistikers ist also der Bayes-Schätzer.
Die reinen Strategien des Statistikers sind alle möglichen Schätzer. Es ist interessant, dass sich gemischte Strategien für den Statistiker gar nicht lohnen. Stellen wir uns z.B. vor, der Statistiker würde eine gemischte Strategie der folgenden Art einsetzen: Er sucht sich n Schätzer $\hat{\theta}_1, \ldots, \hat{\theta}_n$ aus und wählt dann den Schätzer $\hat{\theta}_j$ mit Wahrscheinlichkeit $y_j \geq 0$, wobei $y_1 + \ldots + y_n = 1$. Sein erwarteter Verlust wäre dann

$$y_1 B_q(\hat{\theta}_1) + \ldots + y_n B_q(\hat{\theta}_n) \geq \min_{j=1,\ldots,n} B_q(\hat{\theta}_j),$$

denn $(y_1, \ldots, y_n)$ ist ein Wahrscheinlichkeitsvektor. Der Statistiker hätte einfach den Schätzer $\hat{\theta}_j$ mit dem kleinsten Bayes-Risiko benutzen können, der dann mindestens genauso gut wie die gemischte Strategie wäre.

Nun betrachten wir das Geschehen aus Sicht der beiden Spieler.

Natur denkt: Ich wähle eine gemischte Strategie q. Der Statistiker, der ja rational handelt, wird diese Strategie nach mehreren Runden erkennen und mit dem entsprechenden Bayes-Schätzer antworten. Mein Gewinn ist dabei das Bayes-Risiko dieses Schätzers, das ich

maximieren möchte. Also muss ich als meine gemischte Strategie die ungünstigste a-priori-Verteilung auswählen!

Statistiker denkt: Gemischte Strategien lohnen sich für mich nicht. Also wähle ich eine reine Strategie, d.h. einen Schätzer $\hat{\theta}$. Die Natur wird diesen Schätzer nach mehreren Runden rekonstruieren können. Sie wird dann mit dem Wert θ antworten, der ihren Gewinn $R(\theta, \hat{\theta})$ maximiert. Mein Verlust ist also gegeben durch das maximale Risiko von $\hat{\theta}$:

$$M(\hat{\theta}) = \sup_{\theta \in \Theta} R(\theta; \hat{\theta}).$$

Somit muss ich denjenigen Schätzer $\hat{\theta}$ wählen, der den Wert $M(\hat{\theta})$ minimiert, also den Minimax-Schätzer!

Handeln beide Seiten rational, so muss das Spiel im Nash-Gleichgewicht verharren: Die Natur wählt θ's zufällig gemäß der ungünstigsten a-priori-Verteilung, während der Statistiker darauf mit dem entsprechenden Bayes-Schätzer antwortet, der gleichzeitig der Minimax-Schätzer ist, s. Abschnitt 6.3.

KAPITEL 7

Dichteschätzer

Angenommen wir beobachten eine Stichprobe $(x_1, \ldots, x_n)$, wie z.B. diese (wobei jeder Punkt der Stichprobe als ein vertikaler Strich dargestellt wird):

Wir gehen davon aus, dass $x_1, \ldots, x_n$ eine Realisierung von unabhängigen und identisch verteilten Zufallsvariablen $X_1, \ldots, X_n$ mit unbekannter Dichte f ist. In diesem Kapitel werden wir die Dichte f schätzen. Das Schwierige an diesem Problem ist, dass wir keine parametrischen Annahmen an die Dichte f machen wollen.

Zunächst einmal kann man die folgende Idee ausprobieren. Wir können die Verteilungsfunktion F unserer Zufallsvariablen durch die empirische Verteilungsfunktion $\widehat{F}_n$ schätzen. Die Dichte f ist die Ableitung der Verteilungsfunktion F. Somit können wir versuchen, die Dichte f durch die Ableitung von $\widehat{F}_n$ zu schätzen. Diese Idee funktioniert allerdings nicht, da die Funktion $\widehat{F}_n$ nicht differenzierbar (und sogar nicht stetig) ist. Man muss also andere Methoden benutzen.

7.1. Histogramm

Wir wollen nun das Histogramm einführen, das als ein sehr primitiver Schätzer für die Dichte aufgefasst werden kann. Sei $(x_1, \ldots, x_n) \in \mathbb{R}^n$ eine Stichprobe. Sei $c_0, \ldots, c_k$ eine aufsteigende Folge reeller Zahlen mit der Eigenschaft, dass die komplette Stichprobe $x_1, \ldots, x_n$ im Intervall (c_0, c_k) liegt. Typischerweise wählt man die Zahlen c_i so, dass die Abstände zwischen den aufeinanderfolgenden Zahlen gleich sind. In diesem Fall nennt man $h := c_i - c_{i-1}$ die *Bandbreite*.

Die Anzahl der Stichprobenvariablen x_j im Intervall $(c_{i-1}, c_i]$ wird mit n_i bezeichnet, somit gilt

$$n_i = \sum_{j=1}^{n} \mathbb{1}_{x_j \in (c_{i-1}, c_i]}, \quad i = 1, \ldots, k.$$

Teilt man n_i durch den Stichprobenumfang n, so führt dies zur *relativen Häufigkeit*

$$f_i = \frac{n_i}{n}.$$

Als *Histogramm* wird die graphische Darstellung dieser relativen Häufigkeiten bezeichnet, siehe Abbildung 1. Man konstruiert nämlich über jedem Intervall $(c_{i-1}, c_i]$ ein Rechteck mit

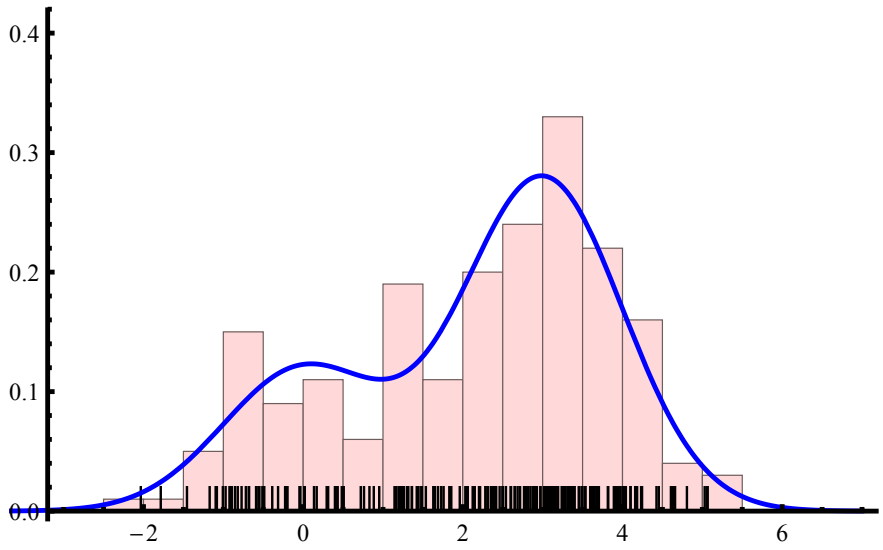


ABBILDUNG 1. Das Histogramm einer Stichprobe vom Umfang $n = 200$. Die glatte blaue Kurve ist die wahre Dichte f, die in der Praxis unbekannt ist.

dem Flächeninhalt f_i. Das Histogramm ist dann die Vereinigung dieser Rechtecke. Es ist offensichtlich, dass die Summe der relativen Häufigkeiten 1 ergibt, d.h.

$$\sum_{i=1}^{k} f_i = 1.$$

Das bedeutet, dass der Flächeninhalt unter dem Histogramm gleich 1 ist. Außerdem gilt $f_i \geq 0$.

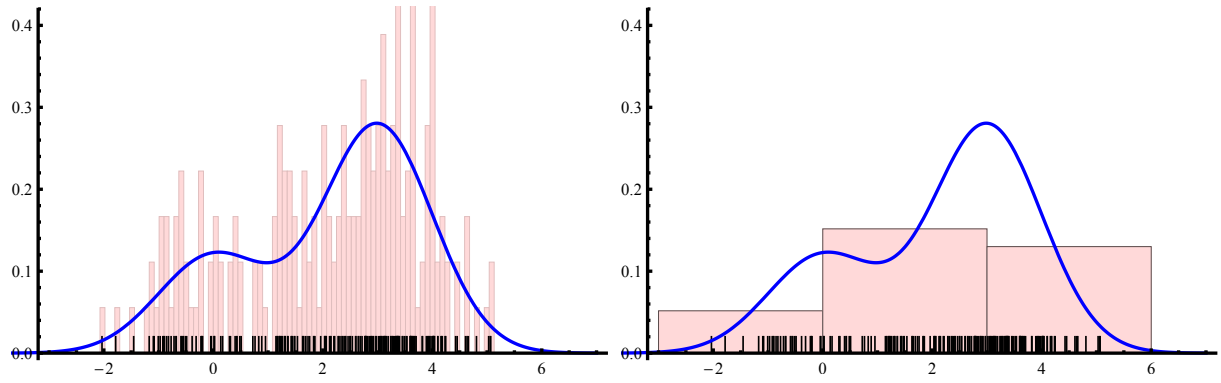


ABBILDUNG 2. Das Histogramm einer Stichprobe vom Umfang $n = 200$ mit einer schlecht gewählten Bandbreite $h = c_i - c_{i-1}$. Links: Die Bandbreite ist zu klein. Rechts: Die Bandbreite ist zu groß. In beiden Fällen zeigt die glatte blaue Kurve die wahre Dichte f.

Das Histogramm hat den Nachteil, dass die Wahl der c_i's bzw. die Wahl der Bandbreite h willkürlich ist. Wird die Bandbreite zu klein oder zu groß gewählt, so kommt es zu Histogrammen, die die Dichte nur schlecht approximieren, siehe Abbildung 2. Außerdem ist

das Histogramm eine lokal konstante, nicht stetige Funktion, obwohl die Dichte f in vielen Beispielen stetig und nicht lokal konstant ist. Im nächsten Abschnitt betrachten wir einen Dichteschätzer, der zumindest von diesem zweiten Nachteil frei ist.

7.2. Kerndichteschätzer

Wir werden nun eine bessere Methode zur Schätzung der Dichte betrachten, den Kerndichteschätzer.

Definition 7.2.1. Ein *Kern* ist eine messbare Funktion $K : \mathbb{R} \to [0, \infty)$, so dass
(1) $K(x) \geq 0$ für alle $x \in \mathbb{R}$ und
(2) $\int_{\mathbb{R}} K(x)\mathrm{d}x = 1$.

Die Bedingungen in der Definition eines Kerns sind somit die gleichen, wie in der Definition einer Dichte.

Definition 7.2.2. Sei $(x_1, \ldots, x_n) \in \mathbb{R}^n$ eine Stichprobe. Sei K ein Kern und $h > 0$ ein Parameter, der die *Bandbreite* heißt. Der *Kerndichteschätzer* ist definiert durch

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

Bemerkung 7.2.3. Jedem Punkt x_i in der Stichprobe wird in dieser Formel ein „Beitrag“ der Form

$$\frac{1}{nh} K\left(\frac{x - x_i}{h}\right)$$

zugeordnet. Der Kerndichteschätzer $\hat{f}_n$ ist die Summe der einzelnen Beiträge. Das Integral jedes einzelnen Beitrags ist gleich $1/n$, denn

$$\int_{\mathbb{R}} \frac{1}{hn} K\left(\frac{x - x_i}{h}\right) \mathrm{d}x = \frac{1}{n} \int_{\mathbb{R}} K(y)\mathrm{d}y = \frac{1}{n}.$$

Um das Integral zu berechnen, haben wir dabei die Variable $y := \frac{x - x_i}{h}$ mit $dy = \frac{dx}{h}$ eingeführt. Somit ist das Integral von $\hat{f}_n$ gleich 1:

$$\int_{\mathbb{R}} \hat{f}_n(x)\mathrm{d}x = 1.$$

Es ist außerdem klar, dass $f_n(x) \geq 0$ für alle $x \in \mathbb{R}$. Somit ist $\hat{f}_n$ tatsächlich eine Dichte.

Bemerkung 7.2.4. Die Idee hinter dem Kerndichteschätzer zeigt Abbildung 3. Auf dieser Abbildung ist der Kerndichteschätzer der Stichprobe

$$(-4, -3, -2.5, 4.5, 5.0, 5.5, 5.75, 6.5)$$

zu sehen. Die Zahlen aus der Stichprobe werden durch rote Kreise auf der x-Achse dargestellt. Die gestrichelten grünen Kurven zeigen die Beiträge der einzelnen Punkte. In diesem Fall

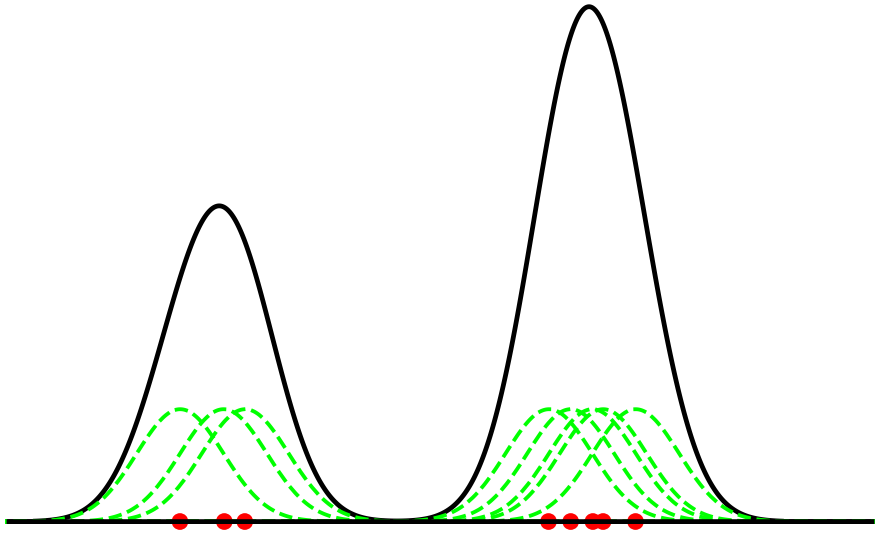


ABBILDUNG 3. Konstruktion des Kerndichteschätzers.

benutzen wir den Gauß-Kern, der unten eingeführt wird. Die Summe der einzelnen Beiträge ist der Kerndichteschätzer $\hat{f}_n$, der durch die schwarze Kurve dargestellt wird.

In der Definition des Kerndichteschätzers kommen zwei noch zu wählende Parameter vor: der Kern K und die Bandbreite h. Für die Wahl des Kerns gibt es z.B. die folgenden Möglichkeiten, s. Abbildung 4.

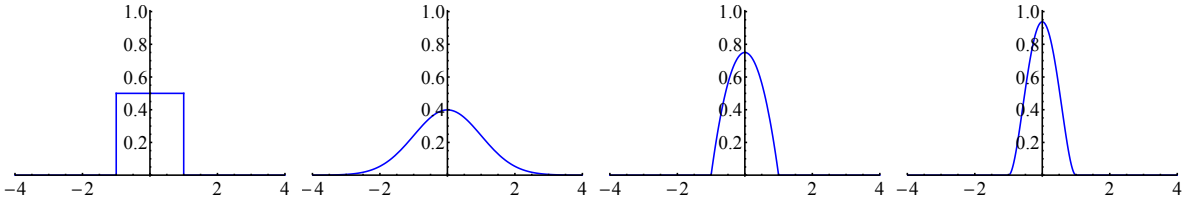


ABBILDUNG 4. Beispiele von Kernen: Rechteckskern, Gauß, Epanechnikov, Bisquare.

Beispiel 7.2.5. Der *Rechteckskern* ist definiert durch

$$K(x) = \frac{1}{2}\mathbb{1}_{x\in[-1,1]}.$$

Der mit dem Rechteckskern assoziierte Kerndichteschätzer ist somit gegeben durch

$$\hat{f}_n(x) = \frac{1}{2nh}\sum_{i=1}^{n}\mathbb{1}_{x_i\in[x-h,x+h]}$$

und wird auch als *gleitendes Histogramm* bezeichnet. Ein Nachteil des Rechteckskerns ist, dass er nicht stetig ist.

Beispiel 7.2.6. Der *Gauß-Kern* ist nichts Anderes, als die Dichte der Standardnormalverteilung:

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

Es gilt dann

$$\frac{1}{h}K\left(\frac{x-x_i}{h}\right) = \frac{1}{\sqrt{2\pi}h}\exp\left(-\frac{(x-x_i)^2}{2h^2}\right),$$

was der Dichte der Normalverteilung $\mathrm{N}(x_i, h^2)$ entspricht. Der Kerndichteschätzer $\hat{f}_n$ ist das arithmetische Mittel solcher Dichten.

Beispiel 7.2.7. Der *Epanechnikov-Kern* ist definiert durch

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

Dieser Kern verschwindet außerhalb des Intervalls $(-1,1)$, hat also einen kompakten Träger.

Beispiel 7.2.8. Der *Bisquare-Kern* ist gegeben durch

$$K(x) = \begin{cases} \frac{15}{16}(1-x^2)^2, & \text{falls } x \in (-1,1), \\ 0, & \text{sonst.} \end{cases}$$

Dieser Kern besitzt ebenfalls einen kompakten Träger und ist glatter als der Epanechnikov-Kern.

Wir müssen noch beschreiben, wie wir den Kern K und die Bandbreite h wählen. Man kann zeigen, dass die Wahl des Kerns nicht kritisch ist: so liefern z.B. der Gauß-Kern und der Epanechnikov-Kern sehr ähnliche Ergebnisse. Viel wichtiger ist die Wahl der Bandbreite, s. Abbildung 5. Bei einer zu kleinen Bandbreite h kommt es zum sogenannten „undersmoothing“: $\hat{f}_n(x)$ hat viele sehr hohe „Peaks“ und ist sehr klein sonst. Bei einer zu großen Bandbreite ist $\hat{f}_n(x)$ sehr flach („oversmoothing“).

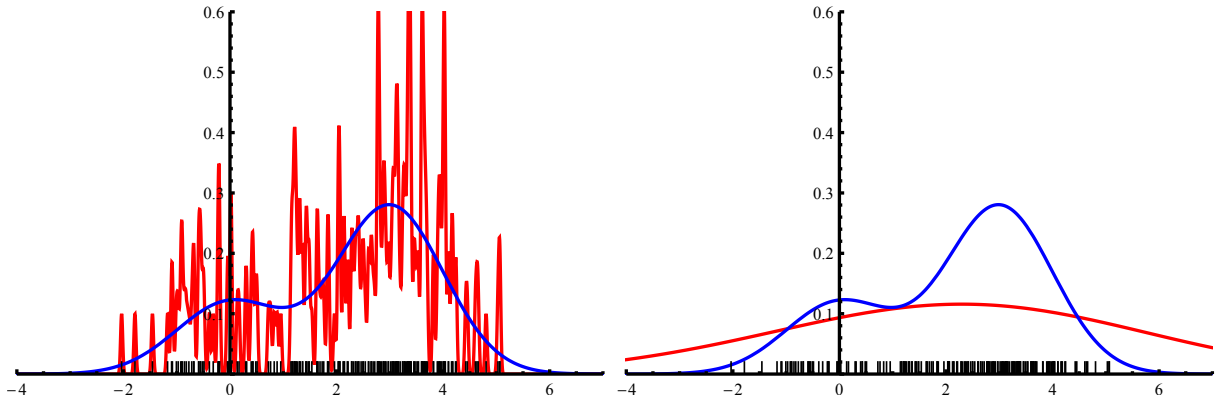


Abbildung 5. Das kann bei einer falschen Wahl der Bandbreite passieren. Blaue Kurve: Die wahre Dichte (eine Mischung aus zwei Normalverteilungen). Schwarze Striche: Die Stichprobe vom Umfang $n = 200$. Rote Kurve: Kerndichteschätzer. Links: Die Bandbreite ist zu klein (undersmoothing). Rechts: Die Bandbreite ist zu groß (oversmoothing).

7.3. Optimale Wahl der Bandbreite

In diesem Abschnitt untersuchen wir den Bias und die Varianz des Kerndichteschätzers. Als Anwendung werden wir eine Regel zur Wahl der Bandbreite herleiten. Unser Ziel ist es, die wichtigsten Ideen zu skizzieren. Wir werden uns hier nicht um eine strenge Begründung der Resultate kümmern.

Seien $X_1, X_2, \dots$ unabhängige identisch verteilte Zufallsvariablen mit einer Dichte f, die nicht bekannt ist und geschätzt werden soll. Im Folgenden nehmen wir an, dass die Dichte f hinreichend oft differenzierbar ist. Sei K ein Kern mit den Eigenschaften

$$\int_{\mathbb{R}} yK(y)\mathrm{d}y = 0, \quad m_2 := \int_{\mathbb{R}} y^2 K(y)\mathrm{d}y < \infty. \tag{7.3.1}$$

Die erste Eigenschaft ist z.B. für symmetrische Kerne mit $K(x) = K(-x)$ erfüllt. Wir betrachten den Kerndichteschätzer

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

Satz 7.3.1. Für den Bias des Kerndichteschätzers gilt

$$\mathrm{Bias}\,\hat{f}_n(x) = \mathbb{E}\hat{f}_n(x) - f(x) = \frac{1}{2}f''(x)m_2h^2 + o(h^2), \quad h \downarrow 0.$$

Beweis. Nach der Definition des Kerndichteschätzers $\hat{f}_n(x)$ gilt

$$\mathbb{E}\hat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n}\mathbb{E}K\left(\frac{x - X_i}{h}\right) = \frac{1}{h}\mathbb{E}K\left(\frac{x - X_1}{h}\right),$$

wobei wir benutzt haben, dass $X_1, \dots, X_n$ identisch verteilt sind. Es sei bemerkt, dass dieser Ausdruck (und somit auch der Bias) nicht von n abhängt. Da die Dichte von X_1 gleich f ist, gilt

$$\mathbb{E}\hat{f}_n(x) = \frac{1}{h}\int_{\mathbb{R}} K\left(\frac{x - z}{h}\right) f(z)\mathrm{d}z = \int_{\mathbb{R}} K(y)f(x - hy)\mathrm{d}y.$$

Nun benutzen wir die Taylor-Reihe für f in der Umgebung von x:

$$f(x - yh) = f(x) - f'(x)hy + \frac{1}{2}f''(x)h^2y^2 + O(h^3y^3), \quad h \downarrow 0.$$

Durch Einsetzen ergibt sich

$$\mathbb{E}\hat{f}_n(x) = f(x)\int_{\mathbb{R}} K(y)\mathrm{d}y - f'(x)h\int_{\mathbb{R}} yK(y)\mathrm{d}y + \frac{1}{2}f''(x)h^2\int_{\mathbb{R}} K(y)y^2\mathrm{d}y + O(h^3).$$

Unter Berücksichtigung von $\int_{\mathbb{R}} K(y)\mathrm{d}y = 1$ (nach Definition eines Kerns) sowie (7.3.1) erhalten wir, dass

$$\mathbb{E}\hat{f}_n(x) = f(x) + \frac{1}{2}f''(x)m_2h^2 + o(h^2).$$

Daraus ergibt sich die Behauptung des Satzes. □

Satz 7.3.2. Es sei zusätzlich $R(K) := \int_{\mathbb{R}} K^2(y)\mathrm{d}y < \infty$. Für die Varianz des Kerndichteschätzers gilt

$$n \operatorname{Var} \hat{f}_n(x) = \frac{f(x)R(K)}{h} + O(1), \quad h \downarrow 0.$$

Beweis. Wegen der Unabhängigkeit von $X_1, \ldots, X_n$ erhalten wir

$$n \operatorname{Var} \hat{f}_n(x) = \frac{1}{h^2} \operatorname{Var} K\left(\frac{x - X_1}{h}\right) = \frac{1}{h^2}\mathbb{E}K^2\left(\frac{x - X_1}{h}\right) + \left(\frac{1}{h}\mathbb{E}K\left(\frac{x - X_1}{h}\right)\right)^2.$$

Der zweite Term ist $(\mathbb{E}\hat{f}_n(x))^2$, was durch C abgeschätzt werden kann. Wir schauen uns den ersten Term an:

$$\frac{1}{h^2}\mathbb{E}K^2\left(\frac{x - X_1}{h}\right) = \frac{1}{h^2}\int_{\mathbb{R}} K^2\left(\frac{x - z}{h}\right) f(z)\mathrm{d}z = \frac{1}{h}\int_{\mathbb{R}} K^2(y) f(x - hy)\mathrm{d}y.$$

Nach der Taylor-Entwicklung gilt $f(x - hy) = f(x) + O(h)$, so dass

$$\frac{1}{h^2}\mathbb{E}K^2\left(\frac{x - X_1}{h}\right) = \frac{f(x)}{h}\int_{\mathbb{R}} K^2(y)\mathrm{d}y + O(1) = \frac{f(x)}{h}R(K) + O(1).$$

Daraus ergibt sich die Behauptung. □

Bemerkung 7.3.3. Idealerweise würden wir die Bandbreite h so wählen, dass sowohl der Bias als auch die Varianz klein sind. Dies führt zum sogenannten *Varianz-Bias-Dilemma*: Für kleine h ist der Bias zwar klein, die Varianz aber groß. Für große h ist die Varianz klein, dafür aber der Bias groß. Man muss einen vernünftigen Kompromiss zwischen dem Bias und der Varianz finden.

Zu diesem Zweck definieren wir den *mittleren quadratischen Fehler* des Kerndichteschätzers an der Stelle $x \in \mathbb{R}$ wie folgt:

$$\operatorname{MSE} \hat{f}_n(x) = \mathbb{E}[(\hat{f}_n(x) - f(x))^2] = \operatorname{Var} \hat{f}_n(x) + (\operatorname{Bias} \hat{f}_n(x))^2.$$

Indem wir nun die Entwicklungen für die Varianz und den Bias einsetzen und dabei die Restterme ignorieren, erhalten wir eine Approximation an MSE, die *asymptotischer mittlerer quadratischer Fehler* genannt wird:

$$\operatorname{AMSE} \hat{f}_n(x) = \frac{f(x)R(K)}{nh} + \left(\frac{1}{2}f''(x)m_2h^2\right)^2.$$

Wir wollen nun die Güte des Schätzers $\hat{f}_n(x)$ durch den *integrierten asymptotischen mittleren quadratischen Fehler* messen:

$$\operatorname{AMISE}(h) = \int_{\mathbb{R}} \operatorname{AMSE} \hat{f}_n(x)\mathrm{d}x = \frac{R(K)}{nh} + \frac{1}{4}m_2^2 R(f'')h^4.$$

Dabei ist $R(f'') = \int_{\mathbb{R}}(f''(x))^2\mathrm{d}x$. Wir wollen nun die Bandbreite h so wählen, dass die Funktion $\operatorname{AMISE}(h)$ minimiert wird. Es sei bemerkt, dass der erste Term für $h \downarrow 0$ gegen $+\infty$ strebt (kleine Bandbreiten sind schlecht), während der zweite Term für $h \uparrow +\infty$ gegen $+\infty$ divergiert (große Bandbreiten sind ebenfalls schlecht), s. Abbildung 6.

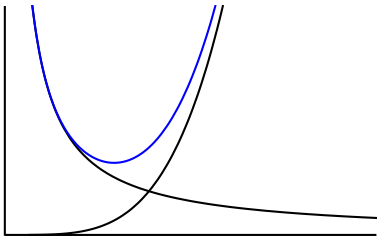


ABBILDUNG 6. Varianz-Bias-Dilemma. Schwarze Kurven: Funktionen $\text{const} \cdot h^{-1}$ und $\text{const} \cdot h^4$. Blaue Kurve: deren Summe.

Indem wir die Ableitung auf 0 setzen, erhalten wir

$$h_{\text{opt}} = n^{-\frac{1}{5}} \left(\frac{R(K)}{m_2^2 R(f'')} \right)^{\frac{1}{5}}.$$

Somit hat die optimale Bandbreite die Größenordnung $\text{const} \cdot n^{-1/5}$, ein ziemlich unerwartetes Ergebnis! Für den entsprechenden Fehler gilt

$$\text{AMISE}(h_{\text{opt}}) = \text{const} \cdot n^{-4/5}.$$

Ähnliche Berechnungen lassen sich übrigens für das Histogramm durchführen und ergeben die optimale Bandbreite $\text{const} \cdot n^{-1/3}$ und einen entsprechenden Fehler von $\text{const} \cdot n^{-2/3}$. Nun gilt $4/5 > 2/3$. Daran sieht man, dass der Kerndichteschätzer für große Werte von n besser als das Histogramm abschneidet.

Leider können wir die obige Formel für h_{opt} in der Praxis nicht für die Wahl der Bandbreite benutzen, denn die Formel enthält eine unbekannte Größe, nämlich $R(f'')$. Um diese Größe zu schätzen, müssten wir f'' schätzen. Dies ist aber mindestens genau so schwer, wie f zu schätzen!

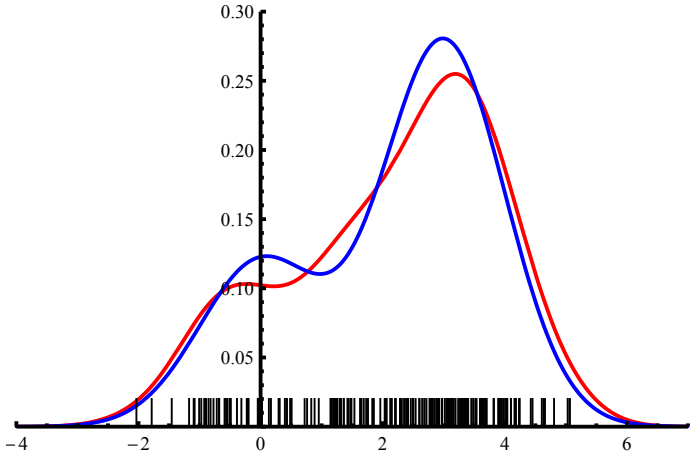


ABBILDUNG 7. Kerndichteschätzer mit einer nach der Silverman-Regel gewählten Bandbreite. Blaue Kurve: Die wahre Dichte (eine Mischung aus zwei Normalverteilungen). Schwarze Striche: Die Stichprobe vom Umfang $n = 200$. Rote Kurve: Kerndichteschätzer.

Silverman-Faustregel.[1] Einen möglichen Ausweg hat Silverman vorgeschlagen. Wir sollten im Ausdruck $R(f'') = \int_{\mathbb{R}}(f''(x))^2 \mathrm{d}x$ die unbekannte Dichte f durch die Dichte einer Normalverteilung ersetzen, die die gleiche Varianz wie f besitzt. Der Erwartungswert spielt dabei keine Rolle, denn $R(f'')$ ändert sich nicht, wenn wir $f(x)$ durch $f(x-a)$ ersetzen. Wir werden also die Bandbreite genauso wählen, wie wir sie für die Normalverteilung mit der gleichen Varianz wählen würden (Normalverteilung als „Goldstandard"). Die Varianz von f ist zwar auch unbekannt, kann aber durch die empirische Varianz S_n^2 der Stichprobe $X_1, \ldots, X_n$ geschätzt werden. Wir bezeichnen die Dichte der Normalverteilung mit Erwartungswert 0 und Varianz σ^2 durch

$$g(x; \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-x^2/(2\sigma^2)}.$$

Es gilt $g(t; S_n^2) = S_n^{-1} g(t/S_n; 1)$. Wir ersetzen $R(f'')$ durch

$$\hat{R} = \int_{\mathbb{R}} (g''(x; S_n^2))^2 \mathrm{d}x = \int_{\mathbb{R}} (S_n^{-3} g''(x/S_n; 1))^2 \mathrm{d}x = S_n^{-5} \int_{\mathbb{R}} (g''(x))^2 \mathrm{d}x = \frac{6}{\sqrt{\pi}} S_n^{-5}.$$

Die Faustregel von Silverman für die Wahl der Bandbreite lautet somit

$$\hat{h}_{\text{Silverman}} = \left(\frac{\sqrt{\pi} R(K)}{6 m_2^2} \right)^{\frac{1}{5}} S_n n^{-\frac{1}{5}}.$$

Im Spezialfall, wenn $K(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}$ der Gauß-Kern ist, ergibt sich die Formel

$$\hat{h}_{\text{Silverman}} = (4/3)^{1/5} S_n n^{-1/5} \approx 1.06 \cdot S_n n^{-1/5}.$$

Die Silverman-Regel ist nicht flexibel genug, da sie auf der Annahme beruht, dass die Dichte „ungefähr die gleiche Form" besitzt, wie die Normalverteilung. Wir werden deshalb eine andere, sehr interessante Methode betrachten, die *Kreuzvalidierung* heißt. Diese Methode ist sehr allgemein und kann in vielen weiteren Problemen der nichtparametrischen Statistik zur Wahl von Regularisierungsparametern angewendet werden, z.B. bei der nichtparametrischen Regression.

Kreuzvalidierung (Kleinste-Quadrate-Version). Wir versuchen, die folgende Funktion durch eine geeignete Wahl der Bandbreite h zu minimieren:

$$J(h) := \int_{\mathbb{R}} (\hat{f}_{n,h}(x) - f(x))^2 \mathrm{d}x = \int_{\mathbb{R}} \hat{f}_{n,h}^2(x) \mathrm{d}x + \int_{\mathbb{R}} f^2(x) \mathrm{d}x - 2 \int_{\mathbb{R}} \hat{f}_{n,h}(x) f(x) \mathrm{d}x \to \min.$$

Wir haben $\hat{f}_{n,h}$ anstelle von $\hat{f}_n$ geschrieben, um die Abhängigkeit von h zu verdeutlichen. Der erste Term kann (als Funktion von h) berechnet werden. Im zweiten Term ist zwar $f(x)$ unbekannt, dies stellt aber kein Problem dar, denn der zweite Term hängt von h nicht ab und kann deshalb weggelassen werden. Wir schauen uns den dritten Term an. Sei also

$$L(h) := \int_{\mathbb{R}} \hat{f}_{n,h}(x) f(x) \mathrm{d}x.$$

[1] Englisch: Silverman's rule of thumb

Leider ist in diesem Term $f(x)$ unbekannt. Sei X eine weitere, von $X_1, \ldots, X_n$ unabhängige Zufallsvariable mit Dichte f. Dann können wir

$$L(h) = \mathbb{E}\hat{f}_{n,h}(X)$$

schreiben, wobei der Erwartungswert $\mathbb{E}$ mittels Integration über alle möglichen Werte von X berechnet wird (während die Stichprobe $x_1, \ldots, x_n$ als deterministisch betrachtet wird). Einen solchen Erwartungswert kann man schätzen, indem man einen Mittelwert über n Realisierungen von $\hat{f}_n(X)$ bildet. Dafür bräuchten wir n Realisierungen von X, die von $\hat{f}_n(x)$, also von $x_1, \ldots, x_n$, unabhängig sind. Wir können zwar $x_1, \ldots, x_n$ als Realisierungen von X betrachten (alle Zufallsvariablen haben die gleiche Dichte f), diese sind aber von sich selbst leider abhängig. Und nun tritt die Idee der Kreuzvalidierung auf die Bühne. Wir wählen ein x_i aus der Stichprobe aus und betrachten es als eine Realisierung von X. Die verbleibenden $n-1$ Elemente der Stichprobe (die ja von x_i unabhängig sind!) benutzen wir um einen „one-leave-out“ Kerndichteschätzer

$$\hat{f}_{n,h}^{(-i)}(x) := \frac{1}{(n-1)h} \sum_{\substack{k=1,\ldots,n \\ k \neq i}} K\left(\frac{x - x_k}{h}\right), \quad x \in \mathbb{R},$$

zu konstruieren, der als Ersatz von $\hat{f}_{n,h}(x)$ benutzt wird. Dabei basiert $\hat{f}_{n,h}$ auf einer Stichprobe vom Umfang n, während sein one-leave-out-Analogon $\hat{f}_{n,h}^{(-i)}$ ein Element weniger benutzt. Dies sollte aber bei einem großen n kein großer Unterschied sein. Wir fassen nun $\hat{f}_{n,h}^{(-i)}(x_i)$ als eine Realisierung von $\hat{f}_{n,h}(X)$ auf und betrachten den folgenden Schätzer von $L(h)$:

$$\hat{L}(h) := \frac{1}{n} \sum_{i=1}^{n} \hat{f}_{n,h}^{(-i)}(x_i).$$

Jedes Element x_i wird benutzt, um die Bandbreite eines auf allen anderen Elementen basierenden Kerndichteschätzers zu „validieren“. Wir können also $J(h)$ (ohne den zweiten, konstanten Term) durch die sogenannte Kreuzvalidierungsfunktion

$$\hat{J}_0(h) := \int_{\mathbb{R}} \hat{f}_{n,h}^2(x)\mathrm{d}x - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{n,h}^{(-i)}(x_i)$$

schätzen. Die Bandbreite h sollte nun so gewählt werden, dass $\hat{J}_0(h)$ minimal wird. Die Berechnung der Funktion $\hat{J}_0(h)$ ist wegen der vielen Summen etwas aufwendig, kann aber für die von uns betrachtete Stichprobe vom Umfang $n = 200$ auf einem Laptop in wenigen Minuten durchgeführt werden, s. Abbildung 8, links. Das Minimum wird für $h \approx 0.407$ erreicht. Der entsprechende Kerndichteschätzer wird auf Abbildung 8, rechts, gezeigt.

Kreuzvalidierung (Maximum-Likelihood-Version). Wir wollen die Bandbreite h so wählen, dass der Kerndichteschätzer $\hat{f}_{n,h}(x)$ die Dichte $f(x)$ möglichst gut approximiert. Die Dichte f ist zwar unbekannt, hat aber Spuren in Form der Stichprobe $x_1, \ldots, x_n$ hinterlassen. Da die Stichprobe $x_1, \ldots, x_n$ gemäß der Dichte f erzeugt wurde, sollte die Likelihood dieser Stichprobe bezüglich $\hat{f}_{n,h}(x)$ möglichst groß sein. Man kann also versuchen, die Bandbreite

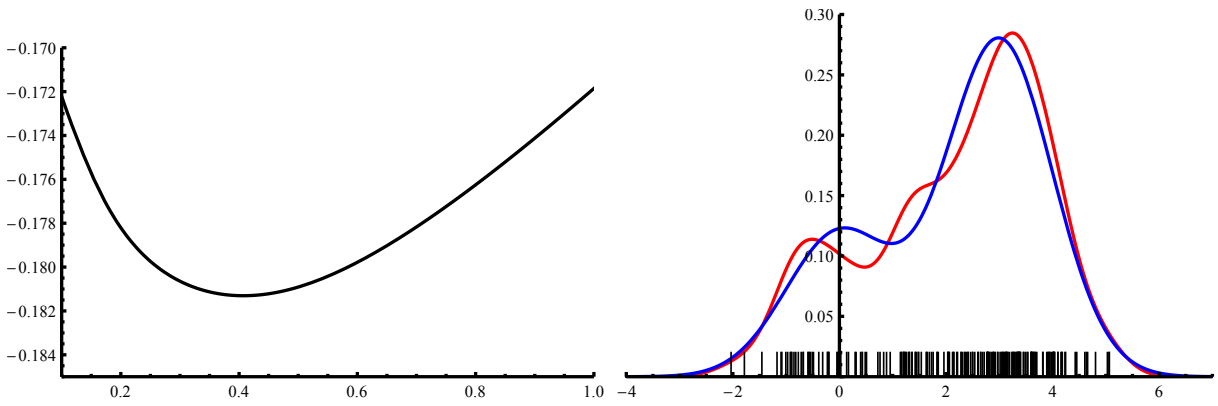


ABBILDUNG 8. Links: Kreuzvalidierungsfunktion $\hat{J}_0(h)$. Rechts: Der Kerndichteschätzer (rote Kurve), dessen Bandbreite $h \approx 0.407$ die Kreuzvalidierungsfunktion minimiert, und die wahre Dichte (blaue Kurve).

so zu wählen, dass die Log-Likelihood-Funktion

$$l(h) = \sum_{i=1}^{n} \log \hat{f}_{n,h}(x_i)$$

maximiert wird. Leider funktioniert dieser Zugang nicht. Bei sehr kleinen Bandbreiten $h \approx 0$ erreicht $\hat{f}_{n,h}$ sehr hohe Werte an den Stellen $x_1, \ldots, x_n$, weshalb das Maximum an der Stelle $h = 0$ erreicht wird! Wir haben aber bereits auf Abbildung 5 gesehen, dass $h = 0$ eine schlechte Wahl ist. Worin bestand nun unser Fehler? Wir haben x_i benutzt, um die Dichte $\hat{f}_{n,h}$ zu validieren, die selbst von x_i abhängt. Der Berater, den wir zur Expertise herangezogen haben, war befangen! Um die Unabhängigkeit zu erreichen, betrachten wir die Log-Likelihoods von x_i bezüglich des one-leave-out-Kerndichteschätzers $\hat{f}_{n,h}^{(-i)}$, also

$$\hat{l}(h) = \sum_{i=1}^{n} \log \hat{f}_{n,h}^{(-i)}(x_i).$$

Wir wählen die Bandbreite h als Maximierer dieser Funktion.

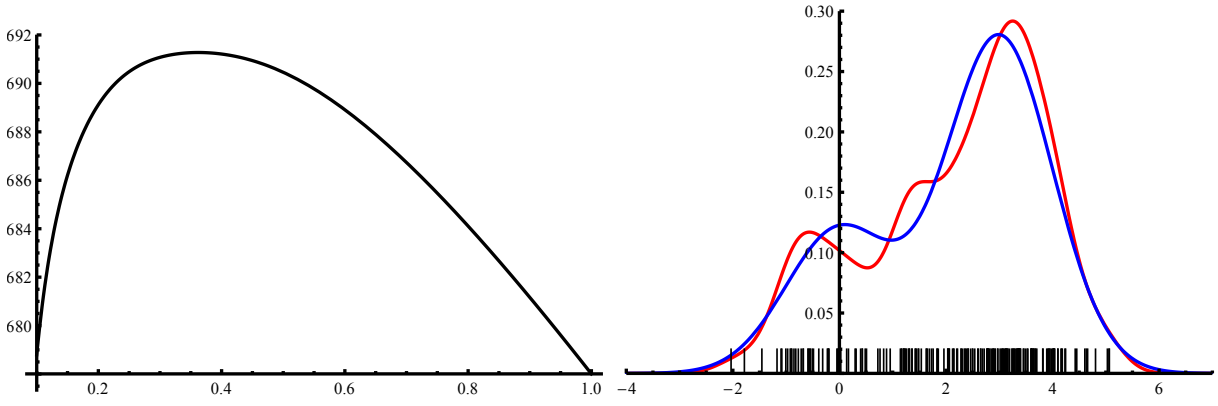


ABBILDUNG 9. Kreuzvalidierungsfunktion $\hat{l}(h)$ und der entsprechende Kerndichteschätzer.

In dem von uns betrachteten Beispiel lässt sich die Funktion $\hat{l}(h)$ schnell numerisch berechnen, s. Abbildung 9. Das Maximum wird an der Stelle $h \approx 0.362$ erreicht.

KAPITEL 8

Wichtige statistische Verteilungen

In diesem Kapitel werden wir die wichtigsten statistischen Verteilungsfamilien einführen. Zu diesen zählen neben der Normalverteilung die folgenden Verteilungsfamilien:

(1) Gammaverteilung (Spezialfälle: Pearson-χ^2-Verteilung und Erlang-Verteilung);
(2) Betaverteilung;
(3) Student-t-Verteilung;
(4) Fisher-Snedecor-F-Verteilung.

Diese Verteilungen werden wir später für die Konstruktion von Konfidenzintervallen und statistischen Tests benötigen.

Abbildung 1. Karl Pearson, William Gosset (Student), Ronald Fisher, George Snedecor

8.1. Gammafunktion und Gammaverteilung

Definition 8.1.1. Die *Gammafunktion* ist gegeben durch

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} \mathrm{e}^{-t} \mathrm{d}t, \quad \alpha > 0.$$

Folgende Eigenschaften der Gammafunktion werden oft benutzt:

(1) $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$.
(2) $\Gamma(n) = (n-1)!$, falls $n \in \mathbb{N}$.
(3) $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Die letzte Eigenschaft kann man wie folgt beweisen: Mit $t = \frac{w^2}{2}$ und $dt = wdw$ gilt

$$\Gamma\left(\frac{1}{2}\right) = \int_0^\infty t^{-\frac{1}{2}} \mathrm{e}^{-t} \mathrm{d}t = \int_0^\infty \frac{\sqrt{2}}{w} \mathrm{e}^{-\frac{w^2}{2}} w \mathrm{d}w = \sqrt{2} \int_0^\infty \mathrm{e}^{-\frac{w^2}{2}} \mathrm{d}w = \sqrt{\pi},$$

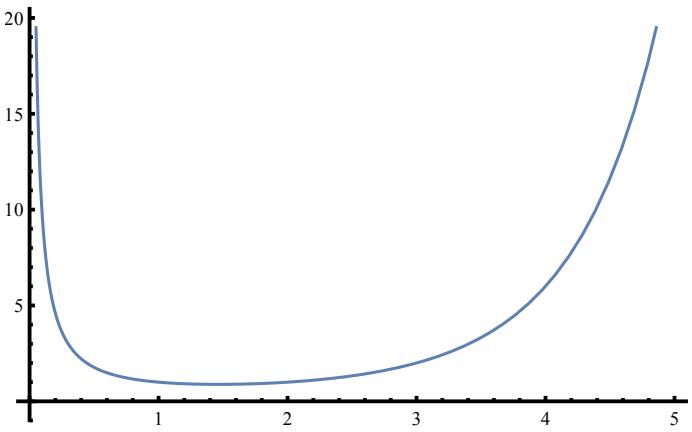


ABBILDUNG 2. Der Graph der Gammafunktion.

denn $\int_0^\infty \mathrm{e}^{-\frac{w^2}{2}}\mathrm{d}w = \frac{1}{2}\sqrt{2\pi}$.

Definition 8.1.2. Eine Zufallsvariable X ist *Gammaverteilt* mit Parametern $\alpha > 0$ und $\lambda > 0$, falls für die Dichte von X gilt

$$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1}\mathrm{e}^{-\lambda x}, \quad x > 0.$$

Notation 8.1.3. $X \sim \mathrm{Gamma}(\alpha, \lambda)$.

Aufgabe 8.1.4. Zeigen Sie, dass $\int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1}\mathrm{e}^{-\lambda x}\mathrm{d}x = 1$.

Bemerkung 8.1.5. Die Gammaverteilung mit Parametern $\alpha = 1$ und $\lambda > 0$ hat Dichte $\lambda\mathrm{e}^{-\lambda t}$, $t > 0$, und stimmt somit mit der Exponentialverteilung mit Parameter λ überein.

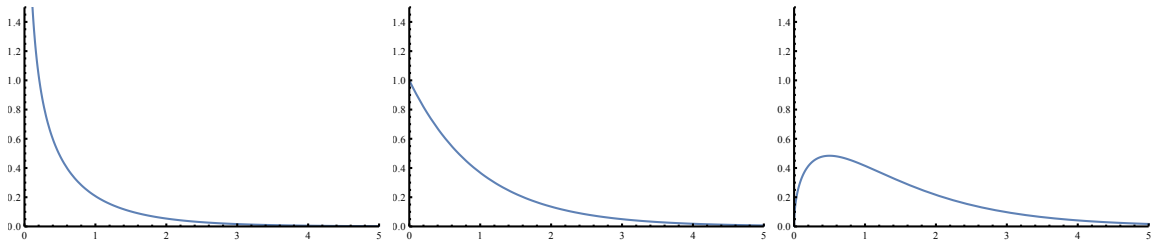


ABBILDUNG 3. Dichten der Gammaverteilungen mit verschiedenen Werten des Parameters α. Links: $\alpha < 1$. Mitte: $\alpha = 1$ (Exponentialverteilung). Rechts: $\alpha > 1$.

Satz 8.1.6. Sei $X \sim \mathrm{Gamma}(\alpha, \lambda)$ eine Gammaverteilte Zufallsvariable. Dann sind die Laplace-Transformierte $m_X(t) := \mathbb{E}\mathrm{e}^{tX}$ und die charakteristische Funktion $\varphi_X(t) :=$

$\mathbb{E}e^{itX}$ gegeben durch

$$m_X(t) = \frac{1}{\left(1-\frac{t}{\lambda}\right)^{\alpha}} \text{ (für } t < \lambda), \quad \varphi_X(t) = \frac{1}{\left(1-\frac{it}{\lambda}\right)^{\alpha}} \text{ (für } t \in \mathbb{R}).$$

Beweis. Für die Laplace-Transformierte ergibt sich

$$m_X(t) = \int_0^\infty e^{tx} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} dx = \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-(\lambda-t)x} dx.$$

Dieses Integral ist für $t < \lambda$ konvergent. Indem wir nun $w = (\lambda - t)x$ einsetzen, erhalten wir, dass

$$m_X(t) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty \left(\frac{w}{\lambda - t}\right)^{\alpha-1} e^{-w} \frac{dw}{\lambda - t} = \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{1}{(\lambda-t)^\alpha} \int_0^\infty w^{\alpha-1} e^{-w} dw = \frac{1}{(1-\frac{t}{\lambda})^\alpha}.$$

Wenn man nun komplexe Werte von t zulässt, dann sind die obigen Integrale in der Halbebene $\operatorname{Re} t < \lambda$ konvergent. Somit stellt $m_X(t)$ eine analytische Funktion in der Halbebene $\operatorname{Re} t < \lambda$ dar und ist für reelle Werte von t gleich $1/(1-\frac{t}{\lambda})^\alpha$. Nach dem Prinzip der analytischen Fortsetzung (Funktionentheorie) muss diese Formel in der ganzen Halbebene gelten. Indem wir nun die Formel für Zahlen der Form $t = is$ benutzen (die in der Halbebene für alle $s \in \mathbb{R}$ liegen), erhalten wir, dass

$$\varphi_X(s) = m_X(is) = \frac{1}{(1-\frac{is}{\lambda})^\alpha}.$$

Somit ist die Formel für die charakteristische Funktion bewiesen. □

Aufgabe 8.1.7. Zeigen Sie, dass für $X \sim \text{Gamma}(\alpha, \lambda)$ gilt

$$\mathbb{E}X = \frac{\alpha}{\lambda}, \quad \operatorname{Var} X = \frac{\alpha}{\lambda^2}.$$

Der nächste Satz heißt die *Faltungseigenschaft der Gammaverteilung*.

Satz 8.1.8. Sind die Zufallsvariablen $X \sim \text{Gamma}(\alpha, \lambda)$ und $Y \sim \text{Gamma}(\beta, \lambda)$ unabhängig, dann gilt für die Summe

$$X + Y \sim \text{Gamma}(\alpha + \beta, \lambda).$$

Beweis. Für die charakteristische Funktion von $X + Y$ gilt wegen der Unabhängigkeit

$$\varphi_{X+Y}(t) = \varphi_X(t) \cdot \varphi_Y(t) = \frac{1}{(1-\frac{it}{\lambda})^\alpha} \cdot \frac{1}{(1-\frac{it}{\lambda})^\beta} = \frac{1}{(1-\frac{it}{\lambda})^{\alpha+\beta}}.$$

Dies ist die charakteristische Funktion einer $\text{Gamma}(\alpha + \beta, \lambda)$-Verteilung. Da die charakteristische Funktion die Verteilung eindeutig bestimmt, muss $X + Y \sim \text{Gamma}(\alpha + \beta, \lambda)$ gelten. □

8.2. Pearsonsche χ^2-Verteilung

Definition 8.2.1. Seien $X_1, \ldots, X_n \sim \mathrm{N}(0,1)$ unabhängige und standardnormalverteilte Zufallsvariablen. Dann heißt die Verteilung von

$$X_1^2 + \ldots + X_n^2$$

die χ^2-*Verteilung* mit n Freiheitsgraden.

Notation 8.2.2. $X_1^2 + \ldots + X_n^2 \sim \chi_n^2$.

Bemerkung 8.2.3. Die Verteilung von $\sqrt{X_1^2 + \ldots + X_n^2}$ heißt die χ-*Verteilung* mit n Freiheitsgraden und wird mit χ_n bezeichnet.

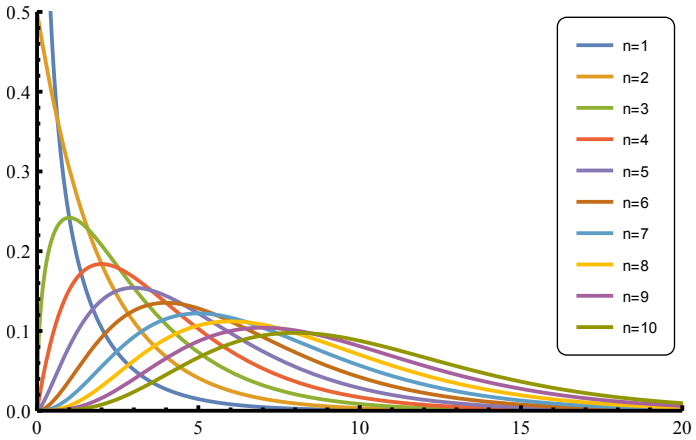


Abbildung 4. Dichten der χ^2-Verteilungen mit $n = 1, \ldots, 10$ Freiheitsgraden.

Wir werden nun zeigen, dass die χ^2-Verteilung ein Spezialfall der Gammaverteilung ist. Zuerst betrachten wir die χ^2-Verteilung mit einem Freiheitsgrad.

Satz 8.2.4. Sei $X \sim \mathrm{N}(0,1)$. Dann ist $X^2 \sim \mathrm{Gamma}(\frac{1}{2}, \frac{1}{2})$. Symbolisch: $\chi_1^2 \stackrel{d}{=} \mathrm{Gamma}(\frac{1}{2}, \frac{1}{2})$.

Beweis. Wir bestimmen die Laplace-Transformierte von X^2:

$$m_{X^2}(t) = \mathbb{E}\mathrm{e}^{tX^2} = \int_{\mathbb{R}} \mathrm{e}^{tx^2} \left(\frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}} \right) \mathrm{d}x = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \mathrm{e}^{-\frac{1-2t}{2}x^2} \mathrm{d}x = \frac{1}{\sqrt{1-2t}}.$$

Das gilt für komplexe t mit $\operatorname{Re} t < \frac{1}{2}$. Insbesondere erhalten wir mit $t = is$, dass für die charakteristische Funktion von X^2 gilt

$$\varphi_{X^2}(s) = \frac{1}{\sqrt{1-2is}}, \quad s \in \mathbb{R}.$$

Dies entspricht der charakteristischen Funktion einer Gammaverteilung mit Parametern $\alpha = 1/2$ und $\lambda = 1/2$. Da die charakteristische Funktion die Verteilung eindeutig festlegt, erhalten wir, dass $X^2 \sim \mathrm{Gamma}(\frac{1}{2}, \frac{1}{2})$. □

Satz 8.2.5. Seien $X_1, \ldots, X_n \sim \mathrm{N}(0,1)$ unabhängige Zufallsvariablen. Dann gilt

$$X_1^2 + \ldots + X_n^2 \sim \mathrm{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right).$$

Symbolisch: $\chi_n^2 \stackrel{d}{=} \mathrm{Gamma}(\frac{n}{2}, \frac{1}{2})$.

Beweis. Wir haben bereits gezeigt, dass $X_1^2, \ldots, X_n^2 \sim \mathrm{Gamma}(\frac{1}{2}, \frac{1}{2})$. Außerdem sind die Zufallsvariablen $X_1^2, \ldots, X_n^2$ unabhängig. Der Satz folgt aus der Faltungseigenschaft der Gammaverteilung. □

Bemerkung 8.2.6. Die Dichte einer χ_n^2-Verteilung ist somit gegeben durch

$$f_{\chi_n^2}(x) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} \mathrm{e}^{-\frac{x}{2}}, \quad x > 0.$$

Für die Dichte einer χ_n-Verteilung erhält man dann z.B. mit dem Transformationssatz

$$f_{\chi_n}(x) = \frac{2}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{n-1} \mathrm{e}^{-\frac{x^2}{2}}, \quad x > 0.$$

Beispiel 8.2.7. Seien $X_1, X_2 \sim \mathrm{N}(0,1)$ unabhängige Zufallsvariablen. Dann gilt

$$X_1^2 + X_2^2 \sim \mathrm{Gamma}\left(1, \frac{1}{2}\right) \sim \mathrm{Exp}\left(\frac{1}{2}\right).$$

Symbolisch: $\chi_2^2 \stackrel{d}{=} \mathrm{Exp}(\frac{1}{2})$.

Aufgabe 8.2.8. Sei $S_n \sim \chi_n^2$.

(a) Zeigen Sie, dass $\frac{S_n - n}{\sqrt{n}} \xrightarrow[n\to\infty]{d} \mathrm{N}(0,1)$.

(b) Zeigen Sie, dass $\sqrt{S_n} - \sqrt{n} \xrightarrow[n\to\infty]{d} \mathrm{N}(0, 1/2)$ (Fisher-Approximation).

Beispiel 8.2.9 (Maxwell-Boltzmann-Verteilung)**.** Es seien (X_1, X_2, X_3) die drei Komponenten der Geschwindigkeit eines Moleküls in einem idealen Gas. Da die Anzahl der Moleküle sehr groß ist und auch die Geschwindigkeit eines Moleküls sich sehr oft chaotisch ändert, fasst man (X_1, X_2, X_3) als einen Zufallsvektor auf, dessen Dichte laut Gibbs-Verteilung proportional zu

$$\exp\left\{-\frac{1}{k_B T} E_{\mathrm{kin}}\right\} = \exp\left\{-\frac{m}{2k_B T}(x_1^2 + x_2^2 + x_3^2)\right\}.$$

sein muss, wobei T die (absolute) Temperatur des Gases, $k_B = 1.38 \cdot 10^{-23} J/K$ die Boltzmann-Konstante und m die Masse des Moleküls ist.
Somit sind X_1, X_2, X_3 unabhängig und normalverteilt mit Parametern 0 und $\sigma^2 = k_B T/m$. Es folgt, dass die kinetische Energie $E_{\mathrm{kin}} = \frac{m}{2}(X_1^2 + X_2^2 + X_3^2)$ bis auf einen Skalierungsfaktor χ_3^2-verteilt ist. Der Betrag der Geschwindigkeit $V := \sqrt{X_1^2 + X_2^2 + X_3^2}$ ist (bis auf einen Faktor) χ_3-verteilt:

$$V/\sigma \sim \chi_3.$$

Die Dichte von V heißt die *Maxwell-Boltzmann-Verteilung* und berechnet sich leicht zu

$$f_V(v) = \left(\frac{m}{2\pi k_B T}\right)^{3/2} 4\pi v^2 \exp\left\{-\frac{mv^2}{2k_B T}\right\}, \quad v > 0.$$

8.3. Poisson-Prozess und die Erlang-Verteilung

Um den Poisson-Prozess einzuführen, betrachten wir folgendes Modell. Ein Gerät, das zum Zeitpunkt 0 installiert wird, habe eine Lebensdauer W_1. Sobald dieses Gerät kaputt geht, wird es durch ein neues baugleiches Gerät ersetzt, das eine Lebensdauer W_2 hat. Sobald dieses Gerät kaputt geht, wird ein neues Gerät installiert, und so weiter. Die Lebensdauer des i-ten Gerätes sei mit W_i bezeichnet. Die Zeitpunkte

$$S_n = W_1 + \ldots + W_n, \quad n \in \mathbb{N},$$

bezeichnet man als *Erneuerungszeiten*, denn zu diesen Zeiten wird ein neues Gerät installiert.

Folgende Annahmen über $W_1, W_2, \ldots$ erscheinen plausibel. Wir nehmen an, dass $W_1, W_2, \ldots$ Zufallsvariablen sind. Da ein Gerät nichts von der Lebensdauer eines anderen mitbekommen kann, nehmen wir an, dass die Zufallsvariablen $W_1, W_2, \ldots$ unabhängig sind. Da alle Geräte die gleiche Bauart haben, nehmen wir an, dass $W_1, W_2, \ldots$ identisch verteilt sind. Welche Verteilung soll es nun sein? Es erscheint plausibel, dass diese Verteilung gedächtnislos sein muss, also werden wir annehmen, dass $W_1, W_2, \ldots$ exponentialverteilt sind.

Definition 8.3.1. Seien $W_1, W_2, \ldots$ unabhängige und mit Parameter $\lambda > 0$ exponentialverteilte Zufallsvariablen. Dann heißt die Folge $S_1, S_2, \ldots$ mit $S_n = W_1 + \ldots + W_n$ ein *Poisson-Prozess* mit Intensität λ.

Abbildung 5. Eine Realisierung des Poisson-Prozesses.

Wie ist nun die n-te Erneuerungszeit S_n verteilt? Da $W_i \sim \mathrm{Exp}(\lambda) \sim \mathrm{Gamma}(1, \lambda)$ ist, ergibt sich aus der Faltungseigenschaft der Gammaverteilung, dass

$$S_n \sim \mathrm{Gamma}(n, \lambda).$$

Diese Verteilung (also die $\mathrm{Gamma}(n, \lambda)$-Verteilung, wobei n eine natürliche Zahl ist), nennt man auch die *Erlang-Verteilung*.

Aufgabe 8.3.2. Zeigen Sie, dass $\mathbb{E}S_n = \frac{n}{\lambda}$ und $\mathrm{Var}\, S_n = \frac{n}{\lambda^2}$.

Wir werden nun kurz auf die Bezeichnung „Poisson-Prozess“ eingehen. Betrachte ein Intervall $I = [a, b] \subset [0, \infty)$ der Länge $l := b - a$. Sei $N(I)$ eine Zufallsvariable, die die Anzahl der Erneuerungen im Intervall I zählt, d.h.:

$$N(I) = \sum_{k=1}^{\infty} \mathbb{1}_{S_k \in I}.$$

Satz 8.3.3. Es gilt $N(I) \sim \mathrm{Poi}(\lambda l)$.

Beweisidee. Betrachte ein sehr kleines Intervall $[t, t+\delta]$, wobei $\delta \approx 0$. Dann gilt aufgrund der Gedächtnislosigkeit der Exponentialverteilung

$$\mathbb{P}[\exists \text{ Erneuerung im Intervall } [t, t+\delta]] = \mathbb{P}[\exists \text{ Erneuerung im Intervall } [0, \delta]].$$

Die Wahrscheinlichkeit, dass es mindestens eine Erneuerung im Intervall $[0, \delta]$ gibt, lässt sich aber folgendermaßen berechnen:

$$\mathbb{P}[\exists \text{ Erneuerung im Intervall } [0, \delta]] = \mathbb{P}[W_1 < \delta] = 1 - \mathrm{e}^{-\lambda\delta} \approx \lambda\delta.$$

Wir können nun ein beliebiges Intervall I der Länge l in $\approx l/\delta$ kleine disjunkte Intervalle der Länge δ zerlegen. Für jedes kleine Intervall der Länge δ ist die Wahrscheinlichkeit, dass es in diesem Intervall mindestens eine Erneuerung gibt, $\approx \lambda\delta$. Außerdem sind verschiedene kleine Intervalle wegen der Gedächtnislosigkeit der Exponentialverteilung unabhängig voneinander. Somit gilt für die Anzahl der Erneuerungen $N(I)$ in einem Intervall I der Länge l

$$N(I) \approx \mathrm{Bin}\left(\frac{l}{\delta}, \lambda\delta\right) \approx \mathrm{Poi}(\lambda l),$$

wobei wir im letzten Schritt den Poisson-Grenzwertsatz benutzt haben. □

8.4. Empirischer Erwartungswert und empirische Varianz einer normalverteilten Stichprobe

Der nächste Satz beschreibt die gemeinsame Verteilung des empirischen Erwartungswerts $\bar{X}_n$ und der empirischen Varianz S_n^2 einer normalverteilten Stichprobe.

Satz 8.4.1. Seien $X_1, \ldots, X_n$ unabhängige und normalverteilte Zufallsvariablen mit Parametern $\mu \in \mathbb{R}$ und $\sigma^2 > 0$. Definiere

$$\bar{X}_n = \frac{X_1 + \ldots + X_n}{n}, \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Dann gelten folgende drei Aussagen:

(1) $\bar{X}_n \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$.
(2) $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$.
(3) Die Zufallsvariablen $\bar{X}_n$ und $\frac{(n-1)S_n^2}{\sigma^2}$ sind unabhängig.

Bemerkung 8.4.2. Teil 3 kann man auch wie folgt formulieren: Die Zufallsvariablen $\bar{X}_n$ und S_n^2 sind unabhängig.

Beweis von Teil 1. Nach Voraussetzung sind $X_1, \ldots, X_n$ normalverteilt mit Parametern (μ, σ^2) und unabhängig. Aus der Faltungseigenschaft der Normalverteilung folgt, dass die

Summe $X_1+\ldots+X_n$ normalverteilt mit Parametern $(n\mu, n\sigma^2)$ ist. Somit ist $\bar{X}_n$ normalverteilt mit Parametern $(\mu, \frac{\sigma^2}{n})$. □

Die in Teil 3 behauptete Unabhängigkeit haben wir bereits aus dem Satz von Basu hergeleitet (s. Korollar 4.14.4). Der Vollständigkeit halber geben wir hier einen unabhängigen, direkten Beweis.

Die folgende Überlegung vereinfacht die Notation im Rest des Beweises. Betrachte die standardisierten Zufallsvariablen

$$X_i' = \frac{X_i - \mu}{\sigma} \sim \mathrm{N}(0,1).$$

Es seien $\bar{X}_n'$ der empirische Mittelwert und $S_n'^2$ die empirische Varianz dieser Zufallsvariablen. Dann gilt

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

Um die Unabhängigkeit von $\bar{X}_n$ und S_n^2 zu zeigen, reicht es, die Unabhängigkeit von $\bar{X}_n'$ und $S_n'^2$ zu zeigen. Außerdem ist

$$\frac{(n-1)S_n^2}{\sigma^2} = \frac{(n-1)S_n'^2}{1}.$$

Für den Rest des Beweises können wir also annehmen, dass $X_1, \ldots, X_n$ standardnormalverteilt sind, ansonsten kann man stattdessen $X_1', \ldots, X_n'$ betrachten.

Beweis von Teil 3. Seien also $X_1, \ldots, X_n \sim \mathrm{N}(0,1)$. Wir zeigen, dass $\bar{X}_n$ und S_n^2 unabhängig sind.

Schritt 1. Wir können S_n^2 als eine Funktion von $X_2 - \bar{X}_n, \ldots, X_n - \bar{X}_n$ auffassen, denn wegen $\sum_{i=1}^n(X_i - \bar{X}_n) = 0$ gilt

$$(n-1)S_n^2 = \left(\sum_{i=2}^n(X_i - \bar{X}_n)\right)^2 + \sum_{i=2}^n(X_i - \bar{X}_n)^2 = \rho(X_2 - \bar{X}_n, \ldots, X_n - \bar{X}_n),$$

wobei

$$\rho(x_2, \ldots, x_n) = \left(\sum_{i=2}^n x_i\right)^2 + \sum_{i=2}^n x_i^2.$$

Somit genügt es zu zeigen, dass die Zufallsvariable $\bar{X}_n$ und der Zufallsvektor $(X_2 - \bar{X}_n, \ldots, X_n - \bar{X}_n)$ unabhängig sind.

Schritt 2. Dazu berechnen wir die gemeinsame Dichte von $(\bar{X}_n, X_2 - \bar{X}_n, \ldots, X_n - \bar{X}_n)$. Die gemeinsame Dichte von $(X_1, \ldots, X_n)$ ist nach Voraussetzung

$$f(x_1, \ldots, x_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^n x_i^2\right).$$

Betrachte nun die Funktion $\psi : \mathbb{R}^n \to \mathbb{R}^n$ mit $\psi = (\psi_1, \ldots, \psi_n)$, wobei

$$\psi_1(x_1, \ldots, x_n) = \bar{x}_n,\ \psi_2(x_1, \ldots, x_n) = x_2 - \bar{x}_n,\ \ldots,\ \psi_n(x_1, \ldots, x_n) = x_n - \bar{x}_n.$$

Die Umkehrfunktion $\phi = \psi^{-1}$ ist somit gegeben durch $\phi = (\phi_1, \ldots, \phi_n)$ mit

$$\phi_1(y_1, \ldots, y_n) = y_1 - \sum_{i=2}^{n} y_i,\ \phi_2(y_1, \ldots, y_n) = y_1 + y_2,\ \ldots,\ \phi_n(y_1, \ldots, y_n) = y_1 + y_n.$$

Die Jacobi-Determinante von ϕ ist konstant (und gleich n, wobei dieser Wert eigentlich nicht benötigt wird).

Schritt 3. Für die Dichte von $(\bar{X}_n, X_2 - \bar{X}_n, \ldots, X_n - \bar{X}_n) = \psi(X_1, \ldots, X_n)$ gilt mit dem Dichtetransformationssatz

$$\begin{aligned} g(y_1, \ldots, y_n) &= nf(x_1, \ldots, x_n) \\ &= \frac{n}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\left(y_1 - \sum_{i=2}^{n} y_i \right)^2 - \frac{1}{2}\sum_{i=2}^{n} (y_1 + y_i)^2 \right) \\ &= \frac{n}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{ny_1^2}{2} \right) \exp\left(-\frac{1}{2}\sum_{i=2}^{n} y_i^2 - \frac{1}{2}\left(\sum_{i=2}^{n} y_i \right)^2 \right). \end{aligned}$$

Somit ist $\bar{X}_n$ unabhängig von $(X_2 - \bar{X}_n, \ldots, X_n - \bar{X}_n)$. □

Beweis von Teil 2. Es gilt die Identität

$$Z := \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 + (\sqrt{n}\bar{X}_n)^2 =: Z_1 + Z_2.$$

Dabei gilt:

(1) $Z \sim \chi_n^2$ nach Definition der χ^2-Verteilung.
(2) $Z_2 \sim \chi_1^2$, denn $\sqrt{n}\bar{X}_n \sim \mathrm{N}(0, 1)$.
(3) Z_1 und Z_2 sind unabhängig (wegen Teil 3 des Satzes).

Damit ergibt sich für die charakteristische Funktion von Z_1:

$$\varphi_{Z_1}(t) = \frac{\varphi_Z(t)}{\varphi_{Z_2}(t)} = \frac{\frac{1}{(1-2it)^{\frac{n}{2}}}}{\frac{1}{(1-2it)^{\frac{1}{2}}}} = \frac{1}{(1-2it)^{(n-1)/2}}.$$

Somit ist $(n-1)S_n^2 = Z_1 \sim \mathrm{Gamma}(\frac{n-1}{2}, \frac{1}{2}) \overset{d}{=} \chi_{n-1}^2$. □

8.5. Student-t-Verteilung

Definition 8.5.1. Seien $X \sim \mathrm{N}(0, 1)$ und $U \sim \chi_r^2$ unabhängige Zufallsvariablen, wobei $r \in \mathbb{N}$. Die Zufallvariable

$$V = \frac{X}{\sqrt{\frac{U}{r}}}$$

heißt *t-verteilt* mit r Freiheitsgraden.

Notation 8.5.2. $V \sim t_r$.

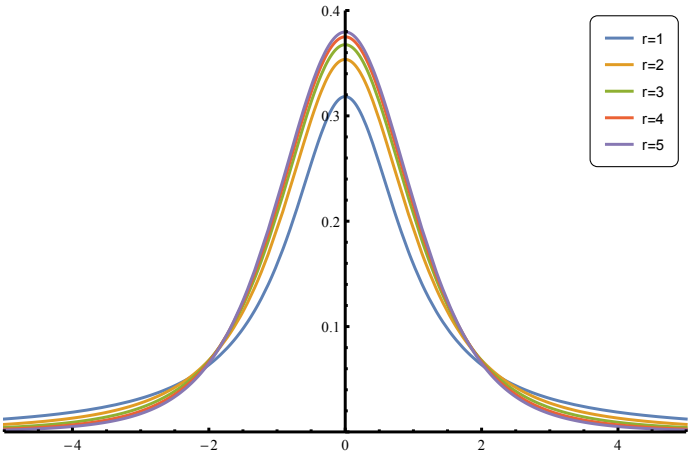


ABBILDUNG 6. Dichten der t_r-Verteilungen mit $r = 1, \ldots, 5$ Freiheitsgraden.

Satz 8.5.3. Die Dichte einer t-verteilten Zufallsvariable $V \sim t_r$ ist gegeben durch

$$f_V(t) = \frac{\Gamma(\frac{r+1}{2})}{\Gamma(\frac{r}{2})} \frac{1}{\sqrt{r\pi}(1+\frac{t^2}{r})^{\frac{r+1}{2}}}, \quad t \in \mathbb{R}.$$

Beweis. Nach Definition der t-Verteilung gilt die Darstellung

$$V = \frac{X}{\sqrt{\frac{U}{r}}},$$

wobei $X \sim \mathrm{N}(0,1)$ und $U \sim \chi^2_r$ unabhängig sind. Die gemeinsame Dichte von X und U ist gegeben durch

$$f_{X,U}(x,u) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}} \cdot \frac{u^{\frac{r-2}{2}} \mathrm{e}^{-\frac{u}{2}}}{2^{\frac{r}{2}}\Gamma(\frac{r}{2})}, \quad x \in \mathbb{R},\ u > 0.$$

Betrachten wir nun die Abbildung $(x,u) \mapsto (v,w)$ mit

$$v = \frac{x}{\sqrt{\frac{u}{r}}},\ w = u.$$

Die Umkehrabbildung ist somit $x = v\sqrt{\frac{w}{r}}$ und $u = w$. Die Jacobi-Determinante der Umkehrabbildung ist $\sqrt{\frac{w}{r}}$. Somit gilt für die gemeinsame Dichte von (V,W)

$$f_{V,W}(v,w) = f_{X,Y}(x,y)\sqrt{\frac{w}{r}} = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{v^2 w}{2r}\right) \frac{w^{\frac{r-2}{2}} \mathrm{e}^{-\frac{w}{2}}}{2^{\frac{r}{2}}\Gamma(\frac{r}{2})} \sqrt{\frac{w}{r}}.$$

Somit kann die Dichte von V wie folgt berechnet werden:

$$f_V(v) = \int_0^\infty f_{V,W}(v,w)\mathrm{d}w = \frac{1}{\sqrt{2\pi r}2^{r/2}\Gamma(\frac{r}{2})} \int_0^\infty \exp\left(-w\left(\frac{v^2}{2r}+\frac{1}{2}\right)\right) w^{\frac{r+1}{2}-1}\mathrm{d}w.$$

Mit der Formel $\int_0^\infty w^{\alpha-1}\mathrm{e}^{-\lambda w}\mathrm{d}w = \frac{\Gamma(\alpha)}{\lambda^\alpha}$ berechnet sich das Integral zu

$$f_V(v) = \frac{1}{\sqrt{2\pi r}2^{r/2}\Gamma(\frac{r}{2})}\frac{\Gamma(\frac{r+1}{2})}{(\frac{v^2}{2r}+\frac{1}{2})^{\frac{r+1}{2}}} = \frac{\Gamma(\frac{r+1}{2})}{\Gamma(\frac{r}{2})}\frac{1}{\sqrt{r\pi}(1+\frac{v^2}{r})^{\frac{r+1}{2}}}.$$

Dies ist genau die gewünschte Formel. □

Beispiel 8.5.4. Die Dichte der t_1-Verteilung ist $f_V(t) = \frac{1}{\pi}\frac{1}{1+t^2}$ und stimmt somit mit der Dichte der Cauchy-Verteilung überein.

Aufgabe 8.5.5. Zeigen Sie, dass für $r \to \infty$ die Dichte der t_r-Verteilung punktweise gegen die Dichte der Standardnormalverteilung konvergiert.

Satz 8.5.6. Seien $X_1, \ldots, X_n$ unabhängige und normalverteilte Zufallsvariablen mit Parametern (μ, σ^2). Dann gilt

$$\sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma} \sim \mathrm{N}(0,1), \quad \sqrt{n}\,\frac{\bar{X}_n - \mu}{S_n} \sim t_{n-1}.$$

Beweis. Die erste Formel folgt aus der Tatsache, dass $\bar{X}_n \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$. Wir beweisen die zweite Formel. Es gilt die Darstellung

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

Da nach Satz 8.4.1 die Zufallsvariablen $\sqrt{n}\,\frac{\bar{X}_n-\mu}{\sigma} \sim \mathrm{N}(0,1)$ und $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$ unabhängig sind, hat $\sqrt{n}\,\frac{\bar{X}_n-\mu}{S_n}$ eine t-Verteilung mit $n-1$ Freiheitsgraden. □

8.6. Fisher-F-Verteilung

Definition 8.6.1. Seien $r, s \in \mathbb{N}$ Parameter. Seien $U_r \sim \chi^2_r$ und $U_s \sim \chi^2_s$ unabhängige Zufallsvariablen. Dann heißt die Zufallsvariable

$$W = \frac{U_r/r}{U_s/s}$$

F-verteilt mit (r,s)-Freiheitsgraden.

Notation 8.6.2. $W \sim F_{r,s}$.

Satz 8.6.3. Die Dichte einer F-verteilten Zufallsvariable $W \sim F_{r,s}$ ist gegeben durch

$$f_W(t) = \frac{\Gamma(\frac{r+s}{2})}{\Gamma(\frac{r}{2})\Gamma(\frac{s}{2})}\left(\frac{r}{s}\right)^{\frac{r}{2}}\frac{t^{\frac{r}{2}-1}}{\left(1+\frac{r}{s}t\right)^{\frac{r+s}{2}}}, \quad t > 0.$$

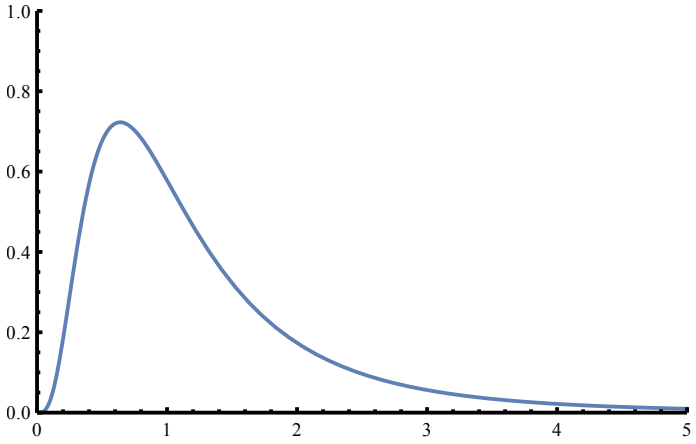


ABBILDUNG 7. Dichte der F-Verteilung mit $(10, 8)$-Freiheitsgraden.

Beweis. Wir werden die Dichte nur bis auf die multiplikative Konstante berechnen. Die Konstante ergibt sich dann aus der Bedingung, dass die Dichte sich zu 1 integriert. Wir schreiben $f(t) \propto g(t)$, falls $f(t) = Cg(t)$, wobei C eine Konstante ist.
Wir haben die Darstellung

$$W = \frac{U_r/r}{U_s/s},$$

wobei $U_r \sim \chi^2_r$ und $U_s \sim \chi^2_s$ unabhängig sind. Für die Dichten von U_r und U_s gilt

$$f_{U_r}(x) \propto x^{\frac{r-2}{2}} \mathrm{e}^{-x/2}, \quad f_{U_s}(x) \propto x^{\frac{s-2}{2}} \mathrm{e}^{-x/2}, \quad x > 0.$$

Somit folgt für die Dichten von U_r/r und U_s/s, dass

$$f_{U_r/r}(x) \propto x^{\frac{r-2}{2}} \mathrm{e}^{-rx/2}, \quad f_{U_s/s}(x) \propto x^{\frac{s-2}{2}} \mathrm{e}^{-sx/2}, \quad x > 0.$$

Für die Dichte von W gilt die Faltungsformel:

$$f_W(t) = \int_{\mathbb{R}} |y| f_{U_r/r}(yt) f_{U_s/s}(y) \mathrm{d}y.$$

Indem wir nun die Dichten von U_r/r und U_s/s einsetzen, erhalten wir, dass

$$f_W(t) \propto \int_0^\infty y(yt)^{\frac{r-2}{2}} \mathrm{e}^{-\frac{ryt}{2}} y^{\frac{s-2}{2}} \mathrm{e}^{-\frac{sy}{2}} \mathrm{d}y \propto t^{\frac{r-2}{2}} \int_0^\infty y^{1+\frac{r-2}{2}+\frac{s-2}{2}} \exp\left(-y\left(\frac{rt}{2}+\frac{s}{2}\right)\right) \mathrm{d}y.$$

Mit der Formel $\int_0^\infty y^{\alpha-1}\mathrm{e}^{-\lambda y}\mathrm{d}y = \frac{\Gamma(\alpha)}{\lambda^\alpha}$ berechnet sich das Integral zu

$$f_W(t) \propto t^{\frac{r-2}{2}} \frac{1}{(rt+s)^{\frac{r+s}{2}}} \propto \frac{t^{\frac{r-2}{2}}}{(1+\frac{r}{s}t)^{\frac{r+s}{2}}}.$$

Dies ist genau die gewünschte Dichte, bis auf eine multiplikative Konstante. □

KAPITEL 9

Konfidenzintervalle

9.1. Definition eines Konfidenzintervalls

Sei $(\mathbb{P}_\theta)_{\theta\in\Theta}$ eine Familie von Wahrscheinlichkeitsmaßen auf dem Stichprobenraum $(\mathfrak{X}, \mathcal{A})$. In diesem Kapitel ist $\Theta = (a, b) \subset \mathbb{R}$ ein Intervall. Seien X eine gemäß $\mathbb{P}_\theta$ verteilte Stichprobe. Wir haben uns bereits mit der Frage beschäftigt, wie man den unbekannten Parameter θ anhand der Stichprobe X schätzen kann. Bei einer solchen Schätzung bleibt aber unklar, wie groß der mögliche Fehler, also die Differenz $\hat{\theta} - \theta$, ist. In der Statistik begnügt man sich normalerweise nicht mit der Angabe eines Schätzers, sondern versucht auch den Schätzfehler zu bestimmen, indem man ein sogenanntes Konfidenzintervall für θ angibt. Das Ziel ist es, das Intervall so zu konstruieren, dass es den wahren Wert des Parameters θ mit einer großen Wahrscheinlichkeit (typischerweise 0.99 oder 0.95) enthält.

Definition 9.1.1. Sei $\alpha \in (0, 1)$ eine kleine Zahl, typischerweise $\alpha = 0.01$ oder $\alpha = 0.05$. Es seien $\underline{\theta} : \mathfrak{X} \to \mathbb{R} \cup \{-\infty\}$ und $\overline{\theta} : \mathfrak{X} \to \mathbb{R} \cup \{+\infty\}$ zwei Stichprobenfunktionen mit

$$\underline{\theta}(x) \leq \overline{\theta}(x) \quad \text{für alle } x \in \mathfrak{X}.$$

Wir sagen, dass $[\underline{\theta}, \overline{\theta}]$ ein *Konfidenzintervall* für θ zum *Konfidenzniveau* $1 - \alpha \in (0, 1)$ ist, falls

$$\mathbb{P}_\theta[\underline{\theta}(X) \leq \theta \leq \overline{\theta}(X)] \geq 1 - \alpha \text{ für alle } \theta \in \Theta.$$

Somit ist die Wahrscheinlichkeit, dass das zufällige Intervall $(\underline{\theta}, \overline{\theta})$ den richtigen Wert θ enthält, mindestens $1 - \alpha$, also typischerweise 0.99 oder 0.95.

Die allgemeine Vorgehensweise bei der Konstruktion der Konfidenzintervalle ist diese: Man versucht, eine sogenannte *Pivot-Statistik* zu finden, d.h. eine Funktion $T(X; \theta)$ der Stichprobe X und des unbekannten Parameters θ mit der Eigenschaft, dass die Verteilung von $T(X; \theta)$ unter $\mathbb{P}_\theta$ nicht von θ abhängt und explizit angegeben werden kann. Das heißt, es soll gelten, dass

$$\mathbb{P}_\theta[T(X; \theta) \leq t] = F(t),$$

wobei $F(t)$ nicht von θ abhängt. Dabei soll die Funktion $T(X; \theta)$ den Parameter θ tatsächlich auf eine nichttriviale Weise enthalten. Für $\alpha \in (0, 1)$ bezeichnen wir mit Q_α das α-Quantil der Verteilungsfunktion F, d. h. die Lösung der Gleichung $F(Q_\alpha) = \alpha$. Dann gilt

$$\mathbb{P}_\theta \left[Q_{\frac{\alpha}{2}} \leq T(X; \theta) \leq Q_{1-\frac{\alpha}{2}} \right] = 1 - \alpha \text{ für alle } \theta \in \Theta.$$

Indem wir nun diese Ungleichung nach θ auflösen, erhalten wir ein Konfidenzintervall für θ zum Konfidenzniveau $1 - \alpha$.

Im Folgenden werden wir verschiedene Beispiele von Konfidenzintervallen betrachten.

Aufgabe 9.1.2. Seien $X_1, \ldots, X_n$ gleichverteilt auf $[0, \theta]$, wobei $\theta > 0$ unbekannt ist. Sei $M_n := \max\{X_1, \ldots, X_n\}$. Bestimmen Sie ein $a > 1$ derart, dass $[M_n, aM_n]$ ein Konfidenzintervall für θ zum Niveau $1 - \alpha$ bildet.

Aufgabe 9.1.3. Seien $X_1, \ldots, X_n$ unabhängige Zufallsvariablen mit der Dichte $h_\theta(x) = \mathrm{e}^{-(x-\theta)} \mathbb{1}_{x \geq \theta}$, wobei $\theta \in \mathbb{R}$ der unbekannte Parameter sei. Als Schätzer für θ betrachten wir $Y = \min\{X_1, \ldots, X_n\}$. Finden Sie für ein gegebenes $\alpha \in (0, 1)$ Zahlen p und q (die nicht von θ abhängen) mit

$$\mathbb{P}_\theta[\theta < Y + p] = \mathbb{P}_\theta[\theta > Y + q] = \frac{\alpha}{2} \text{ für alle } \theta \in \mathbb{R}.$$

Konstruieren Sie ein Konfidenzintervall zum Niveau $1 - \alpha$ für θ.

9.2. Konfidenzintervalle für die Parameter der Normalverteilung

In diesem Abschnitt seien $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ unabhängige und mit Parametern (μ, σ^2) normalverteilte Zufallsvariablen. Unser Ziel ist es, Konfidenzintervalle für μ und σ^2 zu konstruieren. Dabei werden wir vier Fälle betrachten:

(1) Konfidenzintervall für μ bei bekanntem σ^2.
(2) Konfidenzintervall für μ bei unbekanntem σ^2.
(3) Konfidenzintervall für σ^2 bei bekanntem μ.
(4) Konfidenzintervall für σ^2 bei unbekanntem μ.

Fall 1: Konfidenzintervall für μ bei bekanntem σ^2. Es seien also $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ unabhängig, wobei μ unbekannt und σ^2 bekannt seien. Wir konstruieren ein Konfidenzintervall für μ. Ein natürlicher Schätzer für μ ist $\bar{X}_n$. Wir haben gezeigt, dass

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

Wir werden nun $\bar{X}_n$ standardisieren:

$$\sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma} \sim \mathrm{N}(0, 1).$$

Für $\alpha \in (0, 1)$ sei z_α das α-Quantil der Standardnormalverteilung. D.h., z_α sei die Lösung der Gleichung $\Phi(z_\alpha) = \alpha$, wobei Φ die Verteilungsfunktion der Standardnormalverteilung bezeichnet. Somit gilt

$$\mathbb{P}_\mu\left[z_{\frac{\alpha}{2}} \leq \sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma} \leq z_{1-\frac{\alpha}{2}}\right] = 1 - \alpha \text{ für alle } \mu \in \mathbb{R}.$$

Nach μ umgeformt führt dies zu

$$\mathbb{P}_\mu\left[\bar{X}_n - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha \text{ für alle } \mu \in \mathbb{R}.$$

Wegen der Symmetrie der Normalverteilung ist $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$. Somit ist ein Konfidenzintervall zum Niveau $1 - \alpha$ für μ gegeben durch

$$\left[\bar{X}_n - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\ \bar{X}_n + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right].$$

Der Mittelpunkt dieses Intervalls ist $\bar{X}_n$.

Bemerkung 9.2.1. Man kann auch „nichtsymmetrische“ Konfidenzintervalle konstruieren. Wähle dazu $\alpha_1 \geq 0$, $\alpha_2 \geq 0$ mit $\alpha = \alpha_1 + \alpha_2$. Dann gilt

$$\mathbb{P}_\mu \left[z_{\alpha_1} \leq \sqrt{n}\, \frac{\bar{X}_n - \mu}{\sigma} \leq z_{1-\alpha_2} \right] = 1 - \alpha \quad \text{für alle } \mu \in \mathbb{R}.$$

Nach μ umgeformt führt dies zu

$$\mathbb{P}_\mu \left[\bar{X}_n - z_{1-\alpha_2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n - z_{\alpha_1} \frac{\sigma}{\sqrt{n}} \right] = 1 - \alpha \quad \text{für alle } \mu \in \mathbb{R}.$$

Wegen $z_{\alpha_1} = -z_{1-\alpha_1}$ führt dies zu folgendem Konfidenzintervall für μ:

$$\left[\bar{X}_n - z_{1-\alpha_2} \frac{\sigma}{\sqrt{n}},\ \bar{X}_n + z_{1-\alpha_1} \frac{\sigma}{\sqrt{n}} \right].$$

Interessiert man sich z. B. nur für eine obere Schranke für μ, so kann man $\alpha_1 = \alpha$ und $\alpha_2 = 0$ wählen. Dann erhält man folgendes Konfidenzintervall für μ:

$$\left[-\infty,\ \bar{X}_n + z_{1-\alpha} \frac{\sigma}{\sqrt{n}} \right].$$

Die Konstruktion der nichtsymmetrischen Konfidenzintervalle lässt sich auch für die nachfolgenden Beispiele durchführen, wird aber hier nicht mehr wiederholt.

Fall 2: Konfidenzintervall für μ bei unbekanntem σ^2. Es seien $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ unabhängig, wobei μ und σ^2 beide unbekannt seien. Wir konstruieren ein Konfidenzintervall für μ. Es gilt zwar nach wie vor, dass $\sqrt{n}\, \frac{\bar{X}_n - \mu}{\sigma} \sim \mathrm{N}(0,1)$, wir können das aber nicht für die Konstruktion eines Konfidenzintervalls für μ benutzen, denn der Parameter σ^2 ist unbekannt. Wir werden deshalb σ^2 durch einen Schätzer, nämlich $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$, ersetzen. Wir haben im vorigen Kapitel gezeigt, dass

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

Sei $t_{n-1,\alpha}$ das α-Quantil der t_{n-1}-Verteilung. Somit gilt

$$\mathbb{P}_{\mu,\sigma^2} \left[t_{n-1,\frac{\alpha}{2}} \leq \sqrt{n}\, \frac{\bar{X}_n - \mu}{S_n} \leq t_{n-1,1-\frac{\alpha}{2}} \right] = 1 - \alpha \quad \text{für alle } \mu \in \mathbb{R}, \sigma^2 > 0.$$

Nach μ umgeformt führt dies zu

$$\mathbb{P}_{\mu,\sigma^2} \left[\bar{X}_n - t_{n-1,1-\frac{\alpha}{2}} \frac{S_n}{\sqrt{n}} \leq \mu \leq \bar{X}_n - t_{n-1,\frac{\alpha}{2}} \frac{S_n}{\sqrt{n}} \right] = 1 - \alpha \quad \text{für alle } \mu \in \mathbb{R}, \sigma^2 > 0.$$

Wegen der Symmetrie der t-Verteilung gilt $t_{n-1,\frac{\alpha}{2}} = -t_{n-1,1-\frac{\alpha}{2}}$. Somit erhalten wir folgendes Konfidenzintervall für μ zum Niveau $1 - \alpha$:

$$\left[\bar{X}_n - t_{n-1,1-\frac{\alpha}{2}} \frac{S_n}{\sqrt{n}},\ \bar{X}_n + t_{n-1,1-\frac{\alpha}{2}} \frac{S_n}{\sqrt{n}} \right].$$

Fall 3: Konfidenzintervall für σ^2 bei bekanntem μ. Seien nun $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$, wobei μ bekannt und σ^2 unbekannt seien. Wir konstruieren ein Konfidenzintervall für σ^2.

Ein natürlicher Schätzer für σ^2 ist

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

Dann gilt

$$\frac{n\widetilde{S}_n^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi_n^2.$$

Sei $\chi^2_{n,\alpha}$ das α-Quantil der χ^2-Verteilung mit n Freiheitsgraden. Dann gilt

$$\mathbb{P}_{\sigma^2}\left[\chi^2_{n,\frac{\alpha}{2}} \leq \frac{n\widetilde{S}_n^2}{\sigma^2} \leq \chi^2_{n,1-\frac{\alpha}{2}}\right] = 1-\alpha \text{ für alle } \sigma^2 > 0.$$

Nach σ^2 umgeformt führt dies zu folgendem Konfidenzintervall für σ^2 zum Niveau $1-\alpha$:

$$\left[\frac{n\widetilde{S}_n^2}{\chi^2_{n,1-\frac{\alpha}{2}}}, \frac{n\widetilde{S}_n^2}{\chi^2_{n,\frac{\alpha}{2}}}\right].$$

Es sei bemerkt, dass die χ^2-Verteilung nicht symmetrisch ist.

Fall 4: Konfidenzintervall für σ^2 bei unbekanntem μ. Seien $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$, wobei μ und σ^2 beide unbekannt seien. Wir konstruieren ein Konfidenzintervall für σ^2. Ein natürlicher Schätzer für σ^2 ist

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

Bekannt ist, dass

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

Somit gilt

$$\mathbb{P}_{\mu,\sigma^2}\left[\chi^2_{n-1,\frac{\alpha}{2}} \leq \frac{(n-1)S_n^2}{\sigma^2} \leq \chi^2_{n-1,1-\frac{\alpha}{2}}\right] = 1-\alpha \text{ für alle } \mu \in \mathbb{R}, \sigma^2 > 0.$$

Nach σ^2 umgeformt führt dies zu

$$\mathbb{P}_{\mu,\sigma^2}\left[\frac{(n-1)S_n^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{(n-1)S_n^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right] = 1-\alpha \text{ für alle } \mu \in \mathbb{R}, \sigma^2 > 0.$$

Somit erhält man folgendes Konfidenzintervall für σ^2 zum Niveau $1-\alpha$

$$\left[\frac{(n-1)S_n^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{(n-1)S_n^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right].$$

Aufgabe 9.2.2. Für jedes $r \in \mathbb{N}$ sei Z_r eine χ^2-verteilte Zufallsvariable mit r Freiheitsgraden. Zeigen Sie, dass

$$\frac{Z_r - r}{\sqrt{2r}} \xrightarrow[r\to\infty]{d} N(0,1).$$

Aufgabe 9.2.3. Für $\alpha \in (0,1)$ seien z_α und $\chi^2_{r,\alpha}$ die α-Quantile der Standardnormalverteilung $N(0,1)$ bzw. der χ^2_r-Verteilung. Zeigen Sie, dass

$$\lim_{r\to\infty} \frac{\chi^2_{r,\alpha} - r}{\sqrt{2r}} = z_\alpha.$$

Aufgabe 9.2.4. Es seien n Geräte gleicher Bauart gegeben. Die Lebensdauer des Geräts i werde mit einer Zufallsvariable X_i modelliert. Dabei seien $X_1, \ldots, X_n$ unabhängig und exponentialverteilt mit Parameter $\theta > 0$. Es soll ein Konfidenzintervall für die erwartete Lebensdauer $\frac{1}{\theta}$ bestimmt werden.

(1) Zeigen Sie, dass $2n\theta\bar{X}_n$ eine χ^2-Verteilung mit $2n$ Freiheitsgraden hat.
(2) Konstruieren Sie ein Konfidenzintervall zum Niveau $1-\alpha$ für $\frac{1}{\theta}$.

9.3. Zweistichprobenprobleme

Bislang haben wir nur sogenannte Einstichprobenprobleme betrachtet. Es gibt aber auch mehrere Probleme, bei denen man zwei Stichproben miteinander vergleichen muss.

Beispiel 9.3.1. Es sollen zwei Futterarten für Masttiere verglichen werden. Dazu betrachtet man zwei Gruppen von Tieren. Die erste, aus n Tieren bestehende Gruppe bekommt Futter 1. Die zweite, aus m Tieren bestehende Gruppe, bekommt Futter 2. Mit $X_1, \ldots, X_n$ wird die Gewichtszunahme der Tiere der ersten Gruppe notiert. Entsprechend bezeichnen wir die Gewichtszunahmen der Tiere aus der zweiten Gruppe mit $Y_1, \ldots, Y_m$. Die Aufgabe besteht nun darin, die beiden Futterarten zu vergleichen, also ein Konfidenzintervall für $\mu_1 - \mu_2$ zu finden, wobei μ_1 bzw. μ_2 der Erwartungswert der ersten bzw. der zweiten Stichprobe ist.

Beispiel 9.3.2. Es wurden zwei Messverfahren zur Bestimmung einer physikalischen Größe entwickelt. Es soll nun ermittelt werden, welches Verfahren eine größere Genauigkeit (also eine kleinere Streuung der Messergebnisse) hat. Dazu wird die physikalische Größe zuerst n Mal mit dem ersten Verfahren gemessen, und dann m Mal mit dem zweiten Verfahren. Es ergeben sich zwei Stichproben $X_1, \ldots, X_n$ und $Y_1, \ldots, Y_m$. Diesmal sollen die Streuungen der beiden Stichproben verglichen werden, also ein Konfidenzintervall für σ_1^2/σ_2^2 konstruiert werden, wobei σ_1^2 bzw. σ_2^2 die Varianz der ersten bzw. der zweiten Stichprobe ist.

Für die obigen Beispiele erscheint folgendes Modell plausibel. Wir betrachten zwei Stichproben $X_1, \ldots, X_n$ und $Y_1, \ldots, Y_m$. Wir nehmen an, dass

(1) $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ unabhängige Zufallsvariablen sind.
(2) $X_1, \ldots, X_n \sim \mathrm{N}(\mu_1, \sigma_1^2)$.
(3) $Y_1, \ldots, Y_m \sim \mathrm{N}(\mu_2, \sigma_2^2)$.

Wir werden Konfidenzintervalle für $\mu_1 - \mu_2$ und σ_1^2/σ_2^2 konstruieren.

Fall 1: Konfidenzintervall für $\mu_1 - \mu_2$ bei bekannten σ_1^2 und σ_2^2. Es seien also σ_1^2 und σ_2^2 bekannt. Da $X_1, \ldots, X_n \sim \mathrm{N}(\mu_1, \sigma_1^2)$ und $Y_1, \ldots, Y_m \sim \mathrm{N}(\mu_2, \sigma_2^2)$, folgt aus der Faltungseigenschaft der Normalverteilung, dass

$$\bar{X}_n := \frac{X_1 + \ldots + X_n}{n} \sim \mathrm{N}\left(\mu_1, \frac{\sigma_1^2}{n}\right), \quad \bar{Y}_m := \frac{X_1 + \ldots + Y_m}{m} \sim \mathrm{N}\left(\mu_2, \frac{\sigma_2^2}{m}\right).$$

Ein natürlicher Schätzer für $\mu_1 - \mu_2$ ist gegeben durch

$$\bar{X}_n - \bar{Y}_m \sim \mathrm{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right).$$

Indem der Erwartungswert subtrahiert und durch die Standardabweichung geteilt wird, erhält man eine standardnormalverteilte Zufallsvariable:

$$\frac{\bar{X}_n - \bar{Y}_m - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim \mathrm{N}(0,1).$$

Es gilt also, dass

$$\mathbb{P}_{\mu_1,\mu_2}\left[z_{\frac{\alpha}{2}} \leq \frac{\bar{X}_n - \bar{Y}_m - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \leq z_{1-\frac{\alpha}{2}}\right] = 1 - \alpha \quad \text{für alle } \mu_1, \mu_2 \in \mathbb{R}.$$

Aufgrund der Symmetrieeigenschaft der Normalverteilung können wir $z = z_{1-\frac{\alpha}{2}} = -z_{\frac{\alpha}{2}}$ definieren. Umgeformt nach $\mu_1 - \mu_2$ erhält man das Konfidenzintervall

$$\left[\bar{X}_n - \bar{Y}_m - z\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}},\ \bar{X}_n - \bar{Y}_m + z\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\right].$$

Fall 2: Konfidenzintervall für $\mu_1 - \mu_2$ bei unbekannten aber gleichen σ_1^2 und σ_2^2. Seien nun σ_1^2 und σ_2^2 unbekannt. Um das Problem zu vereinfachen, werden wir annehmen, dass σ_1^2 und σ_2^2 gleich sind, d. h. $\sigma^2 := \sigma_1^2 = \sigma_2^2$.

Schritt 1. Genauso wie in Fall 1 gilt

$$\frac{\bar{X}_n - \bar{Y}_m - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim \mathrm{N}(0,1).$$

Leider können wir das nicht zur Konstruktion eines Konfidenzintervalls für $\mu_1 - \mu_2$ direkt verwenden, denn σ ist unbekannt. Wir werden deshalb σ schätzen.

Schritt 2. Ein Schätzer für σ^2, der nur auf der ersten Stichprobe basiert, ist gegeben durch

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

Analog gibt es einen Schätzer für σ^2, der nur auf der zweiten Stichprobe basiert:

$$S_Y^2 = \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \bar{Y}_m)^2.$$

Für diese Schätzer gilt

$$\frac{(n-1)S_X^2}{\sigma^2} \sim \chi_{n-1}^2, \quad \frac{(m-1)S_Y^2}{\sigma^2} \sim \chi_{m-1}^2.$$

Bemerke, dass diese zwei χ^2-verteilten Zufallsvariablen unabhängig sind. Somit folgt

$$\frac{(n-1)S_X^2}{\sigma^2} + \frac{(m-1)S_Y^2}{\sigma^2} \sim \chi_{n+m-2}^2.$$

Betrachte nun folgenden Schätzer für σ^2, der auf beiden Stichproben basiert:

$$S^2 = \frac{1}{n+m-2}\left(\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 + \sum_{j=1}^{m}(Y_j - \bar{Y}_m)^2\right) = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}.$$

Somit gilt

$$\frac{(n+m-2)S^2}{\sigma^2} \sim \chi^2_{n+m-2}.$$

Der Erwartungswert einer χ^2_{n+m-2}-verteilten Zufallsvariable ist $n+m-2$. Daraus folgt insbesondere, dass S^2 ein erwartungstreuer Schätzer für σ^2 ist, was die Wahl der Normierung $1/(n+m-2)$ erklärt.

Schritt 3. Aus Schritt 1 und Schritt 2 folgt, dass

$$\frac{\bar{X}_n - \bar{Y}_m - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n}+\frac{1}{m}}} = \frac{\frac{\bar{X}_n - \bar{Y}_m - (\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}}{\sqrt{\frac{1}{n+m-2}\frac{(n+m-2)S^2}{\sigma^2}}} \sim t_{n+m-2}.$$

Dabei haben wir benutzt, dass der Zähler und der Nenner des obigen Bruchs unabhängig voneinander sind. Das folgt aus der Tatsache, dass S_X^2 und $\bar{X}_n$ sowie S_Y^2 und $\bar{Y}_n$ unabhängig voneinander sind, sowie aus der Tatsache, dass die Vektoren (X, S_X^2) und (Y, S_Y^2) unabhängig voneinander sind. Somit gilt

$$\mathbb{P}_{\mu_1,\mu_2,\sigma^2}\left[t_{n+m-2,\frac{\alpha}{2}} \le \frac{\bar{X}_n - \bar{Y}_m - (\mu_1-\mu_2)}{S\sqrt{\frac{1}{n}+\frac{1}{m}}} \le t_{n+m-2,1-\frac{\alpha}{2}}\right] = 1-\alpha \text{ für alle } \mu_1, \mu_2, \sigma^2.$$

Wegen der Symmetrie der t-Verteilung gilt $t := t_{n+m-2,1-\frac{\alpha}{2}} = -t_{n+m-2,\frac{\alpha}{2}}$. Umgeformt nach $\mu_1 - \mu_2$ ergibt sich folgendes Konfidenzintervall für $\mu_1 - \mu_2$ zum Konfidenzniveau $1-\alpha$:

$$\left[\bar{X}_n - \bar{Y}_m - S\sqrt{\frac{1}{n}+\frac{1}{m}}\,t,\ \bar{X}_n - \bar{Y}_m + S\sqrt{\frac{1}{n}+\frac{1}{m}}\,t\right].$$

Fall 3: Konfidenzintervall für σ_1^2/σ_2^2 bei unbekannten μ_1 und μ_2. Seien also μ_1 und μ_2 unbekannt. Wir konstruieren ein Konfidenzintervall für σ_1^2/σ_2^2. Die natürlichen Schätzer für σ_1^2 und σ_2^2 sind gegeben durch

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2, \quad S_Y^2 = \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \bar{Y}_m)^2.$$

Es gilt

$$\frac{(n-1)S_X^2}{\sigma_1^2} \sim \chi^2_{n-1}, \quad \frac{(m-1)S_Y^2}{\sigma_2^2} \sim \chi^2_{m-1}$$

und diese beiden Zufallsvariablen sind unabhängig. Es folgt, dass

$$\frac{S_X^2/\sigma_1^2}{S_Y^2/\sigma_2^2} = \frac{\frac{(n-1)S_X^2}{\sigma_1^2}\cdot\frac{1}{n-1}}{\frac{(m-1)S_Y^2}{\sigma_2^2}\cdot\frac{1}{m-1}} \sim F_{n-1,m-1}.$$

Wir bezeichnen mit $F_{n-1,m-1,\alpha}$ das α-Quantil der $F_{n-1,m-1}$-Verteilung. Deshalb gilt, dass

$$\mathbb{P}_{\mu_1,\mu_2,\sigma_1^2,\sigma_2^2}\left[F_{n-1,m-1,\frac{\alpha}{2}} \leq \frac{S_X^2/\sigma_1^2}{S_Y^2/\sigma_2^2} \leq F_{n-1,m-1,1-\frac{\alpha}{2}}\right] = 1-\alpha \text{ für alle } \mu_1,\mu_2,\sigma_1^2,\sigma_2^2 > 0.$$

Somit ergibt sich folgendes Konfidenzintervall für σ_1^2/σ_2^2 zum Konfidenzniveau $1-\alpha$:

$$\left[\frac{1}{F_{n-1,m-1,1-\frac{\alpha}{2}}} \cdot \frac{S_X^2}{S_Y^2},\ \frac{1}{F_{n-1,m-1,\frac{\alpha}{2}}} \cdot \frac{S_X^2}{S_Y^2}\right].$$

Fall 4: Konfidenzintervall für σ_1^2/σ_2^2 bei bekannten μ_1 und μ_2. Ähnlich wie in Fall 3 (Übungsaufgabe).

Zum Schluss betrachten wir ein Beispiel, bei dem es sich nur scheinbar um ein Zweistichprobenproblem handelt.

Beispiel 9.3.3 (Verbundene Stichproben)**.** Bei einem Psychologietest füllen n Personen jeweils einen Fragebogen aus. Die Fragebögen werden ausgewertet und die Ergebnisse der Personen mit $X_1,\dots,X_n$ notiert. Nach der Therapiezeit werden von den gleichen Personen die Ergebnisse mit $Y_1,\dots,Y_n$ festgehalten. In diesem Modell gibt es zwei Stichproben, allerdings sind die Annahmen des Zweistichprobenmodells hier nicht plausibel. Es ist nämlich klar, dass X_1 und Y_1 abhängig sind, denn beide Ergebnisse gehören zu derselben Person. Allgemeiner sind X_i und Y_i abhängig. Eine bessere Vorgehensweise bei diesem Problem ist diese. Wir betrachten die Zuwächse $Z_i = Y_i - X_i$. Diese können wir als unabhängige Zufallsvariablen $Z_1,\dots,Z_n \sim \mathrm{N}(\mu,\sigma^2)$ modellieren. Dabei spiegelt μ den mittleren Therapieerfolg wider. Das Konfidenzintervall für μ wird wie bei einem Einstichprobenproblem gebildet.

9.4. Asymptotische Konfidenzintervalle für die Erfolgswahrscheinlichkeit bei Bernoulli-Experimenten

Seien $X_1,\dots,X_n$ unabhängige und Bernoulli-verteilte Zufallsvariablen mit Parameter $\theta \in (0,1)$. Wir wollen ein Konfidenzintervall für die Erfolgswahrscheinlichkeit θ konstruieren. Ein natürlicher Schätzer für θ ist $\bar{X}_n$. Diese Zufallsvariable hat eine reskalierte Binomialverteilung. Es ist nicht einfach, mit den Quantilen dieser Verteilung umzugehen. Somit ist es schwierig, ein exaktes Konfidenzintervall für θ zu einem vorgegebenen Niveau zu konstruieren. Auf der anderen Seite, können wir nach dem Zentralen Grenzwertsatz die Verteilung von $\bar{X}_n$ für großes n durch eine Normalverteilung approximieren. Man kann also versuchen, ein Konfidenzintervall zu konstruieren, das zumindest bei einem sehr großen Stichprobenumfang n das vorgegebene Niveau approximativ erreicht. Dafür benötigen wir die folgende allgemeine Definition.

Definition 9.4.1. Eine Folge $[\underline{\theta}_1,\overline{\theta}_1],[\underline{\theta}_2,\overline{\theta}_2],\dots$ von Konfidenzintervallen, wobei $\underline{\theta}_n : \mathbb{R}^n \to \mathbb{R}\cup\{-\infty\}$ und $\overline{\theta}_n : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$, heißt *asymptotisches Konfidenzintervall* zum Niveau $1-\alpha$, falls

$$\liminf_{n\to\infty} \mathbb{P}_\theta[\underline{\theta}_n(X_1,\dots,X_n) \leq \theta \leq \overline{\theta}_n(X_1,\dots,X_n)] \geq 1-\alpha \text{ für alle } \theta\in\Theta.$$

Nun kehren wir zu unserem Problem mit den Bernoulli-Experimenten zurück. Nach dem Zentralen Grenzwertsatz gilt

$$\frac{X_1 + \ldots + X_n - n\theta}{\sqrt{n\theta(1-\theta)}} \xrightarrow[n\to\infty]{d} \mathrm{N}(0,1),$$

denn $\mathbb{E}X_i = \theta$ und $\operatorname{Var} X_i = \theta(1-\theta)$. Durch Umformung ergibt sich

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

Sei z_α das α-Quantil der Standardnormalverteilung. Somit gilt

$$\lim_{n\to\infty} \mathbb{P}_\theta \left[z_{\frac{\alpha}{2}} \le \sqrt{n}\frac{\bar{X}_n - \theta}{\sqrt{\theta(1-\theta)}} \le z_{1-\frac{\alpha}{2}} \right] = 1-\alpha \quad \text{für alle } \theta \in (0,1).$$

Aufgrund der Symmetrieeigenschaft der Standardnormalverteilung ist $z_{1-\frac{\alpha}{2}} = -z_{\frac{\alpha}{2}}$. Definiere deshalb $z := z_{1-\frac{\alpha}{2}} = -z_{\frac{\alpha}{2}}$. Somit müssen wir θ bestimmen, so dass folgende Ungleichung erfüllt ist:

$$\sqrt{n}\,|\bar{X}_n - \theta| \le z\sqrt{\theta(1-\theta)}.$$

Quadrierung führt zu

$$n(\bar{X}_n^2 + \theta^2 - 2\bar{X}_n\theta) \le z^2\theta(1-\theta).$$

Dies lässt sich umschreiben zu

$$g(\theta) := \theta^2\left(1 + \frac{z^2}{n}\right) - \theta\left(2\bar{X}_n + \frac{z^2}{n}\right) + \bar{X}_n^2 \le 0.$$

Die Funktion $g(\theta)$ ist quadratisch und hat (wie wir gleich sehen werden) zwei verschiedene reelle Nullstellen. Somit ist $g(\theta) \le 0$ genau dann, wenn θ zwischen diesen beiden Nullstellen liegt. Indem wir nun die Nullstellen mit der p-q-Formel berechnen, erhalten wir folgendes Konfidenzintervall[1] zum Niveau $1-\alpha$ für θ:

$$\left[\frac{\bar{X}_n + \frac{z^2}{2n} - \frac{z}{\sqrt{n}}\sqrt{\bar{X}_n(1-\bar{X}_n) + \frac{z^2}{4n}}}{1 + \frac{z^2}{n}}, \frac{\bar{X}_n + \frac{z^2}{2n} + \frac{z}{\sqrt{n}}\sqrt{\bar{X}_n(1-\bar{X}_n) + \frac{z^2}{4n}}}{1 + \frac{z^2}{n}} \right].$$

Indem wir alle Terme mit $1/\sqrt{n}$ stehen lassen und alle Terme mit $1/n$ ignorieren, erhalten wir für großes n die folgende Approximation:[2]

$$\left[\bar{X}_n - \frac{z}{\sqrt{n}}\sqrt{\bar{X}_n(1-\bar{X}_n)},\ \bar{X}_n + \frac{z}{\sqrt{n}}\sqrt{\bar{X}_n(1-\bar{X}_n)} \right].$$

Später werden wir diese Approximation mit dem Satz von Slutsky begründen.

Die Wahrscheinlichkeit, dass das jeweilige Intervall den richtigen Wert von θ überdeckt, konvergiert zwar gegen $1-\alpha$ für $n \to \infty$, kann aber für ein endliches n kleiner oder größer als $1-\alpha$ sein. Im ersten Fall wird das angekündigte Konfidenzniveau nicht erreicht. Abbildung 1 zeigt die Überdeckungswahrscheinlichkeit als Funktion von θ für $n = 100$. Man sieht, dass das Wilson-Intervall wesentlich besser abschneidet. Verschiedene Konfidenzintervalle für die Erfolgswahrscheinlichkeit werden in der Arbeit von L. D. Brown, T. T. Cai and A. DasGupta

[1] Dieses Konfidenzintervall wurde 1927 von E.B. Wilson vorgeschlagen.

[2] Dieses Konfidenzintervall wird „Standard-Intervall“ oder „Wald-Intervall“ genannt.

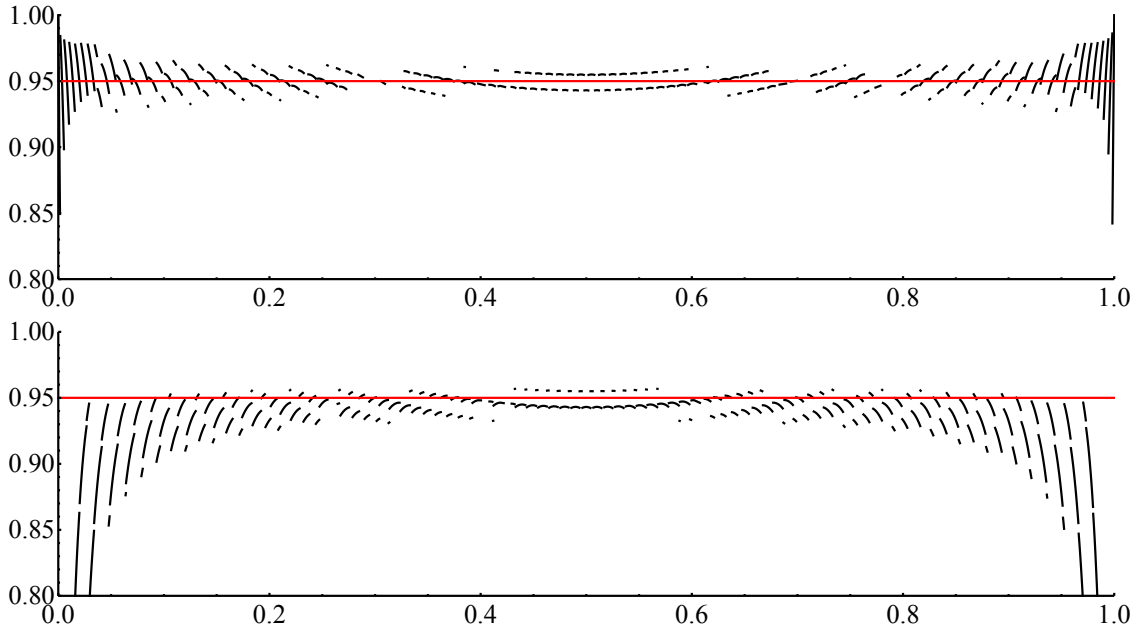


ABBILDUNG 1. Wahrscheinlichkeit, dass das Konfidenzintervall den richtigen Wert von θ überdeckt, als Funktion von θ. Oben: Wilson-Intervall. Unten: Standard-Intervall. Das Konfidenzniveau ist 0.95 (rote Linie) und der Stichprobenumfang $n = 100$.

[Interval Estimation for a Binomial Proportion, *Statistical Science*, 2001, Vol. 16(2), pp. 101-133, Link] besprochen.

Beispiel 9.4.2. Bei einer Wahlumfrage werden n Personen befragt, ob sie eine Partei A wählen. Es soll ein Konfidenzintervall zum Niveau 0.95 für den Stimmenanteil θ konstruiert werden und die Länge dieses Intervalls soll höchstens 0.02 sein. Wie viele Personen müssen dafür befragt werden?

Lösung. Wir betrachten die Wahlumfrage als ein n-faches Bernoulli-Experiment. Die Länge des Konfidenzintervalls für θ soll höchstens 0.02 sein, also erhalten wir die Ungleichung

$$\frac{2z}{\sqrt{n}}\sqrt{\bar{X}_n(1-\bar{X}_n)} \leq 0.02.$$

Quadrieren und nach n Umformen ergibt die Ungleichung

$$n \geq \frac{4z^2\bar{X}_n(1-\bar{X}_n)}{0.02^2}.$$

Der Mittelwert $\bar{X}_n$ ist zwar unbekannt, allerdings gilt $0 \leq \bar{X}_n \leq 1$ und somit $\bar{X}_n(1-\bar{X}_n) \leq 1/4$. Es reicht also auf jeden Fall, wenn

$$n \geq \frac{z^2}{0.02^2}.$$

Nun erinnern wir uns daran, dass z das $(1-\frac{\alpha}{2})$-Quantil der Standardnormalverteilung ist. Das Konfidenzniveau soll $1-\alpha = 0.95$ sein, also ist $1-\frac{\alpha}{2} = 0.975$. Das 0.975-Quantil der

Standardnormalverteilung errechnet sich (z. B. aus einer Tabelle) als Lösung von $\Phi(z) = 0.975$ zu $z = 1.96$. Es müssen also $n \geq \frac{1.96^2}{0.02^2} = 9604$ Personen befragt werden.

9.5. Satz von Slutsky

Bei der Konstruktion von Konfidenzintervallen findet der folgende Satz sehr oft Anwendung.

Satz 9.5.1 (Satz von Slutsky)**.** Seien $X, X_1, X_2, \ldots$ und $Y_1, Y_2, \ldots$ Zufallsvariablen, die auf einem gemeinsamen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ definiert sind. Gilt

$$X_n \xrightarrow[n\to\infty]{d} X \text{ und } Y_n \xrightarrow[n\to\infty]{d} c,$$

wobei c eine Konstante ist, so folgt, dass

$$X_n Y_n \xrightarrow[n\to\infty]{d} cX.$$

Bemerkung 9.5.2. Es lässt sich auch zeigen, dass $X_n + Y_n \xrightarrow[n\to\infty]{d} c + X$. Allgemeiner, für jede stetige Funktion $f : \mathbb{R}^2 \to \mathbb{R}$ gilt, dass

$$f(X_n, Y_n) \xrightarrow[n\to\infty]{d} f(X_n, c).$$

Beweis. Schritt 1. Es genügt, die punktweise Konvergenz der charakteristischen Funktionen zu zeigen. D.h., wir müssen zeigen, dass

$$\lim_{n\to\infty} \mathbb{E}\mathrm{e}^{itX_nY_n} = \mathbb{E}\mathrm{e}^{itcX} \text{ für alle } t \in \mathbb{R}.$$

Sei $\varphi(s) = \mathrm{e}^{its}$. Diese Funktion ist gleichmäßig stetig auf $\mathbb{R}$ und betragsmäßig durch 1 beschränkt. Wir zeigen, dass

$$\lim_{n\to\infty} \mathbb{E}\varphi(X_nY_n) = \mathbb{E}\varphi(cX).$$

Schritt 2. Sei $\varepsilon > 0$ fest. Wegen der gleichmäßigen Stetigkeit von φ gibt es ein $\delta > 0$ mit der Eigenschaft, dass

$$|\varphi(x) - \varphi(y)| \leq \varepsilon \text{ für alle } x, y \in \mathbb{R} \text{ mit } |x - y| \leq \delta.$$

Schritt 3. Sei $A > 0$ so groß, dass $\mathbb{P}[|X| > A] \leq \varepsilon$. Wir können annehmen, dass A und $-A$ Stetigkeitspunkte der Verteilungsfunktion von X sind, ansonsten kann man A vergrößern. Da $X_n \xrightarrow[n\to\infty]{d} X$ und $A, -A$ keine Atome von X sind, folgt, dass

$$\lim_{n\to\infty} \mathbb{P}[|X_n| > A] = \mathbb{P}[|X| > A] \leq \varepsilon.$$

Also gilt $\mathbb{P}[|X_n| > A] \leq 2\varepsilon$ für große n.

Schritt 4. Es gilt

$$\begin{aligned}|\mathbb{E}\varphi(X_nY_n) - \mathbb{E}\varphi(cY)| &\leq \mathbb{E}|\varphi(X_nY_n) - \varphi(cX_n)| + |\mathbb{E}\varphi(cX_n) - \mathbb{E}\varphi(cX)| \\ &\leq E_1 + E_2 + E_3 + E_4\end{aligned}$$

mit

$$\begin{aligned}E_1 &= \mathbb{E}\left[|\varphi(X_nY_n) - \varphi(cX_n)|\mathbb{1}_{|Y_n - c| > \frac{\delta}{A}}\right], \\ E_2 &= \mathbb{E}\left[|\varphi(X_nY_n) - \varphi(cX_n)|\mathbb{1}_{|Y_n - c| \leq \frac{\delta}{A}, |X_n| > A}\right], \\ E_3 &= \mathbb{E}\left[|\varphi(X_nY_n) - \varphi(cX_n)|\mathbb{1}_{|Y_n - c| \leq \frac{\delta}{A}, |X_n| \leq A}\right], \\ E_4 &= |\mathbb{E}\varphi(cX_n) - \mathbb{E}\varphi(cX)|.\end{aligned}$$

Schritt 5. Wir werden nun $E_1, \ldots, E_4$ abschätzen.

E_1: Da $|\varphi(t)| \leq 1$ ist, folgt, dass $E_1 \leq 2\mathbb{P}[|Y_n - c| > \delta/A]$. Dieser Term konvergiert gegen 0 für $n \to \infty$, da Y_n gegen c in Verteilung (und somit auch in Wahrscheinlichkeit) konvergiert.

E_2: Für E_2 gilt die Abschätzung $E_2 \leq 2\mathbb{P}[|X_n| > A] \leq 4\varepsilon$ nach Schritt 3, wenn n groß genug ist.

E_3: Es gilt $E_3 \leq \varepsilon$, da $|X_nY_n - cX_n| \leq \delta$ falls $|Y_n - c| \leq \delta/A$ und $|X_n| \leq A$. Aus Schritt 2 folgt dann, dass $|\varphi(X_nY_n) - \varphi(cX_n)| \leq \varepsilon$.

E_4: Der Term E_4 konvergiert für $n \to \infty$ gegen 0, denn $\lim_{n\to\infty} \mathbb{E}\varphi(cX_n) = \mathbb{E}\varphi(cX)$, denn nach Voraussetzung konvergiert X_n in Verteilung gegen X.

Indem wir nun alles zusammenfassen, erhalten wir, dass

$$\limsup_{n\to\infty} |\mathbb{E}\varphi(X_nY_n) - \mathbb{E}\varphi(cY)| \leq 5\varepsilon.$$

Da $\varepsilon > 0$ beliebig klein gewählt werden kann, folgt, dass $\lim_{n\to\infty} |\mathbb{E}\varphi(X_nY_n) - \mathbb{E}\varphi(cY)| = 0$. Somit ist $\lim_{n\to\infty} \mathbb{E}\varphi(X_nY_n) = \mathbb{E}\varphi(cY)$. □

Beispiel 9.5.3. Seien $X_1, \ldots, X_n$ unabhängig und Bernoulli-verteilt mit Parameter $\theta \in (0, 1)$. Wir konstruieren ein asymptotisches Konfidenzintervall für θ. Nach dem Zentralen Grenzwertsatz gilt

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

Leider kommt hier θ sowohl im Zähler als auch im Nenner vor. Deshalb hat sich bei unserer früheren Konstruktion eine quadratische Gleichung ergeben. Wir werden nun θ im Nenner eliminieren, indem wir es durch einen Schätzer, nämlich $\bar{X}_n$, ersetzen. Nach dem Satz von Slutsky gilt nämlich, dass

$$\sqrt{n}\,\frac{\bar{X}_n - \theta}{\sqrt{\bar{X}_n(1-\bar{X}_n)}} = \sqrt{n}\,\frac{\bar{X}_n - \theta}{\sqrt{\theta(1-\theta)}}\,\sqrt{\frac{\theta(1-\theta)}{\bar{X}_n(1-\bar{X}_n)}} \xrightarrow[n\to\infty]{d} \mathrm{N}(0, 1),$$

denn nach dem Gesetz der großen Zahlen konvergiert $\sqrt{\frac{\theta(1-\theta)}{\bar{X}_n(1-\bar{X}_n)}}$ fast sicher (und somit auch in Verteilung) gegen 1. Es gilt also

$$\lim_{n\to\infty} \mathbb{P}_\theta \left[z_{\frac{\alpha}{2}} \leq \sqrt{n}\, \frac{\bar{X}_n - \theta}{\sqrt{\bar{X}_n(1-\bar{X}_n)}} \leq z_{1-\frac{\alpha}{2}} \right] = 1-\alpha \quad \text{für alle } \theta \in (0,1).$$

Sei $z := -z_{\frac{\alpha}{2}} = z_{1-\frac{\alpha}{2}}$. Daraus ergibt sich folgendes aysmptotisches Konfidenzintervall für θ zum Konfidenzniveau $1-\alpha$:

$$\left[\bar{X}_n - \frac{z}{\sqrt{n}} \sqrt{\bar{X}_n(1-\bar{X}_n)},\ \bar{X}_n + \frac{z}{\sqrt{n}} \sqrt{\bar{X}_n(1-\bar{X}_n)} \right].$$

Dieses Intervall haben wir oben mit einer anderen Methode hergeleitet.

Aufgabe 9.5.4. Zeigen Sie mit dem Satz von Slutsky, dass $t_n \xrightarrow[n\to\infty]{d} \mathrm{N}(0,1)$. Dabei ist t_n die t-Verteilung mit n Freiheitsgraden.

9.6. Konfidenzintervall für den Erwartungswert der Poissonverteilung

Seien $X_1, \ldots, X_n$ unabhängige Zufallsvariablen mit $X_i \sim \mathrm{Poi}(\theta)$, wobei $\theta > 0$. Gesucht ist ein Konfidenzintervall für θ zum Konfidenzniveau $1-\alpha$. Ein natürlicher Schätzer für θ ist $\bar{X}_n$. Da für die Poisson-Verteilung $\mathbb{E}X_i = \operatorname{Var} X_i = \theta$ gilt, folgt durch den zentralen Grenzwertsatz, dass

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

Es sei z_α das α-Quantil der Standardnormalverteilung. Somit gilt

$$\lim_{n\to\infty} \mathbb{P}_\theta \left[z_{\frac{\alpha}{2}} \leq \sqrt{n}\, \frac{\bar{X}_n - \theta}{\sqrt{\theta}} \leq z_{1-\frac{\alpha}{2}} \right] = 1-\alpha \quad \text{für alle } \theta > 0.$$

Aufgrund der Symmetrieeigenschaft der Standardnormaverteilung gilt $z := z_{1-\frac{\alpha}{2}} = -z_{\frac{\alpha}{2}}$. Wir erhalten also folgende Ungleichung für θ:

$$\sqrt{n}\, |\bar{X}_n - \theta| \leq \sqrt{\theta} z.$$

Dies lässt sich durch Quadrierung umschreiben zu

$$g(\theta) := \theta^2 - \theta \left(2\bar{X}_n + \frac{z^2}{n} \right) + \bar{X}_n^2 \leq 0.$$

Die Ungleichung $g(\theta) \leq 0$ gilt genau dann, wenn θ zwischen den beiden Nullstellen der quadratischen Gleichung $g(\theta) = 0$ liegt. Diese lassen durch Verwendung der p-q-Formel berechnen. Es ergibt sich folgendes asymptotisches Konfidenzintervall für θ zum Konfidenzniveau $1-\alpha$:

$$\left[\bar{X}_n + \frac{z^2}{2n} - \frac{z}{\sqrt{n}} \sqrt{\bar{X}_n + \frac{z^2}{2n}},\ \bar{X}_n + \frac{z^2}{2n} + \frac{z}{\sqrt{n}} \sqrt{\bar{X}_n + \frac{z^2}{2n}} \right].$$

Indem man nun alle Terme mit $1/\sqrt{n}$ stehen lässt und alle Terme mit $1/n$ ignoriert, erhält man die Approximation

$$\left[\bar{X}_n - \frac{z}{\sqrt{n}}\sqrt{\bar{X}_n},\ \bar{X}_n + \frac{z}{\sqrt{n}}\sqrt{\bar{X}_n}\right].$$

Das Argument mit der quadratischen Gleichung lässt sich mit dem Satz von Slutsky vermeiden. Nach dem Zentralen Grenzwertsatz gilt nach wie vor

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

Leider kommt hier der Parameter θ sowohl im Zähler als auch im Nenner vor, was im obigen Argument zu einer quadratischen Gleichung führte. Wir können allerdings θ durch einen Schätzer für θ, nämlich durch $\bar{X}_n$, ersetzen. Nach dem starken Gesetz der großen Zahlen konvergiert $\sqrt{\theta/\bar{X}_n}$ fast sicher (und somit auch in Verteilung) gegen 1. Nach dem Satz von Slutsky gilt dann

$$\sqrt{n}\,\frac{\bar{X}_n - \theta}{\sqrt{\bar{X}_n}} = \sqrt{n}\,\frac{\bar{X}_n - \theta}{\sqrt{\theta}}\,\sqrt{\frac{\theta}{\bar{X}_n}} \xrightarrow[n\to\infty]{d} \mathrm{N}(0,1).$$

Somit folgt

$$\lim_{n\to\infty} \mathbb{P}_\theta\left[-z \le \sqrt{n}\,\frac{\bar{X}_n - \theta}{\sqrt{\bar{X}_n}} \le z\right] = 1-\alpha \ \text{ für alle } \theta > 0.$$

Es ergibt sich also wieder einmal das asymptotische Konfidenzintervall

$$\left[\bar{X}_n - \frac{z}{\sqrt{n}}\sqrt{\bar{X}_n},\ \bar{X}_n + \frac{z}{\sqrt{n}}\sqrt{\bar{X}_n}\right].$$

9.7. Asymptotisches Konfidenzintervall um den ML-Schätzer

Seien $X_1, X_2, \ldots$ unabhängige und identisch verteilte Zufallsvariablen mit Dichte/Zähldichte h_θ. Es sei $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ eine asymptotisch normalverteilte Folge von Schätzern von θ, d.h. für alle $\theta \in \Theta$ gelte

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{d} \mathrm{N}(0, \sigma^2(\theta)) \text{ unter } \mathbb{P}_\theta.$$

Z.B. ist das unter Regularitätsbedingungen für den ML-Schätzer erfüllt, wobei $\sigma^2(\theta) = 1/I(\theta)$ und $I(\theta)$ die Fisher-Information ist. Wir können die obige Bedingung wie folgt schreiben:

$$\sqrt{n}\frac{\hat{\theta}_n - \theta}{\sigma(\theta)} \xrightarrow[n\to\infty]{d} \mathrm{N}(0,1) \text{ unter } \mathbb{P}_\theta.$$

Wir wollen nun die quadratische Abweichung $\sigma(\theta)$ durch deren geschätzte Version $\sigma(\hat{\theta}_n)$ ersetzen. Dazu nehmen wir zusätzlich an, dass $\hat{\theta} \xrightarrow[n\to\infty]{d} \theta$ (schwache Konsistenz) und dass $\sigma^2(\theta)$ stetig ist. Aus dem Satz von der stetigen Abbildung folgt, dass

$$\sigma(\hat{\theta}_n) \xrightarrow[n\to\infty]{d} \sigma(\theta).$$

Mit dem Satz von Slutsky ergibt sich nun

$$\sqrt{n}\frac{\hat{\theta}_n - \theta}{\sigma(\hat{\theta}_n)} = \sqrt{n}\frac{\hat{\theta}_n - \theta}{\sigma(\theta)} \cdot \frac{\sigma(\theta)}{\sigma(\hat{\theta}_n)} \xrightarrow[n\to\infty]{d} \mathrm{N}(0,1) \text{ unter } \mathbb{P}_\theta.$$

Somit gilt für alle $\theta \in \Theta$

$$\lim_{n\to\infty} \mathbb{P}_\theta \left[\sqrt{n}\frac{|\hat{\theta}_n - \theta|}{\sigma(\hat{\theta}_n)} \le z_{1-\frac{\alpha}{2}} \right] = 1 - \alpha.$$

Es ergibt sich das folgende Konfidenzintervall[3] zum Niveau $1-\alpha$ für θ:

$$\left[\hat{\theta}_n - z_{1-\frac{\alpha}{2}} \frac{\sigma(\hat{\theta}_n)}{\sqrt{n}}, \hat{\theta}_n + z_{1-\frac{\alpha}{2}} \frac{\sigma(\hat{\theta}_n)}{\sqrt{n}} \right].$$

9.8. Konfidenzband für die Verteilungsfunktion

Es seien $X_1, \dots, X_n$ u.i.v. Zufallsvariablen mit einer unbekannten Verteilungsfunktion F. Wir können F durch die empirische Verteilungsfunktion

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \le t\}}, \quad t \in \mathbb{R},$$

schätzen, aber wie groß ist der Fehler? Im Folgenden werden wir zwei Funktionen $F_n^-(t) = F_n^-(t; X_1, \dots, X_n)$ und $F_n^+(t) = F_n^+(t; X_1, \dots, X_n)$ konstruieren derart, dass F mit einer großen Wahrscheinlichkeit zwischen F_n^- und F_n^+ liegt. Dabei verlangen wir, dass die Ungleichungen $F_n^-(t) \le F(t) \le F_n^+(t)$ für alle $t \in \mathbb{R}$ *gleichzeitig* gelten, weshalb wir von einem Konfidenzband und nicht einem Konfidenzintervall sprechen.

Definition 9.8.1. Ein Paar von Funktionen $F_n^-, F_n^+ : \mathbb{R} \times \mathbb{R}^n \to [0,1]$ heißt *Konfidenzband* zum Niveau $1-\alpha$ für F, falls

$$\mathbb{P}[\forall t \in \mathbb{R} \colon F_n^-(t; X_1, \dots, X_n) \le F(t) \le F_n^+(t; X_1, \dots, X_n)] \ge 1 - \alpha$$

für jede Verteilungsfunktion F.

Um ein Konfidenzband zu konstruieren, benötigen wir die folgende Konzentrationsungleichung.

Satz 9.8.2 (Dvoretzky-Kiefer-Wolfowitz-Massart)**.** Für alle $\varepsilon > 0$ gilt

$$\mathbb{P}\left[\sup_{t\in\mathbb{R}} |\hat{F}_n(t) - F(t)| \ge \varepsilon \right] \le 2\mathrm{e}^{-2n\varepsilon^2}.$$

Ohne Beweis.

[3] Dieses Konfidenzintervall wird oft als Wald-Intervall bezeichnet.

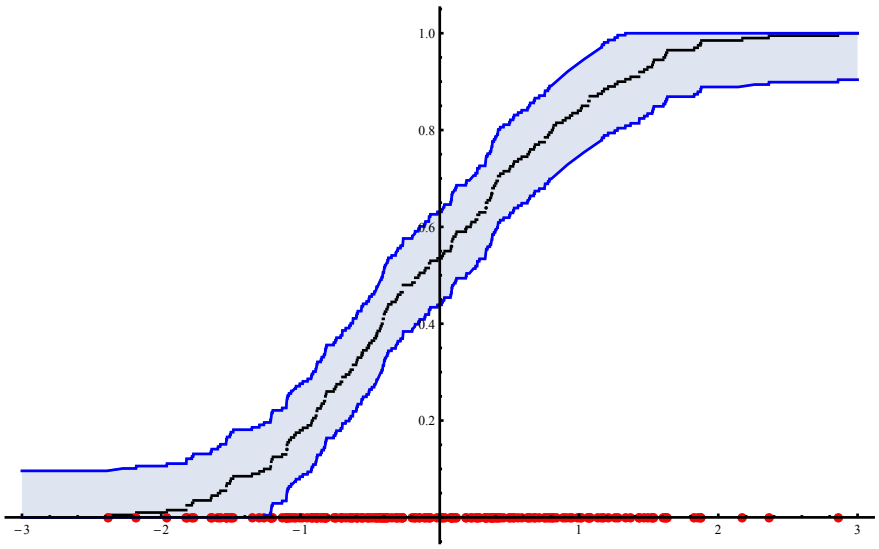


ABBILDUNG 2. Konfidenzband für die Verteilungsfunktion. Rote Kreise: Stichprobe. Schwarze Kurve: Empirische Verteilungsfunktion. Blaue Kurven: Funktionen $F_n^-(t)$ und $F_n^+(t)$. Das Konfidenzniveau ist 0.95 und der Stichprobenumfang $n = 200$.

Aufgabe 9.8.3. Leiten Sie aus der obigen Ungleichung den Satz von Glivenko-Cantelli her: $\sup_{t\in\mathbb{R}} |\hat{F}_n(t) - F(t)| \xrightarrow[n\to\infty]{f.s.} 0$.

Sei nun ein Konfidenzniveau $1-\alpha$ vorgegeben. Wir setzen $2\mathrm{e}^{-2n\varepsilon^2} = \alpha$, woraus sich ergibt, dass

$$\varepsilon = \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}.$$

Es folgt aus Satz 9.8.2, dass

$$\mathbb{P}[\forall t \in \mathbb{R}\colon \hat{F}_n(t) - \varepsilon \le F(t) \le \hat{F}_n(t) + \varepsilon] = 1 - \mathbb{P}\left[\sup_{t\in\mathbb{R}} |\hat{F}_n(t) - F(t)| > \varepsilon\right] \ge 1 - \alpha.$$

Wir können also das folgende Konfidenzband für F konstruieren (s. Abbildung 2):

$$F_n^-(t) = \max\left\{\hat{F}_n(t) - \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}, 0\right\}, \quad F_n^+(t) = \min\left\{\hat{F}_n(t) + \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}, 1\right\}.$$

9.9. Konfidenzintervalle für Quantile

Seien $X_1, \dots, X_n$ unabhängig und identisch verteilt mit Verteilungsfunktion F. Wir nehmen an, dass F stetig und streng monoton steigend ist, und definieren das β-Quantil Q_β von F als die Lösung der Gleichung

$$F(Q_\beta) = \beta, \quad \beta \in (0,1).$$

Ein natürlicher Schätzer von Q_β ist die Ordnungsstatistik $X_{([\beta n])}$, wobei $X_{(i)}$ die i-te Ordnungsstatistik der Stichprobe $X_1, \dots, X_n$ ist. Im Folgenden konstruieren wir ein Konfidenzintervall zum Niveau $1-\alpha$ für Q_β. Wir werden versuchen, Indizes $1 \le i \le j \le n$ so zu

bestimmen, dass $[X_{(i)}, X_{(j)}]$ ein Konfidenzintervall ist. Es soll die Forderung

$$\mathbb{P}[X_{(i)} \leq Q_\beta \leq X_{(j)}] \geq 1 - \alpha$$

erfüllt werden. Es gilt

$$\begin{aligned}\mathbb{P}[X_{(i)} \leq Q_\beta \leq X_{(j)}] &= \mathbb{P}\left[\sum_{k=1}^{n} \mathbb{1}_{\{X_k \leq Q_\beta\}} \in \{i, i+1, \ldots, j-1\}\right] \\ &= \mathbb{P}[i \leq S < j] \\ &= \sum_{k=i}^{j-1} \binom{n}{k} \beta^k (1-\beta)^{n-k},\end{aligned}$$

denn die Zufallsvariable $S := \sum_{k=1}^{n} \mathbb{1}_{\{X_k \leq Q_\beta\}}$ ist $\mathrm{Bin}(n, \beta)$-verteilt. Es reicht also, i und j so zu wählen, dass die Summe auf der rechten Seite $\geq 1 - \alpha$ ist. Zum Beispiel können wir

$$\begin{aligned}i &:= \max\left\{m : \mathbb{P}[S < m] = \sum_{k=0}^{m-1} \binom{n}{k} \beta^k (1-\beta)^{n-k} \leq \frac{\alpha}{2}\right\}, \\ j &:= \min\left\{m : \mathbb{P}[S \geq m] = \sum_{k=m}^{n} \binom{n}{k} \beta^k (1-\beta)^{n-k} \leq \frac{\alpha}{2}\right\}\end{aligned}$$

wählen.

Beispiel 9.9.1. Sei eine Stichprobe $X_1, \ldots, X_{200}$ vom Umfang $n = 200$ gegeben. Wir konstruieren ein Konfidenzintervall zum Niveau 0.95 für den theoretischen Median. Für die Zufallsvariable $S \sim \mathrm{Bin}(200, 1/2)$ rechnet man nach, dass

$$\begin{aligned}\mathbb{P}[S < 86] = \mathbb{P}[S \geq 115] &\approx 0.0200, \\ \mathbb{P}[S < 87] = \mathbb{P}[S \geq 114] &\approx 0.0279.\end{aligned}$$

Es gilt also

$$\mathbb{P}[86 \leq S < 114] \approx 1 - 0.0479 \geq 0.95.$$

Also wissen wir, dass der theoretische Median mit einer Wahrscheinlichkeit von mindestens 0.95 zwischen $X_{(86)}$ und $X_{(114)}$ liegt.

KAPITEL 10

Tests statistischer Hypothesen

In der Statistik muss man oft Hypothesen testen, z.B. muss man anhand einer Stichprobe entscheiden, ob ein unbekannter Parameter einen vorgegebenen Wert annimmt. Zuerst betrachten wir ein Beispiel.

10.1. Ist eine Münze fair?

Es sei eine Münze gegeben. Wir wollen testen, ob diese Münze fair ist, d.h. ob die Wahrscheinlichkeit von „Kopf“, die wir mit θ bezeichnen, gleich $1/2$ ist. Dazu werfen wir die Münze z.B. $n = 200$ Mal. Sei S die Anzahl der Würfe, bei denen die Münze Kopf zeigt. Nun betrachten wir zwei Hypothesen:

- Nullhypothese H_0: Die Münze ist fair, d.h., $\theta = 1/2$.
- Alternativhypothese H_1: Die Münze ist nicht fair, d.h., $\theta \neq 1/2$.

Wir müssen entscheiden, ob wir die Nullhypothese H_0 verwerfen oder beibehalten. Die Entscheidung muss anhand des Wertes von S getroffen werden. Unter der Nullhypothese gilt, dass $\mathbb{E}_{H_0} S = 200 \cdot \frac{1}{2} = 100$. Die Idee besteht nun darin, die Nullhypothese zu verwerfen, wenn S stark von 100 abweicht. Die Größe $|S-100|$ bezeichnet man in diesem Fall als *Teststatistik*. Dabei sind große Werte von $|S - 100|$ ein Hinweis darauf, dass die Nullhypothese verworfen werden muss, d.h. große Werte sind *signifikant*.
Wir wählen also eine Konstante $c \in \{0, 1, \ldots\}$ und verwerfen H_0, falls $|S - 100| > c$. Andernfalls behalten wir die Hypothese H_0 bei. Bei diesem Vorgehen können wir zwei Arten von Fehlern machen:

- *Fehler* 1. *Art*: H_0 wird verworfen, obwohl H_0 richtig ist.
- *Fehler* 2. *Art*: H_0 wird nicht verworfen, obwohl H_0 falsch ist.

Wie sollte nun die Konstante c (der sogenannte *kritische Wert*) gewählt werden? Man möchte natürlich die Wahrscheinlichkeiten der beiden Arten von Fehlern klein halten. In diesem Beispiel ist es allerdings nicht möglich, die Wahrscheinlichkeit eines Fehlers 2. Art zu bestimmen. Der Grund dafür ist, dass man für die Berechnung dieser Wahrscheinlichkeit den Wert von θ kennen muss, bei einem Fehler 2. Art ist allerdings nur bekannt, dass $\theta \neq 1/2$ ist. Die Wahrscheinlichkeit eines Fehlers 1. Art kann aber sehr wohl bestimmt werden und ist

$$\mathbb{P}_{H_0}[|S - 100| > c] = 2\mathbb{P}_{H_0}[S > 100 + c] = 2 \sum_{k=100+c+1}^{200} \binom{200}{k} \frac{1}{2^{200}},$$

da $S \sim \mathrm{Bin}(200, 1/2)$ unter H_0. Wir wollen nun c so wählen, dass die Wahrscheinlichkeit eines Fehlers 1. Art nicht größer als ein kleines vorgegebenes Niveau $\alpha \in (0, 1)$ ist. Normalerweise wählt man $\alpha = 0.01$ oder 0.05. Hier wählen wir das Niveau $\alpha = 0.05$. Nun rechnet man nach,

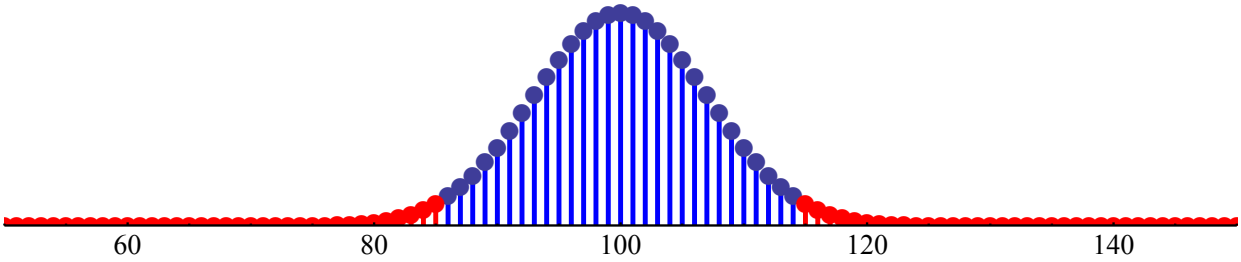


ABBILDUNG 1. Zähldichte der Binomialverteilung mit Parametern $n = 200$ und $\theta = 1/2$. Rot: Ablehnungsbereich. Blau: Annahmebereich.

dass

$$\mathbb{P}_{H_0}[|S - 100| > c] = \begin{cases} 0.05596, & \text{für } c = 13, \\ 0.04003, & \text{für } c = 14. \end{cases}$$

Damit die Wahrscheinlichkeit eines Fehlers 1. Art kleiner als $\alpha = 0.05$ ist, müssen wir also $c \geq 14$ wählen. Dabei ist es sinnvoll, c möglichst klein zu wählen, denn sonst vergrößert man die Wahrscheinlichkeit eines Fehlers 2. Art. Also wählen wir $c = 14$. Unsere Entscheidungsregel lautet nun wie folgt:

- Wir verwerfen H_0, falls $|S - 100| > 14$.
- Sonst behalten wir die Hypothese H_0 bei.

Das Beibehalten von H_0 bedeutet nicht, dass H_0 „bewiesen“ wurde. Es kann ja immer noch sein, dass die Münze unfair mit einem $\theta = 1/2 + 10^{-10}$ ist und ein dermaßen kleiner Unterschied kann bei 200 Würfen sowieso nicht erkannt werden. Das Beibehalten von H_0 bedeutet lediglich, dass in den vorhandenen Daten keine ausreichenden Hinweise gegen H_0 gefunden wurden.

Eine wichtige Größe zur Auswertung von statistischen Tests ist der p-Wert.

Definition 10.1.1. Als *p-Wert* bezeichnet man die Wahrscheinlichkeit (unter der Nullhypothese), dass die Teststatistik einen mindestens so extremen Wert annimmt, wie der in der Stichprobe beobachtete Wert.

Hat man z.B. bei 200 Würfen 150 Mal „Kopf“ beobachtet, so ist der p-Wert gegeben durch

$$\mathbb{P}_{H_0}[|S - 100| \geq |150 - 100|] = 2\mathbb{P}_{H_0}[S \geq 150] = 2 \sum_{k=150}^{200} \binom{200}{k} \frac{1}{2^{200}} \approx 8.393 \cdot 10^{-13}.$$

Der Wert 150 weicht vom unter der Nullhypothese erwarteten Wert 100 um 50 ab. Bei einer richtigen Nullhypothese H_0 hat eine Abweichung von mindestens 50 eine Wahrscheinlichkeit von lediglich $8.393 \cdot 10^{-13}$. Deshalb muss in diesem Fall die Nullhypothese ohne große Zweifel verworfen werden.

Der p-Wert liegt immer zwischen 0 und 1. Ein kleiner p-Wert ist ein Hinweis darauf, dass die Nullhypothese verworfen werden muss.

Asymptotischer Test. Im obigen Beispiel kann man für die Berechnung der Wahrscheinlichkeiten die Approximation durch die Normalverteilung benutzen. Es soll ein c mit

$$\mathbb{P}_{H_0}[S - 100 < -c] \leq \frac{\alpha}{2}$$

bestimmt werden. Um die Güte der Approximation zu verbessern, benutzen wir den $\frac{1}{2}$-Trick. Da c ganz ist, ist die obige Ungleichung äquivalent zu

$$\mathbb{P}_{H_0}[S - 100 \leq -c - 0.5] \leq \frac{\alpha}{2}.$$

Unter H_0 gilt $S \sim \mathrm{Bin}(200, 1/2)$ und somit $\mathbb{E}_{H_0} S = 100$, $\mathrm{Var}_{H_0} S = 200 \cdot \frac{1}{2} \cdot \frac{1}{2} = 50$. Die obige Ungleichung ist äquivalent zu

$$\mathbb{P}_{H_0}\left[\frac{S - 100}{\sqrt{50}} \leq -\frac{c + 0.5}{\sqrt{50}}\right] \leq \frac{\alpha}{2}.$$

Nun können wir die Normalverteilungsapproximation benutzen und die obige Ungleichung durch die folgende ersetzen:

$$\Phi\left(-\frac{c + 0.5}{\sqrt{50}}\right) \leq \frac{\alpha}{2}$$

Somit muss für c die folgende Ungleichung gelten:

$$\frac{c + 0.5}{\sqrt{50}} \geq -z_{\frac{\alpha}{2}},$$

wobei $z_{\frac{\alpha}{2}}$ das $\frac{\alpha}{2}$-Quantil der Standardnormalverteilung ist. Wegen der Symmetrie der Standardnormalverteilung gilt $-z_{\frac{\alpha}{2}} = z_{1-\frac{\alpha}{2}}$. Für $\alpha = 0.05$ ist $z_{1-\frac{\alpha}{2}} = z_{0.975} = 1.96$ und somit ist die obige Ungleichung äquivalent zu $c \geq 13.36$. Somit müssen wir $c = 14$ wählen. Die Entscheidungsregel bleibt genauso wie oben.

Aufgabe 10.1.2. Anlässlich der Show „Wetten, dass..?“ wettet Herr Müller, dass er die Farben zweier Filzstifte durch Ablecken eines mit ihnen gemalten Strichs auf einem Blatt Papier erkennen kann. Er weiß, dass von den 10 bemalten Papieren genau 5 mit dem roten und 5 mit dem blauen Filzstift bemalt wurden. Testen Sie die Hypothese H_0 : „er kann es nicht“ gegen die Alternativhypothese H_1 : „er kann es“. Verwenden Sie dazu die Statistik X, die die Anzahl der richtig zugeordneten roten Striche bei 5 Zügen angibt. Bestimmen Sie damit einen kritischen Wert c, sodass $\varphi(x) = \mathbb{1}_{\{x \geq c\}}$ ein Test zum Niveau 0.01 ist, d.h. eine Anzahl richtig zugeordneter roter Striche, ab der man bereit ist zum Niveau 0.01 zu akzeptieren, dass Herr Müller diese Gabe tatsächlich besitzt. *Hinweis*: Bestimmen Sie die Verteilung von X unter der Hypothese, dass Herr Müller die Farben in Wirklichkeit nicht auseinanderhalten kann und nur rät.

10.2. Tests für die Parameter der Normalverteilung

Seien $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ unabhängige und mit Parametern (μ, σ^2) normalverteilte Zufallsvariablen. Wir wollen Hypothesen über die Parameter μ und σ^2 testen. Wir werden die folgenden vier Fälle betrachten:

(1) Tests für μ bei bekanntem σ^2.

(2) Tests für μ bei unbekanntem σ^2.
(3) Tests für σ^2 bei bekanntem μ.
(4) Tests für σ^2 bei unbekanntem μ.

Fall 1: Tests für μ bei bekanntem σ^2 (Gauß-z-Test).

Seien $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma_0^2)$ unabhängig, wobei die Varianz σ_0^2 bekannt sei. Wir wollen verschiedene Hypothesen über μ testen, z.B. $\mu = \mu_0$, $\mu \geq \mu_0$ oder $\mu \leq \mu_0$, wobei μ_0 ein vorgegebener Wert ist. Wir betrachten die Teststatistik

$$T := \sqrt{n}\,\frac{\bar{X}_n - \mu_0}{\sigma_0}.$$

Unter $\mu = \mu_0$ gilt $T \sim \mathrm{N}(0,1)$. Wir betrachten drei Fälle in Abhängigkeit davon, wie die zu testende Hypothese formuliert wird.

Fall 1A. $H_0 : \mu = \mu_0$; $H_1 : \mu \neq \mu_0$. Die Nullhypothese H_0 sollte verworfen werden, wenn $|T|$ groß ist. Dabei sollte die Wahrscheinlichkeit eines Fehlers 1. Art höchstens α sein. Dies führt zu der Entscheidungsregel, dass die Nullhypothese H_0 verworfen wird, falls $|T| > z_{1-\frac{\alpha}{2}}$, s. Abbildung 2 (links).

Fall 1B. $H_0 : \mu \geq \mu_0$; $H_1 : \mu < \mu_0$. Die Nullhypothese H_0 sollte verworfen werden, wenn T klein ist. Also verwerfen wir H_0, falls $T < z_\alpha$, s. Abbildung 2 (Mitte). Unter $\mu = \mu_0$ ist die Wahrscheinlichkeit eines Fehlers erster Art gleich α. Man kann zeigen, dass für $\mu > \mu_0$ (was auch zu H_0 gehört), die Wahrscheinlichkeit, H_0 irrtümlich zu verwerfen, kleiner als α ist.

Fall 1C. $H_0 : \mu \leq \mu_0$; $H_1 : \mu > \mu_0$. Hier sollte H_0 verworfen werden, wenn T groß ist. In diesem Fall wird H_0 verworfen, wenn $T > z_{1-\alpha}$, , s. Abbildung 2 (rechts).

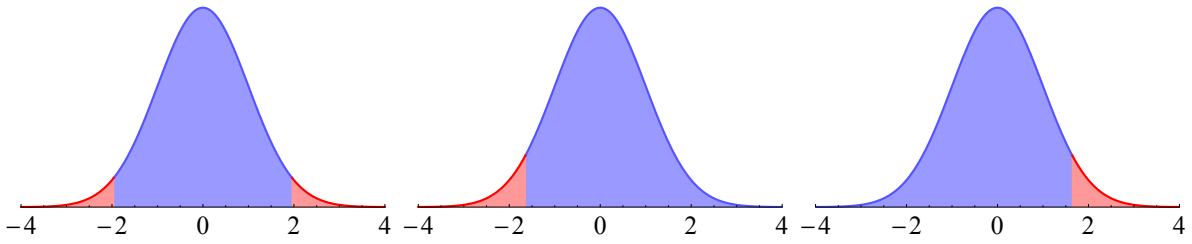


Abbildung 2. Vorgehensweise beim Gauß-z-Test. Rot: Ablehnungsbereich. Blau: Annahmebereich. Links: Zweiseitiger Test (Fall 1A). Mitte: Einseitiger Test (Fall 1B). Rechts: Einseitiger Test (Fall 1C).

Fall 2: Tests für μ bei unbekanntem σ^2 (Student-t-Test).

Seien $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$, wobei μ und σ^2 unbekannt seien. Wir möchten Hypothesen über μ testen, z. B. $\mu = \mu_0$, $\mu \geq \mu_0$ oder $\mu \leq \mu_0$, wobei μ_0 vorgegeben ist. Die Teststatistik aus Fall 1 können wir dafür nicht verwenden, denn sie enthält den unbekannten Parameter σ^2. Deshalb schätzen wir zuerst σ^2 durch

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

Wir betrachten die Teststatistik

$$T := \sqrt{n}\,\frac{\bar{X}_n - \mu_0}{S_n}.$$

Dann gilt unter $\mu = \mu_0$, dass $T \sim t_{n-1}$.

Fall 2A. $H_0 : \mu = \mu_0$; $H_1 : \mu \neq \mu_0$. Die Nullhypothese H_0 sollte verworfen werden, wenn $|T|$ groß ist. Dabei sollte die Wahrscheinlichkeit eines Fehlers 1. Art höchstens α sein. Wegen der Symmetrie der t-Verteilung erhalten wir die folgende Entscheidungsregel: H_0 wird verworfen, falls $|T| > t_{n-1,1-\frac{\alpha}{2}}$.

Fall 2B. $H_0 : \mu \geq \mu_0$; $H_1 : \mu < \mu_0$. Die Nullhypothese H_0 wird verworfen, wenn $T < t_{n-1,\alpha}$.

Fall 2C. $H_0 : \mu \leq \mu_0$; $H_1 : \mu > \mu_0$. Die Nullhypothese H_0 wird verworfen, wenn $T > t_{n-1,1-\alpha}$.

Fall 3: Tests für σ^2 bei bekanntem μ (χ^2-Streuungstest).

Seien $X_1, \ldots, X_n \sim \mathrm{N}(\mu_0, \sigma^2)$ unabhängig, wobei der Erwartungswert μ_0 bekannt sei. Wir wollen verschiedene Hypothesen über die quadratische Streuung σ^2 der Stichprobe testen, wie z. B. $\sigma^2 = \sigma_0^2$, $\sigma^2 \geq \sigma_0^2$ oder $\sigma^2 \leq \sigma_0^2$, wobei σ_0^2 vorgegeben ist. Ein natürlicher Schätzer für σ^2 ist

$$\widetilde{S}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_0)^2.$$

Unter $\sigma^2 = \sigma_0^2$ gilt

$$T := \frac{n\widetilde{S}_n^2}{\sigma_0^2} = \sum_{i=1}^{n}\left(\frac{X_i - \mu_0}{\sigma_0}\right)^2 \sim \chi_n^2.$$

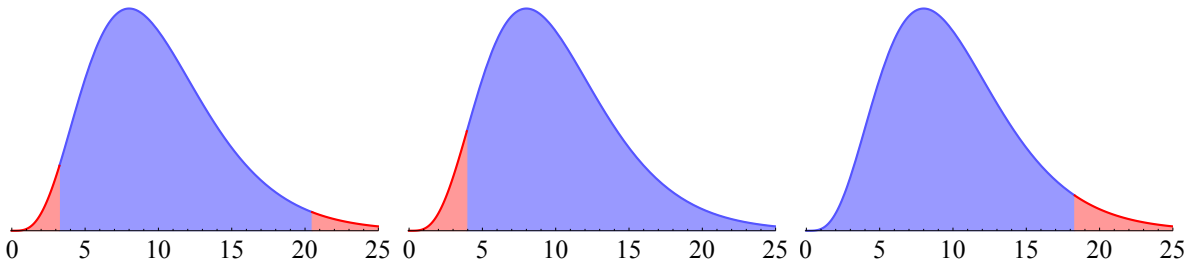


Abbildung 3. Vorgehensweise beim χ^2-Test. Rot: Ablehnungsbereich. Blau: Annahmebereich. Links: Zweiseitiger Test (Fall 1A). Mitte: Einseitiger Test (Fall 1B). Rechts: Einseitiger Test (Fall 1C).

Fall 3A. $H_0 : \sigma^2 = \sigma_0^2$; $H_1 : \sigma^2 \neq \sigma_0^2$. Die Nullhypothese H_0 sollte abgelehnt werden, wenn T zu groß oder zu klein ist. Die χ^2-Verteilung ist nicht symmetrisch. Dies führt zu folgender Entscheidungsregel: H_0 wird verworfen, wenn $T < \chi^2_{n,\frac{\alpha}{2}}$ oder $T > \chi^2_{n,1-\frac{\alpha}{2}}$.

Fall 3B. $H_0 : \sigma^2 \geq \sigma_0^2$; $H_1 : \sigma^2 < \sigma_0^2$. Die Nullhypothese H_0 sollte verworfen werden, wenn T zu klein ist. Dies führt zu folgender Entscheidungsregel: H_0 wird verworfen, wenn $T < \chi^2_{n,\alpha}$ ist.

Fall 3C. $H_0 : \sigma^2 \leq \sigma_0^2$; $H_1 : \sigma^2 > \sigma_0^2$. Die Nullhypothese H_0 sollte verworfen werden, wenn T zu groß ist. Dies führt zu folgender Entscheidungsregel: H_0 wird verworfen, wenn $T > \chi^2_{n,1-\alpha}$ ist.

Fall 4: Tests für σ^2 bei unbekanntem μ (χ^2-Streuungstest).

Seien $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$, wobei μ und σ^2 unbekannt seien. Wir wollen Hypothesen über σ^2 testen, z. B. $\sigma^2 = \sigma_0^2$, $\sigma^2 \geq \sigma_0^2$ oder $\sigma^2 \leq \sigma_0^2$, wobei σ_0^2 vorgegeben ist. Ein natürlicher Schätzer für σ^2 ist in diesem Fall

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

Unter $\sigma^2 = \sigma_0^2$ gilt

$$T := \frac{(n-1)S_n^2}{\sigma_0^2} \sim \chi^2_{n-1}.$$

Die Entscheidungsregeln sind also die gleichen wie in Fall 3, lediglich muss man die Anzahl der Freiheitsgrade der χ^2-Verteilung durch $n-1$ ersetzen.

10.3. Zweistichprobentests für die Parameter der Normalverteilung

Nun betrachten wir zwei Stichproben $(X_1, \ldots, X_n)$ und $(Y_1, \ldots, Y_m)$. Wir wollen verschiedene Hypothesen über die Lage und die Streuung dieser Stichproben testen. Z. B. kann man sich für die Hypothese interessieren, dass die Erwartungswerte (bzw. Streuungen) der beiden Stichproben gleich sind. Wir machen folgende Annahmen:

(1) $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ sind unabhängige Zufallsvariablen.
(2) $X_1, \ldots, X_n \sim \mathrm{N}(\mu_1, \sigma_1^2)$.
(3) $Y_1, \ldots, Y_m \sim \mathrm{N}(\mu_2, \sigma_2^2)$.

Wir wollen nun Hypothesen über $\mu_1 - \mu_2$ und σ_1^2/σ_2^2 testen. Dabei werden wir uns auf die Nullhypothesen der Form $\mu_1 = \mu_2$ bzw. $\sigma_1^2 = \sigma_2^2$ beschränken. Nullhypothesen der Form $\mu_1 \geq \mu_2$, $\mu_1 \leq \mu_2$, $\sigma_1^2 \geq \sigma_2^2$, $\sigma_1^2 \leq \sigma_2^2$ können analog betrachtet werden.

Fall 1: Test für $\mu_1 = \mu_2$ bei bekannten σ_1^2 und σ_2^2 (Zweistichproben-z-Test).

Es seien also σ_1^2 und σ_2^2 bekannt. Wir können $\mu_1 - \mu_2$ durch $\bar{X}_n - \bar{Y}_m$ schätzen. Unter der Nullhypothese $H_0 : \mu_1 = \mu_2$ gilt, dass

$$T := \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim \mathrm{N}(0,1).$$

Die Nullhypothese H_0 wird verworfen, wenn $|T|$ groß ist, also wenn $|T| > z_{1-\frac{\alpha}{2}}$.

Fall 2: Test für $\mu_1 = \mu_2$ bei unbekannten aber gleichen σ_1^2 und σ_2^2 (Zweistichproben-t-Test).

Es seien nun σ_1^2 und σ_2^2 unbekannt. Um das Problem zu vereinfachen, werden wir annehmen, dass die Varianzen gleich sind, d.h. $\sigma^2 := \sigma_1^2 = \sigma_2^2$. Wir schätzen σ^2 durch

$$S = \frac{1}{n+m-2}\left(\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 + \sum_{j=1}^{m}(Y_j - \bar{Y}_m)^2\right).$$

Wir betrachten die folgende Teststatistik:

$$T := \frac{\bar{X}_n - \bar{Y}_m}{S\sqrt{\frac{1}{n} + \frac{1}{m}}}.$$

Wir haben bei der Konstruktion der Konfidenzintervalle gezeigt, dass $T \sim t_{n+m-2}$ unter $\mu_1 = \mu_2$. Somit wird die Nullhypothese H_0 verworfen, wenn $|T| > t_{n+m-2,1-\frac{\alpha}{2}}$.

Fall 3: Test für $\sigma_1^2 = \sigma_2^2$ bei unbekannten μ_1 und μ_2 (F-Test).

Seien also μ_1 und μ_2 unbekannt. Wir wollen die Nullhypothese $H_0 : \sigma_1^2 = \sigma_2^2$ testen. Natürliche Schätzer für σ_1^2 und σ_2^2 sind gegeben durch

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2, \quad S_Y^2 = \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \bar{Y}_m)^2.$$

Bei der Konstruktion der Konfidenzintervalle haben wir gezeigt, dass für $\sigma_1^2 = \sigma_2^2$

$$T := \frac{S_X^2}{S_Y^2} \sim F_{n-1,m-1}.$$

Die Hypothese H_0 sollte verworfen werden, wenn T zu klein oder zu groß ist. Dabei ist die F-Verteilung nicht symmetrisch. Die Nullhypothese wird also verworfen, wenn $T < F_{n-1,m-1,\frac{\alpha}{2}}$ oder $T > F_{n-1,m-1,1-\frac{\alpha}{2}}$.

Fall 4: Test für $\sigma_1^2 = \sigma_2^2$ bei bekannten μ_1 und μ_2 (F-Test).

Analog zu Fall 3 (Übung).

10.4. Allgemeine Modellbeschreibung

Wir beschreiben nun allgemein das statistische Testproblem. Sei $(\mathbb{P}_\theta)_{\theta\in\Theta}$ eine Familie von Wahrscheinlichkeitsmaßen auf dem Stichprobenraum $(\mathfrak{X}, \mathcal{A})$. Der Parameterraum Θ sei in zwei disjunkte Teilmengen Θ_0 und Θ_1 aufgeteilt, d.h.

$$\Theta = \Theta_0 \cup \Theta_1, \quad \Theta_0 \cap \Theta_1 = \varnothing.$$

Sei X eine Stichprobe, die zufällig aus $\mathfrak{X}$ gemäß $\mathbb{P}_\theta$ gezogen wird, wobei $\theta \in \Theta$ unbekannt sei. Wir betrachten nun zwei Hypothesen:

(1) Die Nullhypothese H_0: $\theta \in \Theta_0$.
(2) Die Alternativhypothese H_1: $\theta \in \Theta_1$.

Wir sollen anhand der Stichprobe X entscheiden, ob wir H_0 verwerfen oder beibehalten.

Definition 10.4.1. Ein *Test* ist eine messbare Funktion $\varphi : \mathfrak{X} \to \{0, 1\}$.

Interpretation:

- H_0 wird verworfen, falls $\varphi(X) = 1$.
- H_0 wird beibehalten, falls $\varphi(X) = 0$.

Die Menge $K := \{x \in \mathfrak{X} \colon \varphi(x) = 1\}$ heißt der *Ablehnungsbereich*, denn H_0 wird verworfen, falls $X \in H_0$.

Wir werden auch einen allgemeineren Begriff benötigen:

Definition 10.4.2. Ein *randomisierter Test* ist eine messbare Funktion $\varphi : \mathfrak{X} \to [0, 1]$.

Interpretation: Um die Entscheidung zu treffen, ob H_0 verworfen werden soll, berechnet man zuerst $p := \varphi(X)$. Danach führt man ein Bernoulli-Experiment mit Erfolgswahrscheinlichkeit p durch. Bei Erfolg verwirft man H_0, bei Misserfolg behält man H_0 bei.

Im folgenden betrachten wir immer randomisierte Tests.

Definition 10.4.3. Die Funktion $G : \mathfrak{X} \to [0, 1]$ mit $G(\theta) = \mathbb{E}_\theta \varphi(X)$ heißt die *Gütefunktion* eines Tests.

Dabei ist $\mathbb{E}_\theta \varphi(X)$ die Wahrscheinlichkeit (unter $\mathbb{P}_\theta$), dass der Test φ die Hypothese H_0 verwirft. Es gilt also:

- Für $\theta \in \Theta_0$ ist $G(\theta)$ die Wahrscheinlichkeit, dass H_0 irrtümlicherweise verworfen wird (Fehler 1. Art).
- Für $\theta \in \Theta_1$ ist $1 - G(\theta)$ die Wahrscheinlichkeit, dass H_0 irrtümlicherweise beibehalten wird (Fehler 2. Art).

Definition 10.4.4. Für $\theta \in \Theta_1$ heißt $G(\theta) = \mathbb{E}_\theta \varphi(X)$ die *Macht* des Tests. Die Macht ist also die Wahrscheinlichkeit, dass die falsche Nullhypothese entlarvt wird.

Beim Testen können wir zwei Arten von Fehlern machen:

- Fehler 1. Art: H_0 wird verworfen, obwohl H_0 richtig ist.
- Fehler 2. Art: H_0 wird nicht verworfen, obwohl H_0 falsch ist.

Normalerweise versucht man φ (bzw. den Ablehnungsbereich K) so zu wählen, dass die Wahrscheinlichkeit eines Fehlers 1. Art durch ein vorgegebenes Niveau $\alpha \in (0, 1)$ beschränkt ist, typischerweise $\alpha = 0.01$ oder 0.05.

Definition 10.4.5. Ein Test $\varphi : \mathfrak{X} \to [0,1]$ hat *Signifikanzniveau* $\alpha \in (0,1)$, falls
$$\mathbb{E}_\theta \varphi(X) \leq \alpha \text{ für alle } \theta \in \Theta_0.$$

Sei Φ_α die Menge aller Tests zum Signifikanzniveau α. Unter allen Tests zum Niveau α möchte man nun denjenigen finden, der eine möglichst große Macht hat.

Definition 10.4.6. Wir sagen, dass $\varphi : \mathfrak{X} \to [0,1]$ *gleichmäßig bester Test* zum Niveau α ist, wenn $\varphi \in \Phi_\alpha$ und
$$\mathbb{E}_\theta \varphi(X) = \sup_{\psi \in \Phi_\alpha} \mathbb{E}_\theta \psi(X) \text{ für alle } \theta \in \Theta_1.$$

Diese Bedingung besagt, dass für alle $\theta \in \Theta_1$ der Test φ eine kleinere Wahrscheinlichkeit eines Fehlers 2. Art unter $\mathbb{P}_\theta$ hat als jeder andere Test $\psi \in \Phi_\alpha$.

Zum Schluss definieren wir noch den p-Wert. Stellen wir uns vor, dass wir unsere Testentscheidung auf dem Wert einer Statistik $T : \mathfrak{X} \to \mathbb{R}$ basieren. Es kann z.B. sein, dass große Werte von T für eine Ablehnung von H_0 sprechen. In diesem Fall hat der Test die Form $\varphi(x) = \mathbb{1}_{T(x) \geq c}$ für einen kritischen Wert c.

Definition 10.4.7. Der p-Wert einer Beobachtung $x \in \mathfrak{X}$ ist gegeben durch
$$p\text{-Wert(x)} = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta[T(X) \geq T(x)].$$

Ein p-Wert, der kleiner als α ist, führt zur Ablehnung von H_0.

Beispiel 10.4.8. Wir berechnen die Gütefunktion des Gauß-z-Tests. Seien $X_1, \ldots, X_n \sim \mathrm{N}(\mu, 1)$ unabhängig. Wir wollen $H_0 : \mu = 0$ gegen $H_1 : \mu \neq 0$ testen, wobei wir hier der Einfachheit halber angenommen haben, dass $\mu_0 = 0$ und $\sigma_0^2 = 1$. Der zweiseitige Gauß-z-Test ist gegeben durch
$$\varphi(x_1, \ldots, x_n) = \mathbb{1}_{|\sqrt{n}\bar{x}_n| > z_{1-\frac{\alpha}{2}}}.$$
Unter $\mathbb{P}_\mu$ gilt $\bar{X}_n \sim \mathrm{N}(\mu, 1/n)$, somit $\sqrt{n}\bar{X}_n - \sqrt{n}\mu \sim \mathrm{N}(0,1)$. Die Gütefunktion berechnet sich zu
$$\begin{aligned} G(\mu) &= \mathbb{E}_\mu \varphi(X_1, \ldots, X_n) \\ &= \mathbb{P}_\mu \left[|\sqrt{n}\bar{X}_n| > z_{1-\frac{\alpha}{2}} \right] \\ &= \mathbb{P}_\mu \left[\sqrt{n}\bar{X}_n - \sqrt{n}\mu < -z_{1-\frac{\alpha}{2}} - \sqrt{n}\mu \right] + \mathbb{P}_\mu \left[\sqrt{n}\bar{X}_n - \sqrt{n}\mu > z_{1-\frac{\alpha}{2}} - \sqrt{n}\mu \right] \\ &= 1 - \Phi(\sqrt{n}\mu + z_{1-\frac{\alpha}{2}}) + \Phi(\sqrt{n}\mu - z_{1-\frac{\alpha}{2}}), \end{aligned}$$
s. Abbildung 4 (links), wobei $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \mathrm{e}^{-t^2/2} \mathrm{d}t$ die Standardnormalverteilungsfunktion bezeichnet. Die Gütefunktion ist gleich α an der Stelle 0 (Wahrscheinlichkeit eines Fehlers 1. Art) und konvergiert gegen 1 für $\mu \to \pm\infty$ (somit wird die Alternative bei großem $|\mu|$ mit großer Wahrscheinlichkeit richtig entlarvt).

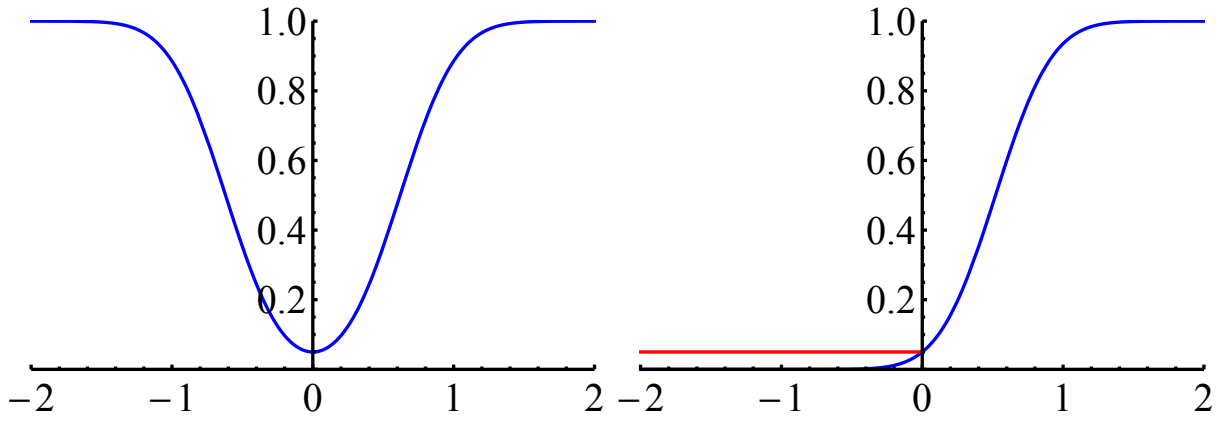


ABBILDUNG 4. Gütefunktion des Gauß-z-Tests für $\mu_0 = 0$, $\sigma_0^2 = 1$ und $n = 10$. Links: Zweiseitiger Test (Fall 1A). Rechts: Einseitiger Test (Fall 1C).

Aufgabe 10.4.9. Seien $X_1, \dots, X_n \sim \mathrm{N}(\mu, 1)$ unabhängig. Betrachten Sie die einseitigen Hypothesen $H_0 : \mu \leq 0$ und $H_1 : \mu > 0$.

(a) Bestimmen Sie die Gütefunktion des einseitigen Gauß-z-Tests

$$\varphi(x_1, \dots, x_n) = \mathbb{1}_{\sqrt{n}\bar{x}_n > z_{1-\alpha}}.$$

(b) Zeigen Sie, dass die Gütefunktion für alle $\mu \leq 0$ unterhalb von α bleibt, s. Abbildung 4 (rechts). D.h. es handelt sich tatsächlich um einen Test zum Niveau α.

10.5. Tests einfacher Hypothesen: Neyman-Pearson-Theorie

In diesem Kapitel betrachten wir den Fall, wenn beide Hypothesen *einfach* sind, d.h.

$$\Theta_0 = \{\theta_0\},\ \Theta_1 = \{\theta_1\},\ \Theta = \{\theta_0, \theta_1\}.$$

Zur Vereinfachung der Notation schreiben wir im Folgenden $\mathbb{P}_0$ bzw. $\mathbb{P}_1$ für $\mathbb{P}_{\theta_0}$ bzw. $\mathbb{P}_{\theta_1}$.

Annahme: Die Wahrscheinlichkeitsmaße $\mathbb{P}_0$ und $\mathbb{P}_1$ besitzen Dichten h_0 und h_1 bzgl. eines σ-endlichen Maßes λ auf $(\mathfrak{X}, \mathcal{A})$.

Wir werden nun zeigen, wie man einen gleichmäßig besten Test zum Niveau α konstruiert. Eine ganz natürliche Vorgehensweise ist diese: man entscheidet sich für H_1 bzw. H_0 wenn der sogenannte *Likelihood-Quotient* $h_1(x)/h_0(x)$ größer bzw. kleiner als ein vorgegebener Wert k ist. Ist der Quotient gleich k, so ist man sich nicht sicher und randomisiert mit Erfolgswahrscheinlichkeit γ.

Definition 10.5.1. Seien $k \in [0,\infty]$ und $\gamma \in [0,1]$. Ein *Likelihood-Quatienten-Test* (oder *LQ-Test*) ist ein Test der Form

$$\varphi(x) = \begin{cases} 1, & \text{falls } \frac{h_1(x)}{h_0(x)} > k, \\ 0, & \text{falls } \frac{h_1(x)}{h_0(x)} < k, \\ \gamma, & \text{falls } \frac{h_1(x)}{h_0(x)} = k. \end{cases} \tag{10.5.1}$$

Mögliche Unbestimmtheiten der Form $0/0$ werden wir im Folgenden ignorieren, denn die Menge $A := \{x \in \mathfrak{X} \colon h_0(x) = h_1(x) = 0\}$ ist eine Nullmenge bzgl. $\mathbb{P}_0$ und $\mathbb{P}_1$. Somit ist die Wahrscheinlichkeit, dass die Stichprobe X in A landet gleich 0 sowohl unter H_0 als auch unter H_1.

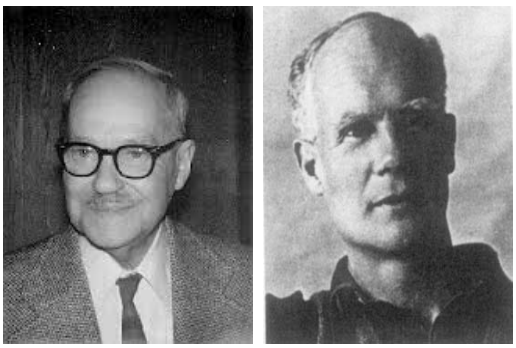


Abbildung 5. Jerzy Neyman und Egon Pearson (der Sohn von Karl Pearson)

Lemma 10.5.2 (Neyman-Pearson, Teil 1)**.** Sei φ ein LQ-Test mit $\mathbb{E}_0\varphi(X) = \alpha$. Dann gilt

$$\mathbb{E}_1\varphi(X) = \sup_{\psi \colon \mathbb{E}_0\psi(X) \le \alpha} \mathbb{E}_1\psi(X),$$

d.h. φ ist gleichmäßig bester Test zum Niveau α.

Beweis. Sei ψ ein Test zum Niveau α, d.h. $\mathbb{E}_0\psi(X) \le \alpha$. Es reicht zu zeigen, dass

$$\mathbb{E}_1\varphi(X) \ge \mathbb{E}_1\psi(X).$$

Wir behaupten, dass

$$(\varphi(x) - \psi(x))(h_1(x) - kh_0(x)) \ge 0 \text{ für alle } x \in \mathfrak{X}. \tag{10.5.2}$$

Um dies zu zeigen, betrachten wir drei Fälle:

Fall 1: $h_1(x) - kh_0(x) > 0$. Dann gilt $\varphi(x) = 1$ und folglich $\varphi(x) - \psi(x) \geq 0$, woraus sich die Behauptung (10.5.2) ergibt.

Fall 2: $h_1(x) - kh_0(x) < 0$. Dann gilt $\varphi(x) = 0$. Es folgt $\varphi(x) - \psi(x) \leq 0$, und (10.5.2) ist richtig.

Fall 3: $h_1(x) - kh_0(x) = 0$. In diesem Fall ist das Produkt auf der linken Seite von (10.5.2) gleich 0 und (10.5.2) stimmt.

Indem wir (10.5.2) bzgl. λ integrieren, erhalten wir

$$\int_{\mathfrak{X}} (\varphi - \psi) h_1 \mathrm{d}\lambda \geq k \int_{\mathfrak{X}} (\varphi - \psi) h_0 \mathrm{d}\lambda.$$

Nachdem wir Integrale als Erwartungswerte darstellen, ergibt sich

$$\mathbb{E}_1[\varphi(X) - \psi(X)] \geq k\, \mathbb{E}_0[\varphi(X) - \psi(X)].$$

Es gilt allerdings $\mathbb{E}_0\varphi(X) = \alpha$, während $\mathbb{E}_0\psi(X) \leq \alpha$. Somit ist der Erwartungswert auf der rechten Seite nichtnegativ und es folgt, dass $\mathbb{E}_1[\varphi(X) - \psi(X)] \geq 0$. Das beweist die Behauptung. □

Lemma 10.5.3 (Neyman-Pearson, Teil 2)**.** Zu jedem $\alpha \in (0, 1)$ lassen sich $k \in [0, \infty)$ und $\gamma \in [0, 1]$ finden, so dass für den durch (10.5.1) definierten Test φ

$$\mathbb{E}_0\varphi(X) = \alpha$$

gilt. Laut Teil 1 des Neyman-Pearson-Lemmas ist φ gleichmäßig bester Test zum Niveau α.

Beweis. Sei φ gegeben durch (10.5.1). Durch eine geeignete Wahl von k und γ wollen wir erreichen, dass

$$\mathbb{E}_0\varphi(X) \overset{\text{def}}{=} \mathbb{P}_0[T > k] + \gamma \mathbb{P}_0[T = k] \overset{!}{=} \alpha.$$

Die Statistik T ist $\mathbb{P}_0$-fast sicher endlich, denn es gilt

$$\mathbb{P}_0[T = +\infty] = \mathbb{P}_0[h_0 = 0] = \int_{\{h_0=0\}} h_0 \mathrm{d}\lambda = \int_{\{h_0=0\}} 0 \mathrm{d}\lambda = 0.$$

Die Funktion $y \mapsto \mathbb{P}_0[T > y]$ ist rechtsstetig, monoton nichtsteigend und es gilt $\lim_{y\to+\infty} \mathbb{P}_0[T > y] = 0$ und $\mathbb{P}_0[T > y] = 1$ falls $y < 0$. Startend mit $y = +\infty$ verkleinern wir den Wert von y solange $\mathbb{P}_0[T > y] \leq \alpha$ gilt. D.h. wir definieren

$$k = \inf\{y > 0 \colon \mathbb{P}_0[T > y] \leq \alpha\} \in [0, \infty).$$

Nun gibt es zwei Fälle.

Fall 1: Der Wert α wird erreicht, d.h. $\mathbb{P}_0[T > k] = \alpha$. Dann können wir $\gamma = 0$ setzen.

Fall 2: Der Wert α wird übersprungen, d.h. $\mathbb{P}_0[T > k] < \alpha$ und $\mathbb{P}_0[T \geq k] \geq \alpha$. Wir definieren dann

$$\gamma := \frac{\alpha - \mathbb{P}_0[T > k]}{\mathbb{P}_0[T = k]} \in [0, 1],$$

wobei wir bemerken, dass $\alpha - \mathbb{P}_0[T > k] \leq \mathbb{P}_0[T \geq k] - \mathbb{P}_0[T > k] = \mathbb{P}_0[T = k]$ und folglich $\gamma \leq 1$. □

Beispiel 10.5.4 (Likelihood-Quotienten-Test für den Parameter der Binomialverteilung)**.** Es sei $X \sim \mathrm{Bin}(n, \theta)$ mit einem unbekannten $\theta \in (0, 1)$. Wir betrachten die einfachen Hypothesen

$$H_0 : \theta = \theta_0 \quad \text{und} \quad H_1 : \theta = \theta_1,$$

wobei $\theta_0, \theta_1 \in (0, 1)$ vorgegeben seien und wir der Einfachheit halber $\theta_0 < \theta_1$ voraussetzen. Der gesunde Menschenverstand sagt, dass wir H_0 verwerfen müssen, wenn X „zu groß“ ist. Diese Intuition wird durch die Neyman-Pearson-Theorie bestätigt.
Der Stichprobenraum ist $\mathfrak{X} = \{0, \ldots, n\}$. Sei λ das Zählmaß auf $\{0, \ldots, n\}$, d.h. $\lambda(\{x\}) = 1$ für alle $x \in \mathfrak{X}$. Die Zähldichten h_0 und h_1 sind gegeben durch

$$h_i(x) = \binom{n}{x} \theta_i^x (1 - \theta_i)^{n-x}, \quad x \in \{0, \ldots, n\}, \quad i \in \{0, 1\}.$$

Für den Likelihood-Quotienten erhalten wir

$$\Lambda(x) := \frac{h_1(x)}{h_0(x)} = \frac{\binom{n}{x}\theta_1^x(1-\theta_1)^{n-x}}{\binom{n}{x}\theta_0^x(1-\theta_0)^{n-x}} = \left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right)^x \cdot \left(\frac{1-\theta_1}{1-\theta_0}\right)^n.$$

Der Likelihood-Quotienten-Test fällt die Entscheidung in Abhängigkeit davon, ob $\Lambda(x)$ größer, kleiner oder gleich einem bestimmten Wert k ist. Da aber $\frac{\theta_1}{\theta_0} > 1$ und $\frac{1-\theta_0}{1-\theta_1} > 1$, ist $\Lambda(x)$ eine monoton steigende Funktion von x. Somit gilt

$$\Lambda(x) > k \iff x > k^*, \quad \Lambda(x) < k \iff x < k^*, \quad \Lambda(x) = k \iff x = k^*,$$

für einen passenden Wert k^*. Wir können also den Likelihood-Quotienten-Test auch folgendermaßen darstellen:

$$\varphi(x) = \begin{cases} 1, & \text{falls } x > k^*, \\ 0, & \text{falls } x < k^*, \\ \gamma^*, & \text{falls } x = k^* \end{cases}$$

für $x \in \{0, 1 \ldots, n\}$. Wir müssen noch die Werte $k^* \in \{0, \ldots, n\}$ und $\gamma^* \in [0, 1]$ so bestimmen, dass die Wahrscheinlichkeit eines Fehlers 1. Art gleich α wird, d.h.

$$\mathbb{P}_0[X > k^*] + \gamma^* \mathbb{P}_0[X = k^*] = \alpha. \tag{10.5.3}$$

Startend mit $k = n$ (in welchem Fall $\mathbb{P}_0[X > k] = 0$ ist) verkleinern wir den Wert k solange die Wahrscheinlichkeit $\mathbb{P}_0[X > k]$ unterhalb von α bleibt. Sei $k = k^*$ der kleinstmögliche Wert, für den noch $\mathbb{P}_0[X > k] \leq \alpha$ gilt, d.h.

$$k^* = \min\left\{ k \in \{0, \ldots, n\} : \mathbb{P}_0[X > k] \stackrel{\text{def}}{=} \sum_{i=k}^{n} \binom{n}{i} \theta_0^i (1 - \theta_0)^{n-i} \leq \alpha \right\}.$$

Würden wir nun H_0 verwerfen, wenn $X > k^*$, und sonst H_0 beibehalten, so wäre das Signifikanzniveau $\leq \alpha$. Bei einer weiteren Verkleinerung von k^* würde das Signifikanzniveau den Wert α übersteigen. Um das Niveau von *exakt* α zu erreichen, müssen wir im Fall $X = k^*$ eine zufällige Entscheidung treffen. Entscheiden wir uns für H_1 mit Wahrscheinlichkeit

$$\gamma^* = \frac{\alpha - \mathbb{P}_0[X > k^*]}{\mathbb{P}_0[X = k^*]} \in [0, 1],$$

so ist (10.5.3) erfüllt und das Niveau ist exakt α.

Beispiel 10.5.5 (Erkennung eines Signals im Rauschen)**.** Es seien $X_1, \dots, X_n \sim N(\mu, \sigma_0^2)$ unabhängig, wobei σ_0^2 bekannt sei. Fasst man $X_1, \dots, X_n$ als Messungen eines Signals zu verschiedenen Zeitpunkten auf, so kann man folgende Hypothesen aufstellen:

$$H_0 : \ \mu = 0, \ \text{d.h. es wurde nur Rauschen empfangen,}$$
$$H_1 : \ \mu = \mu_1, \ \text{d.h. es wurde ein verrauschtes Signal der Stärke } \mu_1 \text{ empfangen.}$$

Dabei sein $\mu_1 > 0$ bekannt. Der Stichprobenraum ist $\mathfrak{X} = \mathbb{R}^n$. Die Dichten von $\mathbb{P}_0$ und $\mathbb{P}_1$ bzgl. des Lebesgue-Maßes λ sind

$$h_0(x_1, \dots, x_n) = \left(\frac{1}{\sqrt{2\pi}\sigma_0}\right)^n \mathrm{e}^{-\frac{1}{2\sigma_0^2}\sum_{i=1}^n x_i^2}, \quad h_1(x_1, \dots, x_n) = \left(\frac{1}{\sqrt{2\pi}\sigma_0}\right)^n \mathrm{e}^{-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (x_i-\mu_1)^2}.$$

Nach dem Lemma von Neyman-Pearson basiert der gleichmäßig beste Test auf dem Wert der LQ-Statistik

$$\Lambda(x_1, \dots, x_n) := \frac{h_1(x_1, \dots, x_n)}{h_0(x_1, \dots, x_n)} = \mathrm{e}^{\frac{\mu_1}{\sigma_0^2}\sum_{i=1}^n x_i - \frac{n}{2\sigma_0^2}\mu_1^2}.$$

Da Λ eine monoton steigende Funktion von $\bar{x}_n$ ist, hat der gleichmäßig beste Test die Form

$$\varphi(x_1, \dots, x_n) = \begin{cases} 1, & \text{falls } \bar{x}_n > k^*, \\ 0, & \text{falls } \bar{x}_n < k^*, \end{cases}$$

wobei der Fall $\bar{x}_n = k^*$ wegen der Stetigkeit der Normalverteilung ignoriert werden kann und k^* so gewählt werden muss, dass $\mathbb{P}_0[\bar{X}_n > k^*] = \alpha$. Da $\sqrt{n}\bar{X}_n$ unter H_0 standardnormalverteilt ist, müssen wir $k^* = z_{1-\alpha}/\sqrt{n}$ wählen. Der gleichmäßig beste Test sieht also folgendermaßen aus:

$$\varphi(x_1, \dots, x_n) = \begin{cases} 1, & \text{falls } \sqrt{n}\bar{x}_n > z_{1-\alpha}, \\ 0, & \text{falls } \sqrt{n}\bar{x}_n < z_{1-\alpha}. \end{cases}$$

Es sei bemerkt, dass der resultierende Test nicht von μ_1 abhängt.

Aufgabe 10.5.6. Bestimmen Sie die Wahrscheinlichkeit eines Fehlers 2. Art für den obigen Test. Zeigen Sie, dass diese Wahrscheinlichkeit für $n \to \infty$ gegen 0 konvergiert.

Aufgabe 10.5.7. Sei φ ein gleichmäßig bester Test zum Niveau α für $H_0 = \{\theta_0\}$ gegen $H_1 = \{\theta_1\}$. Zeigen Sie, dass $\mathbb{E}_{\theta_1}\varphi(X) \geq \alpha$.

Aufgabe 10.5.8. Betrachten Sie ein statistisches Modell $(\mathfrak{X}, \mathcal{A}, \{\mathbb{P}_0, \mathbb{P}_1\})$, wobei $\mathbb{P}_0$ und $\mathbb{P}_1$ die Dichten h_0 und h_1 bzgl. eines σ-endlichen Maßes λ auf $\mathfrak{X}$ besitzen. Ein Test φ für $H_0 = \{0\}$ gegen $H_1 = \{1\}$ heißt ein *Minimax-Test*, wenn das Maximum der Irrtumswahrscheinlichkeiten erster und zweiter Art minimal unter allen Tests ist. Zeigen Sie:

(a) Es gibt einen Likelihood-Quotienten Test φ mit $\mathbb{E}_0[\varphi] = \mathbb{E}_1[1 - \varphi]$.
(b) Der Test aus (a) ist ein Minimax-Test.

10.6. Tests für einseitige Hypothesen bei monotonen Dichtequotienten

Sei $(\mathbb{P}_\theta)_{\theta\in\Theta}$ eine Familie von Wahrscheinlichkeitsmaßen auf dem Stichprobenraum $(\mathfrak{X},\mathcal{A})$. In diesem Abschnitt nehmen wir an, dass $\Theta = (\theta_-,\theta_+) \subset \mathbb{R}$ ein (möglicherweise unendliches) Intervall ist und betrachten die einseitigen Hypothesen

$$H_0 : \theta \le \theta_0 \quad \text{und} \quad H_1 : \theta > \theta_0,$$

wobei $\theta_0 \in \Theta$ vorgegeben sei.

Die Aufgabe, H_0 gegen H_1 zu testen, ist sicherlich schwieriger, als die Aufgabe, die einfachen Hypothese $\theta = \theta_1$ und $\theta = \theta_2$ gegeneinander zu testen, wobei $\theta_1 \le \theta_0 < \theta_2$. Die in Beispielen 10.5.4 und 10.5.5 konstruierten gleichmäßig besten Tests für einfache Hypothesen basieren jeweils auf einer Statistik T, die unabhängig von der Wahl von θ_1 und θ_2 ist. Diese Kohärenzeigenschaft lässt hoffen, dass auch der gleichmäßig beste Test für die einseitigen Hypothesen auf derselben Statistik basieren muss. Wir werden nun eine allgemeine Eigenschaft von statistischen Modellen formulieren, die garantiert, dass alle LQ-Tests von einfachen Hypothesen auf derselben Statistik basieren.

Annahme: $(\mathbb{P}_\theta)_{\theta\in\Theta}$ ist eine *dominierte* Familie von Wahrscheinlichkeitsmaßen auf $(\mathfrak{X},\mathcal{A})$, d.h. es gibt ein σ-endliches Maß λ auf $(\mathfrak{X},\mathcal{A})$, sodass für jedes $\theta\in\Theta$ das Wahrscheinlichkeitsmaß $\mathbb{P}_\theta$ eine Dichte h_θ bzgl. λ besitzt.

Definition 10.6.1. Die Familie $(\mathbb{P}_\theta)_{\theta\in\Theta}$ besitzt *monotone Dichtequotienten* in einer Statistik $T : \mathfrak{X} \to \mathbb{R}$, wenn für alle $\theta_1 < \theta_2$ eine monoton steigende Funktion $H_{\theta_1,\theta_2} : \mathbb{R} \to [0,\infty]$ existiert mit

$$\frac{h_{\theta_2}(x)}{h_{\theta_1(x)}} = H_{\theta_1,\theta_2}(T(x)) \quad \mathbb{P}_{\theta_1}\text{- und } \mathbb{P}_{\theta_2}\text{-f.s.}$$

Beispiel 10.6.2. Sei $\mathbb{P}_\theta$ die Binomialverteilung $\mathrm{Bin}(n,\theta)$ auf $\mathfrak{X} = \{0,\dots,n\}$. Der Parameterraum ist dabei $\Theta = (0,1)$. Als λ nehmen wir das Zählmaß auf $\{0,\dots,n\}$, d.h. $\lambda(\{x\}) = 1$ für alle $x \in \mathfrak{X}$. Die Zähldichte h_θ ist dann gegeben durch

$$h_\theta(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \quad x \in \{0,\dots,n\}.$$

Die Dichtequotienten sehen für $\theta_1 < \theta_2$ folgendermaßen aus:

$$\frac{h_{\theta_2}(x)}{h_{\theta_1(x)}} = \frac{\binom{n}{x}\theta_1^x(1-\theta_1)^{n-x}}{\binom{n}{x}\theta_2^x(1-\theta_2)^{n-x}} = \left(\frac{\theta_2(1-\theta_1)}{\theta_1(1-\theta_2)}\right)^x \cdot \left(\frac{1-\theta_1}{1-\theta_0}\right)^n,$$

was eine monoton steigende Funktion von $T(x) := x$ ist, denn $\frac{\theta_2}{\theta_1} > 1$ und $\frac{1-\theta_1}{1-\theta_2} > 1$. Somit liegen hier monotone Dichtequotienten vor.

Beispiel 10.6.3. Seien $X_1, \dots, X_n$ unabhängige Zufallsvariablen mit Dichte/Zähldichte $h_\theta(x)$. Bildet $h_\theta(x) = a(\theta)b(x)\mathrm{e}^{c(\theta)d(x)}$ eine Exponentialfamilie, so gilt für den Likelihood-Quotienten

$$\frac{h_{\theta_2}(x_1)\dots h_{\theta_2}(x_n)}{h_{\theta_1}(x_1)\dots h_{\theta_1}(x_n)} = \exp\left\{(c(\theta_2) - c(\theta_1))\sum_{i=1}^n d(x_i)\right\}.$$

Ist die Funktion $c(\theta)$ monoton steigend, so liegen monotone Dichtequotienten in der Statistik $T(x_1, \dots, x_n) := \sum_{i=1}^n d(x_i)$ vor. Diese Beobachtung liefert mehrere Beispiele von statistischen Modellen mit monotonen Dichtequotienten, etwa

(a) $X_1, \dots, X_n \sim \mathrm{N}(\mu, \sigma_0^2)$ unabhängig mit bekanntem $\sigma_0^2 > 0$. Dabei ist $T(x_1, \dots, x_n) = \sum_{i=1}^n x_i$.
(b) $X_1, \dots, X_n \sim \mathrm{N}(\mu_0, \sigma^2)$ unabhängig mit bekanntem $\mu_0 \in \mathbb{R}$. Dabei ist $T(x_1, \dots, x_n) = \sum_{i=1}^n x_i^2$.
(c) $X_1, \dots, X_n \sim \mathrm{Poi}(\theta)$ unabhängig. Dabei ist $T(x_1, \dots, x_n) = \sum_{i=1}^n x_i$.
(d) $X_1, \dots, X_n \sim \mathrm{Exp}(1/\theta)$ unabhängig. Dabei ist $T(x_1, \dots, x_n) = \sum_{i=1}^n x_i$.

Der nächste Satz beschreibt den gleichmäßig besten Test zum Niveau α für einseitige Hypothesen bei monotonen Dichtequotienten in einer Statistik T. Dieser Test basiert auf dem Wert der Statistik T.

Satz 10.6.4 (Gleichmäßig bester Test für einseitige Hypothesen)**.** Sei $(\mathbb{P}_\theta)_{\theta\in\Theta}$ mit $\Theta = (\theta_-, \theta_+) \subset \mathbb{R}$ eine Familie von Wahrscheinlichkeitsmaßen auf $(\mathfrak{X}, \mathcal{A})$ mit monotonen Dichtequotienten in einer Statistik $T : \mathfrak{X} \to \mathbb{R}$. Sei $\theta_0 \in \Theta$ vorgegeben und betrachte die einseitigen Hypothesen $H_0 : \theta \le \theta_0$ und $H_1 : \theta > \theta_0$.

(a) Zu jedem $\alpha \in (0, 1)$ existieren $k^* \in \mathbb{R}$ und $\gamma^* \in [0, 1]$ mit

$$\mathbb{P}_{\theta_0}[T > k^*] + \gamma^*\mathbb{P}_{\theta_0}[T = k^*] = \alpha.$$

(b) Der durch

$$\varphi^*(x) = \begin{cases} 1, & \text{falls } T(x) > k^*, \\ 0, & \text{falls } T(x) < k^*, \\ \gamma^*, & \text{falls } T(x) = k^*. \end{cases}$$

definierte Test φ^* ist gleichmäßig bester Test für H_0 gegen H_1 zum Niveau α.

Beweis. Wir zeigen nur Teil (b), denn der Beweis von Teil (a) verläuft genauso wie im zweiten Teil des Neyman-Pearson-Lemmas.

Schritt 1: Die Macht von φ^*. Sei $\theta > \theta_0$. Wir zeigen, dass für jeden Test φ mit $\mathbb{E}_{\theta_0}\varphi(X) \le \alpha$ gilt

$$\mathbb{E}_\theta\varphi^*(X) \ge \mathbb{E}_\theta\varphi(X).$$

Das bedeutet, dass φ^* eine kleinere Wahrscheinlichkeit eines Fehlers 2. Art besitzt als jeder Test φ zum Niveau α. Wir behaupten, dass

$$\varphi^*(x) = \begin{cases} 1, & \text{falls } \frac{h_\theta(x)}{h_{\theta_0}(x)} > k, \\ 0, & \text{falls } \frac{h_\theta(x)}{h_{\theta_0}(x)} < k, \\ \gamma^*, & \text{falls } \frac{h_\theta(x)}{h_{\theta_0}(x)} = k, \end{cases} \tag{10.6.1}$$

wobei $k = H_{\theta_0,\theta}(k^*)$. In der Tat,

$$\frac{h_\theta(x)}{h_{\theta_0}(x)} > k \iff H_{\theta_0,\theta}(T(x)) > k = H_{\theta_0,\theta}(k^*) \iff T(x) > k^*,$$

wobei wir die Monotonie von $H_{\theta_0,\theta}$ benutzt haben. Analoge Äquivalenzen gelten für „<“ und „=“.
Es folgt aus (10.6.1), dass φ^* ein LQ-Test für die einfache Hypothese $\tilde{H}_0 := \{\theta_0\}$ gegen $\tilde{H}_1 := \{\theta_1\}$ ist. Das Niveau dieses Tests ist

$$\mathbb{E}_{\theta_0}\varphi^*(X) = \mathbb{P}_{\theta_0}[T > k^*] + \gamma^*\mathbb{P}_{\theta_0}[T = k^*] = \alpha.$$

Es folgt aus dem ersten Teil des Neyman-Pearson-Lemmas, dass φ^* eine nicht kleinere Macht besitzt, als jeder andere Test zum Niveau $\leq \alpha$, also z.B. $\mathbb{E}_\theta\varphi^*(X) \geq \mathbb{E}_\theta\varphi(X)$. Das beweist die Behauptung von Schritt 1.

Schritt 2: Das Niveau von φ^*. Wir zeigen, dass φ^* ein Test zum Niveau α ist, d.h.

$$\mathbb{E}_\theta\varphi^*(X) \leq \alpha \text{ für alle } \theta \leq \theta_0.$$

Für $\theta = \theta_0$ fogt das aus Teil (a) des Satzes. Sei also $\theta < \theta_0$. Wir behaupten, dass

$$\varphi^*(x) = \begin{cases} 1, & \text{falls } \frac{h_\theta(x)}{h_{\theta_0}(x)} < k', \\ 0, & \text{falls } \frac{h_\theta(x)}{h_{\theta_0}(x)} > k', \\ \gamma^*, & \text{falls } \frac{h_\theta(x)}{h_{\theta_0}(x)} = k', \end{cases} \tag{10.6.2}$$

wobei $k' = 1/H_{\theta,\theta_0}(k^*)$. In der Tat,

$$\frac{h_\theta(x)}{h_{\theta_0}(x)} < k' \iff \frac{h_{\theta_0}(x)}{h_\theta(x)} > \frac{1}{k'} \iff H_{\theta,\theta_0}(T(x)) > \frac{1}{k'} = H_{\theta,\theta_0}(k^*) \iff T(x) > k^*,$$

wegen der Monotonie von H_{θ,θ_0}. Ähnliche Äquivalenzen gelten für „>“ und „=“, was die Behauptung (10.6.2) beweist. Wir betrachten nun

$$\psi^*(x) := 1 - \varphi^*(x) = \begin{cases} 0, & \text{falls } \frac{h_\theta(x)}{h_{\theta_0}(x)} < k', \\ 1, & \text{falls } \frac{h_\theta(x)}{h_{\theta_0}(x)} > k', \\ 1-\gamma^*, & \text{falls } \frac{h_\theta(x)}{h_{\theta_0}(x)} = k', \end{cases}$$

Somit ist ψ^* ein LQ-Test für $\tilde{H}_0 := \{\theta_0\}$ gegen $\tilde{H}_1 := \{\theta\}$ zum Niveau

$$\mathbb{E}_{\theta_0}\psi^*(X) = 1 - \mathbb{E}_{\theta_0}\varphi^*(X) = 1 - \alpha.$$

Aus dem ersten Teil des Neyman-Pearson-Lemmas folgt, dass ψ^* gleichmäßig bester Test zum Niveau $1-\alpha$ ist, d.h.

$$\mathbb{E}_\theta\psi^*(X) = \sup_{\psi\colon \mathbb{E}_{\theta_0}\psi(X)=1-\alpha} \mathbb{E}_\theta\psi(X).$$

Somit gilt für φ^*, dass

$$\begin{aligned}\mathbb{E}_\theta\varphi^*(X) = 1 - \mathbb{E}_\theta\psi^*(X) &= 1 - \sup_{\psi\colon \mathbb{E}_{\theta_0}\psi(X)=1-\alpha} \mathbb{E}_\theta\psi(X) \\ &= \inf_{\psi\colon \mathbb{E}_{\theta_0}\psi(X)=1-\alpha} \mathbb{E}_\theta[1-\psi(X)] = \inf_{\varphi\colon \mathbb{E}_{\theta_0}\varphi(X)=\alpha} \mathbb{E}_\theta\varphi(X),\end{aligned}$$

wobei wir beim letzten Übergang $\varphi := 1 - \psi$ gesetzt haben. Nun ist aber $\varphi(X) := \alpha$ ein gültiger Test mit $\mathbb{E}_{\theta_0}\varphi(X) = \mathbb{E}_\theta\varphi(X) = \alpha$, somit ergibt sich

$$\mathbb{E}_\theta\varphi^*(X) = \inf_{\varphi\colon \mathbb{E}_{\theta_0}\varphi(X)=\alpha} \mathbb{E}_\theta\varphi(X) \leq \alpha,$$

was die Behauptung beweist. □

Beispiel 10.6.5 (Einseitiger Binomial-Test ist gleichmäßig bester Test)**.** Ein alterprobtes Medikament führe zu einer Besserung mit bekannter Wahrscheinlichkeit $\theta_0 \in (0,1)$. Ein neues Medikament wurde in n Fällen erprobt und führte zu einer Besserung in $x \in \{0,\dots,n\}$ Fällen. Wir betrachten die Hypothesen

$$H_0 : \text{neues Medikament ist nicht besser als das alte,}$$
$$H_1 : \text{neues Medikament ist besser.}$$

Das folgende statistische Modell erscheint natürlich: Die Beobachtung $X \sim \mathrm{Bin}(n,\theta)$ ist binomialverteilt mit einem unbekanntem $\theta \in (0,1)$, der Stichprobenraum ist $\mathfrak{X} = \{0,\dots,n\}$. Dann lauten unsere Hypothesen $H_0 : \theta \leq \theta_0$ und $H_1 : \theta > \theta_0$. Die Bedingung der monotonen Dichtequotienten gilt mit $T(x) = x$, wie in Beispiel 10.6.2 gezeigt wurde. Somit hat der gleichmäßig bester Test die Form

$$\varphi(x) = \begin{cases} 1, & \text{falls } x > k^*, \\ 0, & \text{falls } x < k^*, \\ \gamma^*, & \text{falls } x = k^* \end{cases}$$

für $x \in \{0,1\dots,n\}$. Es bleibt nur noch, die Werte k^* und γ^* so zu wählen, dass die Wahrscheinlichkeit H_0 unter $\theta = \theta_0$ irrtümlich zu verwerfen gleich α wird, d.h.

$$\mathbb{E}_{\theta_0}\varphi(X) = \mathbb{P}_{\theta_0}[T > k^*] + \gamma^*\mathbb{P}_{\theta_0}[T = k^*] = \alpha.$$

Die Parameter k^* und γ^* können nun genauso wie in Beispiel 10.5.4 bestimmt werden, nämlich

$$k^* = \min\left\{k \in \{0,\dots,n\} : \mathbb{P}_{\theta_0}[X > k] \overset{\text{def}}{=} \sum_{i=k}^{n} \binom{n}{i}\theta_0^i(1-\theta_0)^{n-i} \leq \alpha\right\},$$
$$\gamma^* = \frac{\alpha - \mathbb{P}_{\theta_0}[X > k^*]}{\mathbb{P}_{\theta_0}[X = k^*]}.$$

Beispiel 10.6.6 (Einseitiger Gauß-z-Test ist gleichmäßig bester Test)**.** Betrachte unabhängige Beobachtungen $X_1,\dots,X_n \sim \mathrm{N}(\mu,\sigma_0^2)$, wobei $\sigma_0^2 > 0$ bekannt sei. Wir interessieren uns für die Hypothesen

$$H_0 : \mu \leq \mu_0 \quad \text{und} \quad H_1 : \mu > \mu_0,$$

wobei μ_0 vorgegeben sei. Der einseitige Gauß-z-Test basiert auf der Teststatistik

$$T := \sqrt{n}\frac{\bar{X}_n - \mu_0}{\sigma_0},$$

die unter $\mu = \mu_0$ standardnormalverteilt ist, und verwirft H_0 falls $T > z_{1-\alpha}$. Um zu zeigen, dass dieser Test gleichmäßig bester Test zum Niveau α ist, müssen wir die Bedingung der

monotonen Dichtequetienten überprüfen. Der Stichprobenraum ist $\mathfrak{X} = \mathbb{R}^n$ und die Dichte von $\mathbb{P}_\mu$ bzgl. des n-dimensionalen Lebesgue-Maßes ist

$$h_\mu(x_1, \dots, x_n) = \left(\frac{1}{\sqrt{2\pi}\sigma_0}\right)^n \exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (x_i - \mu)^2\right\}.$$

Für beliebige $\mu_1 < \mu_2$ ist der Dichtequotient gegeben durch

$$\begin{aligned}\frac{h_{\mu_2}(x_1, \dots, x_n)}{h_{\mu_1}(x_1, \dots, x_n)} &= \exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^n \left((x_i - \mu_2)^2 - (x_i - \mu_1)^2\right)\right\} \\ &= \exp\left\{\frac{(\mu_2 - \mu_1)\sum_{i=1}^n x_i}{\sigma_0^2} + \frac{n}{2\sigma_0^2}(\mu_1^2 - \mu_2^2)\right\} \\ &= \exp\left\{\frac{\mu_2 - \mu_1}{\sigma_0^2} n\left(\frac{T\sigma_0}{\sqrt{n}} + \mu_0\right) + \frac{n}{2\sigma_0^2}(\mu_1^2 - \mu_2^2)\right\},\end{aligned}$$

wobei wir beim letzten Übergang die Identität $\sum_{i=1}^n x_i = n(\frac{T\sigma_0}{\sqrt{n}} + \mu_0)$ benutzt haben. Die rechte Seite ist eine monoton steigende Funktion von T, folglich liegen monotone Dichtequotienten vor.

Aufgabe 10.6.7. Seien $X_1, \dots, X_n \sim \mathrm{N}(0, \sigma^2)$ unabhängig, wobei der Erwartungswert gleich 0 ist und die Varianz $\sigma^2 > 0$ unbekannt sei. Zeigen Sie, dass die Bedingung der monotonen Dichtequotienten mit $T(x_1, \dots, x_n) = \sum_{i=1}^n x_i^2$ gilt und konstruieren Sie den gleichmäßig besten Test zum Niveau α für $H_0 : \sigma^2 \leq \sigma_0^2$ gegen $H_1 : \sigma^2 > \sigma_0^2$.

Aufgabe 10.6.8. Seien $X_1, \dots, X_n \sim \mathrm{Poi}(\theta)$ unabhängig, wobei $\theta > 0$. Zeigen Sie, dass die Bedingung der monotonen Dichtequetienten mit $T(x_1, \dots, x_n) = \sum_{i=1}^n x_i$ gilt und beschreiben Sie den gleichmäßig besten Test von $H_0 : \theta \leq \theta_0$ gegen $H_1 : \theta > \theta_0$ zum Niveau α.

10.7. Verallgemeinerter Likelihood-Quotienten-Test

Für einfache Hypothesen ist der Likelihood-Quotienten-Test ein gleichmäßig bester Test. Für nicht-einfache Hypothesen lässt sich der LQ-Test wie folgt verallgemeinern.

Annahme: $(\mathbb{P}_\theta)_{\theta\in\Theta}$ ist eine dominierte Familie von Wahrscheinlichkeitsmaßen auf dem Stichprobenraum $(\mathfrak{X}, \mathcal{A})$, d.h. es gibt ein σ-endliches Maß λ auf $(\mathfrak{X}, \mathcal{A})$, sodass für jedes $\theta \in \Theta$ das Wahrscheinlichkeitsmaß $\mathbb{P}_\theta$ eine Dichte h_θ bzgl. λ besitzt.

Wir betrachten die Hypothesen

$$H_0 : \theta \in \Theta_0 \quad \text{und} \quad H_1 : \theta \in \Theta_1,$$

wobei $\Theta = \Theta_0 \cup \Theta_1$ eine disjunkte Zerlegung des Parameterraumes Θ sei.

Definition 10.7.1. Die *verallgemeinerte Likelihood-Quotienten-Statistik* ist definiert durch

$$\Lambda(x) = \frac{\sup_{\theta\in\Theta_0} h_\theta(x)}{\sup_{\theta\in\Theta} h_\theta(x)} \in [0,1], \quad x \in \mathfrak{X}.$$

Kleine Werte von Λ sprechen für eine Ablehnung von H_0.

Definition 10.7.2. Der *verallgemeinerte LQ-Test* ist definiert durch

$$\varphi(x) = \begin{cases} 1, & \text{falls } \Lambda(x) \le c, \\ 0, & \text{falls } \Lambda(x) > c, \end{cases}$$

Dabei soll die Wahl von $c \in [0,1]$ sicherstellen, dass der Test Niveau α besitzt, d.h.

$$\sup_{\theta\in\Theta_0} \mathbb{E}_\theta \varphi(X) = \alpha.$$

Beispiel 10.7.3 (Zweiseitiger Student-t-Test als verallgemeinerter LQ-Test)**.** Betrachte unabhängige normalverteilte Beobachtungen $X_1, \dots, X_n \sim \mathrm{N}(\mu, \sigma^2)$. Der Stichprobenraum ist $\mathfrak{X} = \mathbb{R}^n$ und der Parameterraum ist eine Halbebene:

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

Wir betrachten die Hypothesen $H_0 : \mu = \mu_0$ und $H_1 : \mu \neq \mu_0$ mit einem vorgegebenen $\mu_0 \in \mathbb{R}$ und einem unbekanntem σ^2, d.h.

$$\Theta_0 = \{(\mu_0, \sigma^2) : \sigma^2 > 0\}, \quad \Theta_1 = \{(\mu, \sigma^2) \colon \mu \neq \mu_0, \sigma^2 > 0\}.$$

Die Dichte von $\mathbb{P}_{\mu,\sigma^2}$ bzgl. des Lebesgue-Maßes auf $\mathbb{R}^n$ ist

$$h_{\mu,\sigma^2}(x_1, \dots, x_n) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right\}.$$

Um das Supremum der Likelihood-Funktion $h_{\mu,\sigma^2}(x_1, \dots)$ über Θ bzw. Θ_0 zu bestimmen, erinnern wir uns an die Maximum-Likelihood-Schätzer

$$\operatorname*{arg\,max}_{(\mu,\sigma^2)\in\Theta} h_{\mu,\sigma^2}(x_1, \dots, x_n) = \left(\bar{x}_n, \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}_n)^2\right)$$

und

$$\operatorname*{arg\,max}_{(\mu,\sigma^2)\in\Theta_0} h_{\mu,\sigma^2}(x_1, \dots, x_n) = \left(\mu_0, \frac{1}{n}\sum_{i=1}^n (x_i - \mu_0)^2\right).$$

Einsetzen der optimalen Werte von (μ, σ^2) in die Dichte h_{μ,σ^2} ergibt

$$\sup_{(\mu,\sigma^2)\in\Theta} h_{\mu,\sigma^2}(x_1, \dots, x_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}_n)^2\right)^{-n/2} \mathrm{e}^{-n/2}.$$

bzw.

$$\sup_{(\mu,\sigma^2)\in\Theta_0} h_{\mu,\sigma^2}(x_1,\dots,x_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{1}{n}\sum_{i=1}(x_i-\mu_0)^2\right)^{-n/2} \mathrm{e}^{-n/2}.$$

Die Likelihood-Quotienten-Statistik ergibt sich somit zu

$$\Lambda(x_1,\dots,x_n) = \left(\frac{\sum_{i=1}(x_i-\bar{x}_n)^2}{\sum_{i=1}(x_i-\mu_0)^2}\right)^{n/2}.$$

Um den Zusammenhang zur Student-t-Statistiki herzustellen, benutzen wir die Steiner-Formel $\sum_{i=1}(x_i-\mu_0)^2 = \sum_{i=1}^n(x_i-\bar{x}_n)^2 + n(\bar{x}_n-\mu_0)^2$ und schreiben Λ wie folgt um:

$$\Lambda(x_1,\dots,x_n) = \left(\frac{1}{1+\frac{n(\bar{x}_n-\mu_0)^2}{\sum_{i=1}(x_i-\bar{x}_n)^2}}\right)^{n/2} = \left(\frac{1}{1+T^2/(n-1)}\right)^{n/2},$$

was eine monoton fallende Funktion von $|T|$ ist, wobei

$$T(x_1,\dots,x_n) := \sqrt{n}\,\frac{\bar{x}_n-\mu_0}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n(x_i-\bar{x}_n)^2}}$$

die Student-t-Statistik bezeichnet. Die Bedingung $\Lambda(x_1,\dots,x_n)\le c$ ist äquivalent zur Bedingung $|T(x_1,\dots,x_n)|>c'$ für ein passendes c'. Wählt man c bzw. c' so, dass beide Tests das gleiche Signifikanzniveau α besitzen, so treffen die beiden Tests die gleiche Entscheidung.

10.8. Asymptotische Tests für die Erfolgswahrscheinlichkeit bei Bernoulli-Experimenten

Manchmal ist es nicht möglich oder schwierig, einen exakten Test zum Niveau α zu konstruieren. In diesem Fall kann man versuchen, einen Test zu konstruieren, der zumindest approximativ (bei großem Stichprobenumfang n) das Niveau α erreicht. Wir werden nun die entsprechende Definition einführen. Seien $X_1, X_2, \dots$ unabhängige und identisch verteilte Zufallsvariablen mit Dichte bzw. Zähldichte h_θ, wobei $\theta\in\Theta$. Es sei außerdem eine Zerlegung des Parameterraumes Θ in zwei disjunkte Teilmengen Θ_0 und Θ_1 gegeben:

$$\Theta = \Theta_0\cup\Theta_1, \quad \Theta_0\cap\Theta_1 = \varnothing.$$

Wir wollen die Nullhypothese $H_0 : \theta\in\Theta_0$ gegen die Alternativhypothese $H_1 : \theta\in\Theta_1$ testen.

Definition 10.8.1. Eine Folge von Borel-Funktionen $\varphi_1, \varphi_2, \dots$ mit $\varphi_n : \mathbb{R}^n \to \{0,1\}$ heißt *asymptotischer Test* zum Niveau $\alpha\in(0,1)$, falls

$$\limsup_{n\to\infty}\sup_{\theta\in\Theta_0} \mathbb{P}_\theta[\varphi_n(X_1,\dots,X_n)=1] \le \alpha.$$

Dabei ist φ_n die zum Stichprobenumfang n gehörende Entscheidungsregel.

Wir werden nun asymptotische Tests für die Erfolgswahrscheinlichkeit θ bei Bernoulli-Experimenten konstruieren. Seien $\mathrm{X}_1,\dots,X_n$ unabhängige und mit Parameter $\theta\in(0,1)$ Bernoulli-verteilte Zufallsvariablen. Wir wollen verschiedene Hypothesen über den Parameter θ testen, z. B.

$\theta = \theta_0$, $\theta \geq \theta_0$ oder $\theta \leq \theta_0$. Ein natürlicher Schätzer für θ ist $\bar{X}_n$. Wir betrachten die Teststatistik

$$T_n := \sqrt{n}\,\frac{\bar{X}_n - \theta_0}{\sqrt{\theta_0(1-\theta_0)}}.$$

Unter der Hypothese $\theta = \theta_0$ gilt nach dem zentralen Grenzwertsatz

$$T_n \xrightarrow[n\to\infty]{d} \mathrm{N}(0,1).$$

Wir betrachten nun drei verschiedene Fälle.

Fall A. $H_0 : \theta = \theta_0$; $H_1 : \theta \neq \theta_0$. In diesem Fall sollte H_0 verworfen werden, wenn $|T_n|$ groß ist. Entscheidungsregel: H_0 wird verworfen, wenn $|T_n| \geq z_{1-\frac{\alpha}{2}}$.

Fall B. $H_0 : \theta \geq \theta_0$; $H_1 : \theta < \theta_0$. Die Nullhypothese H_0 sollte verworfen werden, wenn T_n klein ist. Entscheidungsregel: H_0 wird verworfen, wenn $T_n \leq z_\alpha$.

Fall C. $H_0 : \theta \leq \theta_0$; $H_1 : \theta > \theta_0$. Die Nullhypothese H_0 sollte verworfen werden, wenn T_n groß ist. Entscheidungsregel: H_0 wird verworfen, wenn $T_n \geq z_{1-\alpha}$.

Nun betrachten wir ein Zweistichprobenproblem, bei dem zwei Parameter θ_1 und θ_2 von zwei Bernoulli-verteilten Stichproben verglichen werden sollen. Wir machen folgende Annahmen:

(1) $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ sind unabhängige Zufallsvariablen.
(2) $X_1, \ldots, X_n \sim \mathrm{Bern}(\theta_1)$.
(3) $Y_1, \ldots, Y_m \sim \mathrm{Bern}(\theta_2)$.

Man kann sich z.B. zwei Gruppen von Patienten vorstellen, die sich zwei verschiedenen Therapien unterzogen haben. Die Zufallsvariablen X_i (bzw. Y_j) spiegeln den Behandlungserfolg in der ersten (bzw. zweiten) Gruppe wider. Es sollen nun Hypothesen über die Erfolgswahrscheinlichkeiten θ_1 und θ_2 getestet werden, z. B. $\theta_1 = \theta_2$, $\theta_1 \geq \theta_2$ oder $\theta_1 \leq \theta_2$. Parallel dazu kann man auch asymptotische Konfidenzintervalle für $\theta_1 - \theta_2$ konstruieren.

Ein natürlicher Schätzer für $\theta_1 - \theta_2$ ist $\bar{X}_n - \bar{Y}_m$. Für seinen Erwartungswert und Varianz gilt

$$\mathbb{E}[\bar{X}_n - \bar{Y}_m] = \theta_1 - \theta_2, \quad \mathrm{Var}[\bar{X}_n - \bar{Y}_m] = \frac{\theta_1(1-\theta_1)}{n} + \frac{\theta_2(1-\theta_2)}{m}.$$

Wir definieren uns die zentrierte und normierte Zufallsvariable

$$\tilde{T}_{n,m} = \frac{\bar{X}_n - \bar{Y}_m - (\theta_1 - \theta_2)}{\sqrt{\frac{\theta_1(1-\theta_1)}{n} + \frac{\theta_2(1-\theta_2)}{m}}}.$$

Satz 10.8.2. Seien $\theta_1, \theta_2 \in (0,1)$. Unter $\mathbb{P}_{\theta_1,\theta_2}$ gilt

$$\tilde{T}_{n,m} \xrightarrow{d} \mathrm{N}(0,1) \text{ für } n, m \to \infty.$$

Beweis. Es seien $n_1, n_2, \ldots$ und $m_1, m_2, \ldots$ zwei Folgen mit $\lim_{k\to\infty} n_k = \lim_{k\to\infty} m_k = +\infty$. Wir zeigen, dass

$$\tilde{T}_{n_k,m_k} \xrightarrow[k\to\infty]{d} \mathrm{N}(0,1).$$

Wir haben die Darstellung $\tilde{T}_{n_k,m_k} = Z_{1;k} + Z_{2;k} + \ldots + Z_{n_k+m_k;k}$, wobei

$$Z_{i;k} = \begin{cases} \frac{X_i - \theta_1}{n_k \sqrt{\frac{\theta_1(1-\theta_1)}{n_k} + \frac{\theta_2(1-\theta_2)}{m_k}}}, & \text{falls } i = 1, \ldots, n_k, \\ -\frac{Y_{i-n_k} - \theta_2}{m_k \sqrt{\frac{\theta_1(1-\theta_1)}{n_k} + \frac{\theta_2(1-\theta_2)}{m_k}}}, & \text{falls } i = n_k + 1, \ldots, n_k + m_k. \end{cases}$$

Wir wollen den zentralen Grenzwertsatz von Ljapunow verwenden. Es gilt:

(1) Die Zufallsvariablen $Z_{1;k}, Z_{2;k}, \ldots, Z_{n_k+m_k;k}$ sind unabhängig.
(2) $\mathbb{E}Z_{i;k} = 0$.
(3) $\sum_{i=1}^{n_k+m_k} \operatorname{Var} Z_{i;k} = 1$.

Die letzte Eigenschaft kann man wie folgt nachweisen:

$$\sum_{i=1}^{n_k+m_k} \operatorname{Var} Z_{i;k} = \frac{1}{\frac{\theta_1(1-\theta_1)}{n_k} + \frac{\theta_2(1-\theta_2)}{m_k}} \left(n_k \cdot \frac{\theta_1(1-\theta_1)}{n_k^2} + m_k \cdot \frac{\theta_2(1-\theta_2)}{m_k^2} \right) = 1.$$

Wir müssen also nur noch die Ljapunow-Bedingung überprüfen. Sei $\delta > 0$ beliebig. Es gilt

$$\begin{aligned} \sum_{i=1}^{n_k+m_k} \mathbb{E}|Z_{i;k}|^{2+\delta} &= \frac{n_k \mathbb{E}|X_1 - \theta_1|^{2+\delta}}{n_k^{2+\delta} \left(\frac{\theta_1(1-\theta_1)}{n_k} + \frac{\theta_2(1-\theta_2)}{m_k} \right)^{\frac{2+\delta}{2}}} + \frac{m_k \mathbb{E}|Y_1 - \theta_2|^{2+\delta}}{m_k^{2+\delta} \left(\frac{\theta_1(1-\theta_1)}{n_k} + \frac{\theta_2(1-\theta_2)}{m_k} \right)^{\frac{2+\delta}{2}}} \\ &\leq \frac{2^{2+\delta}}{n_k^{\delta/2} \left(\theta_1(1-\theta_1) + n_k \frac{\theta_2(1-\theta_2)}{m_k} \right)^{\frac{2+\delta}{2}}} + \frac{2^{2+\delta}}{m_k^{\delta/2} \left(m_k \frac{\theta_1(1-\theta_1)}{n_k} + \theta_2(1-\theta_2) \right)^{\frac{2+\delta}{2}}} \\ &\leq \frac{2^{2+\delta}}{n_k^{\delta/2} (\theta_1(1-\theta_1))^{\frac{2+\delta}{2}}} + \frac{2^{2+\delta}}{m_k^{\delta/2} (\theta_2(1-\theta_2))^{\frac{2+\delta}{2}}} \end{aligned}$$

was für $k \to \infty$ gegen 0 konvergiert, da $n_k, m_k \to +\infty$. Aus dem zentralen Grenzwertsatz von Ljapunow folgt nun die Behauptung des Satzes. □

Die Größe $\tilde{T}_{n,m}$ beinhaltet die unbekannten Parameter θ_1 und θ_2 sowohl im Zähler als auch im Nenner. Deshalb ersetzen wir θ_1 und θ_2 im Nenner durch die entsprechenden Schätzer $\bar{X}_n$ und $\bar{Y}_m$. Nach dem Gesetz der großen Zahlen gilt $\bar{X}_n \to \theta_1$ und $\bar{Y}_m \to \theta_2$ fast sicher für $n, m \to \infty$. Aus dem Satz von Slutsky kann man dann herleiten (Übungsaufgabe), dass

$$\frac{\bar{X}_n - \bar{Y}_m - (\theta_1 - \theta_2)}{\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n} + \frac{\bar{Y}_m(1-\bar{Y}_m)}{m}}} \xrightarrow{d} \mathrm{N}(0,1) \text{ für } n, m \to \infty.$$

Wir betrachten nun drei verschiedene Nullhypothesen.

Fall A. $H_0 : \theta_1 = \theta_2$; $H_1 : \theta_1 \neq \theta_2$. Unter H_0 gilt dann

$$T_{n,m} := \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n} + \frac{\bar{Y}_m(1-\bar{Y}_m)}{m}}} \xrightarrow{d} \mathrm{N}(0,1) \text{ für } n, m \to \infty.$$

Wir verwerfen H_0, wenn $|T_{n,m}|$ groß ist. Entscheidungsregel: H_0 wird verworfen, wenn $|T_{n,m}| \geq z_{1-\frac{\alpha}{2}}$.

Fall B. $H_0 : \theta_1 \geq \theta_2$; $H_1 : \theta_1 < \theta_2$. Die Nullhypothese H_0 sollte verworfen werden, wenn $T_{n,m}$ klein ist. Entscheidungsregel: H_0 wird verworfen, wenn $T_{n,m} \leq z_\alpha$.

Fall C. $H_0 : \theta_1 \leq \theta_2$; $H_1 : \theta_1 > \theta_2$. Die Nullhypothese H_0 sollte verworfen werden, wenn $T_{n,m}$ groß ist. Entscheidungsregel: H_0 wird verworfen, wenn $T_{n,m} \geq z_{1-\alpha}$.

Aufgabe 10.8.3 (Punktweise Konsistenz)**.** Zeigen Sie, dass für den im Fall A konstruierten Test gilt:

$$\lim_{n,m\to\infty} \mathbb{P}_{\theta_1,\theta_2}[|T_{n,m}| > z_{1-\frac{\alpha}{2}}] = 1 \text{ für alle } \theta_1, \theta_2 \in (0,1) \text{ mit } \theta_1 \neq \theta_2,$$

d.h. die Alternative wird mit einer Wahrscheinlichkeit erkannt, die gegen 1 konvergiert.

Aufgabe 10.8.4. Konstruieren Sie ein asymptotisches Konfidenzintervall für $\theta_1 - \theta_2$ zum Konfidenzniveau $1 - \alpha$.

10.9. Pearson-χ^2-Test

Beispiel 10.9.1. In seinen klassischen Versuchen hat Gregor Mendel Vererbung von Merkmalen bei Erbsen untersucht. In einem Experiment aus dem Jahre 1865 züchtete er 556 Erbsen der folgenden 4 Typen:

- 315 rund und gelb,
- 101 kantig und gelb,
- 108 rund und grün,
- 32 kantig und grün.

Theoretisch sollten diese Typen im Verhältnis $9 : 3 : 3 : 1$ stehen. Die theoretischen Erwartungswerte der 4 Typen sind $312.75, 104.25, 104.25, 34.75$, was von Mendels Werten abweicht. Sind diese Abweichungen nun so groß, dass Mendelsche Theorie verworfen werden kann? Diese Frage kann man mit dem Pearson-χ^2-Test beantworten.

Wir beginnen mit der Definition der Multinomialverteilung. Man betrachte n Bälle, die unabhängig voneinander in Behälter geworfen werden. Die Anzahl der Behälter sei d und die Wahrscheinlichkeit, dass ein Ball im Behälter $i \in \{1, \ldots, d\}$ landet, sei $p_i \geq 0$, wobei $p_1 + \ldots + p_d = 1$. Bezeichnen wir mit X_i die Anzahl der Bälle, die im Behälter i landen, so ist der Zufallsvektor $(X_1, \ldots, X_d)$ multinomialverteilt mit Parametern $(n; p_1, \ldots, p_d)$. Es gilt

$$\mathbb{P}[X_1 = x_1, \ldots, X_d = x_d] = \binom{n}{x_1, \ldots, x_d} p_1^{x_1} \ldots p_d^{x_d}$$

für alle $x_1, \ldots, x_d \in \mathbb{N}_0$ mit $x_1 + \ldots + x_d = n$. Notation: $(X_1, \ldots, X_d) \sim \text{Mult}(n; p_1, \ldots, p_d)$.

Aufgabe 10.9.2. Zeigen Sie, dass für die Marginalverteilungen $X_i \sim \text{Bin}(n, p_i)$ gilt.

Beispiel 10.9.3. Wir werfen einen fairen Würfel n Mal. Es sei X_1 die Anzahl der Einsen, X_2 die Anzahl der Zweien, usw. Dann ist $(X_1, \ldots, X_6)$ multinomialverteilt mit Parametern $(n; \frac{1}{6}, \ldots, \frac{1}{6})$.

Mit dem Pearson-χ^2-test kann man Hypothesen über die Parameter $(p_1, \ldots, p_d)$ testen. Wir betrachten eine multinomialverteilte Stichprobe

$$(X_1, \ldots, X_d) \sim \text{Mult}(n; p_1, \ldots, p_d),$$

wobei n, d bekannt und $p_1, \ldots, p_d$ unbekannt seien. Für einen vorgegebenen Wahrscheinlichkeitsvektor $(p_1^*, \ldots, p_d^*)$ betrachten wir die Hypothesen

$$H_0 : (p_1, \ldots, p_d) = (p_1^*, \ldots, p_d^*) \text{ und } H_1 : (p_1, \ldots, p_d) \neq (p_1^*, \ldots, p_d^*).$$

Wir stellen das dazugehörige statistische Modell auf. Der Stichprobenraum ist die endliche Menge

$$\mathfrak{X} = \{(x_1, \ldots, x_d) \in \mathbb{N}_0^d \colon x_1 + \ldots + x_d = n\}.$$

Der Parameterraum ist ein Simplex

$$\Theta = \{(p_1, \ldots, p_d) : p_1, \ldots, p_d \geq 0,\ p_1 + \ldots + p_d = 1\}.$$

Für ein $(p_1, \ldots, p_d) \in \Theta$ ist $\mathbb{P}_{p_1,\ldots,p_d}$ ein Wahrscheinlichkeitsmaß auf $\mathfrak{X}$ mit

$$\mathbb{P}_{p_1,\ldots,p_d}[A] = \sum_{(x_1,\ldots,x_d)\in A} \binom{n}{x_1, \ldots, x_d} p_1^{x_1} \ldots p_d^{x_d}, \quad A \subset \mathfrak{X}.$$

Um die oben formulierte Hypothese H_0 zu testen, werden wir die quadratischen Abweichungen der beobachteten Werte $x_1, \ldots, x_d$ von den erwarteten Werten $np_1^*, \ldots, np_d^*$ mit speziellen Gewichten summieren.

Definition 10.9.4. Die *Pearson-Statistik* ist definiert durch

$$T_n(x_1, \ldots, x_d) = \sum_{i=1}^{d} \frac{(x_i - np_i^*)^2}{np_i^*}.$$

Der Pearson-χ^2-Test verwirft H_0, wenn T_n größer als ein kritischer Wert ist. Um den kritischen Wert zu bestimmen, müssen wir die Verteilung von T_n unter der Nullhypothese kennen.

Satz 10.9.5 (Karl Pearson, 1900)**.** Unter H_0 gilt $T_n \xrightarrow[n\to\infty]{d} \chi^2_{d-1}$.

Also verwerfen wir H_0, wenn $T_n > \chi^2_{d-1,1-\alpha}$. Dies ist ein asymptotischer Test zum Niveau α, denn nach dem Satz von Pearson gilt für die Wahrscheinlichkeit eines Fehlers 1. Art

$$\lim_{n\to\infty} \mathbb{P}_{H_0}[T_n > \chi^2_{d-1,1-\alpha}] = \alpha.$$

Bemerkung 10.9.6. Man beachte, dass die Anzahl der Freiheitsgrade der χ^2-Verteilung gleich $d-1$ und nicht d ist. Ein Freiheitsgrad ist durch die Relation $\sum_{i=1}^{d}(x_i - np_i^*) = n - n = 0$ verlorengegangen.

Beweis von Satz 10.9.5. Seien $\xi_1, \xi_2, \ldots$ unabhängige identisch verteilte Zufallsvariablen mit Werten in der Menge $\{1, \ldots, d\}$ und Wahrscheinlichkeiten

$$\mathbb{P}[\xi_k = 1] = p_1^*, \quad \ldots, \quad \mathbb{P}[\xi_k = d] = p_d^*, \quad k \in \mathbb{N}.$$

Wir können ξ_k als die Nummer des Behälters interpretieren, in dem der k-te Ball landet. Somit ist der Vektor $(X_1, \ldots, X_d)$ mit

$$X_1 = \sum_{k=1}^{n} \mathbb{1}_{\{\xi_k = 1\}}, \quad \ldots, \quad X_d = \sum_{k=1}^{n} \mathbb{1}_{\{\xi_k = d\}}.$$

multinomialverteilt mit Parametern $(n; p_1^*, \ldots, p_d^*)$. Wir wollen zeigen, dass

$$T_n(X_1, \ldots, X_d) \xrightarrow[n\to\infty]{d} \chi^2_{d-1}.$$

Schritt 1: Kovarianzmatrix. Definitionsgemäß gilt

$$(X_1, \ldots, X_d) = \sum_{k=1}^{n} \left(\mathbb{1}_{\{\xi_k = 1\}}, \ldots, \mathbb{1}_{\{\xi_k = d\}} \right).$$

Also ist $(X_1, \ldots, X_d)$ eine Summe von n unabhängigen identisch verteilten Zufallsvektoren. Um auf diese Summe den multidimensionalen zentralen Grenzwertsatz anzuwenden, müssen wir den Erwartungswert und die Kovarianzmatrix der Summanden ausrechnen. Für alle $i \in \{1, \ldots, d\}$ gilt

$$\mathbb{E}\mathbb{1}_{\{\xi_1 = i\}} = \mathbb{P}[\xi_1 = i] = p_i^*.$$

Außerdem gilt für alle $i, j \in \{1, \ldots, d\}$, dass

$$\begin{aligned} \operatorname{Cov}\left(\mathbb{1}_{\{\xi_1 = i\}}, \mathbb{1}_{\{\xi_1 = j\}} \right) &= \mathbb{P}[\xi_1 = i, \xi_1 = j] - \mathbb{P}[\xi_1 = i]\mathbb{P}[\xi_1 = j] \\ &= \begin{cases} -p_i^* p_j^*, & \text{falls } i \neq j, \\ p_i^*(1 - p_i^*), & \text{falls } i = j. \end{cases} \end{aligned}$$

wobei wir beim letzten Übergang benutzt haben, dass die Ereignisse $\{\xi_1 = i\}$ und $\{\xi_1 = j\}$ für $i \neq j$ disjunkt sind.

Schritt 2: Zentraler Grenzwertsatz. Mit dem $(d-1)$-dimensionalen zentralen Grenzwertsatz ergibt sich

$$U_n := \left(\frac{X_1 - np_1^*}{\sqrt{n}}, \ldots, \frac{X_{d-1} - np_{d-1}^*}{\sqrt{n}} \right) \xrightarrow[n\to\infty]{d} (Z_1, \ldots, Z_{d-1}),$$

wobei $Z = (Z_1, \ldots, Z_{d-1})$ ein $(d-1)$-dimensionaler Gauß-verteilter Zufallsvektor ist mit $\mathbb{E}Z_1 = \ldots = \mathbb{E}Z_{d-1} = 0$ und der Kovarianzmatrix

$$r_{ij} := \operatorname{Cov}(Z_i, Z_j) = \begin{cases} -p_i^* p_j^*, & \text{falls } i \neq j, \\ p_i^*(1 - p_i^*), & \text{falls } i = j. \end{cases}$$

Es sei bemerkt, dass wir die letzte Koordinate weggelassen haben, denn sonst würden wir wegen der Relation $X_1 + \ldots + X_d = n$ im Grenzwert einen Gauß-Vektor $(Z_1, \ldots, Z_d)$ mit

$Z_1 + \ldots + Z_d = 0$ bekommen. Dieser Vektor ist degeneriert, wir müssen aber im nächsten Schritt die Kovarianzmatrix invertieren.

Schritt 3: Normierung auf die Standardnormalverteilung. Das Inverse der Kovarianzmatrix $\Sigma = (r_{ij})_{1\le i,j\le d-1}$ berechnet sich zu (Übung)

$$\Sigma^{-1} = (s_{ij})_{1\le i,j\le d-1} \quad \text{mit} \quad s_{ij} = \begin{cases} 1/p_d^*, & \text{falls } i \neq j, \\ 1/p_d^* + 1/p_j^*, & \text{falls } i = j. \end{cases}$$

Der Zufallsvektor $\Sigma^{-1/2} Z$ ist standardnormalverteilt auf $\mathbb{R}^{d-1}$. Mit dem Satz von der stetigen Abbildung ergibt sich, dass

$$\Sigma^{-1/2} U_n = \Sigma^{-1/2} \left(\frac{X_1 - np_1^*}{\sqrt{n}}, \ldots, \frac{X_{d-1} - np_{d-1}^*}{\sqrt{n}} \right) \xrightarrow[n\to\infty]{d} (N_1, \ldots, N_{d-1}),$$

wobei $(N_1, \ldots, N_{d-1})$ standardnormalverteilt auf $\mathbb{R}^{d-1}$ ist. Durch nochmalige Anwendung desselben Satzes folgt

$$\left(\Sigma^{-1/2} U_n\right)^\top \left(\Sigma^{-1/2} U_n\right) \xrightarrow[n\to\infty]{d} N_1^2 + \ldots + N_{d-1}^2 \sim \chi^2_{d-1}.$$

Auf der anderen Seite gilt

$$\begin{aligned}
\left(\Sigma^{-1/2} U_n\right)^\top \left(\Sigma^{-1/2} U_n\right) &= U_n^\top \Sigma^{-1} U_n \\
&= \sum_{i=1}^{d-1} \sum_{j=1}^{d-1} s_{ij} \left(\frac{X_i - np_i^*}{\sqrt{n}} \right) \left(\frac{X_j - np_j^*}{\sqrt{n}} \right) \\
&= \sum_{j=1}^{d-1} \frac{1}{p_j^*} \frac{(X_j - np_j^*)^2}{n} + \sum_{i=1}^{d-1} \sum_{j=1}^{d-1} \frac{1}{p_d^*} \frac{(X_i - np_i^*)(X_j - np_j^*)}{n} \\
&= \sum_{j=1}^{d-1} \frac{(X_j - np_j^*)^2}{np_j^*} + \frac{1}{np_d^*} \left(\sum_{i=1}^{d-1} (X_i - np_i^*) \right)^2 \\
&= \sum_{j=1}^{d} \frac{(X_j - np_j^*)^2}{np_j^*} \\
&= T_n(X_1, \ldots, X_d),
\end{aligned}$$

wobei wir beim letzten Übergang die Relation $\sum_{i=1}^d (X_i - np_i^*) = 0$ benutzt haben. Fasst man alles zusammen, so ergibt sich die zu beweisende Aussage

$$T_n(X_1, \ldots, X_d) = \left(\Sigma^{-1/2} U_n\right)^\top \left(\Sigma^{-1/2} U_n\right) \xrightarrow[n\to\infty]{d} \chi^2_{d-1}.$$

□

Aufgabe 10.9.7. Zeigen Sie, dass $T_n(x_1, \ldots, x_d) = \sum_{i=1}^d \frac{x_i^2}{np_i^*} - n$.

Der nächste Satz besagt, dass der Pearson-Test jeden festen Parameterwert aus der Alternativhypothese mit einer für $n \to \infty$ gegen 1 konvergierenden Wahrscheinlichkeit erkennt.

Satz 10.9.8 (Punktweise Konsistenz des Pearson-χ^2-Tests)**.** Für eine Stichprobe

$$(X_1, \ldots, X_d) \sim \text{Mult}(n; p_1, \ldots, p_d)$$

betrachten wir die Hypothesen

$$H_0 : (p_1, \ldots, p_d) = (p_1^*, \ldots, p_d^*), \quad \tilde{H}_1 : (p_1, \ldots, p_d) = (p_1', \ldots, p_d'),$$

wobei $(p_1^*, \ldots, p_d^*) \neq (p_1', \ldots, p_d')$ vorgegeben seien. Dann gilt

$$\lim_{n\to\infty} \mathbb{P}_{\tilde{H}_1}[T_n(X_1, \ldots, X_d) > \chi^2_{d-1,1-\alpha}] = 1.$$

Beweis. Nach Voraussetzung gibt es mindestens ein $i \in \{1, \ldots, d\}$ mit $p_i' \neq p_i^*$. Unter $\tilde{H}_1$ gilt nach dem starken Gesetz der großen Zahlen

$$\frac{X_i}{n} = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{\{\xi_k = i\}} \xrightarrow[n\to\infty]{f.s.} p_i' \neq p_i^*.$$

Es folgt, dass unter $\tilde{H}_1$

$$T_n = T_n(X_1, \ldots, X_d) \geq \frac{(X_i - np_i^*)^2}{np_i^*} = \frac{n}{p_i^*} \left(\frac{X_i}{n} - p_i^* \right)^2 \xrightarrow[n\to\infty]{f.s.} +\infty.$$

Daraus folgt, dass T_n gegen $+\infty$ auch in Wahrscheinlichkeit konvergiert, d.h. $\lim_{n\to\infty} \mathbb{P}_{\tilde{H}_1}[T_n > c] = 1$ für jedes feste $c \in \mathbb{R}$. □

Beispiel 10.9.9. Für den oben beschriebenen Versuch von Mendel mit $n = 556$ Erbsen lautet die Nullhypothese

$$(p_1^*, p_2^*, p_3^*, p_4^*) = \left(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16} \right).$$

Die von Mendel beobachteten Werte sind

$$(x_1, x_2, x_3, x_4) = (315, 101, 108, 32).$$

Der Wert der Pearson-Statistik berechnen sich zu $T_n(x_1, x_2, x_3, x_4) = 0.47$. Nach dem Satz von Pearson sollte T_n unter der Nullhypothese approximativ χ^2_3-verteilt sein. Das 0.95-Quantil der χ^2_3-Verteilung ist laut Tabelle $\chi^2_{3,0.95} = 7.81$, was viel größer als der beobachtete Wert von T_n ist. Also kann die Nullhypothese nicht verworfen werden. Die Daten zeigen keine signifikante Abweichung von der Mendelschen Theorie.

Der p-Wert ist die Wahrscheinlichkeit, dass bei einer unabhängigen Wiederholung des Experiments ein Wert der Pearson-Statistik T_n beobachtet wird, der ≥ 0.47 ist, und beträgt in unserem Fall 0.92. Von 10 Biologen, die das Experiment von Mendel unabhängig wiederholen, würde im Durchschnitt nur einer eine bessere Übereinstimmung mit der Theorie beobachten, als Mendel. Auch in anderen Experimenten von Mendel war die Übereinstimmung mit den theoretischen Werten „zu gut“. Aus diesem Grund warf Fisher in einer Arbeit aus dem Jahre 1936 Mendel vor, seine Ergebnisse beschönigt zu haben, was zur sogenannten Mendel-Fisher-Kontroverse führte. An dieser Stelle verweisen wir auf das Buch von A. Franklin, A. W. F. Edwards, D. J. Fairbanks, D. L. Hartl, and T. Seidenfeld, „*Ending the Mendel-Fisher Controversy*“, Univ. Pittsburgh Press, 2008.

Beispiel 10.9.10. Man kann sich fragen, ob in der Dezimaldarstellung der Zahl

$$\pi = 3.141592653589793238\ldots$$

alle 10 Ziffern ungefähr gleich oft vorkommen. Unter den ersten $n = 10^7$ Dezimalstellen von π finden sich so viele verschiedene Ziffern:

x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
999.440	999.333	1.000.306	999.965	1.001.093	1.000.466	999.337	1.000.206	999.814	1.000.040

Wir möchten nun die Hypothese, dass jede Ziffer mit einer Häufigkeit von 1/10 vorkommt, testen. Die Pearson-Statistik berechnet sich zu

$$T_n(x_0, \ldots, x_9) = 2.7834.$$

Nach dem Satz von Pearson sollte T_n unter der Nullhypothese approximativ χ^2_9-verteilt sein. Das 0.95-Quantil der χ^2_9-Verteilung ist laut Tabelle $\chi^2_{9,0.95} = 16.92$, was viel größer als der beobachtete Wert von T_n ist. Somit können wir die Nullhypothese nicht verwerfen.

Wir berechnen noch den p-Wert des Pearson-Tests. Dieser ist definiert als die Wahrscheinlichkeit, dass eine χ^2_9-verteilte Zufallsvariable $\geq$ als der beobachtete Wert 2.7834 ist. Mit der entsprechenden Software erhält man einen p-Wert von 0.9723, was auffallend hoch ist!

Wie kann man nun diesen extrem kleinen Wert der Pearson-Statistik (bzw. den extrem hohen p-Wert) interpretieren? Die Abweichungen der beobachteten Werte $x_0, \ldots, x_9$ (s. die obige Tabelle) von dem erwarteten Wert 10^6 sind viel kleiner, als das, was man bei einer Folge von u.i.v. Zufallsvariablen mit Gleichverteilung auf $\{0, 1, \ldots, 9\}$ erwarten würde. Der kleine Wert der Pearson-Statistik ist ein Hinweis darauf, dass die von uns untersuchte Folge von Ziffern nicht zufällig ist.

Aufgabe 10.9.11. Sei $(X_1, \ldots, X_d) \sim \text{Mult}(n; p_1, \ldots, p_d)$. Zeigen Sie, dass der Maximum-Likelihood-Schätzer für $(p_1, \ldots, p_d)$ durch $(X_1/n, \ldots, X_d/n)$ gegeben ist.

10.10. Exakter Test nach Fisher

Betrachten wir eine Studie, in der die Wirksamkeit eines Medikaments mit der Wirksamkeit eines Placebos verglichen werden soll. Die für die Studie $n = 200$ ausgewählten Patienten werden zufällig in zwei Gruppen mit $n_1 = 100$ und $n_2 = 100$ Personen eingeteilt. Die Einteilung in die Gruppen sollte per Zufallsgenerator geschehen (solche Studien heißen *randomisiert*) um die gleichmäßige Verteilung der bekannten und unbekannten Einflussfaktoren auf die beiden Gruppen sicherzustellen. Die erste Gruppe (Behandlungsgruppe) wird mit dem Medikament behandelt, die zweite Gruppe (Kontrollgruppe) bekommt ein Placebo. Dabei sollten die Patienten nicht wissen, ob sie ein Medikament oder ein Placebo bekommen (solche Studien heißen *einfachblind*). Am besten sollten auch die Ärzte nicht wissen, welche Patienten in welche Gruppe eingeteilt sind (solche Studien heißen *doppelblind*). Die Ergebnisse der Studie fassen wir in einer sogenannten Kontingenztabelle zusammen, die z.B. wie folgt aussehen könnte:

	Erfolg	Misserfolg	Summe
Medikament	40	60	$n_1 = 100$
Placebo	35	65	$n_2 = 100$
Summe	$x = 75$	125	$n = 200$

Man sieht, dass die Erfolgsquote in der Behandlungsgruppe höher ist, als in der Placebogruppe: $40/100 > 35/100$. Aber ist dieser Unterschied signifikant genug, um auf die Wirksamkeit des Medikaments zu schließen? Selbst wenn das Medikament genauso wirken würde, wie das Placebo, wäre die Wahrscheinlichkeit, in der Behandlungsgruppe eine bessere Quote zu erzielen, ungefähr $1/2$.

Diese Frage lässt sich mit dem exakten Test nach Fisher beantworten. Für die Anzahl der Erfolge in der ersten bzw. zweiten Gruppe gilt

$$X_1 \sim \mathrm{Bin}(n_1, p_1), \quad X_2 \sim \mathrm{Bin}(n_2, p_2), \quad X_1 \text{ und } X_2 \text{ unabhängig},$$

wobei $p_1, p_2 \in [0, 1]$ unbekannt sind. Im obigen Experiment haben wir $X_1 = 40$ und $X_2 = 35$ beobachtet.

Wir stellen die folgenden Hypothesen auf:

- H_0: Das Medikament hat die gleiche Wirkung wie das Placebo, also $p_1 = p_2 = p$.
- H_1: Das Medikament hat eine bessere Wirkung als das Placebo, also $p_1 > p_2$.

Es sei bemerkt, dass die Hypothese, die wir nachweisen möchten, als Alternative formuliert wird. Der Nachweis der Alternative gelingt dann, wenn wir ein sehr ungewöhnliches (signifikantes) Ergebnis beobachten, das unter der Nullhypothese extrem unwahrscheinlich wäre. Als Teststatistik wollen wir X_1, also die Anzahl der Erfolge in der Behandlungsgruppe, betrachten. Bei einem „zu großen“ X_1 sollte H_0 verworfen werden.

Stellen wir uns nun vor, die Gesamtzahl der Erfolge in beiden Gruppen, also $X_1 + X_2 = x = 75$, wird festgehalten. Wie sieht dann die bedingte Verteilung von X_1 unter H_0 aus? Für beliebige $x_1, x_2 \in \{0, 1, \ldots, x\}$ mit $x_1 + x_2 = x$ erhalten wir

$$\begin{aligned}
\mathbb{P}_{H_0}[X_1 = x_1 | X_1 + X_2 = x] &= \frac{\mathbb{P}_{H_0}[X_1 = x_1, X_1 + X_2 = x]}{\mathbb{P}_{H_0}[X_1 + X_2 = x]} \\
&= \frac{\mathbb{P}_{H_0}[X_1 = x_1, X_2 = x_2]}{\mathbb{P}_{H_0}[X_1 + X_2 = x]} \\
&= \frac{\mathbb{P}_{H_0}[X_1 = x_1]\mathbb{P}_{H_0}[X_2 = x_2]}{\mathbb{P}_{H_0}[X_1 + X_2 = x]} \\
&= \frac{\binom{n_1}{x_1} p^{x_1}(1-p)^{n_1 - x_1} \binom{n_2}{x_2} p^{x_2}(1-p)^{n_2 - x_2}}{\binom{n}{x} p^x (1-p)^{n-x}} \\
&= \frac{\binom{n_1}{x_1}\binom{n_2}{x_2}}{\binom{n}{x}}.
\end{aligned}$$

Auf dieses Ergebnis kann man auch rein probabilistisch kommen. Man stellt sich die Patienten als Bälle in einer Urne vor, davon seien $n_1 = 100$ schwarz (Behandlungsgruppe) und $n_2 = 100$ weiß (Kontrollgruppe). Nun ist die Gesamtzahl der Erfolge auf $x = 75$ festgelegt.

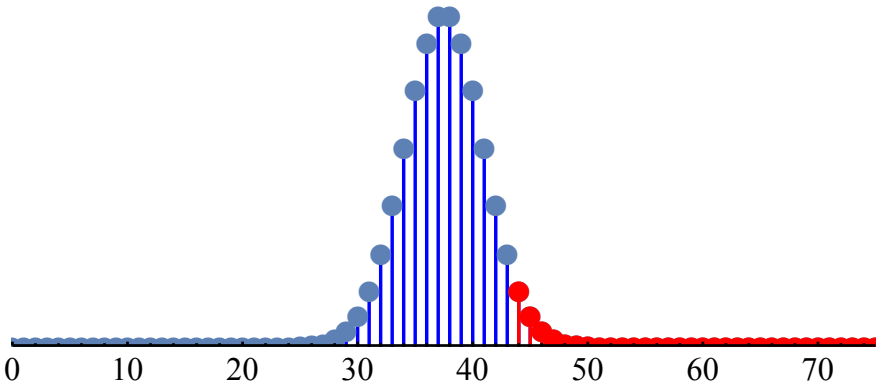


ABBILDUNG 6. Die bedingte Zähldichte von X_1 unter H_0 gegeben, dass $X_1+X_2 = 75$. Rot: Ablehnungsbereich. Blau: Annahmebereich.

Es werden also aus der Urne 75 Bälle gezogen, die den Patienten entsprechen, dessen Behandlung erfolgreich war. Da wir H_0 voraussetzen, haben alle Bälle (unabhängig von der Farbe) die gleiche Chance, gezogen zu werden. Die Zufallsvariable X_1 gibt die Anzahl der schwarzen Bälle unter den gezogenen wider und ist somit hypergeometrisch verteilt.

Die bedingte Verteilung von X_1 gegeben, dass $X_1+X_2 = 75$ und H_0 gilt, wird auf Abbildung 6 gezeigt.

Wir wollen nun H_0 verwerfen, wenn der beobachtete Wert von X_1 ungewöhnlich groß ist. Wir definieren deshalb den p-Wert des Experiments als die Wahrscheinlichkeit (unter H_0), dass bei einer unabhängigen Wiederholung der Studie, in der die gleiche Summe $x = 75$ beobachtet wird, der Wert von X_1 mindestens 40 ist, also

$$\mathbb{P}_{H_0}[X_1 \geq 40 | X_1+X_2 = 75] = \sum_{x_1=40}^{75} \mathbb{P}_{H_0}[X_1 = x_1 | X_1+X_2 = 75] = \sum_{x_1=40}^{75} \frac{\binom{100}{x_1}\binom{100}{75-x_1}}{\binom{200}{75}} \approx 0.2796.$$

Der p-Wert ist somit wesentlich größer als 0.05. Also ist das Medikament nicht signifikant besser als das Placebo.

Man kann auch einen Ablehnungsbereich konstruieren. Wir wählen $\alpha = 0.05$ als Niveau und rechnen leicht nach, dass

$$\mathbb{P}_{H_0}[X_1 \geq 43 | X_1 + X_2 = 75] \approx 0.0719, \quad \mathbb{P}_{H_0}[X_1 \geq 44 | X_1 + X_2 = 75] \approx 0.0396.$$

Also ist der Ablehnungsbereich die Menge $\{44, 45, \ldots, 75\}$. Es sei bemerkt, dass das nur unter der Annahme, dass $X_1 + X_2 = 75$, gilt. Bei einer anderen Gesamtzahl der Erfolge würde sich der Ablehnungsbereich entsprechend ändern.

10.11. Der Anpassungstest von Kolmogorow-Smirnow

Stellen wir uns vor, jemand behauptet, dass die folgende Stichprobe gleichverteilt auf dem entsprechenden Intervall sei:

Wie können wir eine solche Behauptung testen?

Seien $X_1, \dots, X_n$ unabhängige Zufallsvariablen mit einer unbekannten Verteilungsfunktion F. Wir nehmen an, dass F stetig ist und betrachten F als unbekannten Parameter, so dass

$$\Theta = \{F : F \text{ ist eine stetige Verteilungsfunktion}\}.$$

Für eine gegebene stetige Verteilungsfunktion F_0 betrachten wir die Hypothesen

$$H_0 : F = F_0, \quad H_1 : F \neq F_0.$$

Um H_0 zu testen, schätzen wir zuerst F durch die empirische Verteilungsfunktion von $X_1, \dots, X_n$:

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq t}.$$

Unterscheidet sich $\hat{F}_n(t)$ stark von F_0, so ist es ein Hinweis darauf, dass H_0 verletzt wird. Wir bilden deshalb die Kolmogorow-Smirnow-Statistik

$$D_n := \sup_{t \in \mathbb{R}} |\hat{F}_n(t) - F_0(t)|.$$

Um einen auf D_n basierenden Test zu konstruieren, benötigen wir die Verteilung von D_n unter der Nullhypothese. Der nächste Satz besagt, dass diese Verteilung nicht von F_0 abhängt.

Satz 10.11.1. Seien $X_1^*, \dots, X_n^*$ uiv Zufallsvariablen, die auf dem Intervall $[0, 1]$ gleichverteilt sind. Betrachte deren empirische Verteilungsfunktion

$$\hat{F}_n^*(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i^* \leq t}$$

und definiere den entsprechenden Kolmogorow-Smirnow-Abstand

$$D_n^* := \sup_{t \in [0,1]} |\hat{F}_n(t) - t|.$$

Dann hat D_n unter H_0 die gleiche Verteilung wie D_n^*.

Die Verteilung von D_n^* kann mit entsprechender Software numerisch berechnet werden. Sei $q_{n,1-\alpha}$ das $(1-\alpha)$-Quantil von D_n^*. Der Kolmogorow-Smirnow-Test verwirft die Nullhypothese $F = F_0$, falls $D_n > q_{n,1-\alpha}$.

Bei einem großen n kann man auch die asymptotische Version des Kolmogorow-Smirnow-Tests verwenden. Dafür benötigt man den folgenden Satz über die Verteilungskonvergenz von $\sqrt{n} D_n^*$.

Satz 10.11.2. Für alle $x \geq 0$ gilt

$$\lim_{n\to\infty} \mathbb{P}[\sqrt{n}D_n^* \leq x] = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}\mathrm{e}^{-2i^2x^2}.$$

Ohne Beweis.

Sei nun $q_{1-\alpha}$ das $(1-\alpha)$-Quantil der Verteilungsfunktion auf der rechten Seite. Der asymptotische Test von Kolmogorow-Smirnow verwirft die Nullhypothese H_0, falls $\sqrt{n}D_n > q_{1-\alpha}$.

KAPITEL 11

Einfache lineare Regression

11.1. Problemstellung

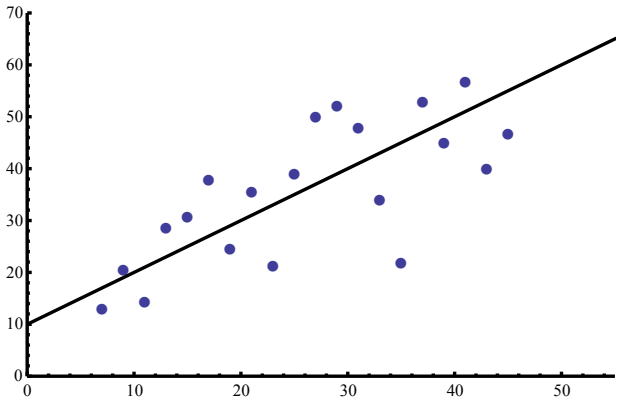
Natur- und Wirtschaftswissenschaften bieten zahlreiche Beispiele von Gesetzen, die eine lineare Abhängigkeit einer Größe y von einer anderen Größe x behaupten:

$$y = \alpha + \beta x.$$

Stellen wir uns zum Beispiel vor, dass zu den Zeitpunkten $x_1, \ldots, x_n$ die Koordinate eines Flugzeugs gemessen wird. Die gemessenen Koordinaten bezeichnen wir mit $y_1, \ldots, y_n$. Wir gehen davon aus, dass sich das Flugzeug mit konstanter Geschwindigkeit bewegt hat, also muss die Relation

$$y_i = \alpha + \beta x_i, \quad i = 1, \ldots, n, \tag{11.1.1}$$

bestehen, wobei β die Geschwindigkeit des Flugzeugs ist und α die Position des Flugzeugs zum Zeitpunkt 0 bezeichnet. Dabei heißen $x_1, \ldots, x_n$ *Ausgangsgrößen* und $y_1, \ldots, y_n$ die *Zielgrößen*.



Leider kann Relation (11.1.1) nicht exakt gelten, denn die Punkte auf dem Bild liegen nicht auf einer Geraden. Das kann z.B. daran liegen, dass die Messungen fehlerbehaftet sind. Wir müssen also das Modell verändern, indem wir die sogenannten *Störgrößen* oder *Residuen* $\varepsilon_1, \ldots, \varepsilon_n$ einführen:

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \ldots, n. \tag{11.1.2}$$

Die Störgrößen $\varepsilon_1, \ldots, \varepsilon_n$ können z.B. als Messfehler interpretiert werden.

In diesem Modell sind $(x_1, y_1), \ldots, (x_n, y_n)$ bekannt und $\alpha, \beta, \varepsilon_1, \ldots, \varepsilon_n$ unbekannt. Das Problem besteht darin, den *Regressionskoeffizienten* α und die *Regressionskonstante* β zu schätzen.

Im Folgenden werden wir drei verschiedene Methoden zur Lösung dieses Problems betrachten: *Methode der kleinsten Quadrate*, den *besten linearen erwartungstreuen Schätzer* und die *Maximum-Likelihood-Methode*.

11.2. Methode der kleinsten Quadrate (MKQ)

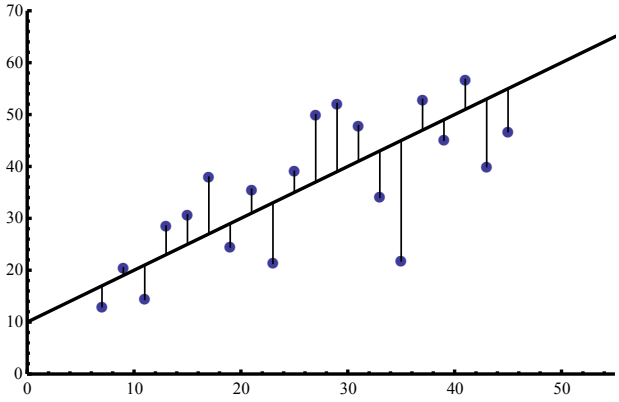
Gegeben sei eine Punktwolke $(x_1, y_1), \ldots, (x_n, y_n)$. Wir wollen diese Punktwolke durch eine Gerade der Form $y = \alpha + \beta x$ approximieren. Als ein Maß dafür, wie gut eine solche Approximation ist, benutzen wir den mittleren quadratischen Fehler.

Definition 11.2.1. Der *mittlere quadratische Fehler* ist die Funktion

$$\mathrm{MSE}(\alpha, \beta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2.$$

Dabei steht MSE für *middle square error*.

Somit ist $\mathrm{MSE}(\alpha, \beta)$ die Summe der Quadrate der vertikalen Abstände zwischen den Punkten (x_i, y_i) und der Geraden $y = \alpha + \beta x$.



Bei der Methode der kleinsten Quadrate sollen α und β so bestimmt werden, dass der mittlere quadratische Fehler minimal wird.

Definition 11.2.2. Der *Kleinste-Quadrate-Schätzer* ist definiert durch

$$(\hat{\alpha}, \hat{\beta}) = \operatorname*{arg\,min}_{(\alpha,\beta)\in\mathbb{R}^2} \mathrm{MSE}(\alpha, \beta).$$

Im Folgenden benutzen wir die Notation

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

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

Dabei ist s_{xx}^2 die empirische Varianz der Stichprobe $(x_1, \ldots, x_n)$, s_{yy}^2 ist die empirische Varianz der Stichprobe $(y_1, \ldots, y_n)$, und s_{xy}^2 ist die empirische Kovarianz der beiden Stichproben.

Im Folgenden werden wir annehmen, dass es unter den Zahlen $x_1, \ldots, x_n$ mindestens zwei verschiedene gibt. Somit ist $s_{xx}^2 \neq 0$.

Satz 11.2.3. Die Funktion $\mathrm{MSE}(\alpha, \beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$ erreicht ihr Minimum für

$$\hat{\beta} = \frac{s_{xy}^2}{s_{xx}^2} \text{ und } \hat{\alpha} = \bar{y}_n - \hat{\beta}\bar{x}_n. \tag{11.2.1}$$

Beweis. Sei zuerst $\beta \in \mathbb{R}$ fest. Wir leiten die Funktion $\mathrm{MSE}(\alpha, \beta)$ nach α ab:

$$\frac{\partial}{\partial\alpha}\mathrm{MSE}(\alpha, \beta) = -\frac{2}{n}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i) = -2\bar{y}_n + 2\alpha + 2\beta\bar{x}_n.$$

Indem wir die Ableitung gleich null setzen und nach α umformen erhalten wir

$$\alpha = \bar{y}_n - \beta\bar{x}_n.$$

Nun sei β wieder variabel. Wir setzen nun $\alpha = \bar{y}_n - \beta\bar{x}_n$ in die Funktion $\mathrm{MSE}(\alpha, \beta)$ ein:

$$\begin{aligned}
\mathrm{MSE}(\bar{y}_n - \beta\bar{x}_n, \beta) &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta x_i - (\bar{y}_n - \beta\bar{x}_n))^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}((y_i - \bar{y}_n) - \beta(x_i - \bar{x}_n))^2 \\
&= \frac{n-1}{n}(s_{yy}^2 - 2\beta s_{xy}^2 + \beta^2 s_{xx}^2).
\end{aligned}$$

Wir minimieren diese Funktion nach β. Zu diesem Zweck leiten wir die Funktion nach β ab:

$$\frac{\partial}{\partial\beta}\mathrm{MSE}(\bar{y}_n - \beta\bar{x}_n, \beta) = \frac{n-1}{n}(-2s_{xy}^2 + 2\beta s_{xx}^2).$$

Setzen wir die Ableitung gleich null, so erhalten wir

$$\beta = \frac{s_{xy}^2}{s_{xx}^2}.$$

□

Bemerkung 11.2.4. Die zweite Gleichung in (11.2.1) besagt, dass der Punkt $(\bar{x}_n, \bar{y}_n)$ auf der Regressionsgeraden $y = \hat{\alpha} + \hat{\beta}x$ liegt.

Bemerkung 11.2.5. Um eine einfache Interpretation der Formeln (11.2.1) zu geben, stellen wir uns vor, dass X und Y zwei Zufallsvariablen mit $Y = \alpha + \beta X$ sind. Dann gilt

$$\operatorname{Cov}(X, Y) = \operatorname{Cov}(X, \alpha + \beta X) = \beta \operatorname{Var} X.$$

Somit können wir α und β wie folgt darstellen:

$$\beta = \frac{\operatorname{Cov}(X, Y)}{\operatorname{Var} X}, \quad \alpha = \mathbb{E}Y - \beta \mathbb{E}X. \tag{11.2.2}$$

Formel (11.2.1) ist eine „empirische“ Version von (11.2.2) in der die Kovarianz, die Varianz und die Erwartungswerte durch die entsprechenden Schätzer ersetzt wurden.

Aufgabe 11.2.6. Betrachten Sie das Modell $y_i = \alpha + \varepsilon_i$, $i = 1, \ldots, n$, und schätzen Sie α mit der Methode der kleinsten Quadrate.

Aufgabe 11.2.7. Betrachten Sie das Modell $y_i = \beta x_i + \varepsilon_i$, $i = 1, \ldots, n$, und schätzen Sie β mit der Methode der kleinsten Quadrate.

Aufgabe 11.2.8. Der mittlere quadratische Fehler ist als Summe der *vertikalen* quadratischen Abstände zwischen den Punkten (x_i, y_i) und der Geraden $y = \alpha + \beta x$ definiert. Wie ändert sich der Schätzer, wenn man stattdessen die Summe der *horizontalen* quadratischen Abstände betrachtet?

11.3. Bester linearer erwartungstreuer Schätzer

In diesem Abschnitt betrachten wir das folgende *Gauß-Markov*-Modell:

(1) Die Ausgangsgrößen $x_1, \ldots, x_n$ sind deterministisch und nicht alle gleich.
(2) Die Zielgrößen $y_1, \ldots, y_n$ sind Realisierungen der Zufallsvariablen $Y_1, \ldots, Y_n$ mit
$$Y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \ldots, n.$$
(3) Für die Störgrößen $\varepsilon_1, \ldots, \varepsilon_n$ gilt:
$$\mathbb{E}\varepsilon_i = 0, \quad \operatorname{Var} \varepsilon_i = \sigma^2, \quad \mathbb{E}[\varepsilon_i \varepsilon_j] = 0, \quad i \neq j.$$

Bedingung (3) bedeutet, dass keine systematischen Fehler vorliegen, die Fehler unkorreliert sind und die gleiche Varianz besitzen. Für die Zufallsvariablen $Y_1, \ldots, Y_n$ ergibt sich, dass

$$\mathbb{E}Y_i = \alpha + \beta x_i, \quad \operatorname{Var} Y_i = \sigma^2, \quad \mathbb{E}[Y_i Y_j] = 0 \text{ für } i \neq j.$$

Das Ziel ist es, die unbekannten Parameter α, β und σ^2 zu schätzen.

Definition 11.3.1. Ein *linearer Schätzer* ist ein Schätzer der Form

$$f_1 Y_1 + \ldots + f_n Y_n,$$

wobei $f_1, \ldots, f_n \in \mathbb{R}$.

Die Koeffizienten $f_1, \ldots, f_n$ dürfen von $x_1, \ldots, x_n$ aber nicht von den unbekannten Parametern α, β, σ^2 abhängen.

Lineare Schätzer bilden einen n-dimensionalen Vektorraum V mit Basis $Y_1, \ldots, Y_n$. Wir bezeichnen diesen Vektorraum mit

$$V := \{f_1 Y_1 + \ldots + f_n Y_n \colon f_1, \ldots, f_n \in \mathbb{R}\}.$$

Wir definieren auf V das Skalarprodukt

$$\langle f_1 Y_1 + \ldots + f_n Y_n, g_1 Y_1 + \ldots + g_n Y_n \rangle := f_1 g_1 + \ldots + f_n g_n.$$

Die Basis $Y_1, \ldots, Y_n$ ist orthonormal. Da die Zufallsvariablen $Y_1, \ldots, Y_n$ nach den Annahmen des Gauß-Markov-Modells unkorreliert sind, gilt

(11.3.1)

$$\begin{aligned}\operatorname{Cov}(f_1 Y_1 + \ldots + f_n Y_n, g_1 Y_1 + \ldots + g_n Y_n) &= \sigma^2 \sum_{i=1}^n f_i g_i \\ &= \sigma^2 \langle f_1 Y_1 + \ldots + f_n Y_n, g_1 Y_1 + \ldots + g_n Y_n \rangle.\end{aligned}$$

Somit stimmt das Skalarprodukt bis auf den Faktor σ^2 mit der Kovarianz überein. Insbesondere gilt

$$\operatorname{Var}(f_1 Y_1 + \ldots + f_n Y_n) = \sigma^2 \sum_{i=1}^n f_i^2, \tag{11.3.2}$$

also stimmt die Varianz (bis auf den Faktor σ^2) mit der quadrierten Länge des Vektors überein. Wir werden Formeln (11.3.1) und (11.3.2) sehr oft benutzen, um Kovarianzen und Varianzen von linearen Schätzern zu berechnen.

Satz 11.3.2. Ein linearer Schätzer $\hat{\beta} = d_1 Y_1 + \ldots + d_n Y_n$ ist genau dann erwartungstreu für β, wenn

$$\sum_{i=1}^n d_i = 0 \text{ und } \sum_{i=1}^n d_i x_i = 1. \tag{11.3.3}$$

Beweis. Wir berechnen den Erwartungswert von $\hat{\beta}$:

$$\mathbb{E}_{\alpha,\beta,\sigma^2} \hat{\beta} = \mathbb{E}_{\alpha,\beta,\sigma^2} \left[\sum_{i=1}^n d_i Y_i \right] = \sum_{i=1}^n d_i (\alpha + \beta x_i) = \alpha \sum_{i=1}^n d_i + \beta \sum_{i=1}^n d_i x_i.$$

Damit $\hat{\beta}$ erwartungstreu ist, muss dieser Ausdruck gleich β für alle α, β sein. Dies ist genau dann der Fall, wenn $\sum_{i=1}^n d_i = 0$ und $\sum_{i=1}^n d_i x_i = 1$. □

Definition 11.3.3. Ein linearer erwartungstreuer Schätzer $\hat{\beta} = d_1Y_1 + \ldots + d_nY_n$ heißt *bester linearer erwartungstreuer Schätzer* für β, falls für jeden anderen linearen erwartungstreuen Schätzer $\tilde{\beta} = \tilde{d}_1Y_1 + \ldots + \tilde{d}_nY_n$ gilt, dass

$$\mathrm{Var}_{\alpha,\beta,\sigma^2} \hat{\beta} \leq \mathrm{Var}_{\alpha,\beta,\sigma^2} \tilde{\beta} \text{ für alle } \alpha, \beta \in \mathbb{R}, \sigma^2 \geq 0.$$

Abkürzung: BLUE (*best linear unbiased estimator*).

Satz 11.3.4. Ein Schätzer $\hat{\beta} = d_1Y_1 + \ldots + d_nY_n$ ist BLUE für β genau dann, wenn

$$d_i = \frac{x_i - \bar{x}_n}{(n-1)s_{xx}^2}, \quad i = 1, \ldots, n. \tag{11.3.4}$$

Beweis. Für die Varianz von $\hat{\beta}$ gilt $\mathrm{Var}_{\alpha,\beta,\sigma^2} \hat{\beta} = \sigma^2 \sum_{i=1}^n d_i^2$, denn $Y_1, \ldots, Y_n$ sind unkorreliert und $\mathrm{Var}\, Y_i = \sigma^2$. Wir möchten also folgendes Optimierungsproblem lösen: Minimiere die Funktion $f := \sum_{i=1}^n d_i^2$ unter Nebenbedingungen $f_1 := \sum_{i=1}^n d_i = 0$ und $f_2 := \sum_{i=1}^n d_i x_i = 1$. Mit der Methode der Lagrange-Multiplikatoren ergeben sich die Gleichungen:

$$\mathrm{grad}(f - \lambda_1 f_1 - f_2\lambda_2) = 0, \quad f_1 = 0, \quad f_2 = 1.$$

Berechnet man den Gradienten, so erhält man

$$2d_i = \lambda_1 + \lambda_2 x_i \text{ für } i = 1, \ldots, n, \quad \sum_{i=1}^n d_i = 0, \quad \sum_{i=1}^n d_i x_i = 1. \tag{11.3.5}$$

Als Lösung von (11.3.5) ergibt sich (11.3.4) sowie $\lambda_1 = \frac{-2\bar{x}_n}{(n-1)s_{xx}^2}$ und $\lambda_2 = \frac{2}{(n-1)s_{xx}^2}$. □

Bemerkung 11.3.5. Die Gleichungen (11.3.5) haben eine transparente geometrische Bedeutung. Die Menge aller erwartungstreuen Schätzer für β ist ein $(n-2)$-dimensionaler affiner Unterraum V_β von V, der durch die Gleichungen $\sum_{i=1}^n d_i = 0$ und $\sum_{i=1}^n d_i x_i = 1$ gegeben ist. Der BLUE für β ist derjenige Punkt von V_β, der den Abstand zum Ursprung 0 minimiert. Geometrisch gesehen, ist der BLUE die orthogonale Projektion von 0 auf V_β.

Es sei V_1 der zweidimensionale Unterraum von V, der von den Schätzern $Y_1 + \ldots + Y_n$ und $x_1Y_1 + \ldots + x_nY_n$ aufgespannt ist. Es ist klar, dass der BLUE der Schnitt von V_1 und V_β ist. Die Gleichungen $2d_i = \lambda_1 + \lambda_2 x_i$, $i = 1, \ldots, n$, bedeuten, dass $\hat{\beta}$ in V_1 liegt. Die beiden anderen Gleichungen in (11.3.5) bedeuten, dass $\hat{\beta}$ in V_β liegt.

Völlig analog kann man auch α schätzen.

Satz 11.3.6. Ein linearer Schätzer $\hat{\alpha} = c_1Y_1 + \ldots + c_nY_n$ ist ein erwartungstreuer Schätzer für α genau dann, wenn

$$\sum_{i=1}^n c_i = 1 \text{ und } \sum_{i=1}^n c_i x_i = 0. \tag{11.3.6}$$

Beweis. Übungsaufgabe. □

Satz 11.3.7. Ein Schätzer $\hat{\alpha} = c_1 Y_1 + \ldots + c_n Y_n$ ist BLUE für α genau dann, wenn

$$c_i = \frac{1}{n} - \bar{x}_n d_i = \frac{1}{n} - \frac{\bar{x}_n (x_i - \bar{x}_n)}{(n-1)s_{xx}^2}.$$

Beweis. Übungsaufgabe. □

Fassen wir Sätze 11.3.4 und 11.3.7 zusammen, so erhalten wir die folgenden Formeln für BLUE von β und α:

$$\hat{\beta} = \sum_{i=1}^n d_i Y_i = \sum_{i=1}^n \frac{x_i - \bar{x}_n}{(n-1)s_{xx}^2} Y_i = \sum_{i=1}^n \frac{(x_i - \bar{x}_n)(Y_i - \bar{Y}_n)}{(n-1)s_{xx}^2} = \frac{s_{xY}^2}{s_{xx}^2},$$
$$\hat{\alpha} = \sum_{i=1}^n \left(\frac{1}{n} - \bar{x}_n d_i \right) Y_i = \bar{Y}_n - \bar{x}_n \hat{\beta}.$$

Das sind aber exakt die Kleinste-Quadrate-Schätzer.

Satz 11.3.8 (Gauß-Markov)**.** Die besten linearen erwartungstreuen Schätzer für α und β sind die Kleinste-Quadrate-Schätzer.

Der nächste Satz beschreibt die Kovarianzmatrix des Zufallsvektors $(\hat{\alpha}, \hat{\beta})$:

Satz 11.3.9. Es gilt

$$\operatorname{Var} \hat{\beta} = \frac{\sigma^2}{(n-1)s_{xx}^2}, \quad \operatorname{Var} \hat{\alpha} = \frac{1}{n} \left(\sum_{i=1}^n x_i^2 \right) \frac{\sigma^2}{(n-1)s_{xx}^2}, \quad \operatorname{Cov}(\hat{\alpha}, \hat{\beta}) = -\frac{\sigma^2 \bar{x}_n}{(n-1)s_{xx}^2}.$$

Beweis. Wegen der Unkorreliertheit von $Y_1, \ldots, Y_n$ gilt

$$\operatorname{Var} \hat{\beta} = \sigma^2 \sum_{i=1}^n d_i^2 = \sigma^2 \sum_{i=1}^n \frac{(x_i - \bar{x}_n)^2}{(n-1)^2 s_{xx}^4} = \sigma^2 \frac{(n-1)s_{xx}^2}{(n-1)^2 s_{xx}^4} = \frac{\sigma^2}{(n-1)s_{xx}^2}.$$

Beweis der beiden anderen Formeln ist eine Übungsaufgabe. □

Satz 11.3.10. Es sei V_1 der von $Y_1 + \ldots + Y_n$ und $x_1 Y_1 + \ldots + x_n Y_n$ aufgespannte zweidimensionale Unterraum von V. Die BLUE-Schätzer $\hat{\alpha}$ und $\hat{\beta}$ liegen in V_1 und es gilt

$$\mathbb{E}[(Y_1 + \ldots + Y_n)\hat{\beta}] = 0, \quad \mathbb{E}[(x_1 Y_1 + \ldots + x_n Y_n)\hat{\alpha}] = 0.$$

Beweis. Dass $\hat{\beta}$ in V_1 liegt (und sogar $V_1 \cap V_\beta = \{\hat{\beta}\}$), haben wir bereits in Bemerkung 11.3.5 gesehen. Für $\hat{\alpha}$ verläuft der Beweis analog. Die Aussagen über die Kovarianzen folgen direkt aus Sätzen 11.3.2 und 11.3.6 □

11.4. Schätzer für Residuen ε_i und Varianz σ^2

Wir benutzen nach wie vor das Gauß-Markov-Modell aus dem vorangegangenen Abschnitt. Nachdem wir α und β geschätzt haben, können wir die Residuen $\varepsilon_i = Y_i - \alpha - \beta x_i$ schätzen:

Definition 11.4.1. Die *geschätzten Residuen* $\hat{\varepsilon}_i$ sind definiert durch

$$\hat{\varepsilon}_i = Y_i - \hat{\alpha} - \hat{\beta} x_i, \quad i = 1, \ldots, n.$$

Es folgt direkt aus der Definition, dass $\hat{\varepsilon}_i$ ein linearer Schätzer ist.

Satz 11.4.2. Die geschätzten Residuen $\hat{\varepsilon}_i$ sind erwartungstreue Schätzer für 0, d.h. es gilt

$$\mathbb{E}\hat{\varepsilon}_1 = \ldots = \mathbb{E}\hat{\varepsilon}_n = 0.$$

Beweis. Wir wissen, dass $\hat{\alpha}$ und $\hat{\beta}$ erwartungstreue Schätzer für α und β sind. Somit gilt

$$\mathbb{E}\hat{\varepsilon}_i = \mathbb{E}[Y_i - \hat{\alpha} - \hat{\beta} x_i] = \mathbb{E}Y_i - \mathbb{E}\hat{\alpha} - (\mathbb{E}\hat{\beta}) x_i = (\alpha + \beta x_i) - \alpha - \beta x_i = 0.$$

□

Eine allgemeine Charakterisierung der linearen und für 0 erwartungstreuen Schätzer liefert der nächste Satz.

Satz 11.4.3. Ein linearer Schätzer $\sum_{i=1}^n f_i Y_i$ ist erwartungstreuer Schätzer für 0 genau dann, wenn

$$\sum_{i=1}^n f_i = 0, \quad \sum_{i=1}^n f_i x_i = 0.$$

Beweis. Übung. □

Die Menge der linearen erwartungstreuen Schätzer für 0 ist ein $(n-2)$-dimensionaler Unterraum von V. Wir bezeichnen diesen Unterraum mit V_0.

Aufgabe 11.4.4. Zeigen Sie, dass die affinen Unterräume V_α und V_β (die aus allen linearen Schätzern bestehen, die erwartungstreu für α bzw. β sind), parallel zum linearen Unterraum V_0 sind, nämlich

$$V_\alpha = \hat{\alpha} + V_0, \quad V_\beta = \hat{\beta} + V_0.$$

Satz 11.4.5. Es gilt $\mathbb{E}[\hat{\varepsilon}_i\hat{\alpha}] = 0$ und $\mathbb{E}[\hat{\varepsilon}_i\hat{\beta}] = 0$ für alle $i = 1, \ldots, n$.

Beweis. Aus Satz 11.4.3 folgt, dass der Unterraum V_0 orthogonal zu dem von $Y_1 + \ldots + Y_n$ und $x_1Y_1 + \ldots + x_nY_n$ erzeugten zweidimensionalen linearen Unterraum V_1 ist. Somit gilt $V_0 = V_1^{\perp}$. Die Schätzer $\hat{\alpha}$ und $\hat{\beta}$ liegen in V_1. Somit ist $\hat{\varepsilon}_i$ orthogonal zu $\hat{\alpha}$ und $\hat{\beta}$. □

Wir werden nun auch σ^2 schätzen.

Definition 11.4.6. Definiere zwei Schätzer $\hat{\sigma}^2$ und S^2 für den Parameter σ^2 durch

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2, \quad S^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2.$$

Die Bedeutung des Faktors $\frac{1}{n-2}$ wird im folgenden Satz klar.

Satz 11.4.7. Für die Erwartungswerte der Schätzer $\hat{\sigma}^2$ und S^2 gilt

$$\mathbb{E}\hat{\sigma}^2 = \frac{n-2}{n}\sigma^2, \quad \mathbb{E}S^2 = \sigma^2.$$

Beweis. Geometrisch gesehen ist $\hat{\varepsilon}_i = Y_i - \hat{\alpha} - \hat{\beta}x_i$ die orthogonale Projektion von Y_i auf den linearen Unterraum V_0. Es bezeichne P die orthogonale Projektion auf V_0. Dann gilt $\hat{\varepsilon}_i = PY_i$. Es ergibt sich

$$\sum_{i=1}^{n}\mathbb{E}\hat{\varepsilon}_i^2 = \sigma^2\sum_{i=1}^{n}\langle PY_i, PY_i\rangle = \sigma^2\sum_{i=1}^{n}\langle P^{\top}PY_i, Y_i\rangle = \sigma^2\sum_{i=1}^{n}\langle PY_i, Y_i\rangle = \sigma^2\operatorname{Sp}P,$$

wobei wir die Eigenschaften des Projektors $P^{\top} = P$ und $P^2 = P$ benutzt haben. Die Spur eines Projektors ist die Dimension von $\operatorname{Im}P$, also in unserem Fall $\dim V_0 = n - 2$. □

Wir geben nun einen alternativen Beweis von Satz 11.4.7.

Lemma 11.4.8. Es gilt

$$\mathbb{E}\hat{\varepsilon}_i^2 = \frac{n-2}{n}\sigma^2 + \frac{\sigma^2}{(n-1)s_{xx}^2}\left(\frac{1}{n}\sum_{j=1}^{n}x_j^2 - x_i^2 + 2x_i\bar{x}_n - 2\bar{x}_n^2\right). \tag{11.4.1}$$

Bemerkung 11.4.9. Zum Vergleich: $\mathbb{E}\varepsilon_i^2 = \sigma^2$. Es sei bemerkt, dass $\mathbb{E}\hat{\varepsilon}_i^2$ nicht von α und β abhängt.

Beweis. Aus der Definition $\hat{\varepsilon}_i = Y_i - \hat{\alpha} - \hat{\beta}x_i$ ergibt sich, dass

$$\begin{aligned}\operatorname{Var}\hat{\varepsilon}_i &= \operatorname{Var}[Y_i - \hat{\alpha} - \hat{\beta}x_i] \\ &= \operatorname{Var} Y_i + \operatorname{Var}\hat{\alpha} + x_i^2 \operatorname{Var}\hat{\beta} - 2\operatorname{Cov}(Y_i, \hat{\alpha}) - 2x_i \operatorname{Cov}(Y_i, \hat{\beta}) + 2\operatorname{Cov}(\hat{\alpha}, \hat{\beta}).\end{aligned}$$

Es gilt $\operatorname{Var} Y_i = \alpha + \beta x_i$. Wir haben $\operatorname{Var}\hat{\alpha}$, $\operatorname{Var}\hat{\beta}$ und $\operatorname{Cov}(\hat{\alpha}, \hat{\beta})$ bereits berechnet. Außerdem gilt

$$\operatorname{Cov}(Y_i, \hat{\beta}) = \operatorname{Cov}\left(Y_i, \sum_{j=1}^{n} d_j Y_j\right) = d_i = \frac{x_i - \bar{x}_n}{(n-1)s_{xx}^2},$$
$$\operatorname{Cov}(Y_i, \hat{\alpha}) = \operatorname{Cov}\left(Y_i, \sum_{j=1}^{n} c_j Y_j\right) = c_i = \frac{1}{n} - \bar{x}_n d_i.$$

Indem man nun all diese Werte einsetzst, erhält man (11.4.1). □

Beweis von Satz 11.4.7. Benutzt man das Ergebnis von Lemma 11.4.8, so erhält man nach einigen Transformationen $\sum_{i=1}^{n} \mathbb{E}\hat{\varepsilon}_i^2 = n - 2$. □

Aufgabe 11.4.10. Bestimmen Sie $\mathbb{E}[\hat{\varepsilon}_i \hat{\varepsilon}_j]$ für $i \neq j$. Zum Vergleich: $\mathbb{E}[\varepsilon_i \varepsilon_j] = 0$ für $i \neq j$.

11.5. Maximum-Likelihood-Methode

Wir betrachten nun ein Modell, in dem die Residuen normalverteilt sind.

(1) Die Ausgangsgrößen $x_1, \ldots, x_n$ sind deterministisch und nicht alle gleich.
(2) Die Zielgrößen $y_1, \ldots, y_n$ sind Realisierungen der Zufallsvariablen $Y_1, \ldots, Y_n$ mit
$$Y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \ldots, n.$$
(3) Die Störgrößen $\varepsilon_1, \ldots, \varepsilon_n$ sind unabhängige und mit Parametern $(0, \sigma^2)$ normalverteilte Zufallsvariablen.

Es sollen die unbekannten Parameter α, β, σ^2 geschätzt werden.

Die Zufallsvariablen $Y_1, \ldots, Y_n$ sind unabhängig mit $Y_i \sim \mathrm{N}(\alpha + \beta x_i, \sigma^2)$. Es sei bemerkt, dass $Y_1, \ldots, Y_n$ nicht identisch verteilt sind.

Satz 11.5.1. Die Maximum-Likelihood-Schätzer für α, β und σ^2 sind

$$\hat{\beta} = \frac{s_{xy}^2}{s_{xx}^2}, \quad \hat{\alpha} = \bar{y}_n - \hat{\beta}\bar{x}_n, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2,$$

wobei $\hat{\varepsilon}_i = y_i - \hat{\alpha} - \hat{\beta}x_i$ die geschätzten Residuen sind.

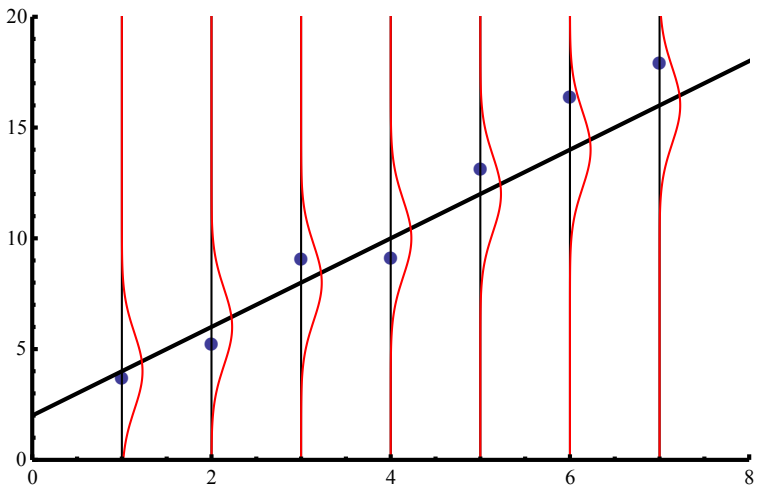


ABBILDUNG 1. Modell mit normalverteilten Residuen. Die schwarze Gerade ist die Regressionsgerade $y = \alpha + \beta x$. Die roten Kurven zeigen die Dichten von $Y_1, \ldots, Y_n$.

Beweis. Die Likelihood-Funktion, also die Dichte des Zufallsvektors $(Y_1, \ldots, Y_n)$ an der Stelle $(y_1, \ldots, y_n)$, ist gegeben durch

$$L(y_1, \ldots, y_n; \alpha, \beta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \mathrm{e}^{-\frac{(y_i - \alpha - \beta x_i)^2}{2\sigma^2}}.$$

Die Idee der Maximum-Likelihood-Methode besteht darin, diese Funktion zu maximieren. Hierbei bietet es sich an, die log-Likelihood-Funktion zu betrachten. Diese ist gegeben durch

$$\log L(\alpha, \beta, \sigma^2) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2.$$

Sei zunächst $\sigma^2 > 0$ fest. Da wir die Funktion $\log L$ nun bezüglich α, β maximieren wollen, müssen wir die Funktion $\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$ minimieren. Durch die Methode der kleinsten Quadrate wissen wir, dass das Minimum für

$$\hat{\beta} = \frac{s_{xy}^2}{s_{xx}^2}, \quad \hat{\alpha} = \bar{y}_n - \hat{\beta} \bar{x}_n$$

erreicht wird. Diese Werte hängen nicht von σ^2 ab.

Nun sei σ^2 wieder variabel. Wir bilden die Ableitung nach σ^2 und setzen diese gleich null:

$$\frac{\partial}{\partial(\sigma^2)} \log L(\hat{\alpha}, \hat{\beta}, \sigma^2) = -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2 \stackrel{!}{=} 0.$$

Als Lösung ergibt sich $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2$. □

Aufgabe 11.5.2. Zeigen Sie, dass die Statistik $(\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2)$ suffizient ist.

11.6. Gemeinsame Verteilung von $(\hat{\alpha}, \hat{\beta}, S^2)$

Wir betrachten das Modell mit normalverteilten Residuen aus Abschnitt 11.5. Der nächste Satz beschreibt die gemeinsame Verteilung von des Zufallsvektors $(\hat{\alpha}, \hat{\beta}, S^2)$.

Satz 11.6.1. Es gilt

(1) Der Vektor $(\hat{\alpha}, \hat{\beta})$ ist bivariat normalverteilt mit

$$\hat{\alpha} \sim \mathrm{N}\left(\alpha, \frac{\sigma^2}{n(n-1)s_{xx}^2} \sum_{i=1}^{n} x_i^2\right), \quad \hat{\beta} \sim \mathrm{N}\left(\beta, \frac{\sigma^2}{(n-1)s_{xx}^2}\right)$$

und

$$\mathrm{Cov}(\hat{\alpha}, \hat{\beta}) = -\frac{\sigma^2 \bar{x}_n}{(n-1)s_{xx}^2}.$$

(2) $(\hat{\alpha}, \hat{\beta})$ und S^2 sind unabhängig.

(3) $\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$.

Beweis von Teil 1. Der Vektor $(\hat{\alpha}, \hat{\beta})$ ist eine lineare Transformation des Vektors $(\varepsilon_1, \ldots, \varepsilon_n)$; siehe Sätze 11.3.4 und 11.3.7. Somit ist $(\hat{\alpha}, \hat{\beta})$ bivariat normalverteilt. Insbesondere sind auch die Komponenten $\hat{\alpha}$ und $\hat{\beta}$ normalverteilt. Wir wissen, dass $\mathbb{E}\hat{\alpha} = \alpha$ und $\mathbb{E}\hat{\beta} = \beta$, denn beide Schätzer sind erwartungstreu. Die Formeln für die Varianzen von $\hat{\alpha}$ und $\hat{\beta}$, sowie die Formel für die Kovarianz, haben wir bereits in Satz 11.3.9 hergeleitet. □

Beweis von Teil 2. Wir werden sogar zeigen, dass $(\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)$ und $(\hat{\alpha}, \hat{\beta})$ unabhängig sind. Dies impliziert die Unabhängigkeit von S^2 und $(\hat{\alpha}, \hat{\beta})$, denn S^2 ist eine Funktion von $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$. Der Vektor $(\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n, \hat{\alpha}, \hat{\beta})$ ist multivariat normalverteilt, denn er kann als eine lineare Transformation von $(\varepsilon_1, \ldots, \varepsilon_n)$ dargestellt werden. Da bei der multivariaten Normalverteilung die Unabhängigkeit und die Unkorreliertheit äquivalent sind, reicht es zu zeigen, dass

$$\mathrm{Cov}(\hat{\varepsilon}_i, \hat{\alpha}) = \mathrm{Cov}(\hat{\varepsilon}_i, \hat{\beta}) = 0, \quad i = 1, \ldots, n.$$

Das haben wir aber bereits in Satz 11.4.5 gezeigt. □

Für den Beweis von Teil 3 benötigen wir ein Lemma.

Lemma 11.6.2. Seien U und V unabhängige Zufallsvariablen mit $U \sim \chi^2_n$ und $U+V \sim \chi^2_{n+m}$, wobei $n, m \in \mathbb{N}$. Dann ist $V \sim \chi^2_m$.

Beweis. Da U und V unabhängig sind, gilt für die charakteristischen Funktionen die Relation

$$\varphi_{U+V}(t) = \varphi_U(t) \cdot \varphi_V(t).$$

Somit ergibt sich

$$\varphi_V(t) = \frac{\varphi_{U+V}(t)}{\varphi_U(t)} = \frac{(1-2it)^{-(n+m)/2}}{(1-2it)^{-n/2}} = (1-2it)^{-m/2}.$$

Dies ist die charakteristische Funktion einer χ^2_m-Verteilung. Somit ist $V \sim \chi^2_m$. □

Beweis von Teil 3. Stellen wir uns vor, wir ersetzen alle x_i durch $x_i - \bar{x}_n$ und ändern die Residuen ε_i nicht. Es ist leicht zu sehen, dass sich dadurch auch die geschätzten Residuen $\hat{\varepsilon}_i$ nicht ändern. Somit können wir im Folgenden annahmen, dass $\bar{x}_n = 0$, ansonsten ersetze x_i durch $x_i - \bar{x}_n$. Dadurch vereinfacht sich erheblich die Notation, denn

$$c_i = \frac{1}{n}, \quad d_i = \frac{x_i}{\sum_{j=1}^n x_j^2}$$

und

$$\hat{\alpha} \sim \mathrm{N}\left(\alpha, \frac{\sigma^2}{n}\right), \quad \hat{\beta} \sim \mathrm{N}\left(\beta, \frac{\sigma^2}{\sum_{j=1}^n x_j^2}\right).$$

Es gilt

$$\begin{aligned}(n-2)S^2 &= \sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta}x_i)^2 \\ &= \sum_{i=1}^n ((Y_i - \alpha - \beta x_i) + (\alpha - \hat{\alpha}) + (\beta - \hat{\beta})x_i)^2 \\ &= \sum_{i=1}^n \varepsilon_i^2 - n(\hat{\alpha} - \alpha)^2 - \left(\sum_{j=1}^n x_j^2\right)(\hat{\beta} - \beta)^2,\end{aligned}$$

wobei die letzte Gleichheit eine Übung ist. Wir haben somit die Darstellung

$$U + \frac{(n-2)S^2}{\sigma^2} = Z,$$

wobei

$$Z = \sum_{i=1}^n \left(\frac{\varepsilon_i}{\sigma}\right)^2 \sim \chi^2_n, \quad U = \left(\frac{1}{\sigma}\sqrt{n}(\hat{\alpha} - \alpha)\right)^2 + \left(\frac{1}{\sigma}\left(\sum_{j=1}^n x_j^2\right)^{1/2}(\hat{\beta} - \beta)\right)^2 \sim \chi^2_2.$$

Außerdem wissen wir, dass S^2 und U unabhängig sind, denn S^2 ist unabhängig von $(\hat{\alpha}, \hat{\beta})$ laut Teil 2 des Satzes. Aus Lemma 11.6.2 folgt, dass $\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$. □

11.7. Konfidenzintervalle für α, β, σ^2

Nun können wir Konfidenzintervalle für die unbekannten Parameter α, β und σ^2 konstruieren. Dabei benutzen wir nach wie vor das Modell mit normalverteilten Residuen aus Abschnitt 11.5.

Konfidenzintervall für α. Wir konstruiren ein Konfidenzintervall für α zu einem vorgegebenem Niveau $1-\xi$. Wir wissen aus Satz 11.6.1, dass

$$T_1 := \frac{\hat{\alpha}-\alpha}{\sigma_\alpha \sigma} \sim \mathrm{N}(0,1),$$

wobei

$$\sigma_\alpha^2 = \frac{1}{n(n-1)s_{xx}^2}\sum_{i=1}^n x_i^2.$$

Da σ jedoch nicht bekannt ist, ersetzen wir es durch einen Schätzer, nämlich S, und betrachten die Statistik

$$T_2 := \frac{\hat{\alpha}-\alpha}{\sigma_\alpha S} = \frac{T_1}{S/\sigma} = \frac{T_1}{\sqrt{\frac{1}{n-2}\frac{(n-2)S^2}{\sigma^2}}} \sim t_{n-2},$$

denn wir wissen aus Satz 11.6.1, dass $T_1 \sim \mathrm{N}(0,1)$, $\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$ und diese beiden Zufallsvariablen unabhängig sind. Somit gilt

$$\mathbb{P}\left[t_{n-2,\xi/2} \le \frac{\hat{\alpha}-\alpha}{\sigma_\alpha\, S} \le t_{n-2,1-\xi/2}\right] = 1-\xi.$$

Dies führt zu folgendem Konfidenzintervall für α zum Niveau $1-\xi$:

$$\left[\hat{\alpha} - t_{n-2,1-\xi/2}\sigma_\alpha S, \hat{\alpha} + t_{n-2,1-\xi/2}\sigma_\alpha S\right].$$

Konfidenzintervall für β. Es sei ein Konfidenzniveau $1-\xi$ vorgegeben. Wir wissen aus Satz 11.6.1, dass

$$T_1 := \frac{\hat{\beta}-\beta}{\sigma_\beta \sigma} \sim \mathrm{N}(0,1),$$

wobei

$$\sigma_\beta^2 = \frac{1}{(n-1)s_{xx}^2}.$$

Wir ersetzen den unbekannten Parameter σ durch S und betrachten somit

$$T_2 := \frac{\hat{\beta}-\beta}{\sigma_\beta S} = \frac{T_1}{S/\sigma} = \frac{T_1}{\sqrt{\frac{1}{n-2}\frac{(n-2)S^2}{\sigma^2}}} \sim t_{n-2},$$

denn $T_1 \sim \mathrm{N}(0,1)$, $\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$ und diese Zufallsvariablen sind unabhängig. Somit gilt

$$\mathbb{P}\left[t_{n-2,\xi/2} \le \frac{\hat{\beta}-\beta}{\sigma_\beta S} \le t_{n-2,1-\xi/2}\right] = 1-\xi.$$

Dies führt zum folgenden Konfidenzintervall für β zum Niveau $1-\xi$:

$$\left[\hat{\beta} - t_{n-2,1-\xi/2}\sigma_\beta S, \hat{\beta} + t_{n-2,1-\xi/2}\sigma_\beta S\right].$$

Konfidenzintervall für σ^2. Es gilt

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}.$$

Somit ist

$$\mathbb{P}\left[\chi^2_{n-2,\xi/2} \leq \frac{(n-2)S^2}{\sigma^2} \leq \chi^2_{n-2,1-\xi/2}\right] = 1-\xi.$$

Dies führt zu folgendem Konfidenzintervall für σ^2 zum Niveau $1-\xi$:

$$\left[\frac{(n-2)S^2}{\chi^2_{n-2,1-\xi/2}}, \frac{(n-2)S^2}{\chi^2_{n-2,\xi/2}}\right].$$

Konfidenzintervall für $\alpha+\beta x_0$. Sei $x_0 \in \mathbb{R}$ gegeben. Wir wollen ein Konfidenzintervall für $\alpha+\beta x_0$ konstruiren. Ein natürlicher Schätzer für diesen Parameter ist $\hat{\alpha}+\hat{\beta}x_0$. Um ein Konfidenzintervall zu konstruieren, müssen wir die Verteilung des Schätzers kennen.

Satz 11.7.1. Der Schätzer $\hat{\alpha}+\hat{\beta}x_0$ ist erwartungstreu für $\alpha+\beta x_0$ und es gilt

$$\hat{\alpha}+\hat{\beta}x_0 \sim \mathrm{N}\left(\alpha+\beta x_0, \sigma^2\left(\frac{1}{n}+\frac{(x_0-\bar{x}_n)^2}{(n-1)s^2_{xx}}\right)\right).$$

Beweis. Da $\hat{\alpha}+\hat{\beta}x_0$ als eine lineare Transformation von $Y_1,\ldots,Y_n$ dargestellt werden kann, ist $\hat{\alpha}+\hat{\beta}x_0$ normalverteilt. Es reicht also, den Erwartungswert und die Varianz von $\hat{\alpha}+\hat{\beta}x_0$ zu berechnen. Für den Erwartungswert gilt

$$\mathbb{E}[\hat{\alpha}+\hat{\beta}x_0] = \mathbb{E}[\hat{\alpha}] + x_0\mathbb{E}[\hat{\beta}] = \alpha+\beta x_0,$$

denn $\mathbb{E}\hat{\alpha}=\alpha$ und $\mathbb{E}\hat{\beta}=\beta$. Für die Varianz von $\hat{\alpha}+\hat{\beta}x_0$ gilt

$$\begin{aligned}\operatorname{Var}[\hat{\alpha}+\hat{\beta}x_0] &= \operatorname{Var}\hat{\alpha} + (\operatorname{Var}\hat{\beta})x_0^2 + 2x_0\operatorname{Cov}(\hat{\alpha},\hat{\beta}) \\ &= \frac{\sigma^2}{n(n-1)s^2_{xx}}\sum_{i=1}^n x_i^2 + \frac{\sigma^2 x_0^2}{(n-1)s^2_{xx}} - \frac{2x_0\sigma^2\bar{x}_n}{(n-1)s^2_{xx}},\end{aligned}$$

wobei wir die Formeln aus Satz 11.3.9 benutzt haben. Nach einigen Transformationen erhalten wir

$$\begin{aligned}\operatorname{Var}[\hat{\alpha}+\hat{\beta}x_0] &= \frac{\sigma^2}{(n-1)s^2_{xx}}\left(\frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}_n^2 + \bar{x}_n^2 + x_0^2 - 2x_0\bar{x}_n\right) \\ &= \frac{\sigma^2}{(n-1)s^2_{xx}}\left(\frac{1}{n}\sum_{i=1}^n (x_i-\bar{x}_n)^2 + (x_0-\bar{x}_n)^2\right) \\ &= \sigma^2\left(\frac{1}{n}+\frac{(x_0-\bar{x}_n)^2}{(n-1)s^2_{xx}}\right),\end{aligned}$$

was genau die gewünschte Formel ist. □

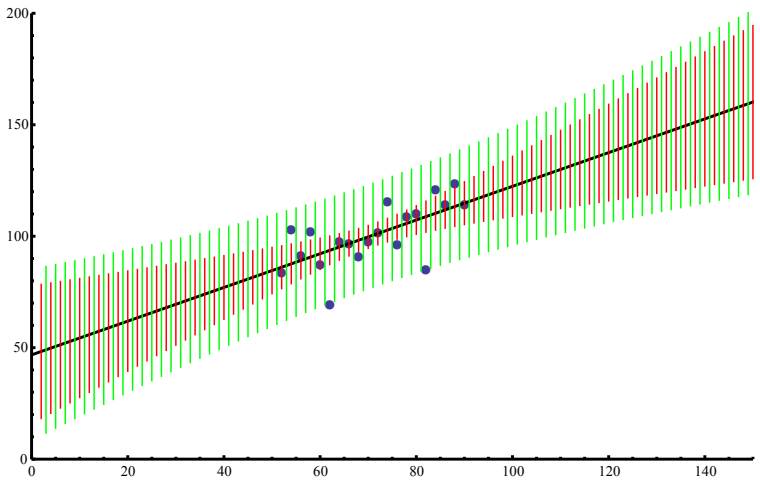


ABBILDUNG 2. Konfidenz- und Vorhersageintervalle in der einfachen linearen Regression. Die schwarze Gerade ist die geschätzte Regressionsgerade $y = \hat{\alpha} + \hat{\beta}x$. Die roten Intervalle sind die Konfidenzintervalle für $\alpha + \beta x_0$ für verschiedene Werte von x_0. Die grünen Intervalle sind die Vorhersageintervalle für $Y(x_0)$. Das Konfidenzniveau ist 0.95.

Notation 11.7.2. Für $x_0 \in \mathbb{R}$ definiere

$$\sigma^2(x_0) := \frac{1}{\sigma^2} \operatorname{Var}[\hat{\alpha} + \hat{\beta}x_0] = \frac{1}{n} + \frac{(x_0 - \bar{x}_n)^2}{(n-1)s_{xx}^2}.$$

Nun sind wir zur Konstruktion eines Konfidenzintervalls für $\alpha + \beta x_0$ bereit. Es gilt

$$T_1 := \frac{\hat{\alpha} + \hat{\beta}x_0 - (\alpha + \beta x_0)}{\sigma\sigma(x_0)} \sim \mathrm{N}(0,1).$$

Allerdings ist σ unbekannt. Somit betrachten wir

$$T_2 := \frac{\hat{\alpha} + \hat{\beta}x_0 - (\alpha + \beta x_0)}{S\sigma(x_0)} = \frac{T_1}{S/\sigma} = \frac{T_1}{\sqrt{\frac{1}{n-2}\frac{(n-2)S^2}{\sigma^2}}} \sim t_{n-2},$$

denn $T_1 \sim \mathrm{N}(0,1)$, $\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$ und diese Zufallsvariablen sind unabhängig. Somit gilt

$$\mathbb{P}\left[-t_{n-2,1-\xi/2} \leq \frac{\hat{\alpha} + \hat{\beta}x_0 - (\alpha + \beta x_0)}{S\sigma(x_0)} \leq t_{n-2,1-\xi/2}\right] = 1 - \xi.$$

Umstellen nach $\alpha + \beta x_0$ führt zum folgenden Konfidenzintervall zum Niveau $1 - \xi$:

$$\left[\hat{\alpha} + \hat{\beta}x_0 - t_{n-2,1-\xi/2}S\sigma(x_0), \hat{\alpha} + \hat{\beta}x_0 + t_{n-2,1-\xi/2}S\sigma(x_0)\right].$$

Aufgabe 11.7.3. Für welchen Wert von x_0 ist die Länge des Konfidenzintervalls für $\alpha + \beta x_0$ am kleinsten?

Vorhersageintervalle. Stellen wir uns vor, wir haben die Punkte $(x_1, y_1), \ldots, (x_n, y_n)$ beobachtet. Nun sei uns ein weiterer Wert x_0 gegeben und wir sollen den entsprechenden Wert von y, der mit $Y(x_0)$ bezeichnet wird, vorhersagen. Es sei bemerkt, dass die Vorhersage von $Y(x_0)$ und das Schätzen von $\alpha + \beta x_0$ unterschiedliche Probleme sind, da sich diese beiden Größen durch ein unbekanntes Residuum unterscheiden. Wir nehmen nämlich an, dass

$$Y(x_0) = \alpha + \beta x_0 + \varepsilon_0 \text{ mit } \varepsilon_0 \sim \mathrm{N}(0, \sigma^2)$$

und dass das Residuum ε_0 von allen anderen Residuen $\varepsilon_1, \ldots, \varepsilon_n$ unabhängig ist.

Definition 11.7.4. Ein *Vorhersageintervall* für $Y(x_0)$ zu einem vorgegebenen Konfidenzniveau $1 - \xi$ ist ein zufälliges und von $Y_1, \ldots, Y_n$ abhängiges Intervall $[\theta_1, \theta_2]$ mit

$$\mathbb{P}[\theta_1 < Y(x_0) < \theta_2] \geq 1 - \xi.$$

Satz 11.7.5. Es gilt $Y(x_0) - \hat{\alpha} - \hat{\beta} x_0 \sim \mathrm{N}(0, \sigma^2 \tilde{\sigma}^2(x_0))$ mit

$$\tilde{\sigma}^2(x_0) := 1 + \sigma^2(x_0) = 1 + \frac{1}{n} + \frac{(x_0 - \bar{x}_n)^2}{(n-1)s_{xx}^2}.$$

Beweis. Die Aussage folgt aus Satz 11.7.1, denn $Y(x_0) \sim \mathrm{N}(\alpha + \beta x_0, \sigma^2)$ und $\hat{\alpha} + \hat{\beta} x_0 \sim (\alpha + \beta x_0, \sigma^2 \sigma^2(x_0))$ sind unabhängig. □

Wir können nun ein Vorhersageintervall für $Y(x_0)$ wie folgt konstruieren. Es gilt

$$\frac{Y(x_0) - \hat{\alpha} - \hat{\beta} x_0}{\sigma \tilde{\sigma}(x_0)} \sim \mathrm{N}(0, 1)$$

und folglich auch

$$\frac{Y(x_0) - \hat{\alpha} - \hat{\beta} x_0}{S \tilde{\sigma}(x_0)} \sim t_{n-2}.$$

Dies führt zu

$$\mathbb{P}\left[\hat{\alpha} + \hat{\beta} x_0 - S\tilde{\sigma}(x_0) t_{n-2,1-\xi/2} \leq Y(x_0) \leq \hat{\alpha} + \hat{\beta} x_0 + S\tilde{\sigma}(x_0) t_{n-2,1-\xi/2}\right] = 1 - \xi.$$

Somit ist ein Vorhersageintervall für $Y(x_0)$ zum Niveau $1 - \xi$ gegeben durch:

$$\left[\hat{\alpha} + \hat{\beta} x_0 - S\tilde{\sigma}(x_0) t_{n-2,1-\xi/2}, \hat{\alpha} + \hat{\beta} x_0 + S\tilde{\sigma}(x_0) t_{n-2,1-\xi/2}\right].$$

Es sei bemerkt, dass das Vorhersageintervall für $Y(x_0)$ etwas länger ist, als das Konfidenzintervall für $\alpha + \beta x_0$. Der Wert $Y(x_0)$ ist schwieriger vorherzusagen als $\alpha + \beta x_0$, denn $Y(x_0)$ beinhaltet eine zusätzliche unbekannte Störgröße ε_0.

Konfidenzband für die Regressionsgerade $y = \alpha + \beta x$. Es seien m Punkte $x_{01}, \ldots, x_{0m} \in \mathbb{R}$ gegeben. Mit den Methoden der vorherigen Abschnitte können wir zu einem vorgegebenen

Konfidenzniveau $1-\xi$ Konfidenzintervalle $I_1, \ldots, I_m$ für $\alpha+\beta x_{01}, \ldots, \alpha+\beta x_{0m}$ konstruieren. Diese haben die Eigenschaft

$$\mathbb{P}[\alpha + \beta x_{0j} \in I_j] = 1 - \xi \text{ für alle } j = 1, \ldots, m.$$

Allerdings ist es im Allgemeinen falsch, dass

$$\mathbb{P}[\alpha + \beta x_{01} \in I_1, \ldots, \alpha + \beta x_{0m} \in I_m] = 1 - \xi.$$

Für die Wahrscheinlichkeit, dass die Ereignisse $\{\alpha + \beta x_{0j} \in I_j\}$ *simultan* eintreten, gilt die *Bonferroni-Abschätzung*

$$\mathbb{P}[\alpha + \beta x_{01} \in I_1, \ldots, \alpha + \beta x_{0m} \in I_m] \geq 1 - m\xi.$$

Möchte man also die Werte $\alpha + \beta x_{01}, \ldots \alpha + \beta x_{0m}$ simultan schätzen, so verschlechtert sich das Konfidenzniveau auf $1 - m\xi$. Für m groß genug ist diese Zahl negativ, so dass wir gar keine vernünftige Aussage machen können.

Wir fragen uns deshalb, ob es möglich ist, zwei von der Stichprobe $Y_1, \ldots, Y_n$ abhängige Funktionen $\theta_1(x) \leq \theta_2(x)$ zu konstruieren, so dass

$$\mathbb{P}[\forall x \in \mathbb{R}\colon \theta_1(x) \leq \alpha + \beta x \leq \theta_2(x)] \geq 1 - \xi.$$

Ein solches Paar von Funktionen heißt ein *Konfidenzband* zum Niveau $1-\xi$ für die Regressionsgerade $y = \alpha+\beta x$. Die obige Bedingung besagt, dass die komplette Gerade $y = \alpha+\beta x$ mit Wahrscheinlichkeit mindestens $1 - \xi$ zwischen $\theta_1(x)$ und $\theta_2(x)$ liegt. Angesichts der obigen Konstruktion der Konfidenzintervalle für den Wert $\alpha + \beta x_0$ ist es natürlich, den folgenden Ansatz für θ_1 und θ_2 zu machen:

$$\theta_1(x) = \hat{\alpha} + \hat{\beta}x - lS\sqrt{\frac{1}{n} + \frac{(x-\bar{x}_n)^2}{(n-1)s_{xx}^2}}, \quad \theta_2(x) = \hat{\alpha} + \hat{\beta}x + lS\sqrt{\frac{1}{n} + \frac{(x-\bar{x}_n)^2}{(n-1)s_{xx}^2}}.$$

Dabei ist l eine Konstante, die größer als $t_{n-2,1-\xi/2}$ sein sollte.

Satz 11.7.6. Sei $\xi \in (0,1)$ und definiere $l := \sqrt{2F_{2,n-2,1-\xi}}$. Dann gilt

$$\mathbb{P}\left[\forall x \in \mathbb{R}\colon \left|\hat{\alpha} + \hat{\beta}x - \alpha - \beta x\right| \leq lS\sqrt{\frac{1}{n} + \frac{(x-\bar{x}_n)^2}{(n-1)s_{xx}^2}}\right] = 1 - \xi.$$

Für den Beweis benötigen wir ein Lemma.

Lemma 11.7.7. Seien $a, b, c, d \in \mathbb{R}$ mit $a \neq 0, c > 0, d > 0$. Dann gilt

$$\max_{x\in\mathbb{R}} \frac{(a+bx)^2}{c+dx^2} = \frac{a^2}{c} + \frac{b^2}{d}.$$

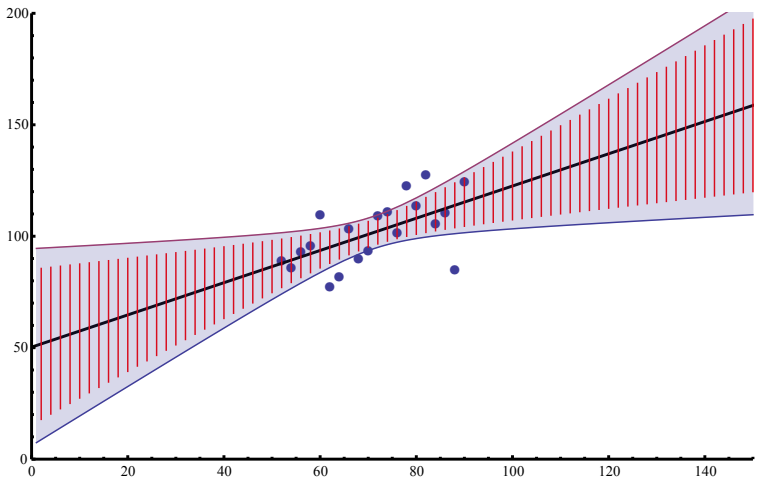


ABBILDUNG 3. Konfidenzintervalle und Konfidenzband in der einfachen linearen Regression. Die schwarze Gerade ist die geschätzte Regressionsgerade $y = \hat{\alpha} + \hat{\beta}x$. Die roten Intervalle sind die Konfidenzintervalle für $\alpha + \beta x$ für verschiedene Werte von x. Das blaue Gebiet ist das Konfidenzband für die Regressionsgerade $y = \alpha + \beta x$. Das Konfidenzniveau ist 0.95.

Beweis. Man kann die Funktion $x \mapsto \frac{(a+bx)^2}{c+dx^2}$ ableiten und zeigen, dass das Maximum für $x = \frac{bc}{ad}$ erreicht wird. Wir geben hier einen anderen Beweis. Wir zeigen, dass für alle $x \in \mathbb{R}$,

$$\frac{(a+bx)^2}{c+dx^2} \leq \frac{a^2}{c} + \frac{b^2}{d}$$

und dass für mindestens ein x die Gleichheit eintritt. Indem wir beide Seiten mit $c + dx^2$ multiplizieren, erhalten wir

$$a^2 + 2abx + b^2x^2 \leq a^2 + \frac{a^2dx^2}{c} + \frac{b^2c}{d} + b^2x^2.$$

Die Terme a^2 und b^2x^2 können gekürzt werden. Multipliziert man beide Seiten mit cd, so erhält man

$$2abcdx \leq a^2d^2x^2 + b^2c^2.$$

Diese Ungleichung ist aber richtig, denn $(adx - bc)^2 \geq 0$. Außerdem tritt die Gleichheit für $x = \frac{bc}{ad}$ ein. □

Beweis von Satz 11.7.6. Wir können o.E.d.A. annehmen, dass $\bar{x}_n = 0$, denn andernfalls könnten wir alle x_i durch $x_i - \bar{x}_n$ ersetzen, was lediglich zu einer parallelen Verschiebung der Regressionsgeraden und des Konfidenzbandes führen würde. Man muss zeigen, dass

$$\mathbb{P}\left[\frac{1}{2}\max_{x\in\mathbb{R}}\frac{(\hat{\alpha}+\hat{\beta}x-(\alpha+\beta x))^2}{S^2\left(\frac{1}{n}+\frac{x^2}{(n-1)s_{xx}^2}\right)} \leq \frac{l^2}{2}\right] = 1-\xi.$$

Da $l^2/2 = F_{2,n-2,1-\xi}$, reicht es zu zeigen, dass

$$\frac{1}{2}\max_{x\in\mathbb{R}}\frac{((\hat{\alpha}-\alpha)+x(\hat{\beta}-\beta))^2}{\frac{S^2}{n}+\frac{S^2x^2}{(n-1)s_{xx}^2}} \sim F_{2,n-2}. \tag{11.7.1}$$

Das Maximum berechnet sich mit Lemma 11.7.7 zu

$$\frac{1}{2}\max_{x\in\mathbb{R}}\frac{((\hat{\alpha}-\alpha)+x(\hat{\beta}-\beta))^2}{\frac{S^2}{n}+\frac{S^2x^2}{(n-1)s_{xx}^2}} = \frac{(\hat{\alpha}-\alpha)^2}{2\frac{S^2}{n}} + \frac{(\hat{\beta}-\beta)^2}{2\frac{S^2}{(n-1)s_{xx}^2}} = \frac{\frac{1}{2}\left(\frac{\hat{\alpha}-\alpha}{\sigma/\sqrt{n}}\right)^2 + \frac{1}{2}\left(\frac{\hat{\beta}-\beta}{\sigma/\sqrt{(n-1)s_{xx}^2}}\right)^2}{\frac{1}{n-2}\cdot\frac{(n-2)S^2}{\sigma^2}}.$$

Unter Berücksichtigiung von $\bar{x}_n = 0$ folgt aus Satz 11.6.1, dass

$$\frac{\hat{\alpha}-\alpha}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1), \quad \frac{\hat{\beta}-\beta}{\sigma/\sqrt{(n-1)s_{xx}^2}} \sim \mathrm{N}(0,1), \quad \frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}.$$

Außerdem sind diese drei Zufallsvariablen unabhängig. Mit der Definition der F-Verteilung ergibt sich (11.7.1). □

KAPITEL 12

Bootstrap

Bootstrap ist eine sehr einfache aber auch sehr nützliche und allgemeine Methode zur Konstruktion von Konfidenzintervallen und Tests. Wir betrachten hier nur einige Beispiele. Für mehr Informationen verweisen wir auf das Buch von B. Efron und R. J. Tibshirani, *„An introduction to the Bootstrap“*, Chapman and Hall, 1994.

12.1. Verteilungsfunktion anhand einer einzigen Realisierung berechnen

Gegeben sei eine Stichprobe $x_1, \dots, x_n$. Wir gehen davon aus, dass diese Stichprobe eine Realisierung von unabhängigen identisch verteilten Zufallsvariablen $X_1, \dots, X_n$ mit einer unbekannten Verteilungsfunktion F ist. Wir wollen nun eine gewisse Charakteristik $\theta = \theta(F)$ der Verteilungsfunktion F schätzen. Man kann etwa an die folgenden Beispiele denken:

- Der Median $\theta = F^{-1}(1/2)$.
- Der Interquartilsabstand $\theta = F^{-1}(3/4) - F^{-1}(1/2)$.
- Der Erwartungswert $\theta = \mathbb{E}X_1 = \int_{\mathbb{R}} x \mathrm{d}F(x)$.
- Die mittlere absolute Abweichung bezüglich des Medians $\varphi = \int_{\mathbb{R}} |x - F^{-1}(1/2)| \mathrm{d}F(x)$, usw.

In diesen Beispielen ist es sehr leicht, einen natürlichen Schätzer T von θ zu konstruieren. So kann man z.B. den Interquartilsabstand durch

$$T(x_1, \dots, x_n) := x_{([3n/4])} - x_{([n/4])}$$

schätzen. Nun wollen wir aber auch ein Konfidenzintervall für θ konstruieren. Dabei wollen wir keine parametrischen Annahmen (etwa Normalverteilung) an F machen. Für den Median gibt es eine sehr schöne nichtparametrische Lösung, s. Abschnitt 9.9. Aber wie würde man z.B. bei Inerquartilsabstand vorgehen?

Für die Konstruktion eines Konfidenzintervalls benötigen wir die Verteilung der Zufallsvariable $T = T(X_1, \dots, X_n)$.

Überlegung 1. Stünden uns B unabhängige Realisierungen $t_1, \dots, t_B$ von T zur Verfügung (wobei B eine sehr große Zahl ist), so könnten wir die Verteilungsfunktion $G(t) = \mathbb{P}[T \le t]$ durch die empirische Verteilungsfunktion

$$\hat{G}(t) := \frac{1}{B} \sum_{i=1}^{B} \mathbb{1}_{t_i \le t}, \quad t \in \mathbb{R},$$

schätzen. Wir haben aber leider nur eine einzige Realisierung, nämlich $T(x_1, \dots, x_n)$!

Überlegung 2. Wäre die Verteilungsfunktion F (von X_i) bekannt, so könnten wir eine neue, gemäß F verteilte Stichprobe $x'_1, \dots, x'_n$ auf dem Rechner simulieren um dann eine weitere Realisierung $t' := T(x'_1, \dots, x'_n)$ von T erzeugen. In einer Schleife könnten wir

natürlich auch beliebig viele unabhängige Realisierungen $t'_1, t'_2, \ldots$ erzeugen. Allerdings ist die Verteilungsfunktion F nicht bekannt!

Überlegung 3. Nach diesen Rückschlägen müssen wir uns fragen, was uns überhaupt bekannt ist. Nur die Stichprobe $x_1, \ldots, x_n$! Diese könnten wir benutzen, um die Verteilungsfunktion F (von X_i) durch die empirische Verteilungsfunktion

$$\hat{F}_n(t) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{x_k \le t}, \quad t \in \mathbb{R},$$

zu schätzen. Für großes n sollte (z.B. nach dem Satz von Glivenko-Cantelli) $\hat{F}_n \approx F$ gelten. Wir könnten dann eine gemäß $\hat{F}_n$ verteilte Stichprobe $x_1^*, \ldots, x_n^*$ simulieren und diese wie in Schritt 2 verwenden, um eine (approximative) Realisierung $t^* = T(x_1^*, \ldots, x_n^*)$ von T zu erzeugen. Diese ebenso einfache wie geniale Methode (B. Efron, 1979) nennt man *„Bootstrap“* (sich selbst am Schopf aus dem Sumpf ziehen[1][2]).

Da die empirische Verteilung jedem Element x_i die Wahrscheinlichkeit $1/n$ zuordnet, sind $x_1^*, \ldots, x_n^*$ unabhängige und gemäß einer Gleichverteilung aus der Stichprobe $x_1, \ldots, x_n$ gezogene Elemente. Hier ist ein Beispiel für eine solche Bootstrap-Stichprobe mit $n = 20$:

$$x_{20}, x_{19}, x_7, x_{13}, x_{19}, x_{14}, x_{16}, x_{17}, x_6, x_{15}, x_{12}, x_9, x_1, x_6, x_{20}, x_8, x_{20}, x_{16}, x_{15}, x_{10}.$$

Wegen der Unabhängigkeit wird *mit Zurücklegen* gezogen, so dass es in der Stichprobe $x_1^*, \ldots, x_n^*$ Wiederholungen geben kann. Beim Ziehen ohne Zurücklegen wäre $x_1^*, \ldots, x_n^*$ eine Permutation von $x_1, \ldots, x_n$. Die von uns betrachteten Funktionale sind invariant unter Permutationen der Stichprobe, so dass Ziehen ohne Zurücklegen keine neue Realisierung von T erzeugen würde.

Bootstrap-Verfahren

Gegeben Sei eine Stichprobe $x_1, \ldots, x_n$ und eine Statistik $T : \mathbb{R}^n \to \mathbb{R}$. Wir wollen die Verteilung von T simulieren.

Schritt 1. Lege in eine Urne n Bälle mit den Aufschriften $x_1, \ldots, x_n$. Ziehe n Bälle mit Zurücklegen und notiere die Aufschriften: $x_1^*, \ldots, x_n^*$.

Schritt 2. Berechne $t^* := T(x_1^*, \ldots, x_n^*)$.

Schritt 3. Sei B eine sehr große Zahl. Führe Schritte 1 und 2 in einer Schleife B Mal durch und bezeichne die in Schritt 2 berechneten Werte mit $t_1^*, \ldots, t_B^*$.

[1] Bootstrap (Substantiv, engl.): Stiefelschlaufe. To bootstrap (Verb, engl.): sich selbst in die Lage versetzen, etwas tun zu können.

[2] Aus „Feldzüge und Abenteuer des Freiherrn von Münchhausen“ in der Fassung von Gottfried August Bürger: „Ein andres Mal wollte ich über einen Morast setzen, der mir anfänglich nicht so breit vorkam, als ich ihn fand, da ich mitten im Sprunge war. Schwebend in der Luft wendete ich daher wieder um, wo ich hergekommen war, um einen größern Anlauf zu nehmen. Gleichwohl sprang ich auch zum zweiten Male noch zu kurz und fiel nicht weit vom andern Ufer bis an den Hals in den Morast. Hier hätte ich unfehlbar umkommen müssen, wenn nicht die Stärke meines eigenen Armes mich an meinem eigenen Haarzopfe, samt dem Pferde, welches ich fest zwischen meine Knie schloß, wieder herausgezogen hätte“.

Nun können wir die Verteilungsfunktion von T durch die empirische Verteilungsfunktion

$$\hat{G}(t) := \frac{1}{B}\sum_{i=1}^{B} \mathbb{1}_{t_i^* \le t}, \quad t \in \mathbb{R},$$

schätzen. Ebenso kann man den Erwartungswert und die Varianz von T durch

$$\hat{\mu} := \frac{1}{B}\sum_{i=1}^{B} t_i^*, \quad \text{bzw.} \quad \hat{\sigma}^2 := \frac{1}{B-1}\sum_{i=1}^{B}(t_i^* - \hat{\mu})^2$$

schätzen. Bezeichnet man mit q_β das β-Quantil der Verteilungsfunktion $\hat{G}$, so ergibt sich für θ das approximative Konfidenzintervall $[q_{\alpha/2}, q_{1-\alpha/2}]$.

12.2. Noch ein Beispiel zum Bootstrap

Das folgende Beispiel wurde dem Buch von Efron und Tibshirani entnommen. In einer Studie wurde die Auswirkung der regelmäßigen Einnahme von Aspirin auf das Risiko eines Herzinfarktes untersucht.

	Herzinfarkt	Personen
Aspirin-Gruppe	104	11037
Placebo-Gruppe	189	11034

Die Herzinfarktraten in den beiden Gruppen betragen 104/11037 bzw. 189/11034. Der Quotient der Raten ist

$$\frac{104/11037}{189/11034} \approx 0.55.$$

Diese Zahl ist kleiner als 1, was suggeriert, dass die Einnahme von Aspirin das Herzinfarktrisiko senkt. Aber ist dieses Ergebnis signifikant? Einen möglichen Zugang bietet der exakte Test nach Fisher, s. Abschnitt 10.10. Hier werden wir die Bootstrap-Methode verwenden. Wir bezeichnen mit p bzw. q die Wahrscheinlichkeit, dass eine Person aus der Aspirin- (bzw. Placebo-Gruppe) innerhalb des gegebenen Zeitraumes einen Herzinfarkt erleidet. Wir werden im Folgenden ein (approximatives) Konfidenzintervall für p/q konstruieren.

Modell: Die Anzahl der Herzinfarkte in den beiden Gruppen wird durch unabhängige Zufallsvariablen $X \sim \mathrm{Bin}(11037, p)$ und $Y \sim \mathrm{Bin}(11034, q)$ modelliert. In der Studie wurde eine Realisierung $(104, 189)$ von (X, Y) beobachtet. Ein natürlicher Schätzer für p/q ist

$$T := \frac{X/11037}{Y/11034}.$$

In der Studie wurde eine Realisierung 0.55 von T beobachtet. Wir wollen weitere Realisierungen von T erzeugen. Dafür benötigen wir weitere Realisierungen von X und Y. Leider sind p und q unbekannt, weshalb wir X und Y nicht simulieren können. An dieser Stelle kommt uns die Bootstrap-Methode zu Hilfe. Wir schätzen p und q durch

$$\hat{p} := 104/11037 \quad \text{und} \quad \hat{q} := 189/11034.$$

Anstelle des Zufallsvektors (X, Y) simulieren wir nun dessen Bootstrap-Version (X^*, Y^*) mit $X^* \sim \mathrm{Bin}(11037, \hat{p})$ und $Y^* \sim \mathrm{Bin}(11034, \hat{q})$. In unserer Simulation haben sich folgende

Paare ergeben:

$(x_1^*, y_1^*) = (108, 221)$, $(x_2^*, y_2^*) = (104, 198)$, $(x_3^*, y_3^*) = (99, 182)$, $(x_4^*, y_4^*) = (89, 189)$,
$(x_5^*, y_5^*) = (105, 187)$, $(x_6^*, y_6^*) = (117, 174)$, $(x_6^*, y_6^*) = (112, 168)$, $(x_7^*, y_7^*) = (117, 175)$,
....

Damit erzeugen wir Bootstrap-Realisierungen von T:

$$t_i^* := \frac{x_i^*/11037}{y_i^*/11034}, \quad i = 1, 2, \ldots.$$

Die Verteilungsfunktion von T kann man nun durch die empirische Verteilungsfunktion von $t_1^*, \ldots, t_B^*$ schätzen:

$$\hat{G}(t) := \frac{1}{B} \sum_{i=1}^{B} \mathbb{1}_{t_i^* \le t}, \quad t \in \mathbb{R}.$$

Wir haben $B = 10^5$ gewählt. Abbildung 1 zeigt die von uns berechnete Funktion $\hat{G}(t)$. Nun berechnen wir die Quantile

$$q_{0.025} \approx 0.43, \quad q_{0.975} \approx 0.69$$

von $\hat{G}(t)$ (rote Linien auf Abbildung 1) und erhalten das folgende (zweiseitige, approximative) Konfidenzintervall zum Niveau 0.95 für p/q:

$$[0.43, 0.69].$$

Dieses Intervall überdeckt den Wert 1 nicht, also unterscheidet sich p/q signifikant von 1.

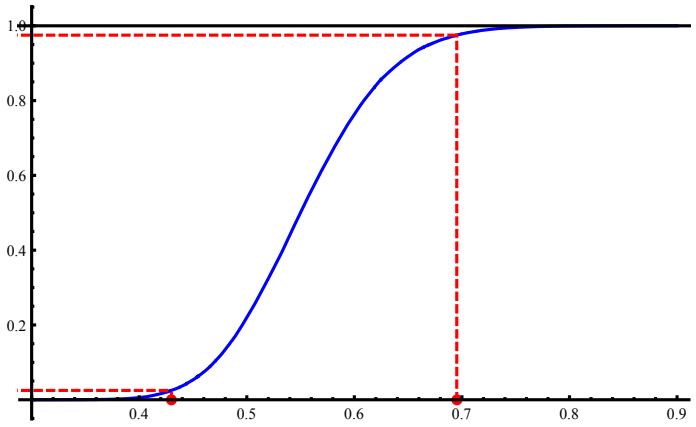


ABBILDUNG 1. Die Funktion $\hat{G}(t)$, ein Schätzer für die Verteilungsfunktion von T. Rote Linien zeigen die Berechnung der Quantile $q_{0.025}$ und $q_{0.975}$.

***