Stochastik I (Statistik)

Skript

Jun.-Prof. Dr. Zakhar Kabluchko

Universität Ulm
Institut für Stochastik

LaTeX-Version von Judith Schmidt

Inhaltsverzeichnis

Vorwort

Dies ist ein Skript zur Vorlesung "Stochastik I (Statistik)", die an der Universität Ulm im Sommersemester 2013 gehalten wurde. Die erste LaTeX-Version des Skripts wurde von Judith Schmidt erstellt. Danach wurde das Skript von mir korrigiert und ergänzt. In Zukunft soll das Skript um ein weiteres Kapitel (Lineare Regression) ergänzt werden.

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

`zakhar DOT kabluchko AT uni-ulm DOT de`

schreiben.

27. September 2013

Zakhar Kabluchko

Literatur

Es gibt sehr viele Lehrbücher über Statistik, z. B.

1. J. Lehn, H. Wegmann. *Einführung in die Statistik.*
2. H. Pruscha. *Vorlesungen über Mathematische Statistik.*
3. H. Pruscha. *Angewandte Methoden der Mathematischen Statistik.*
4. V. Rohatgi. *Statistical Inference.*
5. G. Casella, R. L. Berger. *Statistical Inference.*
6. K. Bosch. *Elementare Einführung in die angewandte Statistik: Mit Aufgaben und Lösungen.*

Folgende Lehrbücher behandeln sowohl Wahrscheinlichkeitstheorie als auch Statistik:

1. H. Dehling und B. Haupt. *Einführung in die Wahrscheinlichkeitstheorie und Statistik.* Springer–Verlag.
2. U. Krengel. *Einführung in die Wahrscheinlichkeitstheorie und Statistik.* Vieweg–Verlag.
3. H.–O. Georgii. *Stochastik: Einführung in die Wahrscheinlichkeitstheorie und Statistik.* De Gruyter.

KAPITEL 1

Stichproben und Stichprobenfunktion

In diesem Kapitel werden wir auf Stichproben und Stichprobenfunktionen eingehen. Als Einstieg beginnen wir mit zwei kleinen Beispielen.

1.1. Stichproben

Beispiel 1.1.1. Wir betrachten ein Experiment, bei dem eine physikalische Konstante (z.B. die Lichtgeschwindigkeit) bestimmt werden soll. Da das Ergebnis des Experiments fehlerbehaftet ist, wird das Experiment mehrmals durchgeführt. Wir bezeichnen die Anzahl der Messungen mit n. Das Resultat der i-ten Messung sei mit $x_i \in \mathbb{R}$ bezeichnet. Fassen wir nun die Resultate aller Messungen zusammen, so erhalten wir eine sogenannte *Stichprobe*

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

Die Anzahl der Messungen (also n) nennen wir den *Stichprobenumfang*. Die Menge aller vorstellbaren Stichproben wird der *Stichprobenraum* genannt und ist in diesem Beispiel $\mathbb{R}^n$.

Beispiel 1.1.2. Wir betrachten eine biometrische Studie, in der ein gewisses biometrisches Merkmal, z.B. die Körpergröße, in einer bestimmten Population untersucht werden soll. Da die Population sehr groß ist, ist es nicht möglich, alle Personen in der Population zu untersuchen. Deshalb werden für die Studie n Personen, die wir mit $1, \ldots, n$ bezeichnen, aus der Population ausgewählt und gewogen. Mit $x_i \in \mathbb{R}$ wird das Gewicht von Person i bezeichnet. Das Ergebnis der Studie kann man dann in einer Stichprobe

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

zusammenfassen. Die Auswahl der n Personen aus der Population erfolgt zufällig und kann somit als ein Zufallsexperiment betrachten werden. Die Grundmenge dieses Experiments sei mit Ω bezeichnet. Die genaue Gestalt von Ω wird im Weiteren keine Rolle spielen. Das Gewicht von Person i kann als eine Zufallsvariable $X_i : \Omega \to \mathbb{R}$ aufgefasst werden. Den Zusammenhang zwischen $(X_1, \ldots, X_n)$ und $(x_1, \ldots, x_n)$ kann man folgendermaßen beschreiben. Jede konkrete Auswahl von n Personen aus der Population entspricht einem Element (Ausgang) ω in der Grundmenge Ω. Das Gewicht der i-ten Person ist dann der Wert der Funktion X_i an der Stelle ω, also $X_i(\omega)$. Es gilt somit

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

Man sagt auch, dass $(x_1, \ldots, x_n)$ eine *Realisierung* des Zufallsvektors $(X_1, \ldots, X_n)$ ist. Oft nennt man $(x_1, \ldots, x_n)$ die *konkrete Stichprobe* und $(X_1, \ldots, X_n)$ die *Zufallsstichprobe*. Es sei noch einmal bemerkt, dass x_i reelle Zahlen, wohingegen $X_i : \Omega \to \mathbb{R}$ Zufallsvariablen (also Funktionen auf einem Wahrscheinlichkeitsraum) sind.

Im Folgenden werden wir sehr oft annehmen, dass $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ unabhängige und identisch verteilte Zufallsvariablen sind. Die Verteilungsfunktion von X_i bezeichnen wir mit

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

1.2. Stichprobenfunktionen, empirischer Mittelwert und empirische Varianz

DEFINITION 1.2.1. Eine beliebige Borel-Funktion $\varphi : \mathbb{R}^n \to \mathbb{R}^m$ heißt *Stichprobenfunktion.*

DEFINITION 1.2.2. Bezeichne mit $X = (X_1, \ldots, X_n) : \Omega \to \mathbb{R}^n$ eine Zufallsstichprobe. Dann heißt die zusammengesetzte Funktion $\varphi \circ X : \Omega \to \mathbb{R}^m$ eine *Statistik*:

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

Im Folgenden werden wir zwei wichtige Beispiele von Stichprobenfunktionen, den empirischen Mittelwert und die empirische Varianz, betrachten. Es sei $(x_1, \ldots, x_n) \in \mathbb{R}^n$ eine Stichprobe.

DEFINITION 1.2.3. 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

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

Dabei ist $\bar{x}_n$ eine Stichprobenfunktion und $\bar{X}_n$ eine Statistik. Im Weiteren werden wir meistens keinen Unterschied zwischen diesen Begriffen machen.

SATZ 1.2.4. *Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit $\mu = \mathbb{E}X_i$ und $\sigma^2 = \operatorname{Var} X_i$. 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 1.2.5. In der Statistik nimmt man an, dass die Stichprobe $(x_1, \ldots, x_n)$ bekannt ist und fragt dann, wie anhand dieser Stichprobe verschiedene Kenngrößen der Zufallsvariablen X_i (etwa der Erwartungswert, die Varianz, die Verteilungsfunktion) "geschätzt" werden können. Zum Beispiel bietet sich der empirische Mittelwert $\bar{x}_n$ (oder $\bar{X}_n$) als ein natürlicher Schätzer für den theoretischen Erwartungswert $\mu = \mathbb{E}X_i$. Der obige Satz zeigt, dass durch

eine solche Schätzung 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.

DEFINITION 1.2.6. Die *empirische Varianz* oder die *Stichprobenvarianz* 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 im folgenden Satz klar.

SATZ 1.2.7. *Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit $\mathbb{E}X_i = \mu$ und $\operatorname{Var} X_i = \sigma^2$. Dann gilt*

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

BEWEIS. Zuerst beweisen wir die Formel

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

Das geht folgendermaßen:

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

Nun 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

$$\mathbb{E}[X_i^2] = \operatorname{Var} X_i + (\mathbb{E}X_i)^2 = \sigma^2 + \mu^2$$

und (mit Satz 1.2.4)

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

□

BEMERKUNG 1.2.8. 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 1.2.9. 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 von $-\frac{1}{n}\sigma^2$.

BEMERKUNG 1.2.10. Die *empirische Standardabweichung* ist definiert durch

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

BEMERKUNG 1.2.11. 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).

BEMERKUNG 1.2.12. Das Stichprobenmittel ist kein robuster Parameter, d.h. es wird stark von Ausreißern beeinflusst. Dies zeigt folgendes Beispiel: 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 20 um, also $(1, 2, 2, 2, 1, 1, 1, 20)$, dann gilt $\bar{x}_n = 3.75$. 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.

KAPITEL 2

Ordnungsstatistiken und Quantile

Um robuste Lage- und Streuungsparameter einführen zu können, benötigen wir Ordnungsstatistiken und Quantile.

2.1. Ordnungsstatistiken und Quantile

DEFINITION 2.1.1. Sei $(x_1, \ldots, x_n) \in \mathbb{R}^n$ eine Stichprobe. Wir können die Elemente der Stichprobe aufsteigend anordnen:

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

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 2.1.2. Der *Stichprobenmedian* ist gegeben durch

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

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

BEISPIEL 2.1.3. 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 $\text{med}_8 = \frac{1+2}{2} = 1.5$.
Als zweite Stichprobe betrachten wir

$$(y_1, \ldots, y_8) = (1, 2, 2, 2, 1, 1, 1, 20).$$

Die Ordnungsstatistiken sind gegeben durch

$$(y_{(1)}, \ldots, y_{(n)}) = (1, 1, 1, 1, 2, 2, 2, 20),$$

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

BEMERKUNG 2.1.4. Im Allgemeinen gilt $\text{med}_n \neq \bar{x}_n$.

Ein weiterer robuster Lageparameter ist das getrimmte Mittel.

DEFINITION 2.1.5. 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 2.1.6. 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 Median ist somit das $\frac{1}{2}$-Quantil.

DEFINITION 2.1.7. 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 2.1.8. 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:

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

BEISPIEL 2.1.9. Weitere Lageparameter, die in der Statistik vorkommen:

(1) Das Bereichsmittel $\frac{x_{(n)}+x_{(1)}}{2}$ (nicht robust).
(2) Das Quartilsmittel $\frac{q_{0,25}+q_{0,75}}{2}$ (robust).

BEISPIEL 2.1.10. 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.

2.2. Verteilung der Ordnungsstatistiken

SATZ 2.2.1. *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:

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

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 $X_1, \ldots, X_n$ sind unabhängig und identisch verteilt mit $\mathbb{P}[X_i \leq t] = F(t)$. Somit ist die Zufallsvariable N binomialverteilt:

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

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

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

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 2.2.2. Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit Dichte f und Verteilungsfunktion F. Man zeige, dass für alle $1 \leq i < j \leq n$ die gemeinsame Dichte der Ordnungsstatistiken $X_{(i)}$ und $X_{(j)}$ durch die folgende Formel gegeben ist:

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

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

SATZ 2.2.3. *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 2.2.4. 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 2.2.5. 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

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

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.

AUFGABE 2.2.6. Seien $X_1, \ldots, X_n$ unabhängig und gleichverteilt auf dem Intervall $[0, 1]$. Man zeige, dass

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

KAPITEL 3

Empirische Verteilungsfunktion

3.1. 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 3.1.1. 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 3.1.2. 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 3.1.3. 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.

DEFINITION 3.1.4. 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 3.1.5. *Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit Verteilungsfunktion F. Dann gilt*

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

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

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:*

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

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

BEMERKUNG 3.1.6. 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 $Y_i = \mathbb{1}_{X_i \le t}$ ein. Diese sind unabhängig und identisch verteilt (da $X_1, X_2, \ldots,$ unabhängig und identisch verteilt sind) mit

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

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

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

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

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

□

3.2. 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 3.2.1. 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 3.2.2. 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

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

Die theoretische Verteilung $\widehat{\mu}_n$ ordnet jeder Menge A die Wahrscheinlichkeit, dass X einen Wert in A annimmt, zu. Die empirische Verteilung 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 3.2.3. 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:

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

3.3. Satz von Gliwenko–Cantelli

Wir haben in Teil 3 von Satz 3.1.5 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 3.3.1. Der *Kolmogorov-Abstand* zwischen der empirischen Verteilungsfunktion $\widehat{F}_n$ und der theoretischen Verteilungsfunktion F wird folgendermaßen definiert:

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

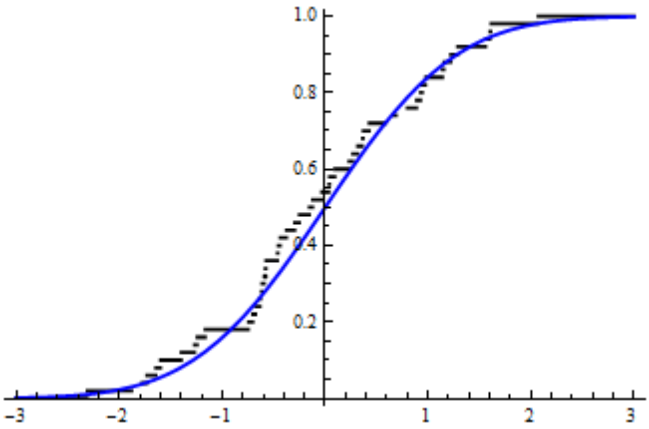

ABBILDUNG 1. Die schwarz dargestellte Funktion ist die empirische Verteilungsfunktion einer Stichprobe vom Umfang $n = 50$ aus der Standardnormalverteilung. Die blaue Kurve ist die Verteilungsfunktion der Normalverteilung. Der Satz von Gliwenko–Cantelli besagt, dass bei steigendem Stichprobenumfang n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

SATZ 3.3.2 (von 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 3.3.3. 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 $\varepsilon > 0$:

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

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

BEMERKUNG 3.3.4. Für jedes $t \in \mathbb{R}$ gilt offenbar

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

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 3.1.5, Teil 3 entspricht. Somit ist der Satz von Gliwenko–Cantelli stärker als Satz 3.1.5, Teil 3.

Beweis von Satz 3.3.2. 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 3.1.5, 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 4

Dichteschätzer

Es seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit Dichte f und Verteilungsfunktion F. Es sei $(x_1, \ldots, x_n)$ eine Realisierung von $(X_1, \ldots, X_n)$. In diesem Kapitel beschäftigen wir uns mit dem folgenden Problem: Man schätze die Dichte f anhand der Stichprobe $(x_1, \ldots, x_n)$.

Zunächst einmal kann man die folgende Idee ausprobieren. Wir können die Verteilungsfunktion F 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.

4.1. 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.

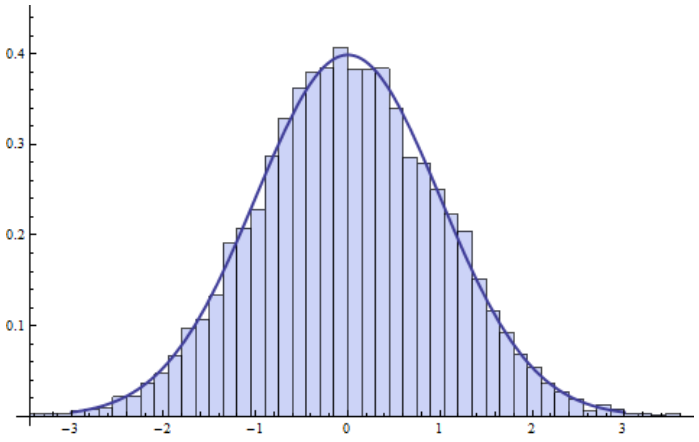

ABBILDUNG 1. Das Histogramm einer standardnormalverteilten Stichprobe vom Umfang $n = 10000$. Die glatte blaue Kurve ist die Dichte der Standardnormalverteilung.

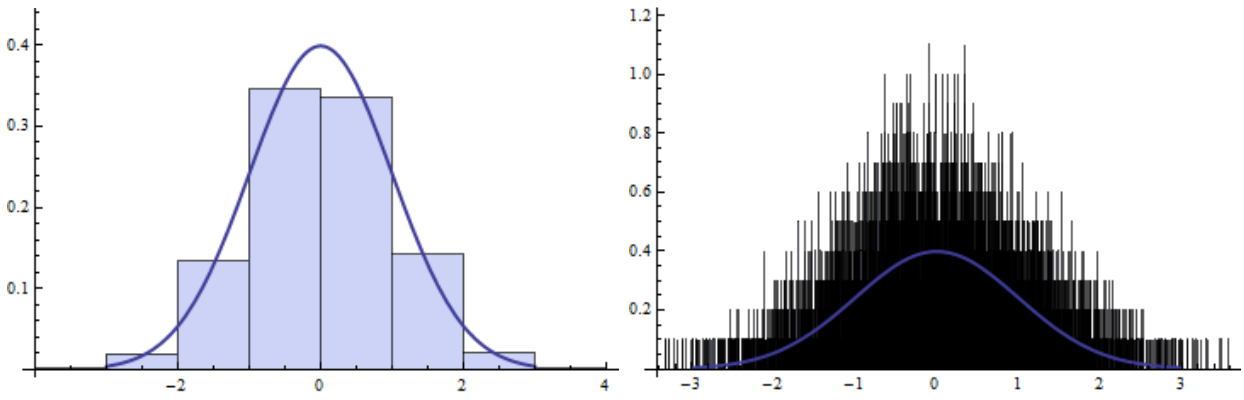

ABBILDUNG 2. Das Histogramm einer standardnormalverteilten Stichprobe vom Umfang 10000 mit einer schlecht gewählten Bandbreite $h = c_i - c_{i-1}$. Links: Die Bandbreite ist zu groß. Rechts: Die Bandbreite ist zu klein. In beiden Fällen zeigt die glatte blaue Kurve die Dichte der Standardnormalverteilung.

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 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 $f_i \geq 0$.

Das Histogramm hat den Nachteil, dass die Wahl der c_i's bzw. die Wahl der Bandbreite h willkürlich ist. Ist 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 meistens weder lokal konstant noch stetig ist. Im nächsten Abschnitt betrachten wir einen Dichteschätzer, der zumindest von diesem zweiten Nachteil frei ist.

4.2. Kerndichteschätzer

Wir werden nun eine bessere Methode zur Schätzung der Dichte betrachten, den Kerndichteschätzer.

DEFINITION 4.2.1. Ein *Kern* ist eine messbare Funktion $K : \mathbb{R} \to [0, \infty)$, so dass

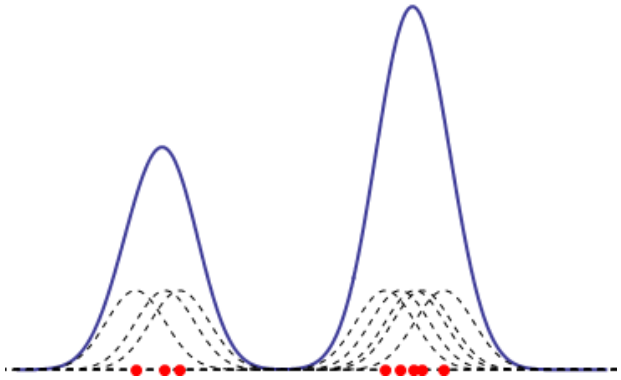

ABBILDUNG 3. Kerndichteschätzer.

(1) $K(x) \geq 0$ für alle $x \in \mathbb{R}$ und
(2) $\int_{\mathbb{R}} K(x)dx = 1$.

Die Bedingungen in der Definition eines Kerns sind somit die gleichen, wie in der Definition einer Dichte.

DEFINITION 4.2.2. 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 4.2.3. 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

$$\int_{\mathbb{R}} \frac{1}{hn} K\left(\frac{x - x_i}{h}\right) dx = \frac{1}{n} \int_{\mathbb{R}} K(y)dy = \frac{1}{n}.$$

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:

$$\int_{\mathbb{R}} \hat{f}_n(x)dx = 1.$$

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 4.2.4. 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 Kurven zeigen die Beiträge der einzelnen Punkte. In diesem Fall benutzen

wir den Gauß–Kern, der unten eingeführt wird. Die Summe der einzelnen Beiträge ist der Kerndichteschätzer $\hat{f}_n$, der durch die blaue 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.

Beispiel 4.2.5. 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 4.2.6. Der *Gauß-Kern* ist nichts Anderes, als die Dichte der Standardnormalverteilung:

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

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 dass das arithmetische Mittel solcher Dichten.

Beispiel 4.2.7. 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 4.2.8. 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.

Die optimale Wahl der Bandbreite h ist ein nichttriviales Problem, mit dem wir uns in dieser Vorlesung nicht beschäftigen werden.

KAPITEL 5

Methoden zur Konstruktion von Schätzern

5.1. Parametrisches Modell

Sei $(x_1, \ldots, x_n)$ eine Stichprobe. In der parametrischen Statistik nimmt man an, dass die Stichprobe $(x_1, \ldots, x_n)$ eine Realisierung von unabhängigen und identisch verteilten Zufallsvariablen $(X_1, \ldots, X_n)$ mit Verteilungsfunktion $F_\theta(x)$ ist. Dabei hängt die Verteilungsfunktion F_θ von einem unbekannten Wert (*Parameter*) θ ab. In den meisten Fällen nimmt man außerdem an, dass entweder die Zufallsvariablen X_i für alle Werte des Parameters θ absolut stetig sind und eine Dichte h_θ besitzen, oder dass sie für alle Werte von θ diskret sind und eine Zähldichte besitzen, die ebenfalls mit h_θ bezeichnet wird. Die Aufgabe der parametrischen Statistik besteht darin, den unbekannten Parameter θ anhand der bekannten Stichprobe $(x_1, \ldots, x_n)$ zu schätzen.

Die Menge aller möglichen Werte des Parameters θ wird der *Parameterraum* genannt und mit Θ bezeichnet. In den meisten Fällen ist $\theta = (\theta_1, \ldots, \theta_p)$ ein Vektor mit Komponenten $\theta_1, \ldots, \theta_p$. In diesem Fall muss der Parameterraum Θ eine Teilmenge von $\mathbb{R}^p$ sein.

Um den Parameter θ anhand der Stichprobe $(x_1, \ldots, x_n)$ zu schätzen, konstruiert man einen Schätzer.

Definition 5.1.1. Ein *Schätzer* ist eine Abbildung

$$\hat{\theta} : \mathbb{R}^n \to \Theta, \quad (x_1, \ldots, x_n) \mapsto \hat{\theta}(x_1, \ldots, x_n).$$

Man muss versuchen, den Schätzer so zu konstruieren, dass $\hat{\theta}(x_1, \ldots, x_n)$ den wahren Wert des Parameters θ möglichst gut approximiert. Wie das geht, werden wir im Weiteren sehen.

Beispiel 5.1.2. Wir betrachten ein physikalisches Experiment, bei dem eine physikalische Konstante (z.B. die Lichtgeschwindigkeit) bestimmt werden soll. Bei n unabhängigen Messungen der Konstanten ergaben sich die Werte $(x_1, \ldots, x_n)$. Normalerweise nimmt man an, dass diese Stichprobe eine Realisierung von n unabhängigen und identisch verteilten Zufallsvariablen $(X_1, \ldots, X_n)$ mit einer Normalverteilung ist:

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

Dabei ist μ der wahre Wert der zu bestimmenden Konstanten und σ^2 die quadratische Streuung des Experiments. Beide Parameter sind unbekannt. Somit besteht das Problem, den Parameter $\theta = (\mu, \sigma^2)$ aus den gegebenen Daten $(x_1, \ldots, x_n)$ zu schätzen. In diesem Beispiel ist der Parameterraum gegeben durch

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

Die Dichte von X_i ist gegeben durch (siehe auch Abbildung 1)

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

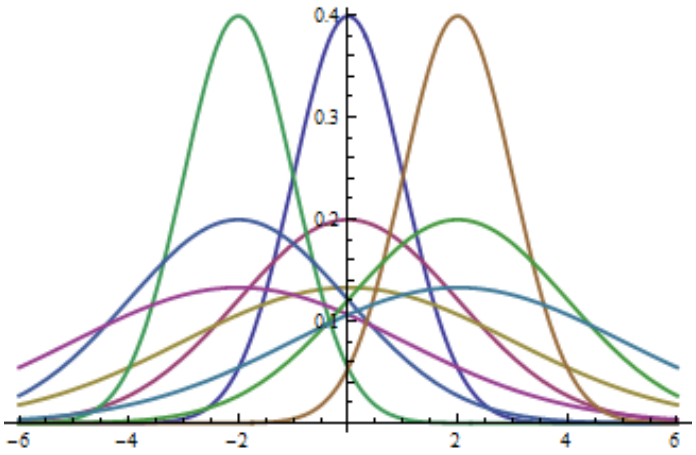

ABBILDUNG 1. Das Bild zeigt die Dichten der Normalverteilungen, die zu verschiedenen Werten der Parameter μ und σ^2 gehören. Die Aufgabe der parametrischen Statistik ist es, zu entscheiden, zu welchen Parameterwerten eine gegebene Stichprobe gehört.

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

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

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*.

An dieser Stelle müssen wir noch eine Notation einführen. Um im parametrischen Modell die Verteilung der Zufallsvariablen $X_1, \ldots, X_n$ eindeutig festzugelegen, muss man den Wert des Parameters θ angeben. Bevor man von der Wahrscheinlichkeit eines mit $X_1, \ldots, X_n$ verbundenen Ereignisses spricht, muss man also sagen, welchen Wert der Parameter θ annehmen soll. Wir werden deshalb sehr oft die folgende Notation verwenden. Mit $\mathbb{P}_\theta[A]$ bezeichnen wir die Wahrscheinlichkeit eines Ereignisses A unter der Annahme, dass die Zufallsvariablen X_i unabhängig und identisch verteilt mit Verteilungsfunktion F_θ (bzw. mit Dichte/ Zähldichte h_θ) sind. Dabei können sich $\mathbb{P}_{\theta_1}[A]$ und $\mathbb{P}_{\theta_2}[A]$ durchaus unterscheiden. Analog bezeichnen wir mit $\mathbb{E}_\theta Z$ und $\mathrm{Var}_\theta Z$ den Erwartungswert bzw. die Varianz einer Zufallsvariable Z unter der Annahme, dass die Zufallsvariablen X_i unabhängig und identisch verteilt mit Verteilungsfunktion F_θ (bzw. mit Dichte/ Zähldichte h_θ) sind.
Die Zufallsvariablen $X_1, \ldots, X_n$ kann man sich als messbare Funktionen auf einem Messraum $(\Omega, \mathcal{A})$ denken. In der Wahrscheinlichkeitstheorie musste man außerdem ein Wahrscheinlichkeitsmaß $\mathbb{P}$ auf diesem Raum angeben. Im parametrischen Modell brauchen wir nicht *ein*

Wahrscheinlichkeitsmaß, sondern eine durch θ parametrisierte *Familie* von Wahrscheinlichkeitsmaßen $\{\mathbb{P}_\theta : \theta \in \Theta\}$ auf $(\Omega, \mathcal{A})$. Je nachdem welchen Wert der Parameter θ annimmt, können wir eines dieser Wahrscheinlichkeitsmaße verwenden.

5.2. Momentenmethode

Wie in der parametrischen Statistik üblich, nehmen wir an, dass die Stichprobe $(x_1, \ldots, x_n)$ eine Realisierung der unabhängigen und identisch verteilten Zufallsvariablen $(X_1, \ldots, X_n)$ mit Verteilungsfunktion F_θ ist. Dabei ist $\theta = (\theta_1, \ldots, \theta_p) \in \mathbb{R}^p$ der unbekannte Parameter. Für die Momentenmethode brauchen wir die folgenden Begriffe.

DEFINITION 5.2.1. 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 5.2.2. 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 $\theta_1, \ldots, \theta_p$. Um p unbekannte Parameter zu finden, brauchen wir normalerweise p Gleichungen. Wir betrachten also ein System aus p Gleichungen mit p Unbekannten:

$$m_1(\theta_1, \ldots, \theta_p) = \hat{m}_1, \quad \ldots, \quad m_p(\theta_1, \ldots, \theta_p) = \hat{m}_p.$$

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

BEISPIEL 5.2.3. 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 5.2.4. 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 5.1.2 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 5.2.5. Momentenmethode für den Parameter der Poisson–Verteilung Poi(θ).
In diesem Beispiel betrachten wir ein Portfolio aus n Versicherungsverträgen. Es sei $x_i \in \{0, 1, \ldots\}$ 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 $(x_1, \ldots, x_n)$ eine Realisierung von n unabhängigen und identisch verteilten Zufallsvariablen $(X_1, \ldots, X_n)$ ist, die eine Poissonverteilung mit einem unbekannten Parameter $\theta \geq 0$ haben.

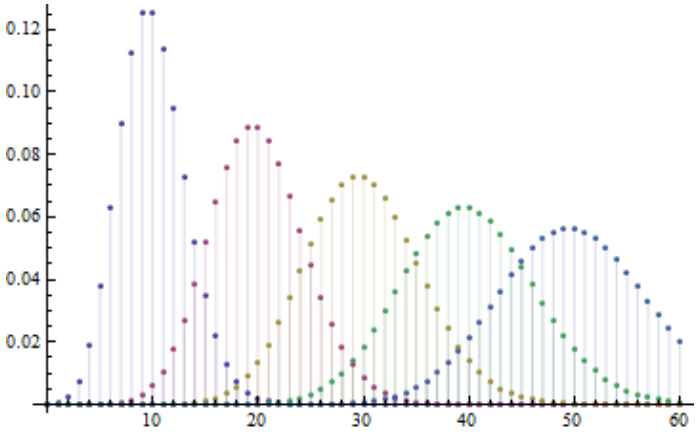

Abbildung 2. Zähldichten der Poissonverteilungen, die zu verschiedenen Werten des Parameters θ gehören.

Wir schätzen θ mit der Momentenmethode. Da der Erwartungswert einer Poi(θ)–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

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

5.3. Maximum–Likelihood–Methode

Die Maximum–Likelihood–Methode wurde von Carl Friedrich Gauß entdeckt und von Ronald Fisher weiterentwickelt. Die Maximum–Likelihood–Methode ist (wie auch die Momentenmethode) ein Verfahren, um Schätzer für die unbekannten Komponenten des Parametervektors $\theta = (\theta_1, \ldots, \theta_p)$ zu gewinnen. Sei $(x_1, \ldots, x_n)$ eine Stichprobe. Wir werden annehmen, dass entweder alle Verteilungen aus der parametrischen Familie $\{F_\theta : \theta \in \Theta\}$ diskret oder alle Verteilungen absolut stetig sind.

Der diskrete Fall. Seien zuerst die Zufallsvariablen X_i für alle Werte des Parameters θ diskret. Wir bezeichnen die Zähldichte von X_i mit h_θ. Dann ist die *Likelihood–Funktion*

gegeben durch

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

Die Likelihood–Funktion hängt sowohl von der Stichprobe, als auch vom Parameterwert θ ab, wir werden sie aber hauptsächlich als Funktion von θ auffassen. Wegen der Unabhängigkeit von $X_1, \ldots, X_n$ gilt

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

Die Likelihood–Funktion ist somit die Wahrscheinlichkeit, die gegebene Stichprobe $(x_1, \ldots, x_n)$ zu beobachten, wobei diese Wahrscheinlichkeit als Funktion des Parameters θ aufgefasst wird.

Der absolut stetige Fall. Seien nun die Zufallsvariablen X_i für alle Werte des Parameters θ absolut stetig. Wir bezeichnen die Dichte von X_i mit h_θ. In diesem Fall definieren wir die Likelihood–Funktion wie folgt:

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

In beiden Fällen besteht die Idee der Maximum–Likelihood–Methode darin, einen Wert von θ zu finden, der die Likelihood–Funktion maximiert:

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

Der *Maximum–Likelihood–Schätzer* (oder der *ML–Schätzer*) ist definiert durch

$$\hat{\theta}_{ML} = \operatorname{argmax}_{\theta \in \Theta} L(\theta).$$

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

Beispiel 5.3.1. Maximum–Likelihood–Schätzer für den Parameter der Bernoulli–Verteilung Bern(θ).
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. Das Ergebnis des Experiments, bei dem die Münze n Mal geworfen wird, können wir in einer Stichprobe $(x_1, \ldots, x_n) \in \{0, 1\}^n$ darstellen, wobei $x_i = 1$ ist, wenn die Münze bei Wurf i "Kopf" gezeigt hat, und $x_i = 0$ ist, wenn die Münze bei Wurf i "Zahl" gezeigt hat. 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 sind mit Parameter θ. Es handelt sich um diskrete Zufallsvariablen und die Zähldichte ist gegeben durch

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

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 3.

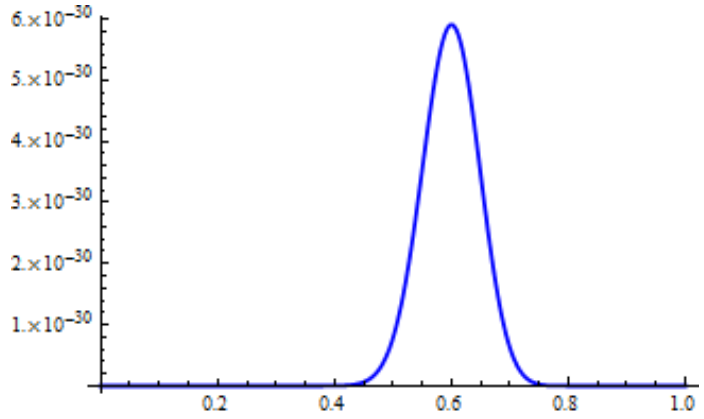

ABBILDUNG 3. Die Likelihood–Funktion $L(\theta) = \theta^{60}(1-\theta)^{40}$, $\theta \in [0,1]$, aus Beispiel 5.3.1, erstes Modell. Das Maximum wird an der Stelle $\theta = 0.6$ erreicht.

Wir benötigen eine Fallunterscheidung.

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

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

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

$$\lim_{\theta \downarrow 0} L(\theta) = \lim_{\theta \uparrow 1} L(\theta) = 0.$$

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

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

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

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

Beispiel 5.3.2. Maximum–Likelihood–Schätzer für den Parameter der Poisson–Verteilung Poi(θ).
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 Poi(θ) ist gegeben durch

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

Dies führt zu folgender Likelihood–Funktion

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

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

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

Nun betrachten wir einige Beispiele zur Maximum–Likelihood–Methode im Falle der absolut stetigen Verteilungen.

Beispiel 5.3.3. Maximum–Likelihood–Schätzer für den Endpunkt der Gleichverteilung 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 4. Wir sollen θ anhand der Stichprobe $(x_1, \ldots, x_n)$ rekonstruieren.

Abbildung 4. 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 Zufallsvariablen X_i sind somit absolut stetig und ihre Dichte ist gegeben durch

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

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 $x_{(n)} = \max\{x_1, \ldots, x_n\}$ die maximale Beobachtung dieser Stichprobe. Der Graph der Likelihood–Funktion ist auf Abbildung 5 zu sehen.

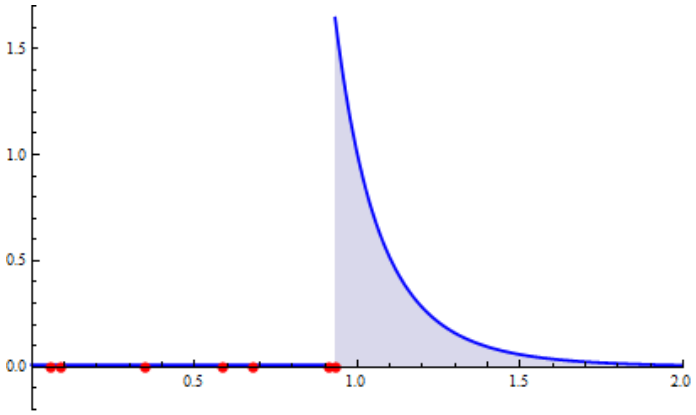

Abbildung 5. Maximum–Likelihood–Schätzung des Endpunktes der Gleichverteilung. Die roten Punkte zeigen die Stichprobe. Die blaue Kurve ist die Likelihood–Funktion $L(\theta)$.

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

$$\hat{\theta}_{ML} = \operatorname*{argmax}_{\theta > 0} L(\theta) = x_{(n)}.$$

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

Beispiel 5.3.5. 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

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

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:

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

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

Beispiel 5.3.6. 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 $\hat{n}$.

5.4. 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 5.4.1. 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 $S \sim \text{Bin}(n, \theta_1)$. Unter $\theta = \theta_2$ (also für Autofahrer vom Typ 2) ist $S \sim \text{Bin}(n, \theta_2)$. 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:

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

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:

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

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 5.4.2. Das Wort "a priori" steht für "vor dem Experiment", das Wort "a posteriori" steht für "nach dem Experiment".

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.
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, dass die Zufallsvariablen $X_1, \ldots, X_n$ diskret sind. Diese Wahrscheinlichkeit berechnet man mit der Bayes–Formel:

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

Wir haben dabei angenommen, dass X_i diskret sind, die Endformel macht 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 $A(t) \propto C \cdot B(t)$ 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: 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 $\mathbb{R}^p$ und einer Dichte $q(\tau)$. Dabei bezeichnen wir mit $\tau \in \mathbb{R}^p$ 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:

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

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

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

Aufgabe 5.4.3. 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 5.4.4. 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:

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

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 5.2.3) oder mit der Maximum–Likelihood–Methode (Beispiel 5.3.1) 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

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

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:

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

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 5.4.5. (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 $\mathrm{Beta}(\alpha, \beta)$. 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:

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

Die a–posteriori–Verteilung stammt also aus derselben Betafamilie, wie die a–priori–Verteilung, bloß die Parameter sind anders. Der Bayes–Schätzer für θ ist der Erwartungswert der a–posteriori–Betaverteilung:

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

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 5.4.6 (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.

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

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 $\hat{\lambda}_{\mathrm{Bayes}}$.

Aufgabe 5.4.7 (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 $\hat{p}_{\mathrm{Bayes}}$.

AUFGABE 5.4.8 (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 5.4.9 (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 $\tau = 1/\sigma^2$.

KAPITEL 6

Güteeigenschaften von Schätzern

Wir erinnern an die Definition des parametrischen Modells. Sei $\{h_\theta : \theta \in \Theta\}$, wobei $\Theta \subset \mathbb{R}^m$, eine Familie von Dichten oder Zähldichten. Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit Dichte oder Zähldichte h_θ. Dabei ist θ der zu schätzende Parameter. Um die Notation zu vereinfachen, werden wir in diesem Kapitel nicht zwischen den Zufallsvariablen $(X_1, \ldots, X_n)$ und deren Realisierung $(x_1, \ldots, x_n)$ unterscheiden. Ein Schätzer ist eine beliebige (Borel–messbare) Funktion

$$\hat{\theta} : \mathbb{R}^n \to \Theta, \quad (X_1, \ldots, X_n) \mapsto \hat{\theta}(X_1, \ldots, X_n).$$

Die Aufgabe eines Schätzers ist es, den richtigen Wert von θ möglichst gut zu erraten. Im Folgenden definieren wir einige Eigenschaften von Schätzern, die uns erlauben, "gute" Schätzer von "schlechten" Schätzern zu unterscheiden.

6.1. Erwartungstreue, Konsistenz, asymptotische Normalverteiltheit

In diesem Kapitel sei der Parameterraum Ω eine Teilmenge von $\mathbb{R}^m$.

Definition 6.1.1. Ein Schätzer $\hat{\theta}$ heißt *erwartungstreu* (oder *unverzerrt*), falls

$$\mathbb{E}_\theta[\hat{\theta}(X_1, \ldots, X_n)] = \theta \text{ für alle } \theta \in \Theta.$$

Bemerkung 6.1.2. Damit diese Definition Sinn macht, muss man voraussetzen, dass die Zufallsvariable (bzw. Zufallsvektor) $\hat{\theta}(X_1, \ldots, X_n)$ integrierbar ist.

Definition 6.1.3. Der *Bias* (die *Verzerrung*) eines Schätzers $\hat{\theta}$ ist

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

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

Bemerkung 6.1.4. Ein Schätzer $\hat{\theta}$ ist genau dann erwartungstreu, wenn $\mathrm{Bias}_\theta(\hat{\theta}) = 0$ für alle $\theta \in \Theta$.

Beispiel 6.1.5. In diesem Beispiel werden wir verschiedene Schätzer für den Endpunkt der Gleichverteilung konstruieren. Es seien $X_1, \ldots, X_n$ 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 $X_1, \ldots, X_n$. Folgende Schätzer für θ erscheinen natürlich.

1. Der 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 wird θ von diesem Schätzer immer unterschätzt.

2. Wir versuchen nun den Schätzer $\hat{\theta}_1$ zu verbessern, indem wir ihn vergrößern. Wir würden ihn gerne um $\theta - X_{(n)}$ vergößern, 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.

$$X_{(1)} \stackrel{!}{=} \theta - X_{(n)}.$$

Lösen wir diese Gleichung bzgl. θ, so erhalten wir den Schätzer

$$\hat{\theta}_2(X_1, \ldots, X_n) = X_{(n)} + X_{(1)}.$$

3. 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:

$$\theta - X_{(n)} \stackrel{!}{=} \frac{1}{n}(X_{(1)} + (X_{(2)} - X_{(1)}) + (X_{(3)} - X_{(2)}) + \ldots + (X_{(n)} - X_{(n-1)})).$$

Da auf der rechten Seite eine Teleskop-Summe steht, erhalten wir die Gleichung

$$\theta - X_{(n)} \stackrel{!}{=} \frac{1}{n} X_{(n)}.$$

Auf diese Weise ergibt sich der Schätzer

$$\hat{\theta}_3(X_1, \ldots, X_n) = \frac{n+1}{n} X_{(n)}.$$

4. 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 6.1.6. 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 erwartungstreue Schätzer geben kann. Die Frage ist nun, welcher Schätzer der beste ist.

DEFINITION 6.1.7. Sei $\Theta = (a, b) \subset \mathbb{R}$ ein Intervall. Der *mittlere quadratische Fehler* (*mean square error*, MSE) eines Schätzers $\hat{\theta} : \mathbb{R}^n \to \mathbb{R}$ ist definiert durch

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

Wir fassen $\mathrm{MSE}_\theta(\hat{\theta})$ als eine Funktion von $\theta \in (a, b)$ auf. Damit die obige Definition Sinn hat, muss man voraussetzen, dass $\hat{\theta}(X_1, \ldots, X_n)$ eine quadratisch integrierbare Zufallsvariable ist.

LEMMA 6.1.8. *Es gilt folgender 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, schreiben wir in diesem Beweis $\hat{\theta}$ für $\hat{\theta}(X_1, \ldots, X_n)$. 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 6.1.9. Ist $\hat{\theta}$ erwartungstreu, so gilt $\mathrm{Bias}_\theta(\hat{\theta}) = 0$ für alle $\theta \in \Theta$ und somit vereinfacht sich Lemma 6.1.8 zu

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

DEFINITION 6.1.10. Seien $\hat{\theta}_1$ und $\hat{\theta}_2$ zwei Schätzer. Wir sagen, dass $\hat{\theta}_1$ *besser* als $\hat{\theta}_2$ ist, falls

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

BEMERKUNG 6.1.11. Falls $\hat{\theta}_1$ und $\hat{\theta}_2$ erwartungstreu sind, dann ist $\hat{\theta}_1$ besser als $\hat{\theta}_2$, wenn

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

BEMERKUNG 6.1.12. In Beispiel 6.1.5 ist $\hat{\theta}_3 = \frac{n+1}{n} X_{(n)}$ der beste Schätzer unter allen erwartungstreuen Schätzern. Der Beweis hierfür folgt später.

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 6.1.13. 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 6.1.14. In Beispiel 6.1.5 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 6.1.15. 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 6.1.16. 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 6.1.17. Eine fast sicher konvergente Folge von Zufallsvariablen konvergiert auch in Wahrscheinlichkeit. Aus der starken Konsistenz folgt somit die schwache Konsistenz.

DEFINITION 6.1.18. 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 6.1.19. Aus der L^2-Konsistenz folgt die schwache Konsistenz.

BEISPIEL 6.1.20. Bei vielen Familien von Verteilungen, z.B. Bern(θ), Poi(θ) oder $\mathrm{N}(\theta, \sigma^2)$ 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 6.1.21. 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 $\theta \in \Theta$

$$\frac{\hat{\theta}(X_1, \ldots, X_n) - a_n(\theta)}{b_n(\theta)} \xrightarrow[n\to\infty]{d} \mathrm{N}(0, 1) \text{ unter } \mathbb{P}_\theta.$$

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 $\theta \in \Theta$

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

BEISPIEL 6.1.22. Es seien $X_1, \ldots, X_n$ unabhängig und Bern(θ)–verteilt, mit $\theta \in (0, 1)$. Dann ist der Schätzer $\hat{\theta} = \bar{X}_n$ asymptotisch normalverteilt, denn nach dem Satz von de Moivre–Laplace gilt

$$\frac{\bar{X}_n - \theta}{\sqrt{\frac{\theta(a-\theta)}{n}}} = \frac{X_1 + \ldots + X_n - n\theta}{\sqrt{n\theta(1-\theta)}} \xrightarrow[n\to\infty]{d} \mathrm{N}(0, 1) \text{ unter } \mathbb{P}_\theta.$$

6.2. 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

$$L_n(X_1, \ldots, X_n; \theta) \xrightarrow[n\to\infty]{f.s.} \mathbb{E}_{\theta_0} \log h_\theta(X) = L_\infty(\theta) = \int_J \log h_\theta(t) h_{\theta_0}(t) dt.$$

LEMMA 6.2.1. *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:

$$\begin{aligned} L_\infty(\theta) - L_\infty(\theta_0) &\leq \mathbb{E}_{\theta_0}\left[\frac{h_\theta(X)}{h_{\theta_0}(X)} - 1\right] \\ &= \int_J \left(\frac{h_\theta(t)}{h_{\theta_0}(t)} - 1\right) h_{\theta_0}(t) dt \\ &= \int_J h_\theta(t) dt - \int_J h_{\theta_0}(t) dt \\ &= 0, \end{aligned}$$

denn $\int_J h_\theta(t)dt = \int_J h_{\theta_0}(t)dt = 1$. □

SATZ 6.2.2. *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

$$\hat{\theta}_{ML}(X_1, \ldots, X_n) = \text{argmax } L_n(X_1, \ldots, X_n; \theta).$$

Indem wir zum Grenzwert für $n \to \infty$ übergehen, erhalten wir

$$\begin{aligned}\lim_{n\to\infty} \hat{\theta}_{ML}(X_1, \ldots, X_n) &= \lim_{n\to\infty} \text{argmax } L_n(X_1, \ldots, X_n; \theta) \\ &= \text{argmax } \lim_{n\to\infty} L_n(X_1, \ldots, X_n; \theta) \\ &= \text{argmax } L_\infty(X_1, \ldots, X_n; \theta) \\ &= \theta_0,\end{aligned}$$

wobei der letzte Schritt aus Lemma 6.2.1 folgt. □

Der obige Beweis ist nicht streng. Insbesondere bedarf der Schritt $\lim_{n\to\infty} \text{argmax} = \text{argmax} \lim_{n\to\infty}$ einer Begründung.

Asymptotische Normalverteiltheit des ML-Schätzers. Wir werden zeigen, dass unter gewissen Regularitätsbedingungen der Maximum–Likelihood–Schätzer asymptotisch normalverteilt ist:

$$\sqrt{n}(\hat{\theta}_{\text{ML}}(X_1, \ldots, X_n) - \theta_0) \xrightarrow[n\to\infty]{d} \text{N}(0, \sigma^2_{\text{ML}}) \text{ unter } \mathbb{P}_{\theta_0},$$

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 6.2.3. 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 6.2.4. *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 $\int_J h_\theta(t)dt = 1$, denn h_θ ist eine Dichte. Wir können nun diese Identität nach θ ableiten:

$$D \int_J h_\theta(t)dt = 0 \text{ und somit } \int_J Dh_\theta(t)dt = 0,$$
$$D^2 \int_J h_\theta(t)dt = 0 \text{ und somit } \int_J D^2 h_\theta(t)dt = 0.$$

Dabei haben wir die Ableitung und das Integral vertauscht, was unter gewissen Regularitätsbedingungen möglich ist. Mit diesen Resultaten erhalten wir, dass

$$\mathbb{E}_\theta Dl_\theta(X) = \int_J \frac{Dh_\theta(X)}{h_\theta(X)} h_\theta(t)dt = \int_J Dh_\theta(X)dt = 0.$$

Somit ist die erste Behauptung des Lemmas bewiesen. Die zweite Behauptung des Lemmas kann man wie folgt zeigen:

$$\begin{aligned}
\mathbb{E}_\theta D^2 l_\theta(X) &= \int_J (D^2 l_\theta(t))h_\theta(t)dt \\
&= \int_J \left(\frac{D^2 h_\theta(t)}{h_\theta(t)} - (Dl_\theta(t))^2 \right) h_\theta(t)dt \\
&= \int_J D^2 h_\theta(t)dt - \mathbb{E}_\theta (Dl_\theta(X))^2 \\
&= -\mathbb{E}_\theta (Dl_\theta(X))^2 \\
&= -I(\theta),
\end{aligned}$$

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 6.2.5. *Unter Regularitätsbedingungen an die Familie $\{h_\theta : \theta \in \Theta\}$ gilt, dass*

(1) $\mathbb{E}_{\theta_0} Dl_{\theta_0}(X) = DL_\infty(\theta_0) = 0.$
(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 $\mathbb{E}_\theta Dl_\theta(X) = D\mathbb{E}_\theta l_\theta(X) = DL_\infty(X)$ und $\mathbb{E}_\theta D^2 l_\theta(X) = D^2 \mathbb{E}_\theta l_\theta(X) = D^2 L_\infty(X)$. □

SATZ 6.2.6. *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

$$\sqrt{n}(\hat{\theta}_{ML}(X_1, \ldots, X_n) - \theta_0) \xrightarrow[n\to\infty]{d} N\left(0, \frac{1}{I(\theta_0)}\right) \text{ unter } \mathbb{P}_{\theta_0}.$$

Beweisidee. Schritt 1. Für den Maximum–Likelihood–Schätzer gilt

$$\hat{\theta}_{ML} = \operatorname{argmax}_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \log h_\theta(X_i) = \operatorname{argmax}_{\theta\in\Theta} L_n(\theta).$$

Somit gilt

$$DL_n(\hat{\theta}_{ML}) = 0.$$

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 6.2.5 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

$$\sqrt{n}DL_n(\theta_0) \xrightarrow[n\to\infty]{d} N \sim \mathrm{N}(0, \operatorname{Var}_{\theta_0} Dl_{\theta_0}(X)) = \mathrm{N}(0, I(\theta_0)),$$

denn $\operatorname{Var}_{\theta_0} Dl_{\theta_0}(X) = I(\theta_0)$ nach Lemma 6.2.5.

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 6.2.2 (Konsistenz) gilt, erhalten wir, dass $\lim_{n\to\infty} \xi_n = \theta_0$. Nach dem Gesetz der großen Zahlen gilt somit

$$D^2L_n(\xi_n) = \frac{1}{n}\sum_{i=1}^{n} D^2 l_{\xi_n}(X_i) \xrightarrow[n\to\infty]{} \mathbb{E}_{\theta_0} D^2 l_{\theta_0}(X) = -I(\theta_0),$$

wobei wir im letzten Schritt Lemma 6.2.5 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 6.2.7. 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 6.2.6 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 6.2.8. Nun betrachten wir ein Beipsiel, 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 6.2.9. *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*

$$n\left(1 - \frac{M_n}{\theta}\right) \xrightarrow[n\to\infty]{d} Exp(1).$$

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

$$\lim_{n\to\infty} \mathbb{P}\left[n\left(1 - \frac{M_n}{\theta}\right) > x\right] = \lim_{n\to\infty}\left(1 - \frac{x}{n}\right)^n = e^{-x}.$$

Somit erhalten wir, dass

$$\lim_{n\to\infty} \mathbb{P}\left[n\left(1 - \frac{M_n}{\theta}\right) \leq x\right] = \begin{cases} e^{-x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Für $x < 0$ ist der Grenzwert gleich 0, denn das Ereignis $M_n > \theta$ ist unmöglich. Daraus ergibt sich die zu beweisende Aussage. □

BEISPIEL 6.2.10. 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 6.2.6 erhalten wir die asymptotische Normalverteiltheit von $\hat{\theta}_{ML}$:

$$\sqrt{n}\left(\frac{1}{\bar{X}_n} - \theta\right) \xrightarrow[n\to\infty]{d} \mathrm{N}(0,\theta^2) \text{ unter } \mathbb{P}_\theta. \tag{6.2.1}$$

Auf der anderen Seite, ergibt sich aus dem zentralen Grenzwertsatz, dass

$$\sqrt{n}\left(\bar{X}_n - \frac{1}{\theta}\right) \xrightarrow[n\to\infty]{d} \mathrm{N}\left(0, \frac{1}{\theta^2}\right) \text{ unter } \mathbb{P}_\theta, \tag{6.2.2}$$

denn $\mathbb{E}X_i = \frac{1}{\theta}$ und $\operatorname{Var} X_i = \frac{1}{\theta^2}$.
Sind nun (6.2.1) und (6.2.2) 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 an die Funktion die Antwort positiv ist.

LEMMA 6.2.11. *Seien $Z_1, Z_2, \ldots$ Zufallsvariablen und $\mu \in \mathbb{R}$ und $\sigma^2 > 0$ Zahlen mit*

$$\sqrt{n}(Z_n - \mu) \xrightarrow[n\to\infty]{d} N(0, \sigma^2).$$

Außerdem sei φ eine differenzierbare Funktion mit $\varphi'(\mu) \neq 0$. Dann gilt:

$$\sqrt{n}(\varphi(Z_n) - \varphi(\mu)) \xrightarrow[n\to\infty]{d} N(0, (\varphi'(\mu)\sigma)^2).$$

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 6.2.12. Als Spezialfall von Lemma 6.2.11 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 (6.2.1) und (6.2.2).

6.3. Cramér–Rao–Ungleichung

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.

SATZ 6.3.1 (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

$$\theta = \mathbb{E}_\theta \hat{\theta}(X_1, \ldots, X_n) = \int_{\mathbb{R}^n} \hat{\theta}(x_1, \ldots, x_n) h_\theta(x_1) \ldots h_\theta(x_n) dx_1 \ldots dx_n.$$

Nun leiten wir nach θ ab:

$$\begin{aligned} 1 &= D \int_{\mathbb{R}^n} \hat{\theta}(x_1, \ldots, x_n) h_\theta(x_1) \ldots h_\theta(x_n) dx_1 \ldots dx_n \\ &= \int_{\mathbb{R}^n} \hat{\theta}(x_1, \ldots, x_n) D[h_\theta(x_1) \cdot \ldots \cdot h_\theta(x_n)] dx_1 \ldots dx_n. \end{aligned}$$

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

$$\begin{aligned} 1 &= \int_{\mathbb{R}^n} \hat{\theta}(x_1, \ldots, x_n) h_\theta(x_1) \ldots h_\theta(x_n) \left(\sum_{i=1}^n D \log h_\theta(x_i) \right) dx_1 \ldots dx_n \\ &= \mathbb{E}_\theta[\hat{\theta}(X_1, \ldots, X_n) U_\theta(X_1, \ldots, X_n)], \end{aligned}$$

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 6.2.4, 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 6.2.4.

Nun erweitern wir mit dem Erwartungswert und wenden die Cauchy–Schwarz–Ungleichung an:

$$\begin{aligned} 1 &= \mathbb{E}_\theta[(\hat{\theta}(X_1,\dots,X_n)-\theta)\cdot U_\theta(X_1,\dots,X_n)] \\ &\le \sqrt{\operatorname{Var}_\theta[\hat{\theta}(X_1,\dots,X_n)]\cdot \operatorname{Var}_\theta[U_\theta(X_1,\dots,X_n)]} \\ &= \sqrt{\operatorname{Var}_\theta[\hat{\theta}(X_1,\dots,X_n)]\cdot nI(\theta)}. \end{aligned}$$

Umgestellt führt dies zu $\operatorname{Var}_\theta \hat{\theta}(X_1,\dots,X_n) \ge \frac{1}{nI(\theta)}$. □

Definition 6.3.2. Ein erwartungstreuer Schätzer $\hat{\theta}: \mathbb{R}^n \to \Theta$ heißt *Cramér–Rao–effizient*, falls für jedes $\theta \in \Theta$

$$\operatorname{Var}_\theta \hat{\theta}(X_1,\dots,X_n) = \frac{1}{nI(\theta)}.$$

Beispiel 6.3.3. Es seien $X_1,\dots,X_n$ unabhängig und Bernoulli–verteilt mit Parameter θ. Der Maximum–Likelihood–Schätzer und gleichzeitig der Momentenschätzer für θ ist der empirische Mittelwert:

$$\hat{\theta}(X_1,\dots,X_n) = \bar{X}_n = \frac{X_1+\dots+X_n}{n}.$$

Die Varianz von $\hat{\theta}$ lässt sich wie folgt berechnen

$$\operatorname{Var}_\theta \hat{\theta}(X_1,\dots,X_n) = \operatorname{Var}_\theta\left[\frac{X_1+\dots+X_n}{n}\right] = \frac{n}{n^2}\operatorname{Var}_\theta X_1 = \frac{\theta(1-\theta)}{n} = \frac{1}{nI(\theta)},$$

denn wir haben in Beispiel 6.2.7 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 6.3.4. 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.

Bemerkung 6.3.5. Der Maximum–Likelihood–Schätzer muss nicht immer Cramér–Rao–effizient (und sogar nicht einmal erwartungstreu) sein. Wir wollen allerdings zeigen, dass bei einem großen Stichprobenumfang n der Maximum–Likelihood–Schätzer $\hat{\theta}_{ML}$ die Cramér–Rao–Schranke asymptotisch erreicht. Nach Satz 6.2.6 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.

6.4. 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 6.4.1. *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*

$$\sqrt{n}(X_{([\alpha n])} - Q_\alpha) \xrightarrow[n\to\infty]{d} N\left(0, \frac{\alpha(1-\alpha)}{h^2(Q_\alpha)}\right).$$

BEWEISIDEE. Sei $t \in \mathbb{R}$. Wir betrachten die Verteilungsfunktion

$$F_n(t) := \mathbb{P}[\sqrt{n}(X_{([\alpha n])} - Q_\alpha) \le t] = \mathbb{P}\left[X_{([\alpha n])} \le Q_\alpha + \frac{t}{\sqrt{n}}\right] = \mathbb{P}[K_n \ge \alpha n],$$

wobei die Zufallsvariable K_n wie folgt definiert wird:

$$K_n = \sum_{i=1}^{n} \mathbb{1}_{X_i \le Q_\alpha + \frac{t}{\sqrt{n}}} = \#\left\{i \in \{1, \ldots, n\} : X_i \le Q_\alpha + \frac{t}{\sqrt{n}}\right\}.$$

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

$$F_n(t) = \mathbb{P}[K_n \ge \alpha n] = \mathbb{P}\left[\frac{K_n - \mathbb{E}[K_n]}{\sqrt{\mathrm{Var}(K_n)}} \ge \frac{\alpha n - \mathbb{E}[K_n]}{\sqrt{\mathrm{Var}(K_n)}}\right].$$

Benutzen wir nun die Entwicklungen von $\mathbb{E}K_n$ und $\mathrm{Var}\, K_n$, so erhalten wir

$$F_n(t) = \mathbb{P}\left[\frac{K_n - \mathbb{E}[K_n]}{\sqrt{\mathrm{Var}(K_n)}} \ge \frac{\alpha n - \alpha n - \sqrt{n}h(Q_\alpha)t - o(\sqrt{n})}{\sqrt{n\alpha(1-\alpha) + o(n)}}\right].$$

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 $\frac{N\sqrt{\alpha(1-\alpha)}}{h(Q_\alpha)} \sim \mathrm{N}\left(0, \frac{\alpha(1-\alpha)}{h^2(Q_\alpha)}\right)$. □

Wir werden nun den obigen Satz benutzen, um den Median und den Mittelwert als Schätzer für den Lageparameter einer Dichte auf Effizienz hin zu untersuchen und zu vergleichen. Sei $h(x)$ 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

$$h_\theta(x) = h(x-\theta), \quad \theta \in \mathbb{R}.$$

Die Aufgabe besteht nun darin, θ zu schätzen. Dabei sei die Dichte h bekannt. Da sowohl der theoretische Median $Q_{1/2}$ 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 $X_{([n/2])}$.
(2) den empirischen Mittelwert $\bar{X}_n$.

Welcher Schätzer ist nun besser und um wieviel? Um diese Frage zu beantworten, benutzen wir den Begriff der relativen Effizienz.

Definition 6.4.2. 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 *relative Effizienz* der beiden Schätzer ist definiert durch

$$e_{\hat{\theta}_n^{(1)}, \hat{\theta}_n^{(2)}}(\theta) = \frac{\sigma_2^2(\theta)}{\sigma_1^2(\theta)}.$$

Bemerkung 6.4.3. Ist z.B. $e_{\hat{\theta}_n^{(1)}, \hat{\theta}_n^{(2)}} > 1$ so heißt es, dass $\hat{\theta}_n^{(1)}$ besser als $\hat{\theta}_n^{(2)}$ ist.

Kommen wir nun wieder zurück zu unserer Frage, welcher Schätzer, $\hat{\theta}_n^{(1)} = X_{([n/2])}$ oder $\hat{\theta}_n^{(2)} = \bar{X}_n$, besser ist. Nach Satz 6.4.1 und nach dem zentralen Grenzwertsatz, sind beide Schätzer asymptotisch normalverteilt mit

$$\sqrt{n}(X_{([n/2])} - \theta) \xrightarrow[n\to\infty]{d} \mathrm{N}\left(0, \frac{1}{4h^2(0)}\right) \text{ und } \sqrt{n}(\bar{X}_n - \theta) \xrightarrow[n\to\infty]{d} \mathrm{N}(0, \mathrm{Var}_\theta X_1) \text{ unter } \mathbb{P}_\theta.$$

Die asymptotischen Varianzen sind also unabhängig von θ und gegeben durch

$$\sigma_1^2 = \frac{1}{4h^2(0)}, \quad \sigma_2^2 = \mathrm{Var}_\theta\, X_1 = \int_{\mathbb{R}} x^2 h^2(x) dx.$$

Nun betrachten wir zwei Beispiele.

Beispiel 6.4.4. Die Gauß–Dichte $h(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. Es gilt $h(0) = 0$, $\mathrm{Var}_\theta\, X_1 = 1$ und somit

$$\sigma_1^2 = \frac{\pi}{2}, \quad \sigma_2^2 = 1, \quad e_{Med,MW} = \frac{2}{\pi} \approx 0.6366 < 1.$$

Somit ist der empirische Mittelwert besser als der empirische Median. Das Ergebnis kann man so 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 6.4.5. Die Laplace–Dichte $h(x) = \frac{1}{2} e^{-|x|}$. Es gilt $h(0) = \frac{1}{2}$, $\mathrm{Var}_\theta\, X_1 = 2$ und somit

$$\sigma_1^2 = 1, \quad \sigma_2^2 = 2, \quad e_{Med,MW} = 2 > 1.$$

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.

KAPITEL 7

Suffizienz und Vollständigkeit

7.1. Definition der Suffizienz im diskreten Fall

BEISPIEL 7.1.1. 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.

Es sei im Weiteren $\{h_\theta : \theta \in \Theta\}$ eine Familie von Zähldichten. Den Fall der Dichten werden wir später betrachten. Ist S ein Zufallsvektor mit Werten in $\mathbb{R}^d$, so bezeichnen wir mit $\operatorname{Im} S = \{z \in \mathbb{R}^d : \mathbb{P}[S = z] \neq 0\}$ die Menge aller Werte z, die der Zufallsvektor S mit einer strikt positiven Wahrscheinlichkeit annehmen kann. Wirr nennen $\operatorname{Im} S$ den *Träger* von S. Im folgenden Abschnitt werden wir annehmen, dass der Träger von $X_1, \ldots, X_n$ nicht von θ abhängt.

DEFINITION 7.1.2. Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte diskrete Zufallsvariablen mit Zähldichte h_θ. Eine Funktion $T : \mathbb{R}^n \to \mathbb{R}^m$ heißt eine *suffiziente Statistik*, wenn für alle $x_1, \ldots, x_n \in \mathbb{R}$ und für alle $t \in \operatorname{Im} T(X_1, \ldots, X_n)$ der Ausdruck

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

eine von θ unabhängige Funktion ist. D.h., für alle $\theta_1, \theta_2 \in \Theta$ soll gelten:

$$\mathbb{P}_{\theta_1}[X_1 = x_1, \ldots, X_n = x_n | T(X_1, \ldots, X_n) = t] = \mathbb{P}_{\theta_2}[X_1 = x_1, \ldots, X_n = x_n | T(X_1, \ldots, X_n) = t].$$

BEISPIEL 7.1.3 (Fortsetzung von Beispiel 7.1.1). Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit $X_i \sim \text{Bern}(\theta)$, wobei $\theta \in (0, 1)$ zu schätzen sei. Die Zähldichte von X_i ist somit gegeben durch

$$h_\theta(x) = \theta^x (1 - \theta)^{1-x}, \quad x \in \{0, 1\}.$$

Wir zeigen nun, dass $T(X_1, \ldots, X_n) = X_1 + \ldots + X_n$ eine suffiziente Statistik ist. Sei $t \in \{0, \ldots, n\}$. Betrachte 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 $x_1 + \ldots + x_n \neq t$ oder $x_i \notin \{0, 1\}$ für mindestens ein i, dann gilt $P(\theta) = 0$. In diesem Fall hängt $P(\theta)$ von θ nicht ab.

Fall 2. Sei nun $x_1 + \ldots + x_n = t$ mit $x_1, \ldots, x_n \in \{0, 1\}$. 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.
Somit ist $T(X_1, \ldots, X_n) = X_1 + \ldots + X_n$ eine suffiziente Statistik. 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. In der Tat, wir haben bereits gezeigt, dass $\bar{X}_n$ die Cramér–Rao–Schranke erreicht und somit der beste erwartungstreue Schätzer ist.

Bemerkung 7.1.4. Im obigen Beispiel haben wir gezeigt, dass für jedes $t \in \{0, 1 \ldots, n\}$ die bedingte Verteilung von $(X_1, \ldots, X_n)$ gegeben, dass $X_1 + \ldots + X_n = t$, eine Gleichverteilung auf der Menge

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

ist. Diese Menge besteht aus $\binom{n}{t}$ Elementen.

Aufgabe 7.1.5. Seien $X_1, \ldots, X_n$ unabhängige und mit Parameter $\theta > 0$ Poisson–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 $X_1 + \ldots + X_n = t$.

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 7.1.6. *Sei T eine suffiziente Statistik und sei $g : \mathrm{Im}\, T(X_1, \ldots, X_n) \to \mathbb{R}^k$ eine injektive Funktion. Dann ist auch*

$$g \circ T : \mathbb{R}^n \to \mathbb{R}^k, \quad (X_1, \ldots, X_n) \mapsto g(T(X_1, \ldots, X_n))$$

eine suffiziente Statistik.

BEWEIS. Sei $t \in \operatorname{Im} g(T(X_1, \ldots, X_n))$. Da T eine suffiziente Statistik ist, hängt

$$\begin{aligned} P(\theta) &:= \mathbb{P}_\theta[X_1 = x_1, \ldots, X_n = x_n | g(T(X_1, \ldots, X_n)) = t] \\ &= \mathbb{P}_\theta[X_1 = x_1, \ldots, X_n = x_n | T(X_1, \ldots, X_n) = g^{-1}(t)] \end{aligned}$$

nicht von θ ab. Dabei ist $g^{-1}(t)$ wohldefiniert, da g injektiv ist. □

7.2. 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 $\{h_\theta : \theta \in \Theta\}$ eine Familie von Zähldichten und $X_1, \ldots, X_n$ unabhängige Zufallsvariablen mit Zähldichte h_θ. Sei $T : \mathbb{R}^n \to \mathbb{R}^m$ eine Funktion. Im nächsten Lemma benutzen wir folgende Notation:

(1) $L(x_1, \ldots, x_n; \theta) = \mathbb{P}_\theta[X_1 = x_1, \ldots, X_n = x_n]$ ist die Likelihood-Funktion.
(2) $q(t; \theta) = \mathbb{P}_\theta[T(X_1, \ldots, X_n) = t]$, wobei $t \in \mathbb{R}^m$, ist die Zähldichte von $T(X_1, \ldots, X_n)$ unter $\mathbb{P}_\theta$.

LEMMA 7.2.1. *Eine Funktion $T : \mathbb{R}^n \to \mathbb{R}^m$ ist genau dann eine suffiziente Statistik, wenn für alle $x_1, \ldots, x_n \in \operatorname{Im} X_1$ die Funktion*

$$\frac{L(x_1, \ldots, x_n; \theta)}{q(T(x_1, \ldots, x_n); \theta)} \tag{7.2.1}$$

nicht von θ abhängt.

BEWEIS. Betrachte den Ausdruck

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

Im Falle $t \neq T(x_1, \ldots, x_n)$ ist die bedingte Wahrscheinlichkeit gleich 0, was unabhängig von θ ist. Sei deshalb $t = T(x_1, \ldots, x_n)$. Dann gilt

$$\begin{aligned} P(\theta) &= \frac{\mathbb{P}_\theta[X_1 = x_1, \ldots, X_n = x_n, T(X_1, \ldots, X_n) = t]}{\mathbb{P}_\theta[T(X_1, \ldots, X_n) = T(x_1, \ldots, x_n)]} \\ &= \frac{\mathbb{P}_\theta[X_1 = x_1, \ldots, X_n = x_n]}{\mathbb{P}_\theta[T(X_1, \ldots, X_n) = T(x_1, \ldots, x_n)]} \\ &= \frac{L(x_1, \ldots, x_n; \theta)}{q(T(x_1, \ldots, x_n); \theta)}. \end{aligned}$$

Somit ist T eine suffiziente Statistik genau dann, wenn (7.2.1) nicht von θ abhängt.

SATZ 7.2.2 (Faktorisierungssatz von Neyman–Fisher). *Eine Funktion $T : \mathbb{R}^n \to \mathbb{R}^m$ ist eine suffiziente Statistik genau dann, wenn es Funktionen $g : \mathbb{R}^m \times \Theta \to \mathbb{R}$ und $h : \mathbb{R}^n \to \mathbb{R}$ gibt, so dass für alle $x_1, \ldots, x_n \in \mathbb{R}$ und alle $\theta \in \Theta$ die folgende Faktorisierung gilt:*

$$L(x_1, \ldots, x_n; \theta) = g(T(x_1, \ldots, x_n); \theta) \cdot h(x_1, \ldots, x_n). \tag{7.2.2}$$

BEWEIS VON "$\Longrightarrow$". Sei T eine suffiziente Statistik. Nach Lemma 7.2.1 ist die Funktion

$$h(x_1, \ldots, x_n) := \begin{cases} \frac{L(x_1, \ldots, x_n; \theta)}{q(T(x_1, \ldots, x_n); \theta)}, & \text{falls } x_1, \ldots, x_n \in \operatorname{Im} X_1, \\ 0, & \text{sonst} \end{cases}$$

unabhängig von θ. Mit diesem h und $g(t;\theta) = q(t;\theta)$ gilt die Faktorisierung (7.2.2). □

Beweis von "$\Longleftarrow$". Es gelte die Faktorisierung (7.2.2). Wir bezeichnen mit $\sum^*$ die Summe über alle $(y_1, \ldots, y_n)$ mit $T(y_1, \ldots, y_n) = T(x_1, \ldots, x_n)$. Dann gilt

$$\begin{aligned}\frac{L(x_1, \ldots, x_n; \theta)}{q(T(x_1, \ldots, x_n); \theta)} &= \frac{g(T(x_1, \ldots, x_n); \theta) h(x_1, \ldots, x_n)}{\sum^* L(y_1, \ldots, y_n; \theta)} \\ &= \frac{g(T(x_1, \ldots, x_n); \theta) h(x_1, \ldots, x_n)}{\sum^* g(T(y_1, \ldots, y_n); \theta) h(y_1, \ldots, y_n)} \\ &= \frac{h(x_1, \ldots, x_n)}{\sum^* h(y_1, \ldots, y_n)}.\end{aligned}$$

Dieser Ausdruck hängt nicht von θ ab. Nach Lemma 7.2.1 ist T suffizient. □

Beispiel 7.2.3. Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit $X_i \sim \text{Bern}(\theta)$, wobei $\theta \in (0, 1)$. Für die Likelihood–Funktion gilt

$$\begin{aligned}L(x_1, \ldots, x_n; \theta) &= h_\theta(x_1) \ldots h_\theta(x_n) \\ &= \theta^{x_1}(1-\theta)^{1-x_1} \mathbb{1}_{x_1 \in \{0,1\}} \cdot \ldots \cdot \theta^{x_n}(1-\theta)^{1-x_n} \mathbb{1}_{x_n \in \{0,1\}} \\ &= \theta^{x_1 + \ldots + x_n}(1-\theta)^{n-(x_1+\ldots+x_n)} \mathbb{1}_{x_1, \ldots, x_n \in \{0,1\}}.\end{aligned}$$

Daraus ist ersichtlich, dass die Neyman–Fisher–Faktorisierung (7.2.2) 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.

7.3. Definition der Suffizienz im absolut stetigen Fall

Bisher haben wir nur diskrete Zufallsvariablen betrachtet. Seien nun $X_1, \ldots, X_n$ absolut stetige Zufallsvariablen mit Dichte h_θ. Die Funktion $T : \mathbb{R}^n \to \mathbb{R}^m$ sei Borel–messbar. Außerdem nehmen wir an, dass $T(X_1, \ldots, X_n)$ ebenfalls eine absolut stetige Zufallsvariable mit Dichte $q(t;\theta)$ ist. Zum Beispiel darf T nicht konstant sein. Die vorherige Definition der Suffizienz macht im absolut stetigen Fall keinen Sinn, denn das Ereignis $T(X_1, \ldots, X_n) = t$ hat Wahrscheinlichkeit 0. Wir benötigen also eine andere Definition. Ein möglicher Zugang besteht darin, den Satz von Neyman–Fisher im absolut stetigen Fall als Definition zu benutzen. Die Likelihood–Funktion sei

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

Definition 7.3.1. Im absolut stetigen Fall heißt eine Statistik T suffizient, wenn es eine Faktorisierung der Form

$$L(x_1, \ldots, x_n; \theta) = g(T(x_1, \ldots, x_n); \theta) \cdot h(x_1, \ldots, x_n)$$

gibt.

BEISPIEL 7.3.2. Seien $X_1, \ldots, X_n$ unabhängige und auf dem Intervall $[0, \theta]$ gleichverteilte Zufallsvariablen, wobei $\theta > 0$ der unbekannte Parameter sei. Somit ist die Dichte von X_i gegeben durch

$$h_\theta(x) = \frac{1}{\theta} \mathbb{1}_{x \in [0,\theta]}.$$

Wir werden nun zeigen, dass $T(X_1, \ldots, X_n) = \max\{X_1, \ldots, X_n\}$ eine suffiziente Statistik ist. Für die Likelihood–Funktion gilt

$$\begin{aligned} L(x_1, \ldots, x_n; \theta) &= h_\theta(x_1) \ldots h_\theta(x_n) \\ &= \frac{1}{\theta^n} \mathbb{1}_{x_1 \in [0,\theta]} \cdot \ldots \cdot \mathbb{1}_{x_n \in [0,\theta]} \\ &= \frac{1}{\theta^n} \mathbb{1}_{\max(x_1,\ldots,x_n) \leq \theta} \cdot \mathbb{1}_{x_1,\ldots,x_n \geq 0} \\ &= g(T(x_1, \ldots, x_n); \theta) \cdot h(x_1, \ldots, x_n), \end{aligned}$$

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, dennn 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 7.3.3. Seien $X_1, \ldots, X_n$ unabhängige und mit Parameter $\theta > 0$ exponentialverteilte Zufallsvariablen. Somit ist die Dichte von X_i gegeben durch

$$h_\theta(x) = \theta \exp(-\theta x) \mathbb{1}_{x \geq 0}.$$

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 7.3.4. 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

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

Somit ist die Likelihood–Funktion gegeben durch

$$\begin{aligned} L(x_1,\ldots,x_n;\mu,\sigma^2) &= h_{\mu,\sigma^2}(x_1)\ldots h_{\mu,\sigma^2}(x_n) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\right) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^n x_i^2 - 2\mu\sum_{i=1}^n x_i + n\mu^2\right]\right). \end{aligned}$$

Nun betrachten wir die Statistik $T:\mathbb{R}^n\to\mathbb{R}^2$ mit

$$(x_1,\ldots,x_n)\mapsto\left(\sum_{i=1}^n x_i^2,\sum_{i=1}^n x_i\right)=(T_1,T_2).$$

Diese Statistik T ist suffizient, denn wir haben die Neyman–Fisher–Faktorisierung

$$L(x_1,\ldots,x_n;\mu,\sigma^2)=g(T_1,T_2;\mu,\sigma^2)h(x_1,\ldots,x_n)$$

mit $h(x_1,\ldots,x_n)=1$ 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 7.3.5. 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 7.3.6. Sei $\{h_\theta:\theta\in\Theta\}$ eine beliebige Familie von Dichten bzw. Zähldichten, dann ist die identische Abbildung $T:\mathbb{R}^n\to\mathbb{R}^n$ mit $(x_1,\ldots,x_n)\mapsto(x_1,\ldots,x_n)$ suffizient. Die Suffizienz folgt aus dem Faktorisierungssatz von Neyman–Fisher, denn für die Likelihood–Funktion gilt

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

Beispiel 7.3.7. Sei $\{h_\theta:\theta\in\Theta\}$ eine beliebige Familie von Dichten oder Zähldichten, dann ist die Statistik

$$T:(x_1,\ldots,x_n)\mapsto(x_{(1)},\ldots,x_{(n)})$$

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 Likelihood–Funktion gilt

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

Diese Funktion ändert sich bei Permutationen von $x_1,\ldots,x_n$ nicht und kann somit als eine Funktion von T und θ dargestellt werden. Somit haben wir eine Neyman–Fisher–Faktorisierung angegeben.

7.4. Vollständigkeit

Sei $\{h_\theta(x) : \theta \in \Theta\}$ eine Familie von Dichten bzw. Zähldichten und seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit Dichte bzw. Zähldichte h_θ.

DEFINITION 7.4.1. Eine Statistik $T : \mathbb{R}^n \to \mathbb{R}^m$ heißt *vollständig*, falls für alle Borel–Funktionen $g : \mathbb{R}^m \to \mathbb{R}$ aus der Gültigkeit von

$$\mathbb{E}_\theta g(T(X_1, \ldots, X_n)) = 0 \text{ für alle } \theta \in \Theta$$

folgt, dass $g(T(X_1, \ldots, X_n)) = 0$ 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 7.4.2. Seien $X_1, \ldots, X_n$ unabhängige und mit Parameter $\theta \in (0, 1)$ Bernoulli–verteilte 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 7.4.3. 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. □

BEISPIEL 7.4.4. 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:

$$q(x;\theta) = nx^{n-1}\theta^{-n}\mathbb{1}_{0\leq x\leq\theta}.$$

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

$$\theta^{-n}\int_0^\theta nx^{n-1}g(x)dx = 0 \text{ für alle } \theta > 0.$$

Wir können durch θ^{-n} teilen:

$$\int_0^\theta nx^{n-1}g(x)dx = 0 \text{ für alle } \theta > 0.$$

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 $[0,\theta]$. □

7.5. Exponentialfamilien

In diesem Abschnitt führen wir den Begriff der Exponentialfamilie ein. 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 7.5.1. Die Familie $\{h_\theta(x) : \theta \in \Theta\}$ heißt *Exponentialfamilie*, falls es Funktionen $a(\theta)$, $b(x)$, $c(\theta)$, $d(x)$ gibt mit

$$h_\theta(x) = a(\theta)b(x)e^{c(\theta)d(x)}.$$

BEISPIEL 7.5.2. 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 $h_\theta(x) = a(\theta)b(x)e^{c(\theta)d(x)}$ 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 7.5.3. 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 $h_{\mu,\sigma^2}(x) = a(\mu)b(x)e^{c(\mu)d(x)}$ 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 $h_{\mu,\sigma^2}(x) = a(\sigma^2)b(x)e^{c(\sigma^2)d(x)}$ mit

$$a(\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right), \quad b(x) = 1, \quad c(\sigma^2) = \frac{1}{2\sigma^2}, \quad d(x) = 2x\mu - x^2.$$

Weitere Beispiele von Exponentialfamilien sind die Familie der Poissonverteilungen und die familie der Exponentialverteilungen. 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 7.5.4. Eine Familie $\{h_\theta : \theta \in \Theta\}$ von Dichten oder Zähldichten heißt eine *m-parametrige Exponentialfamilie*, falls es eine Darstellung der Form

$$h_\theta(x) = a(\theta)b(x)e^{c_1(\theta)d_1(x)+\ldots+c_m(\theta)d_m(x)}$$

gibt.

BEISPIEL 7.5.5. 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

$$h_{\mu,\sigma^2}(x) = a(\mu, \sigma^2)b(x)e^{c_1(\mu,\sigma^2)d_1(x)+c_2(\mu,\sigma^2)d_2(x)}$$

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.

7.6. Vollständige und suffiziente Statistik für Exponentialfamilien

Für eine Exponentialfamilie lässt sich sehr leicht eine suffiziente und vollständige Statistik angeben. Nämlich ist die Statistik $(T_1, \ldots, T_m)$ mit

$$T_1(X_1, \ldots, X_n) = \sum_{j=1}^{n} d_1(X_j), \quad \ldots, \quad T_m(X_1, \ldots, X_n) = \sum_{j=1}^{n} d_m(X_j)$$

suffizient. Um dies zu zeigen, schreiben wir die Likelihood–Funktion wie folgt um:

$$\begin{aligned} L(x_1,\dots,x_n;\theta) &= h_\theta(x_1)\dots h_\theta(x_n) \\ &= (a(\theta))^n b(x_1)\dots b(x_n)\exp\left(\sum_{i=1}^m c_i(\theta)d_i(x_1)\right)\dots\exp\left(\sum_{i=1}^m c_i(\theta)d_i(x_n)\right) \\ &= (a(\theta))^n \exp\left(T_1c_1(\theta)+\dots+T_mc_m(\theta)\right). \end{aligned}$$

Die Suffizienz von $(T_1,\dots,T_m)$ folgt aus dem Faktorisierungssatz von Neyman–Fisher mit

$$h(x_1,\dots,x_n) = b(x_1)\dots b(x_n), \quad g(T_1,\dots,T_m;\theta) = (a(\theta))^n \exp\left(T_1c_1(\theta)+\dots+T_mc_m(\theta)\right).$$

Man kann zeigen, dass diese Statistik auch vollständig ist (ohne Beweis).

Beispiel 7.6.1. 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.

7.7. Der beste erwartungstreue Schätzer

Sei $\Theta=(a,b)\subset\mathbb{R}$. Wir betrachten eine Familie von Dichten bzw. Zähldichten $\{h_\theta:\theta\in\Theta\}$. Seien $X_1,\dots,X_n$ unabhängige und identisch verteilte Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega,\mathcal{A},\mathbb{P}_\theta)$ mit Dichte bzw. Zähldichte h_θ.
Wir bezeichnen mit L^2 die Menge aller Stichprobenfunktionen (Schätzer) $f:\mathbb{R}^n\to\mathbb{R}$ mit

$$\mathbb{E}_\theta[f(X_1,\dots,X_n)^2]<\infty \text{ für alle } \theta\in\Theta.$$

Im Folgenden werden wir nur Schätzer aus L^2 betrachten. Außerdem benutzen wir die folgende Abkürzung: Wir benutzen $\hat\theta$ als eine Bezeichnung für die Zufallsvariable $\hat\theta(X_1,\dots,X_n)$.

Definition 7.7.1. Ein Schätzer $\hat\theta:\mathbb{R}^n\to\mathbb{R}$ heißt *erwartungstreu* (für θ), falls

$$\mathbb{E}_\theta\hat\theta=\theta \text{ für alle } \theta\in\Theta.$$

Wir bezeichnen mit H die Menge der erwartungstreuen Schätzer, d.h.

$$H=\{\hat\theta\in L^2:\hat\theta \text{ ist erwartungstreu}\}.$$

Aufgabe 7.7.2. Zeigen Sie, dass H ein affiner Unterraum ist, d.h. für alle $\hat\theta_1,\hat\theta_2\in H$ und alle $t\in\mathbb{R}$ ist auch $t\hat\theta_1+(1-t)\hat\theta_2\in H$.

Wir haben bereits gesehen, dass es in parametrischen Modellen typischerweise mehrere erwartungstreue Schätzer für den Parameter existieren. Wie wählt man unter diesen Schätzer den besten?

Definition 7.7.3. Ein Schätzer $\hat{\theta}$ heißt *bester erwartungstreuer Schätzer* (für θ), falls er erwartungstreu ist und wenn für jeden anderen erwartungstreuen Schätzer $\tilde{\theta} \in H$ gilt, dass

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

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

Satz 7.7.4. *Seien $\hat{\theta}_1, \hat{\theta}_2 : \mathbb{R}^n \to \Theta$ zwei beste erwartungstreue Schätzer, dann gilt*

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

Bemerkung 7.7.5. Der beste erwartungstreue Schätzer muss nicht in jedem parametrischen Modell existieren. Es kann nämlich durchaus passieren, dass der erwartungstreue Schätzer, der die kleinste Varianz unter $\mathbb{P}_\theta$ hat, nicht die kleinste Varianz unter einem anderen Wahrscheinlichkeitsmaß $\mathbb{P}_{\theta'}$ hat.

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

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

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

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

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 $a(\theta) = 1$.

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 $\theta \in \Theta$. □

DEFINITION 7.7.6. Ein Stichprobenfunktion $\varphi : \mathbb{R}^n \to \Theta$ heißt *erwartungstreuer Schätzer für* 0, falls

$$\mathbb{E}_\theta \varphi = 0 \text{ für alle } \theta \in \Theta.$$

SATZ 7.7.7. *Sei $\hat{\theta}$ ein erwartungstreuer Schätzer für θ. Dann sind die folgenden Bedingungen äquivalent:*

(1) *$\hat{\theta}$ ist der beste erwartungstreue Schätzer für θ.*
(2) *Für alle erwartungstreuen Schätzer φ für 0 gilt, dass $\operatorname{Cov}_\theta(\hat{\theta}, \varphi) = 0$ 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

$$\operatorname{Cov}_\theta(\hat{\theta}, \varphi) = 0 \text{ für alle } \theta \in \Theta.$$

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

$$\operatorname{Var}_\theta \tilde{\theta} = \operatorname{Var}_\theta \hat{\theta} + a^2 \operatorname{Var}_\theta \varphi + 2a \operatorname{Cov}_\theta(\hat{\theta}, \varphi) = \operatorname{Var}_\theta \hat{\theta} + g(a),$$

wobei $g(a) = a^2 \operatorname{Var}_\theta \varphi + 2a \operatorname{Cov}_\theta(\hat{\theta}, \varphi)$. Wäre nun $\operatorname{Cov}_\theta(\hat{\theta}, \varphi) \neq 0$, dann hätte die quadratische Funktion g zwei verschiedene Nullstellen bei 0 und $-2 \operatorname{Cov}_\theta(\hat{\theta}, \varphi) / \operatorname{Var}_\theta \varphi$. (Wir dürfen hier annehmen, dass $\operatorname{Var}_\theta \varphi \neq 0$, denn andernfalls wäre φ fast sicher konstant unter $\mathbb{P}_\theta$ und dann würde $\operatorname{Cov}_\theta(\hat{\theta}, \varphi) = 0$ trivialerweise gelten). Zwischen diesen Nullstellen gäbe es ein $a \in \mathbb{R}$ mit $g(a) < 0$ und es würde folgen, dass $\operatorname{Var}_\theta \tilde{\theta} < \operatorname{Var}_\theta \hat{\theta}$. Das widerspricht aber der Annahme, dass $\hat{\theta}$ der beste erwartungstreue Schätzer für θ ist. Somit muss $\operatorname{Cov}_\theta(\hat{\theta}, \varphi) = 0$ gelten. □

BEWEIS VON "$\Longleftarrow$". Sei $\hat{\theta}$ ein erwartungstreuer Schätzer für θ. Sei außerdem $\operatorname{Cov}_\theta(\varphi, \hat{\theta}) = 0$ für alle erwartungstreuen Schätzer φ für 0. Jetzt werden wir zeigen, dass $\hat{\theta}$ der beste 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. □

7.8. Bedingter Erwartungswert

DEFINITION 7.8.1. Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Seien $A \in \mathcal{A}$ und $B \in \mathcal{A}$ zwei Ereignisse. 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 macht nur dann Sinn, wenn $\mathbb{P}[B] \neq 0$.

Analog kann man den bedingten Erwartungswert definieren.

DEFINITION 7.8.2. 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. 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 macht nur dann Sinn, wenn $\mathbb{P}[B] \neq 0$. Zwischen den beiden Begriffen (bedingte Wahrscheinlichkeit und bedingter Erwartungswert) besteht der folgende Zusammenhang:

$$\mathbb{P}[A|B] = \mathbb{E}[\mathbb{1}_A|B].$$

Wir haben aber gesehen (z.B. bei der Definition der Suffizienz im absolut stetigen Fall), dass man oft bedingte Wahrscheinlichkeiten oder Erwartungswerte 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.

Sei $X : \Omega \to \mathbb{R}$ eine Zufallsvariable, definiert auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ mit $\mathbb{E}|X| < \infty$. 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. 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 7.8.3. 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 7.8.4. 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.

Beispiel 7.8.5. Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Betrachte eine disjunkte Zerlegung $\Omega = \Omega_1 \cup \ldots \cup \Omega_n$, wobei $\Omega_i \in \mathcal{A}$ und $\mathbb{P}[\Omega_i] \neq 0$. Sei $\mathcal{B}$ die σ–Algebra, die von $\{\Omega_1, \ldots, \Omega_n\}$ erzeugt wird. Somit ist

$$\mathcal{B} = \{\Omega_1^{\varepsilon_1} \cup \ldots \cup \Omega_n^{\varepsilon_n} : \varepsilon_1, \ldots, \varepsilon_n \in \{0, 1\}\},$$

wobei $\Omega_i^1 = \Omega_i$ und $\Omega_i^0 = \emptyset$. 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:

$$Z(\omega) := \mathbb{E}[X|\mathcal{B}](\omega) = \frac{\mathbb{E}[X\mathbb{1}_{\Omega_i}]}{\mathbb{P}[\Omega_i]}.$$

Beweis. Beachte, dass $Z := \mathbb{E}[X|\mathcal{B}]$ $\mathcal{B}$–messbar sein muss. Also ist Z konstant auf jeder Menge Ω_i. Sei also $Z(\omega) = c_i$ für $\omega \in \Omega_i$. Es muss außerdem gelten, dass

$$\mathbb{E}[X\mathbb{1}_{\Omega_i}] = \mathbb{E}[Z\mathbb{1}_{\Omega_i}] = c_i\mathbb{P}[\Omega_i].$$

Daraus folgt, dass $c_i = \mathbb{E}[X\mathbb{1}_{\Omega_i}]/\mathbb{P}[\Omega_i]$ sein muss. □

BEISPIEL 7.8.6. 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:

$$Z(s,t) := \mathbb{E}[X|\mathcal{B}](s,t) = \int_0^1 X(s,t)dt, \quad (s,t) \in [0,1]^2.$$

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

$$\mathbb{E}[Z\mathbb{1}_{C\times[0,1]}] = \int_{C\times[0,1]} Z(s,t)dsdt = \int_C \left(\int_0^1 Z(s,t)dt\right) ds = \mathbb{E}[X\mathbb{1}_{C\times[0,1]}].$$

Somit ist auch die zweite Bedingung erfüllt.

BEISPIEL 7.8.7. 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) $\mathbb{E}[X|\{0,\emptyset\}] = \mathbb{E}X$.
(2) $\mathbb{E}[X|\mathcal{A}] = X$.

BEWEIS. Übung. □

SATZ 7.8.8. *Sei $X, Y : \Omega \to \mathbb{R}$ 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. □

Besonders oft wird die Definition der bedingten Erwartung im folgenden Spezialfall benutzt.

DEFINITION 7.8.9. Seien X und Y zwei $\mathcal{A}$–messbare Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$. Sei

$$\sigma(Y) = \{Y^{-1}(C) : C \subset \mathbb{R} \text{ Borel}\}$$

die *von Y erzeugte σ–Algebra*. Der bedingte Erwartungswert von X gegeben Y ist definiert durch

$$\mathbb{E}[X|Y] = \mathbb{E}[X|\sigma(Y)].$$

BEMERKUNG 7.8.10. Aus der Messbarkeit des bedingten Erwartungswertes bzgl. $\sigma(Y)$ kann man herleiten, dass $\mathbb{E}[X|Y]$ eine Borel–Funktion von Y sein muss. Es gibt also eine Borel–Funktion $g : \mathbb{R} \to \mathbb{R}$ mit

$$\mathbb{E}[X|Y] = g(Y) \text{ fast sicher.}$$

Wir schreiben dann $\mathbb{E}[X|Y = t] = g(t)$. Dabei darf $\mathbb{P}[Y = t]$ auch 0 sein.

BEMERKUNG 7.8.11. Seien X, Y zwei diskrete Zufallsvariablen mit gemeinsamer Zähldichte $f_{X,Y}(s,t) = \mathbb{P}[X = s, Y = t]$ und die Zähldichte von Y sei $f_Y(t)$. Dann gilt für den bedingten Erwartungswert

$$\mathbb{E}[X|Y](\omega) = \mathbb{E}[X|Y = t] = \sum_s \mathbb{P}[X = s|Y = t]s = \frac{\sum_s f_{X,Y}(s,t)s}{f_Y(t)}, \text{ wenn } Y(\omega) = t.$$

Seien X, Y zwei absolut stetige 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:

$$\mathbb{E}[X|Y](\omega) = \mathbb{E}[X|Y = t] = \frac{\int_{\mathbb{R}} f_{X,Y}(s,t)sds}{f_Y(t)}, \text{ wenn } Y(\omega) = t.$$

Diese Formel hat Sinn, wenn $f_Y(t) \neq 0$.

BEISPIEL 7.8.12. In diesem Beispiel betrachten wir zwei faire Münzen. Seien X_1 und X_2 zwei Zufallsvariablen, die den Wert 1 annehmen, wenn die erste bzw. die zweite Münze Kopf zeigt, und den Wert 0 sonst. Sei $Y = X_1 + X_2$. Wir bestimmen $\mathbb{E}[X_1|Y]$.

LÖSUNG. Die Grundmenge ist $\Omega = \{0,1\}^2 = \{00, 01, 10, 11\}$. Wir zerlegen die Grundmenge Ω mit Hilfe von Y in

$$\Omega_0 = Y^{-1}(0) = \{00\}, \quad \Omega_1 = Y^{-1}(1) = \{01, 10\}, \quad \Omega_2 = Y^{-1}(2) = \{11\}.$$

Für $\omega \in \Omega_0$ gilt

$$\mathbb{E}[X_1|Y](\omega) = \frac{\mathbb{E}[X_1 \mathbb{1}_{\Omega_0}]}{\mathbb{P}[\Omega_0]} = \frac{0}{1/4} = 0.$$

Für $\omega \in \Omega_1$ gilt

$$\mathbb{E}[X_1|Y](\omega) = \frac{\mathbb{E}[X_1 \mathbb{1}_{\Omega_1}]}{\mathbb{P}[\Omega_1]} = \frac{1/4}{1/2} = \frac{1}{2}.$$

Für $\omega \in \Omega_1$ gilt

$$\mathbb{E}[X_1|Y](\omega) = \frac{\mathbb{E}[X_1 \mathbb{1}_{\Omega_2}]}{\mathbb{P}[\Omega_2]} = \frac{1/4}{1/4} = 1.$$

Zusammenfassend gilt $\mathbb{E}[X_1|Y] = \mathbb{E}[X_1|X_1 + X_2] = \frac{1}{2}(X_1 + X_2)$.

7.9. Satz von Lehmann–Scheffé

Sei $\{h_\theta : \theta \in \Theta\}$ mit $\Theta = (a, b)$ eine Familie von Dichten bzw. Zähldichten und $X_1, \dots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit Dichte bzw. Zähldichte h_θ.

SATZ 7.9.1 (Lehmann–Scheffé). *Sei $\hat{\theta} \in L^2$ ein erwartungstreuer, suffizienter und vollständiger Schätzer für θ. Dann ist $\hat{\theta}$ der beste erwartungstreue Schätzer für θ.*

Bevor wir den Satz beweisen, betrachten wir eine Reihe von Beispielen.

BEISPIEL 7.9.2. Seien $X_1, \dots, X_n$ unabhängig und gleichverteilt auf $[0, \theta]$, wobei $\theta > 0$ geschätzt werden soll. Wir haben bereits gezeigt, dass $X_{(n)} = \max\{X_1, \dots, X_n\}$ 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

$$\hat{\theta}(X_1, \dots, X_n) := \frac{n+1}{n} X_{(n)} = \frac{n+1}{n} \max\{X_1, \dots, X_n\}.$$

Dieser Schätzer ist erwartungstreu, suffizient und vollständig. Nach dem Satz von Lehmann–Scheffé ist somit $\frac{n+1}{n} X_{(n)}$ der beste erwartungstreue Schätzer für θ.

BEISPIEL 7.9.3. Seien $X_1, \dots, X_n$ unabhängige, mit Parameter $\theta \in [0, 1]$ Bernoulli–verteilte Zufallsvariablen. Der Schätzer $\bar{X}_n$ ist erwartungstreu, suffizient und vollständig und somit 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.

BEISPIEL 7.9.4. Seien $X_1, \dots, X_n$ 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.

BEWEIS VON SATZ 7.9.1. Sei $\hat{\theta}$ ein erwartungstreuer, suffizienter und vollständiger Schätzer für θ. Wir wollen zeigen, dass $\hat{\theta}$ bester erwartungstreuer Schätzer für θ ist. Sei $\varphi : \mathbb{R}^n \to \mathbb{R}$ ein erwartungstreuer Schätzer für 0, d.h. $\mathbb{E}_\theta \varphi = 0$ für alle $\theta \in \Theta$. Wir zeigen, dass

$$\mathbb{E}_\theta[\hat{\theta}\varphi] = 0.$$

Dieser Erwartungswert kann mit Hilfe der bedingten Erwartung umgeschrieben werden zu

$$\mathbb{E}_\theta[\hat{\theta}\varphi] = \mathbb{E}_\theta[\mathbb{E}_\theta[\hat{\theta}\varphi|\hat{\theta}] = \mathbb{E}_\theta[\hat{\theta}\mathbb{E}_\theta[\varphi|\hat{\theta}]] = \mathbb{E}_\theta[\hat{\theta}g(\hat{\theta})],$$

wobei $g(\hat{\theta}) = \mathbb{E}_\theta[\varphi|\hat{\theta}]$. Die Funktion g ist unabhängig von θ, da $\hat{\theta}$ suffizient ist. Es gilt

$$\mathbb{E}_\theta[g(\hat{\theta})] = \mathbb{E}_\theta[\mathbb{E}_\theta[\varphi|\hat{\theta}]] = \mathbb{E}_\theta[\varphi] = 0,$$

da φ ein erwartungstreuer Schätzer für 0 ist. Somit folgt durch die Vollständigkeit von $\hat{\theta}$, dass $g = 0$ fast sicher unter $\mathbb{P}_\theta$ ist. Also gilt

$$\mathbb{E}_\theta[\hat{\theta}\varphi] = \mathbb{E}_\theta[\hat{\theta} \cdot 0] = 0.$$

Wir haben gezeigt, dass $\hat{\theta}$ orthogonal zu jedem erwartungstreuen Schätzer von 0 ist. Laut Satz 7.7.7 ist $\hat{\theta}$ der beste erwartungstreue Schätzer für θ. □

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: χ^2–Verteilung und Erlang–Verteilung);
(2) Student–t–Verteilung;
(3) Fisher–Snedecor–F–Verteilung.

Diese Verteilungen werden wir später für die Konstruktion von Konfidenzintervallen und statistischen Tests benötigen.

8.1. Gammafunktion und Gammaverteilung

DEFINITION 8.1.1. Die *Gammafunktion* ist gegeben durch

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt, \quad \alpha > 0.$$

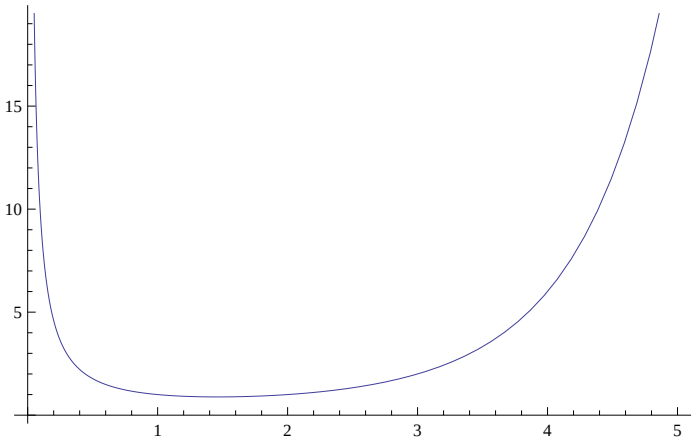

ABBILDUNG 1. Der Graph der Gammafunktion.

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

$$\Gamma\left(\frac{1}{2}\right) = \int_0^\infty t^{-\frac{1}{2}} e^{-t} dt = \int_0^\infty \frac{\sqrt{2}}{w} e^{-\frac{w^2}{2}} w dw = \sqrt{2} \int_0^\infty e^{-\frac{w^2}{2}} dw = \sqrt{\pi},$$

denn $\int_0^\infty e^{-\frac{w^2}{2}} dw = \frac{1}{2}\sqrt{2\pi}$.

DEFINITION 8.1.2. Eine Zufallsvariable X ist *Gammaverteilt* mit Parametern $\alpha > 0$ und $\lambda > 0$, falls für die Dichte von X gilt

$$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \quad x > 0.$$

NOTATION 8.1.3. $X \sim \text{Gamma}(\alpha, \lambda)$.

AUFGABE 8.1.4. Zeigen Sie, dass $\int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} dx = 1$.

BEMERKUNG 8.1.5. Die Gammaverteilung mit Parametern $\alpha = 1$ und $\lambda > 0$ hat Dichte $\lambda e^{-\lambda t}$, $t > 0$, und stimmt somit mit der Exponentialverteilung mit Parameter λ überein.

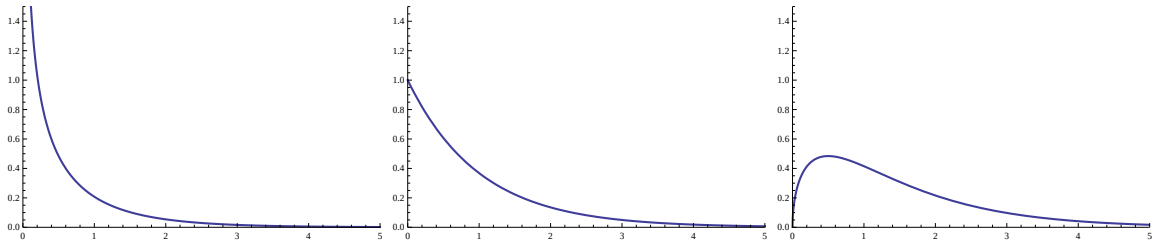

ABBILDUNG 2. Dichten der Gammaverteilungen mit verschiedenen Werten des Parameters α. Links: $\alpha < 1$. Mitte: $\alpha = 1$ (Exponentialverteilung). Rechts: $\alpha > 1$.

SATZ 8.1.6. *Sei $X \sim$ Gamma(α, λ) eine Gammaverteilte Zufallsvariable. Dann sind die Laplace–Transformierte $m_X(t) := \mathbb{E}e^{tX}$ 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 $\text{Re}\, t < \lambda$ konvergent. Somit stellt $m_X(t)$ eine analytische Funktion in der Halbebene $\text{Re}\, t < \lambda$ 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

$$\mathbb{E}X = \frac{\alpha}{\lambda}, \quad \text{Var}\, X = \frac{\alpha}{\lambda^2}.$$

Der nächste Satz heißt die *Faltungseigenschaft der Gammaverteilung.*

SATZ 8.1.8. *Sind die Zufallsvariablen $X \sim Gamma(\alpha, \lambda)$ und $Y \sim Gamma(\beta, \lambda)$ unabhängig, dann gilt für die Summe*

$$X + Y \sim Gamma(\alpha + \beta, \lambda).$$

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 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. χ^2–Verteilung

DEFINITION 8.2.1. Seien $X_1, \ldots, X_n \sim \text{N}(0, 1)$ 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. $X_1^2 + \ldots + X_n^2 \sim \chi_n^2$.

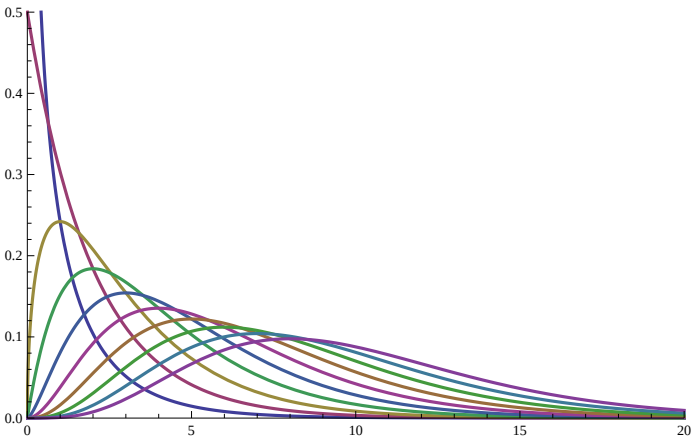

ABBILDUNG 3. Dichten der χ^2–Verteilungen mit $n = 1, \ldots, 10$ Freiheitsgraden.

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 8.2.3. *Sei $X \sim N(0,1)$. Dann ist $X^2 \sim Gamma(\frac{1}{2}, \frac{1}{2})$. Symbolisch: $\chi_1^2 \stackrel{d}{=} Gamma(\frac{1}{2}, \frac{1}{2})$.*

BEWEIS. Wir bestimmen die Laplace–Transformierte von X^2:

$$m_{X^2}(t) = \mathbb{E}e^{tX^2} = \int_{\mathbb{R}} e^{tx^2} \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \right) dx = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{1-2t}{2}x^2} dx = \frac{1}{\sqrt{1-2t}}.$$

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 $X^2 \sim \text{Gamma}(\frac{1}{2}, \frac{1}{2})$. □

SATZ 8.2.4. *Seien $X_1, \ldots, X_n \sim N(0,1)$ unabhängige Zufallsvariablen. Dann gilt*

$$X_1^2 + \ldots + X_n^2 \sim Gamma\left(\frac{n}{2}, \frac{1}{2}\right).$$

Symbolisch: $\chi_n^2 \stackrel{d}{=} Gamma(\frac{n}{2}, \frac{1}{2})$.

BEWEIS. Wir haben bereits gezeigt, dass $X_1^2, \ldots, X_n^2 \sim \text{Gamma}(\frac{1}{2}, \frac{1}{2})$. Außerdem sind die Zufallsvariablen $X_1^2, \ldots, X_n^2$ unabhängig. Der Satz folgt aus der Faltungseigenschaft der Gammaverteilung. □

BEMERKUNG 8.2.5. Die Dichte einer χ_n^2–Verteilung ist somit gegeben durch

$$f(x) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, \quad x > 0.$$

BEISPIEL 8.2.6. Seien $X_1, X_2 \sim \text{N}(0,1)$ unabhängige Zufallsvariablen. Dann gilt

$$X_1^2 + X_2^2 \sim \text{Gamma}\left(1, \frac{1}{2}\right) \sim \text{Exp}\left(\frac{1}{2}\right).$$

Symbolisch: $\chi_2^2 \stackrel{d}{=} \text{Exp}(\frac{1}{2})$.

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 4. Eine Realisierung des Poisson–Prozesses.

Wie ist nun die n–te Erneuerungszeit S_n verteilt? Da $W_i \sim \text{Exp}(\lambda) \sim \text{Gamma}(1, \lambda)$ ist, ergibt sich aus der Faltungseigenschaft der Gammaverteilung, dass

$$S_n \sim \text{Gamma}(n, \lambda).$$

Diese Verteilung (also die $\text{Gamma}(n, \lambda)$–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 $\text{Var}\, S_n = \frac{n}{\lambda^2}$.

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* $N(I) \sim Poi(\lambda l)$.

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:

$$\mathbb{P}[\exists \text{ Erneuerung im Intervall } [0, \delta]] = \mathbb{P}[W_1 < \delta] = 1 - e^{-\lambda\delta} \approx \lambda\delta.$$

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 voneinender. Somit gilt für die Anzahl der Erneuerungen $N(I)$ in einem Intervall I der Länge l

$$N(I) \approx \text{Bin}\left(\frac{l}{\delta}, \lambda\delta\right) \approx \text{Poi}(\lambda l),$$

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

$$\begin{aligned} g(y_1, \ldots, y_n) &= n f(x_1, \ldots, x_n) \\ &= \frac{n}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\left(y_1 - \sum_{i=2}^n y_i\right)^2 - \frac{1}{2}\sum_{i=2}^n (y_1 + y_i)^2\right) \\ &= \frac{n}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{n y_1^2}{2}\right) \exp\left(-\frac{1}{2}\sum_{i=2}^n y_i^2 - \frac{1}{2}\left(\sum_{i=2}^n y_i\right)^2\right). \end{aligned}$$

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 $(n-1)S_n^2 = Z_1 \sim \mathrm{Gamma}(\frac{n-1}{2}, \frac{1}{2}) \stackrel{d}{=} \chi_{n-1}^2$. □

8.5. 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. $V \sim t_r$.

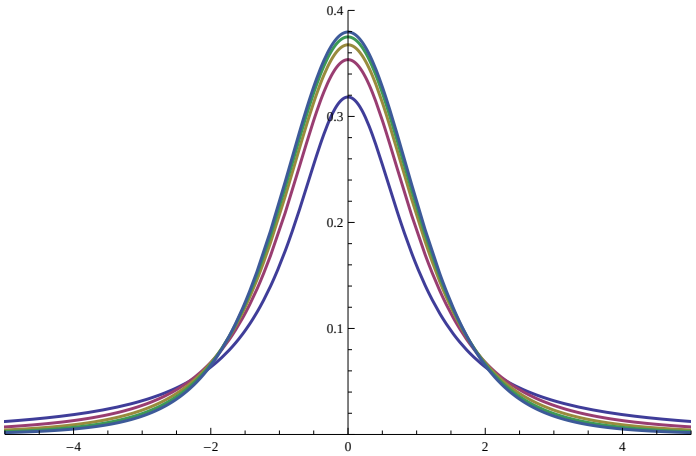

Abbildung 5. Dichten der t_r–Verteilungen mit $r = 1, \ldots, 5$ Freiheitsgraden.

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

$$f_{X,U}(x,u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \cdot \frac{u^{\frac{r-2}{2}} e^{-\frac{u}{2}}}{2^{\frac{r}{2}}\Gamma(\frac{r}{2})}, \quad x \in \mathbb{R},\ u > 0.$$

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)

$$f_{V,W}(v,w) = f_{X,Y}(x,y)\sqrt{\frac{w}{r}} = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{v^2 w}{2r}\right) \frac{w^{\frac{r-2}{2}} e^{-\frac{w}{2}}}{2^{\frac{r}{2}}\Gamma(\frac{r}{2})} \sqrt{\frac{w}{r}}.$$

Somit kann die Dichte von V wie folgt berechnet werden:

$$f_V(v) = \int_0^\infty f_{V,W}(v,w)dw = \frac{1}{\sqrt{2\pi r}2^{r/2}\Gamma(\frac{r}{2})} \int_0^\infty \exp\left(-w\left(\frac{v^2}{2r}+\frac{1}{2}\right)\right) w^{\frac{r+1}{2}-1} dw.$$

Mit der Formel $\int_0^\infty w^{\alpha-1}e^{-\lambda w}dw = \frac{\Gamma(\alpha)}{\lambda^\alpha}$ 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*

$$\sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma} \sim N(0,1), \quad \sqrt{n}\,\frac{\bar{X}_n - \mu}{S_n} \sim t_{n-1}.$$

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. 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. $W \sim F_{r,s}$.

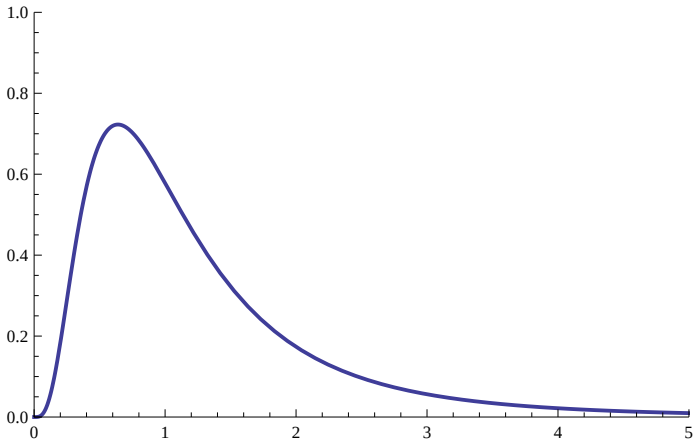

ABBILDUNG 6. Dichte der F–Verteilung mit $(10, 8)$–Freiheitsgraden.

SATZ 8.6.3. *Die Dichte einer F–verteilten Zufallsvariable $W \sim F_{r,s}$ ist gegeben durch*

$$f_W(t) = \frac{\Gamma(\frac{r+s}{2})}{\Gamma(\frac{r}{2})\Gamma(\frac{s}{2})} \left(\frac{r}{s}\right)^{\frac{r}{2}} \frac{t^{\frac{r}{2}-1}}{\left(1 + \frac{r}{s}t\right)^{\frac{r+s}{2}}}, \quad t > 0.$$

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

$$f_{U_r}(x) \propto x^{\frac{r-2}{2}} e^{-x/2}, \quad f_{U_s}(x) \propto x^{\frac{s-2}{2}} e^{-x/2}, \quad x > 0.$$

Somit folgt für die Dichten von U_r/r und U_s/s, dass

$$f_{U_r/r}(x) \propto x^{\frac{r-2}{2}} e^{-rx/2}, \quad f_{U_s/s}(x) \propto x^{\frac{s-2}{2}} e^{-sx/2}, \quad x > 0.$$

Für die Dichte von W gilt die Faltungsformel:

$$f_W(t) = \int_{\mathbb{R}} |y| f_{U_r/r}(yt) f_{U_s/s}(y) dy.$$

Indem wir nun die Dichten von U_r/r und U_s/s einsetzen, erhalten wir, dass

$$f_W(t) \propto \int_0^\infty y(yt)^{\frac{r-2}{2}} e^{-\frac{ryt}{2}} y^{\frac{s-2}{2}} e^{-\frac{sy}{2}} dy \propto t^{\frac{r-2}{2}} \int_0^\infty y^{1+\frac{r-2}{2}+\frac{s-2}{2}} \exp\left(-y\left(\frac{rt}{2}+\frac{s}{2}\right)\right) dy.$$

Mit der Formel $\int_0^\infty y^{\alpha-1} e^{-\lambda y} dy = \frac{\Gamma(\alpha)}{\lambda^\alpha}$ 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

Sei $\{h_\theta(x) : \theta \in \Theta\}$ eine Familie von Dichten bzw. Zähldichten. In diesem Kapitel ist $\Theta = (a, b) \subset \mathbb{R}$ ein Intervall. Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit Dichte bzw. Zähldichte h_θ. Wir haben uns bereits mit der Frage beschäftigt, wie man den Parameter θ anhand der Stichprobe 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 abzuschätzen, 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 9.0.4. Sei $\alpha \in (0, 1)$ eine kleine Zahl, typischerweise $\alpha = 0.01$ oder $\alpha = 0.05$. Es seien $\underline{\theta} : \mathbb{R}^n \to \mathbb{R} \cup \{-\infty\}$ und $\overline{\theta} : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ zwei Stichprobenfunktionen mit

$$\underline{\theta}(x_1, \ldots, x_n) \leq \overline{\theta}(x_1, \ldots, x_n) \quad \text{für alle } x_1, \ldots, x_n \in \mathbb{R}.$$

Wir sagen, dass $[\underline{\theta}, \overline{\theta}]$ ein *Konfidenzintervall* für θ zum Konfidenzniveau $1 - \alpha \in (0, 1)$ ist, falls

$$\mathbb{P}_\theta[\underline{\theta}(X_1, \ldots, X_n) \leq \theta \leq \overline{\theta}(X_1, \ldots, X_n)] \geq 1 - \alpha \text{ für alle } \theta \in \Theta.$$

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 $T(X_1, \ldots, X_n; \theta)$ der Stichprobe $(X_1, \ldots, X_n)$ und des unbekannten Parameters θ mit der Eigenschaft, dass die Verteilung von $T(X_1, \ldots, X_n; \theta)$ unter $\mathbb{P}_\theta$ nicht von θ abhängt und explizit angegeben werden kann. Das heißt, es soll gelten, dass

$$\mathbb{P}_\theta[T(X_1, \ldots, X_n; \theta) \leq t] = F(t),$$

wobei $F(t)$ nicht von θ abhängt. Dabei soll die Funktion $T(X_1, \ldots, X_n; \theta)$ 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

$$\mathbb{P}_\theta \left[Q_{\frac{\alpha}{2}} \leq T(X_1, \ldots, X_n; \theta) \leq Q_{1-\frac{\alpha}{2}} \right] = 1 - \alpha \text{ für alle } \theta \in \Theta.$$

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.

9.1. Konfidenzintervalle für die Parameter der Normalverteilung

In diesem Abschnitt seien $X_1, \dots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ 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 $X_1, \dots, X_n \sim \mathrm{N}(\mu, \sigma^2)$ 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

$$\mathbb{P}_\mu\left[z_{\frac{\alpha}{2}} \le \sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma} \le z_{1-\frac{\alpha}{2}}\right] = 1-\alpha \quad \text{für alle } \mu \in \mathbb{R}.$$

Nach μ umgeformt führt dies zu

$$\mathbb{P}_\mu\left[\bar{X}_n - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X}_n - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] = 1-\alpha \quad \text{für alle } \mu \in \mathbb{R}.$$

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 9.1.1. Man kann auch "nichtsymmetrische" Konfidenzintervalle konstruieren. Wähle dazu $\alpha_1 \ge 0$, $\alpha_2 \ge 0$ mit $\alpha = \alpha_1 + \alpha_2$. Dann gilt

$$\mathbb{P}_\mu\left[z_{\alpha_1} \le \sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma} \le z_{1-\alpha_2}\right] = 1-\alpha \quad \text{für alle } \mu \in \mathbb{R}.$$

Nach μ umgeformt führt dies zu

$$\mathbb{P}_\mu\left[\bar{X}_n - z_{1-\alpha_2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X}_n - z_{\alpha_1}\frac{\sigma}{\sqrt{n}}\right] = 1-\alpha \quad \text{für alle } \mu \in \mathbb{R}.$$

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

$$\mathbb{P}_{\mu,\sigma^2}\left[t_{n-1,\frac{\alpha}{2}} \le \sqrt{n}\,\frac{\bar{X}_n - \mu}{S_n} \le t_{n-1,1-\frac{\alpha}{2}}\right] = 1-\alpha \ \text{ für alle } \mu \in \mathbb{R}, \sigma^2 > 0.$$

Nach μ umgeformt führt dies zu

$$\mathbb{P}_{\mu,\sigma^2}\left[\bar{X}_n - t_{n-1,1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}} \le \mu \le \bar{X}_n - t_{n-1,\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}\right] = 1-\alpha \ \text{ für alle } \mu \in \mathbb{R}, \sigma^2 > 0.$$

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

$$\mathbb{P}_{\sigma^2}\left[\chi^2_{n,\frac{\alpha}{2}} \leq \frac{n\widetilde{S}^2_n}{\sigma^2} \leq \chi^2_{n,1-\frac{\alpha}{2}}\right] = 1-\alpha \text{ für alle } \sigma^2 > 0.$$

Nach σ^2 umgeformt führt dies zu folgendem Konfidenzintervall für σ^2 zum Niveau $1-\alpha$:

$$\left[\frac{n\widetilde{S}^2_n}{\chi^2_{n,1-\frac{\alpha}{2}}}, \frac{n\widetilde{S}^2_n}{\chi^2_{n,\frac{\alpha}{2}}}\right].$$

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

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

Bekannt ist, dass

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

Somit gilt

$$\mathbb{P}_{\mu,\sigma^2}\left[\chi^2_{n-1,\frac{\alpha}{2}} \leq \frac{(n-1)S^2_n}{\sigma^2} \leq \chi^2_{n-1,1-\frac{\alpha}{2}}\right] = 1-\alpha \text{ für alle } \mu \in \mathbb{R}, \sigma^2 > 0.$$

Nach σ^2 umgeformt führt dies zu

$$\mathbb{P}_{\mu,\sigma^2}\left[\frac{(n-1)S^2_n}{\chi^2_{n-1,1-\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{(n-1)S^2_n}{\chi^2_{n-1,\frac{\alpha}{2}}}\right] = 1-\alpha \text{ für alle } \mu \in \mathbb{R}, \sigma^2 > 0.$$

Somit erhält man folgendes Konfidenzintervall für σ^2 zum Niveau $1-\alpha$

$$\left[\frac{(n-1)S^2_n}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{(n-1)S^2_n}{\chi^2_{n-1,\frac{\alpha}{2}}}\right].$$

9.2. Asymptotisches Konfidenzintervall für die Erfolgswahrscheinlichkeit bei Bernoulli–Experimenten

Seien $X_1, \ldots, X_n$ 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 9.2.1. Eine Folge $[\underline{\theta}_1, \overline{\theta}_1], [\underline{\theta}_2, \overline{\theta}_2], \ldots$ 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

$$\liminf_{n\to\infty} \mathbb{P}_\theta[\underline{\theta}_n(X_1, \ldots, X_n) \leq \theta \leq \overline{\theta}_n(X_1, \ldots, X_n)] \geq 1 - \alpha \text{ für alle } \theta \in \Theta.$$

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

$$\lim_{n\to\infty} \mathbb{P}_\theta \left[z_{\frac{\alpha}{2}} \leq \sqrt{n}\frac{\bar{X}_n - \theta}{\sqrt{\theta(1-\theta)}} \leq z_{1-\frac{\alpha}{2}} \right] = 1 - \alpha \text{ für alle } \theta \in (0,1).$$

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:

$$\sqrt{n}\,|\bar{X}_n - \theta| \leq z\sqrt{\theta(1-\theta)}.$$

Quadrierung führt zu

$$n(\bar{X}_n^2 + \theta^2 - 2\bar{X}_n\theta) \leq z^2\theta(1-\theta).$$

Dies lässt sich umschreiben zu

$$g(\theta) := \theta^2\left(1 + \frac{z^2}{n}\right) - \theta\left(2\bar{X}_n + \frac{z^2}{n}\right) + \bar{X}_n^2 \leq 0.$$

Die Funktion $g(\theta)$ ist quadratisch und hat (wie wir gleich sehen werden) zwei verschiedene reelle Nullstellen. Somit ist $g(\theta) \leq 0$ genau dann, wenn θ zwischen diesen beiden Nullstellen liegt. Indem wir nun die Nullstellen mit der p–q–Formel berechnen, erhalten wir folgendes Konfidenzintervall zum Niveau $1 - \alpha$ für θ:

$$\left[\frac{\bar{X}_n + \frac{z^2}{2n} - \frac{z}{\sqrt{n}}\sqrt{\bar{X}_n(1-\bar{X}_n) + \frac{z^2}{4n}}}{1 + \frac{z^2}{n}}, \frac{\bar{X}_n + \frac{z^2}{2n} + \frac{z}{\sqrt{n}}\sqrt{\bar{X}_n(1-\bar{X}_n) + \frac{z^2}{4n}}}{1 + \frac{z^2}{n}} \right].$$

Für großes n erhalten wir die folgende Approximation (indem wir alle Terme mit $1/\sqrt{n}$ stehen lassen und alle Terme mit $1/n$ ignorieren):

$$\left[\bar{X}_n - \frac{z}{\sqrt{n}}\sqrt{\bar{X}_n(1-\bar{X}_n)},\ \bar{X}_n + \frac{z}{\sqrt{n}}\sqrt{\bar{X}_n(1-\bar{X}_n)} \right].$$

Später werden wir diese Approximation mit dem Satz von Slutsky begründen.

BEISPIEL 9.2.2. 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 Mitelwert $\bar{X}_n$ 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.

9.3. Satz von Slutsky

Bei der Konstruktion von Konfidenzintervallen findet der folgende Satz sehr oft Anwendung.

SATZ 9.3.1 (Satz von Slutsky). *Seien $X, X_1, X_2, \ldots$ und $Y, Y_1, Y_2, \ldots$ 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

$$X_nY_n \xrightarrow[n\to\infty]{d} cX.$$

BEWEIS. SCHRITT 1. Es genügt, die punktweise Konvergenz der charakteristischen Funktionen zu zeigen. D.h., wir müssen zeigen, dass

$$\lim_{n\to\infty} \mathbb{E}e^{itX_nY_n} = \mathbb{E}e^{itcX} \text{ für alle } t \in \mathbb{R}.$$

Sei $\varphi(s) = e^{its}$. 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 9.3.2. 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

$$\lim_{n\to\infty}\mathbb{P}_\theta\left[z_{\frac{\alpha}{2}}\le\sqrt{n}\,\frac{\bar{X}_n-\theta}{\sqrt{\bar{X}_n(1-\bar{X}_n)}}\le z_{1-\frac{\alpha}{2}}\right]=1-\alpha\ \text{ für alle }\theta\in(0,1).$$

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 9.3.3. 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.

9.4. Konfidenzintervall für den Erwartungswert der Poissonverteilung

Seien $X_1,\dots,X_n$ 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

$$\lim_{n\to\infty}\mathbb{P}_\theta\left[z_{\frac{\alpha}{2}}\le\sqrt{n}\,\frac{\bar{X}_n-\theta}{\sqrt{\theta}}\le z_{1-\frac{\alpha}{2}}\right]=1-\alpha\ \text{ für alle }\theta>0.$$

Aufgrund der Symmetrieeigenschaft der Standardnormaverteilung gilt $z:=z_{1-\frac{\alpha}{2}}=-z_{\frac{\alpha}{2}}$. Wir erhalten also folgende Ungleichung für θ:

$$\sqrt{n}\,|\bar{X}_n-\theta|\le\sqrt{\theta}z.$$

Dies lässt sich durch Quadrierung umschreiben zu

$$g(\theta):=\theta^2-\theta\left(2\bar{X}_n+\frac{z^2}{n}\right)+\bar{X}_n^2\le 0.$$

Die Ungleichung $g(\theta)\le 0$ 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

$$\left[\bar{X}_n - \frac{z}{\sqrt{n}}\sqrt{\bar{X}_n},\ \bar{X}_n + \frac{z}{\sqrt{n}}\sqrt{\bar{X}_n}\right].$$

9.5. Zweistichprobenprobleme

Bislang haben wir nur sogenannte Einstichprobenprobleme betrachtet. Es gibt aber auch mehrere Probleme, bei denen man zwei Stichproben miteinander vergleichen muss.

Beispiel 9.5.1. 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 9.5.2. 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

$$\mathbb{P}_{\mu_1,\mu_2}\left[z_{\frac{\alpha}{2}} \leq \frac{\bar{X}_n - \bar{Y}_m - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \leq z_{1-\frac{\alpha}{2}}\right] = 1 - \alpha \text{ für alle } \mu_1, \mu_2 \in \mathbb{R}.$$

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

$$\frac{(n-1)S_X^2}{\sigma^2} \sim \chi^2_{n-1}, \quad \frac{(m-1)S_Y^2}{\sigma^2} \sim \chi^2_{m-1}.$$

Bemerke, dass diese zwei χ^2–verteilten Zufallsvariablen unabhängig sind. Somit folgt

$$\frac{(n-1)S_X^2}{\sigma^2} + \frac{(m-1)S_Y^2}{\sigma^2} \sim \chi^2_{n+m-2}.$$

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

$$\mathbb{P}_{\mu_1,\mu_2,\sigma^2}\left[t_{n+m-2,\frac{\alpha}{2}} \leq \frac{\bar{X}_n - \bar{Y}_m - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n} + \frac{1}{m}}} \leq t_{n+m-2,1-\frac{\alpha}{2}}\right] = 1 - \alpha \text{ für alle } \mu_1, \mu_2, \sigma^2.$$

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

$$\mathbb{P}_{\mu_1,\mu_2,\sigma_1^2,\sigma_2^2}\left[F_{n-1,m-1,\frac{\alpha}{2}} \le \frac{S_X^2/\sigma_1^2}{S_Y^2/\sigma_2^2} \le F_{n-1,m-1,1-\frac{\alpha}{2}}\right] = 1-\alpha \ \text{ für alle } \mu_1,\mu_2,\sigma_1^2,\sigma_2^2 > 0.$$

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 9.5.3 (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.

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:

(1) Nullhypothese H_0: Die Münze ist fair, d.h., $\theta = 1/2$.
(2) Alternativhypothese H_1: Die Münze ist nicht fair, d.h., $\theta \neq 1/2$.

Wir müssen uns 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. Dazu wählen wir 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:

(1) Fehler 1. Art: H_0 wird verworfen, obwohl H_0 richtig ist.
(2) Fehler 2. Art: H_0 wird nicht verworfen, obwohl H_0 falsch ist.

Wie sollte nun die Konstante c 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

$$\mathbb{P}_{H_0}[|S - 100| > c] = 2\mathbb{P}_{H_0}[S > 100 + c] = 2 \sum_{k=100+c+1}^{200} \binom{n}{k} \frac{1}{2^n},$$

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, 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. Somit wählen wir $c = 14$. Unsere Entscheidungsregel lautet nun wie folgt:

(1) Wir verwerfen H_0, falls $|S - 100| > 14$.
(2) Sonst behalten wir die Hypothese H_0 bei.

In diesem 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:

$$\frac{c + 0.5}{\sqrt{50}} \geq -z_{\frac{\alpha}{2}}.$$

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.

10.2. Allgemeine Modellbeschreibung

Wir beschreiben nun allgemein das statistische Testproblem. Sei $X = (X_1, \ldots, X_n)$ eine Stichprobe von unabhängigen und identisch verteilten Zufallsvariablen mit Dichte bzw. Zähldichte h_θ, wobei $\theta \in \Theta$ unbekannt sei. Es sei außerdem eine Zerlegung des Parameterraums Θ in zwei disjunkte Teilmengen Θ_0 und Θ_1 gegeben, d.h.

$$\Theta = \Theta_0 \cup \Theta_1, \quad \Theta_0 \cap \Theta_1 = \emptyset.$$

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 $X_1, \ldots, X_n$ entscheiden, ob wir H_0 verwerfen oder beibehalten. Dazu wählen wir eine Borel–Menge $K \subset \mathbb{R}^n$, die Ablehnungsbereich genannt wird. Die Entscheidung wird nun wie folgt getroffen:

(1) Wir verwerfen H_0, falls $(X_1, \ldots, X_n) \in K$.
(2) Wir behalten H_0 bei, falls $(X_1, \ldots, X_n) \notin K$.

Diese Entscheidungsregel kann auch mit Hilfe einer Funktion $\varphi : \mathbb{R}^n \to \{0, 1\}$ formuliert werden, wobei

$$\varphi(x) = \begin{cases} 1, & \text{falls } x \in K, \\ 0, & \text{falls } x \notin K. \end{cases}$$

Die Nullhypothese wird verworfen, falls $\varphi(X_1, \ldots, X_n) = 1$ und wird beibehalten, falls $\varphi(X_1, \ldots, X_n) = 0$. Nun können zwei Arten von Fehlern machen:

(1) Fehler 1. Art: H_0 wird verworfen, obwohl H_0 richtig ist.
(2) Fehler 2. Art: H_0 wird nicht verworfen, obwohl H_0 falsch ist.

Normalerweise versucht man K bzw. φ 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 10.2.1. Eine Borel–Funktion $\varphi : \mathbb{R}^n \to \{0, 1\}$ heißt *Test* zum Niveau $\alpha \in (0, 1)$, falls

$$\mathbb{P}_\theta[\varphi(X_1, \ldots, X_n) = 1] \leq \alpha \quad \text{für alle } \theta \in \Theta_0.$$

Im Folgenden werden wir zahlreiche Beispiele von Tests konstruieren.

10.3. 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 folgende 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 nun verschiedene Hypothesen für μ testen, z. B. $\mu = \mu_0$, $\mu \geq \mu_0$ oder $\mu \leq \mu_0$, wobei μ_0 vorgegeben 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 $|T| > z_{1-\frac{\alpha}{2}}$.

Fall 1B. $H_0 : \mu \geq \mu_0$; $H_1 : \mu < \mu_0$. Die Nullhypothese H_0 sollte verworfen werden, wenn T klein ist. Dies führt zu der Entscheidungsregel, dass H_0 verworfen wird, falls $T < z_\alpha$.

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 $T > z_{1-\alpha}$.

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

$$T := \frac{n\widetilde{S}_n^2}{\sigma_0^2} = \sum_{i=1}^{n} \left(\frac{X_i - \mu_0}{\sigma_0} \right)^2 \sim \chi_n^2.$$

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.

10.4. 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).

10.5. 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 $X_1, X_2, \ldots$ 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:

$$\Theta = \Theta_0 \cup \Theta_1, \quad \Theta_0 \cap \Theta_1 = \emptyset.$$

Wir wollen die Nullhypothese $H_0 : \theta \in \Theta_0$ gegen die Alternativhypothese $H_1 : \theta \in \Theta_1$ testen.

DEFINITION 10.5.1. Eine Folge von Borel–Funktionen $\varphi_1, \varphi_2, \ldots$ mit $\varphi_n : \mathbb{R}^n \to \{0, 1\}$ heißt *asymptotischer Test* zum Niveau $\alpha \in (0, 1)$, falls

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta_0} \mathbb{P}_\theta[\varphi_n(X_1, \ldots, X_n) = 1] \leq \alpha.$$

Dabei ist φ_n die zum Stichprobenumfang n gehörende Entscheidungsregel.

Wir werden nun asymptotische Tests für die Erfolgswahrscheinlichkeit θ bei Bernoulli–Experimenten konstruieren. Seien $X_1, \ldots, X_n$ 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) $Y_1, \ldots, Y_m \sim \mathrm{Bern}(\theta_2)$.

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 $\theta_1 \leq \theta_2$. Ein natürlicher Schätzer für $\theta_1 - \theta_2$ ist $\bar{X}_n - \bar{Y}_m$. Wir definieren uns folgende Größe

$$\tilde{T}_{n,m} = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{\theta_1(1-\theta_1)}{n} + \frac{\theta_2(1-\theta_2)}{m}}}.$$

SATZ 10.5.2. *Unter* $\theta := \theta_1 = \theta_2$ *gilt*

$$\tilde{T}_{n,m} \xrightarrow{d} N(0, 1) \text{ für } n, m \to \infty.$$

BEWEIS. Wir haben die Darstellung $\tilde{T}_{n,m} = Z_{1;n,m} + \ldots + Z_{n+m;n,m}$, wobei

$$Z_{k;n,m} = \begin{cases} \frac{X_k - \theta}{n\sqrt{\theta(1-\theta)}\sqrt{\frac{1}{n}+\frac{1}{m}}}, & \text{falls } k = 1, \ldots, n, \\ -\frac{Y_{k-n} - \theta}{m\sqrt{\theta(1-\theta)}\sqrt{\frac{1}{n}+\frac{1}{m}}}, & \text{falls } k = n+1, \ldots, n+m. \end{cases}$$

Wir wollen den Zentralen Grenzwertsatz von Ljapunow verwenden. Es gilt:

(1) Die Zufallsvariablen $Z_{1;n,m}, \ldots, Z_{n+m;n,m}$ sind unabhängig.
(2) $\mathbb{E}Z_{k;n,m} = 0$.
(3) $\sum_{k=1}^{n+m} \mathbb{E}Z_{k;n,m}^2 = 1$.

Die letzte Eigenschaft kann man folgendermaßen beweisen:

$$\sum_{k=1}^{n+m} \mathbb{E}Z_{k;n,m}^2 = \frac{1}{\theta(1-\theta)\left(\frac{1}{n}+\frac{1}{m}\right)} \left(n \cdot \frac{\theta(1-\theta)}{n^2} + m \cdot \frac{\theta(1-\theta)}{m^2} \right) = 1.$$

Wir müssen also nur noch die Ljapunow–Bedingung überprüfen. Sei $\delta > 0$ beliebig. Es gilt

$$\begin{aligned}
\sum_{k=1}^{n+m} \mathbb{E}|Z_{k;n,m}|^{2+\delta} &= \frac{1}{\sqrt{\theta(1-\theta)}^{2+\delta}\left(\frac{1}{n}+\frac{1}{m}\right)^{\frac{2+\delta}{2}}} \left\{ \frac{n}{n^{2+\delta}} \mathbb{E}|X_1-\theta|^{2+\delta} + \frac{m}{m^{2+\delta}} \mathbb{E}|Y_1-\theta|^{2+\delta} \right\} \\
&\leq \frac{C(\theta)}{\left(\frac{1}{n}+\frac{1}{m}\right)^{\frac{2+\delta}{2}}} \left\{ \frac{1}{n^{1+\delta}} + \frac{1}{m^{1+\delta}} \right\} \\
&= \frac{C(\theta)}{n^{\frac{\delta}{2}}\left(1+\frac{n}{m}\right)^{\frac{2+\delta}{2}}} + \frac{C(\theta)}{m^{\frac{\delta}{2}}\left(\frac{m}{n}+1\right)^{\frac{2+\delta}{2}}},
\end{aligned}$$

was für $n, m \to \infty$ gegen 0 konvergiert. Dabei ist $C(\theta)$ eine von n, m unabhängige Größe. Nach dem Zentralen Grenzwertsatz von Ljapunow folgt die Behauptung des Satzes. □

Die Größe $\tilde{T}_{n,m}$ konvergiert zwar gegen die Standardnormalverteilung, wir können diese Größe allerdings nicht direkt zur Konstruktion von asymptotischen Tests verwenden, denn $\tilde{T}_{n,m}$ beinhaltet die unbekannten Parameter θ_1 und θ_2. Deshalb betrachten wir eine Modifizierung von $\tilde{T}_{n,m}$, in der θ_1 und θ_2 durch die entsprechenden Schätzer $\bar{X}_n$ und $\bar{Y}_m$ ersetzt wurden:

$$T_{n,m} = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n} + \frac{\bar{Y}_m(1-\bar{Y}_m)}{m}}}.$$

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

$$T_{n,m} \xrightarrow{d} \mathrm{N}(0,1) \text{ für } n, m \to \infty.$$

Wir betrachten nun drei verschiedene Nullhypothesen.

Fall A. $H_0 : \theta_1 = \theta_2$; $H_1 : \theta_1 \neq \theta_2$. In diesem Fall sollte H_0 verworfen werden, 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 $T_{n,m} \geq z_{1-\alpha}$.