Einführung in die Stochastik

Vorlesung von Wolfgang Wefelmeyer
Universität zu Köln
Wintersemester 2010/2011
Skript von Markus Schulz

Inhaltsverzeichnis

Vorbemerkung

Die Vorlesung richtet sich an zwei Gruppen von Hörern. Einerseits ist sie eine in sich geschlossene Einführung in einige Begriffe und Methoden der Wahrscheinlichkeitstheorie und Statistik. Sie wendet sich insbesondere an Lehramtsstudenten; viele behandelte Beispiele sind auch für den Unterricht brauchbar. Sie gehört zum Bereich D (Angewandte Mathematik). Andererseits dient die Vorlesung der Einstimmung auf weiterführende Vorlesungen zur Stochastik. In den Bachelor- und Masterstudiengängen gehört die Vorlesung zum Bereich Stochastik und Versicherungsmathematik. Zusammen mit der Vorlesung Wahrscheinlichkeitstheorie I deckt sie das Grundwissen über Stochastik für die Zulassung zur Aktuarausbildung ab. Es ist ratsam, die Einführung schon im dritten Semester zu hören.

Stichworte zum Inhalt der Einführung: Kombinatorik, bedingte Wahrscheinlichkeiten, Bayes-Regel, Gesetz der großen Zahl, zentraler Grenzwertsatz; empirische Schätzer, Maximum-Likelihood-Schätzer, Kernschätzer, lineare Regression, Tests, Konfidenzbereiche.

1 Diskrete Experimente

Definition 1.1. Ein *diskretes Experiment* besteht aus einer endlichen (später auch abzählbaren) Menge $\Omega = \{\omega_1, \omega_2, \ldots\}$ von *Ergebnissen*, denen *Wahrscheinlichkeiten* $P(\{\omega_i\}) = p_i$ zugeordnet sind. Dabei gilt $p_i \geq 0$ und $\sum_i p_i = 1$.

Beispiel 1.1 (Werfen zweier (unterscheidbarer) Münzen). Wir schreiben die Ergebnismenge als $\Omega = \{KK, KZ, ZK, ZZ\}$. Die Ergebnisse KK, KZ, ZK, ZZ sind gleich wahrscheinlich; ihre Wahrscheinlichkeit ist also jeweils $1/4$.

Beispiel 1.2 (Werfen zweier (nicht unterscheidbarer) Münzen). Als Ergebnismenge betrachten wir die möglichen Anzahlen von "Kopf": $\Omega = \{0, 1, 2\}$. Dann gilt $P(\{0\}) = P(\{2\}) = \frac{1}{4}$ und $P(\{1\}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$.

Definition 1.2. Ein *Ereignis* ist eine Menge von Ergebnissen. Die Menge aller möglichen Ereignisse ist die *Potenzmenge* $\mathcal{P}(\Omega)$. Die *Wahrscheinlichkeit* eines Ereignisses A ist $P(A) = \sum_{\omega \in A} P(\{\omega\})$.

Erinnerung: Für $A, B \subset \Omega$ sind folgende *Mengenoperationen* definiert:

$$\begin{aligned}
A^c &= \{\omega \in \Omega : \omega \notin A\} \quad \textit{Komplement},\\
A \cup B &= \{\omega \in \Omega : \omega \in A \vee \omega \in B\} \quad \textit{Vereinigung},\\
A \cap B &= \{\omega \in \Omega : \omega \in A \wedge \omega \in B\} \quad \textit{Durchschnitt},\\
A - B = A \backslash B &= \{\omega \in \Omega : \omega \in A \wedge \omega \notin B\} = A \cap B^c \quad \textit{Differenz},\\
A \triangle B &= (A \cup B) - (A \cap B) \quad \textit{symmetrische Differenz}.
\end{aligned}$$

Definition 1.3. Zwei Ereignisse A und B heißen *disjunkt*, wenn $A \cap B = \emptyset$. Dann schreiben wir $A \cup B = A + B$. Ereignisse A_i, $i \in \mathbb{N}$, heißen *paarweise disjunkt*, wenn je zwei Ereignisse A_i und A_j, $j \neq i$, disjunkt sind. Wir schreiben dann auch $\bigcup_{i \in \mathbb{N}} A_i = \sum_{i=1}^{\infty} A_i$.

Regeln von de Morgan: Zwischen Vereinigung, Durchschnitt und Komplement bestehen folgende Beziehungen für beliebige Ereignisse A_i:

$$\Big(\bigcap_i A_i\Big)^c = \bigcup_i A_i^c, \qquad \Big(\bigcup_i A_i\Big)^c = \bigcap_i A_i^c.$$

Bemerkung 1.1. Durch die Wahrscheinlichkeiten der Ereignisse ist eine Abbildung $P : \mathcal{P}(\Omega) \to [0, 1]$ definiert, für die gilt:

$$\begin{aligned}
P(\Omega) &= 1,\\
P(A + B) &= P(A) + P(B).
\end{aligned}$$

Beweis. Wegen $p_i \geq 0$ ist $P(A) \geq 0$. Es gilt $P(\Omega) = 1$, da $\sum_i p_i = 1$. Man rechnet

$$P(A) + P(B) = \sum_{\omega \in A} P(\{\omega\}) + \sum_{\omega \in B} P(\{\omega\}) = \sum_{\omega \in A+B} P(\{\omega\}) = P(A + B).$$

□

Beispiel 1.3. Bei einer medizinischen Studie könnte man die Ergebnisse "Zustand verbessert", "Zustand gleich" und "Zustand verschlechtert" als $\Omega = \{-1, 0, 1\}$ modelieren.

Beispiel 1.4. An einer Klausur nehmen n Personen teil. Wir numerieren sie durch. Erreichen alle unterschiedliche Punktzahlen, so läßt sich eine Rangliste als Permutation $\{\omega_1, \ldots, \omega_n\}$ von $\{1, \ldots, n\}$ bilden. Teilnehmer 1 hat den Rang ω_1, usw.
Gibt es möglicherweise Teilnehmer mit gleicher Punktzahl, ist die Beschreibung komplizierter. Haben a_1 Teilnehmer die höchste erreichte Punktzahl, so haben sie alle Rang 1. Die Rangliste läßt sich durch eine Permutation des Vektors

$$(1, \ldots, 1, 2, \ldots, 2, \ldots, k, \ldots, k)$$

ausdrücken, in dem a_1-mal 1, a_2-mal 2 usw. steht. Es gilt $\sum_{j=1}^{k} a_j = n$ mit $a_j \geq 1$ und $k \geq 1$. Teilnehmer 1 hat wieder den Rang, der an der ersten Stelle der Permutation steht.

Definition 1.4. Eine Abbildung $X : \Omega \to \mathbb{R}$ heißt *Zufallsvariable*.

Beispiel 1.5 (Zwei (unterscheidbare) Münzen). Der Ergebnisraum für den Wurf zweier Münzen ist $\Omega = \{(i, j) : i, j \in \{0, 1\}\}$, wenn 0 für "Zahl" und 1 für "Kopf" steht. Jedes Ergebnis besitzt die Wahrscheinlichkeit $\frac{1}{4}$. Die Zufallsvariable X gebe nun die Anzahl der Münzen mit "Kopf" an, also $X(i, j) = i + j$. Es gilt

$$\begin{aligned} P(\text{mindestens einmal Kopf}) &= P(X \geq 1) = P(\{\omega : X(\omega) \geq 1\}) \\ &= P(\{(0, 1), (1, 0), (1, 1)\}) = \frac{3}{4}. \end{aligned}$$

Definition 1.5. Die *Verteilung* von X ist gegeben durch

$$P^X(\{b\}) = P(X = b) = P(\{\omega : X(\omega) = b\}) = \sum_{\omega : X(\omega) = b} P(\{\omega\}).$$

Nach Bemerkung 1.1 ist P^X ein Wahrscheinlichkeitsmaß auf dem Wertebereich von X (vgl. Kapitel 8).

Beispiel 1.6 (Zwei (unterscheidbare) Münzen). Wie oben sei $\Omega = \{(i, j) : i, j \in \{0, 1\}\}$ und $P(\{(i, j)\}) = \frac{1}{4}$. Als Zufallsvariablen betrachten wir $X_1(i, j) = i$ und $X_2(i, j) = j$, die *Projektionen*. Dann gilt

$$\begin{aligned} P(\text{erste Münze Kopf}) &= P(X_1 = 1) = 1/2 \\ P(\text{mindestens einmal Kopf}) &= P(X_1 + X_2 \geq 1) = 3/4. \end{aligned}$$

Bemerkung 1.2. Die Verteilung einer Zufallsvariablen X läßt sich mit Hilfe der *Urbildabbildung* ausdrücken:

$$X^{-1}(B) = \{\omega : X(\omega) \in B\} = (X \in B).$$

Insbesondere haben wir $X^{-1}(\{b\}) = \{\omega : X(\omega) = b\}$, also $P^X(\{b\}) = P(X^{-1}(\{b\}))$.

Laplacesche Experimente: Sei Ω endlich. Die *Laplace-Verteilung* (oder *Gleichverteilung*) über Ω ist definiert durch $P(\{\omega\}) = 1/|\Omega|$. Hier und im folgenden bezeichnet der Betrag einer Menge die Anzahl ihrer Elemente, also $|A| = \#\{\omega : \omega \in A\}$. Ein *Laplacesches Experiment* ist ein diskretes Experiment, in dem die Verteilung durch eine Laplace-Verteilung gegeben ist. Für Ereignisse A gilt in diesem Fall

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{Anzahl der günstigen Ergebnisse}}{\text{Anzahl der möglichen Ergebnisse}}.$$

Wahrscheinlichkeiten in Laplaceschen Experimenten berechnet man mit Hilfe der *Kombinatorik*, die im nächsten Abschnitt behandelt wird.

2 Kombinatorik

A. Permutationen. Sei $k \in \{1, \ldots, n\}$. *Permutationen* sind k-Tupel aus verschiedenen Zahlen $1, \ldots, n$. Ihre Anzahl ist

$$(n)_k = n(n-1)\ldots(n-k+1) = \frac{n!}{(n-k)!}.$$

Denn für die Belegung der ersten Stelle des k-Tupels gibt es n Möglichkeiten, für die der zweiten Stelle nur noch $n-1$, usw.
Eine äquivalente Beschreibung der Anzahl der Permutationen: Die Anzahl der Möglichkeiten, k unterscheidbare Kugeln auf n Fächer *ohne* Mehrfachbelegung zu verteilen. Denn für die erste Kugel stehen n Fächer zur Verfügung, für die zweite nur noch $n-1$, usw.
Eine weitere äquivalente Beschreibung: Die Anzahl der Möglichkeiten, aus einer Urne mit n numerierten Kugeln k unter Beachtung der Reihenfolge und ohne Zurücklegen zu ziehen.
Eine weitere äquivalente Beschreibung: Die Anzahl der Abbildungen $\{1, \ldots, k\} \to \{1, \ldots, n\}$, die *injektiv* sind. Denn 1 kann auf einen von n Werten abgebildet werden, 2 kann wegen der Injektivität nur noch auf einen der übrigen $n-1$ Werte abgebildet werden, usw.
Es hängt von der Anwendung ab, welche Beschreibung die bequemste ist.
Für $k = n$ gilt speziell $(n)_k = n(n-1)\ldots 1 = n!$. Unter "Permutationen" versteht man häufig diesen Spezialfall.

Beispiel 2.1. Fünf Gäste sollen auf zehn Zimmer verteilt werden. Dafür gibt es $(10)_5 = 30.240$ Möglichkeiten.
Fünf Kinder sollen sich für ein Gruppenbild in einer Reihe aufstellen. Es gibt 5! mögliche Anordnungen.

B. Kombinationen. Sei $k \in \{0, 1, \ldots, n\}$. *Kombinationen* sind k-elementige Teilmengen von $\{1, \ldots, n\}$. Ihre Anzahl ist

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Begründung: Es gibt $(n)_k$ k-Tupel. Für jedes k-Tupel gibt es $k!$ Anordnungen, die alle aber die gleiche Teilmenge bilden, also nicht unterschieden werden. Also ist die gesuchte Anzahl

$$\frac{(n)_k}{k!} = \frac{n!}{k!(n-k)!} = \binom{n}{k}.$$

Zweite Begründung: Induktion über n. Bezeichne $C_{n,k}$ die gesuchte Zahl. Es gilt $C_{1,0} = 1$ und $C_{1,1} = 1$. Sei $k \in \{0, \ldots, n\}$. An den k-elementigen Teilmengen von $1, \ldots, n+1$ ist die Zahl $n+1$ entweder nicht beteiligt oder beteiligt. Es gilt also $C_{n+1,k} = C_{n,k} + C_{n,k-1}$. Diese Beziehung gilt tatsächlich für die Binomialkoeffizienten. Der Schritt von $k = n$ auf $k = n+1$ ist trivial wegen $C_{n+1,n+1} = 1$.
Eine äquivalente Beschreibung der Anzahl der Kombinationen: Die Anzahl der Möglichkeiten, k *nicht* unterscheidbare Kugeln *ohne* Mehrfachbelegung auf n Fächer zu verteilen. Denn die belegten Fächer beschreiben eine Teilmenge von $\{1, \ldots, n\}$.

Eine weitere äquivalente Beschreibung: Die Anzahl der Möglichkeiten, aus einer Urne mit n numerierten Kugeln k ohne Beachtung der Reihenfolge und ohne Zurücklegen zu ziehen.
Eine weitere äquivalente Beschreibung: Die Anzahl der Abbildungen $\{1,\dots,n\} \to \{0,1\}$ mit $\#f^{-1}(\{1\}) = k$. Denn die Zahlen, die auf 1 abgebildet werden, bilden eine k-elementige Menge.

Beispiel 2.2. Beim Lotto "6 aus 49" gibt es $\binom{49}{6} = 13.983.816$ mögliche Ziehungsergebnisse.

Beispiel 2.3. Ordne a Nullen und b Einsen in einer Reihe an. Dafür hat man $\binom{a+b}{a} = \binom{a+b}{b}$ Möglichkeiten.

Beispiel 2.4. Gegeben seien n_j Zeichen der Sorte j, und k Sorten, also insgesamt $n = n_1 + \cdots + n_k$ Zeichen. Die Anzahl der Anordnungen ist gegeben durch den *Multinomialkoeffizienten*

$$\binom{n}{n_1 \dots n_k} = \frac{n!}{n_1! \cdots n_k!}.$$

Denn für Sorte 1 gibt es $\binom{n}{n_1}$ Möglichkeiten, die Zeichen auf n Stellen zu verteilen; für Sorte 2 hat man anschließend $\binom{n-n_1}{n_2}$ Möglichkeiten. Für beide zusammen gibt es also

$$\binom{n}{n_1}\binom{n-n_1}{n_2} = \frac{n!}{n_1! n_2! (n-n_1-n_2)!}$$

mögliche Anordnungen. Jetzt Sorte 3 hinzunehmen, usw.

C. Variationen. Sei $k \in \mathbb{N}$. *Variationen* sind k-Tupel aus Zahlen $1,\dots,n$. Ihre Anzahl ist n^k. Denn für jede Stelle des k-Tupels gibt es n Möglichkeiten.
Eine äquivalente Beschreibung der Anzahl der Anzahl der Variationen: Die Anzahl der Möglichkeiten, k unterscheidbare Kugeln auf n Fächer zu verteilen.
Eine weitere äquivalente Beschreibung: Die Anzahl der Möglichkeiten, aus einer Urne mit n numerierten Kugeln k mit Beachtung der Reihenfolge und mit Zurücklegen zu ziehen.
Eine weitere äquivalente Beschreibung: Die Anzahl aller Abbildungen $\{1,\dots,k\} \to \{1,\dots,n\}$.

Beispiel 2.5. "Kombinations"-Schloß.

3 Anwendungen der Kombinatorik

Beispiel 3.1. Würden Sie darauf wetten, mit 4 Würfeln mindestens eine 6 zu würfeln? Die Wahrscheinlichkeit dafür kann man über das Gegenereignis "keine 6" errechnen. (Diesen Trick, zunächst das Komplement des gesuchten Ereignisses zu betrachten, werden wir noch häufig anwenden.) Numeriere die Würfel durch. Mit dem Abschnitt über Variationen erkennt man, daßes 6^4 mögliche Wurfergebnisse gibt, von denen 5^4 für das Komplementärereignis günstig sind. Mit dem Laplace-Ansatz erhalten wir

$$P(\text{keine 6}) = \frac{5^4}{6^4}, \quad \text{also} \quad P(\text{mindestens eine 6}) = 1 - \frac{5^4}{6^4} \doteq 0.52 > 1/2.$$

(In der Frage waren die Würfel nicht numeriert. Aber nur durch Numerieren erhalten wir lauter gleichwahrscheinliche Ergebnisse und können Kombinatorik anwenden. Auch dieser Trick wird uns noch häufig helfen.)

Beispiel 3.2 (Olympialotterie von 1971). In einer Trommel befinden sich je 7 Kugeln mit den Ziffern $0, \ldots, 9$. Wir entnehmen nacheinander 7 Kugeln und erhalten eine 7-stellige Zahl. Die Lose mit diesen Zahlen waren gleich teuer. War das fair?
Es ist übersichtlicher, das Problem etwas allgemeiner zu fassen. (Der Grund ist, daß die 7 in zwei Bedeutungen auftritt.) Es sollen in der Trommel je m Kugeln der Ziffern $0, \ldots, 9$ sein, von denen wir $k \leq m$ Kugeln ziehen. Wir numerieren im Geiste die Kugeln mit der gleichen Ziffer jeweils durch. Dann gibt es insgesamt$(10m)_k$ mögliche Ziehungsergebnisse. Für die Ziehung $(0, \ldots, 0)$ gibt es $(m)_k$ günstige Ergebnisse, also hat dieses Ereignis die Wahrscheinlichkeit $(m)_k/(10m)_k$. Für $k = m = 7$ ist dies ungefähr $8.3 \cdot 10^{-10}$. Für die Ziehung $(0, 1, \ldots, k-1)$ gibt es hingegen m^k günstige Ergebnisse, also hat dieses Ereignis die Wahrscheinlichkeit $m^k/(10m)_k$. Für $k = m = 7$ ist dies ungefähr $1.3 \cdot 10^{-7}$. Die Ziehung $(0, 1, \ldots, k-1)$ ist also ungefähr 150-mal so wahrscheinlich wie die Ziehung $(0, \ldots, 0)$, und ein Los mit der letzteren Nummer hätte entsprechend billiger sein sollen.

Beispiel 3.3. Eine Gesellschaft bestehe aus S Personen. Wie wahrscheinlich ist es, daß mindestens eine Person am selben Tag Geburtstag hat wie der Gastgeber? Wir numerieren die Tage eines Jahres durch. Die Gäste numerieren wir mit $1, \ldots, S-1$. Der Gastgeber habe am Tag i_S Geburtstag. Der Ergebnisraum ist $\Omega = \{(i_1, \ldots, i_{S-1}) : i_j \in \{1, \ldots, 365\}\}$. Auf dem Umweg über das Gegenereignis erhalten wir die Wahrscheinlichkeit $1 - 364^{S-1}/365^{S-1}$. In der folgenden Tabelle sind einige Zahlenwerte angegeben:

$S-1$	5	10	15	20	40	$\cdots$	253
Wahrsch.	0.01	0.03	0.04	0.05	0.1	$\cdots$	0.5

Erst bei mehr als 253 Gästen wird die gesuchte Wahrscheinlichkeit größer als 1/2.

Beispiel 3.4. Wiederum bestehe die Gesellschaft aus S Personen. Wie wahrscheinlich ist es, daß mindestens zwei Personen am gleichen Tag Geburtstag haben? Wir numerieren die Gesellschaft mit $1, \ldots, S$. Es gibt $(365)_S$ ungünstige Ereignisse. Die gesuchte Wahrscheinlichkeit ist also $1 - (365)_S/365^S$. Hier einige Zahlenwerte:

S	5	10	15	20	23	40	55
Wahrsch.	0.03	0.12	0.25	0.41	0.51	0.84	0.99

Schon ab 23 Gästen ist die gesuchte Wahrscheinlichkeit größer als 1/2.

Einschluß-Ausschluß-Formel (Siebformel von Poincaré–Sylvester). Seien $A_1, \ldots, A_n$ endliche Teilmengen von Ω. Wir erhalten

$$|A_1 \cup A_2| = |A_1| + |A_2| - |A_1 \cap A_2|$$

und

$$|A_1 \cup A_2 \cup A_3| = |A_1| + |A_2| + |A_3| - |A_1 \cap A_2| - |A_1 \cap A_3| - |A_2 \cap A_3| + |A_1 \cap A_2 \cap A_3|.$$

Allgemein gilt

$$\begin{aligned} \Big|\bigcup_{i=1}^{n} A_i\Big| &= \sum_{i=1}^{n} |A_i| - \sum_{\{i,j\} \subset \{1,\ldots,n\}} |A_i \cap A_j| + \cdots + (-1)^{n-1} |A_1 \cap \ldots \cap A_n| \\ &= \sum_{m=1}^{n} (-1)^{m-1} \sum_{\{i_1,\ldots,i_m\} \subset \{1,\ldots,n\}} \Big|\bigcap_{j=1}^{m} A_{i_j}\Big|. \end{aligned}$$

Oft hängt die Größe der Durchschnitte nur von der *Anzahl* der beteiligten Mengen ab, das heißt, für jedes m gibt es eine Zahl $c(m)$, so daß

$$\Big|\bigcap_{j=1}^{m} A_{i_j}\Big| = c(m)$$

für alle m-elementigen Teilmengen $\{i_1, \ldots, i_m\} \subset \{1, \ldots, n\}$. Dann vereinfacht sich die Einschluß-Ausschluß-Formel zu

$$\Big|\bigcup_{i=1}^{n} A_i\Big| = \sum_{m=1}^{n} (-1)^{m-1} \binom{n}{m} c(m).$$

Beispiel 3.5. Verteile k Kugeln auf n Fächer mit möglicher Mehrfachbelegung. Wie groß ist die Wahrscheinlichkeit, daß mindestens ein Fach leer bleibt?
Für $k < n$ ist die Wahrscheinlichkeit 1. Sei also $k \geq n$. Numeriere die Kugeln durch. Für $i = 1, \ldots, n$ sei A_i die Menge der Zuordnungen, bei denen Fach i leer bleibt. Nach der Formel für die Anzahl der Variationen gilt

$$|A_{i_1} \cap \ldots \cap A_{i_m}| = (n-m)^k = c(m).$$

Die spezielle Einschluß-Ausschluß-Formel liefert für die Anzahl der günstigen Zuordnungen

$$\Big|\bigcup_{i=1}^{n} A_i\Big| = \sum_{m=1}^{n} (-1)^{m-1} \binom{n}{m} (n-m)^k.$$

Die Anzahl aller möglichen Zuordnungen ist n^k. Also ist die gesuchte Wahrscheinlichkeit

$$\frac{1}{n^k}\Big|\bigcup_{i=1}^{n} A_i\Big| = \sum_{m=1}^{n} (-1)^{m-1} \binom{n}{m} \Big(1 - \frac{m}{n}\Big)^k.$$

Beispiel 3.6 (Rencontre, matching, Koinzidenzen. Montmort 1713). Wir betrachten Permutationen der Zahlen $1, \ldots, n$. Wie wahrscheinlich ist es, daß (mindestens) eine Zahl unverändert bleibt?
Sei A_i die Menge der Permutationen, bei denen die Stelle i unverändert bleibt. Es gilt

$$|A_{i_1} \cap \ldots \cap A_{i_m}| = (n-m)! = c(m).$$

Die spezielle Einschluß-Ausschluß-Formel liefert für die Anzahl der günstigen Zuordnungen

$$\Big|\bigcup_{i=1}^{n} A_i\Big| = \sum_{m=1}^{n} (-1)^{m-1} \binom{n}{m} (n-m)!.$$

Die Anzahl aller Permutationen ist $n!$. Die gesuchte Wahrscheinlichkeit ist also

$$\frac{1}{n!}\Big|\bigcup_{i=1}^{n} A_i\Big| = \sum_{m=1}^{n} (-1)^{m-1} \binom{n}{m} \frac{(n-m)!}{n!} = \sum_{m=1}^{n} (-1)^{m-1} \frac{1}{m!}.$$

Für großes n ist das ungefähr $1 - e^{-1} \doteq 0.63$, also fast $2/3$.
Bringt jeder in einer Gruppe ein Geschenk mit, und werden die Geschenke dann zufällig an die Gruppenmitglieder verteilt, so ist es auch bei sehr großen Gruppen recht wahrscheinlich, daß einer sein eigenes Geschenk erhält.

Beispiel 3.7 (Hypergeometrische Verteilung). Eine Lieferung der Größe N enthält K defekte Stücke. Entnehme n Stücke. Mit welcher Wahrscheinlichkeit enthält diese Stichprobe genau k defekte Stücke?
(Diesmal numerieren wir *nicht*.) Die Anzahl aller möglichen Stichproben ist $\binom{N}{n}$. Wir haben $\binom{K}{k}$ Möglichkeiten, k aus K defekten Stücken auszuwählen, und $\binom{N-K}{n-k}$ Möglichkeiten, $n-k$ aus $N-K$ nicht defekten Stücken auszuwählen. Die Anzahl der Stichproben mit k defekten Stücken ist also $\binom{K}{k}\binom{N-K}{n-k}$. Dabei muß k zwischen $\max\{0, K+n-N\}$ und $\min\{n, K\}$ liegen. Die gesuchte Wahrscheinlichkeit ist

$$H_{N,K,n}\{k\} = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

Beispiel 3.8. Eine Lieferung der Größe $N = 1000$ soll nur höchstens 2% Ausschuss enthalten. Der Empfänger darf vertragsgemäß die Lieferung zurückweisen, wenn eine Stichprobe der Größe $n = 10$ mindestens ein defektes Stück enthält. Ist das vernünftig?
Bei K defekten Stücken weist der Empfänger die Lieferung mit Wahrscheinlichkeit

$$H_{1000,K,10}\{1, \ldots, 10\} = 1 - H_{1000,K,10}\{0\}$$

zurück. Für $K = 20$ hat dies ungefähr den Wert 0.18. Von den gerade noch akzeptablen Lieferungen wird also fast jede fünfte zurückgewiesen. (Das wäre etwas zu streng.)

4 Gekoppelte Experimente

Beispiel 4.1 (Irrfahrt). In einem Spiel gewinnen wir 1 Euro mit Wahrscheinlichkeit p und verlieren 1 Euro mit Wahrscheinlichkeit $q = 1-p$. Wir wiederholen das Spiel mehrmals. Unser Startvermögen vor dem ersten Spiel sei 0 Euro. Unser Vermögen nach Spiel 1 ist dann 1 Euro mit Wahrscheinlichkeit $p_1(1) = p$ und -1 Euro mit Wahrscheinlichkeit $p_1(-1) = q$.
Falls wir das erste Spiel gewonnen haben, ist unser Vermögen nach dem zweiten Spiel 2 Euro mit Wahrscheinlichkeit $p_2(2|1) = p$ und 0 Euro mit Wahrscheinlichkeit $p_2(0|1) = q$. Wir nennen solche Wahrscheinlichkeiten "bedingt" (auf das Ergebnis des ersten Spiels).
Entsprechend gilt: Falls wir das erste Spiel verloren haben, ist unser Vermögen nach dem zweiten Spiel 0 Euro mit Wahrscheinlichkeit $p_2(0|-1) = p$ und -2 Euro mit Wahrscheinlichkeit $p_2(-2|-1) = q$.
Das Experiment heißt "gekoppelt", weil unser Vermögen nach dem zweiten Spiel von dem nach dem ersten Spiel abhängt. Das können wir mit einem "Baum" illustrieren. Nach zwei Spielen sei die Entwicklung unseres Vermögens (ein "Pfad") durch $(1, 0)$ gegeben. (Wir haben das erste Spiel gewonnen, das zweite verloren.) Die Wahrscheinlichkeit dafür ist dann

$$p(1, 0) = p_1(1)p_2(0|1) = pq.$$

Analog gilt $p(-1, 0) = qp$, $p(1, 2) = pp$ und $p(-1, -2) = qq$.

Definition 4.1. Seien $\Omega_1, \ldots, \Omega_n$ endlich, und seien p_1 Wahrscheinlichkeiten auf Ω_1, $p_2(\cdot|\omega_1)$ auf Ω_2, ..., $p_n(\cdot|\omega_1, \ldots, \omega_n)$ auf Ω_n. Das *gekoppelte Experiment* (n-stufiges Experiment) gibt einem *Pfad* $(\omega_1, \ldots, \omega_n)$ die Wahrscheinlichkeit

$$p(\omega_1, \ldots, \omega_n) = p_1(\omega_1) \cdot p_2(\omega_2|\omega_1) \cdots p_n(\omega_n|\omega_1, \ldots, \omega_{n-1}).$$

Ein solches Experiment heißt *stochastischer Prozeß*. Für das zugehörige Wahrscheinlichkeitsmaß gilt:

$$P(A_1 \times \cdots \times A_n) = \sum_{\omega_1 \in A_1} p_1(\omega_1) \sum_{\omega_2 \in A_2} p_2(\omega_2|\omega_1) \cdots \sum_{\omega_n \in A_n} p_n(\omega_n|\omega_1, \ldots, \omega_{n-1}).$$

Hängen die p_j nur von ω_{j-1} ab, so heißt der Prozeß *Markov-Kette*.
Hängen die p_j *nicht* von $\omega_1, \ldots, \omega_{j-1}$ ab, dann heißt das Experiment ein *unabhängiges Produkt* von Experimenten. In diesem Fall erhalten für das zugehörige Wahrscheinlichkeitsmaß

$$P(A_1 \times \ldots \times A_n) = P_1(A_1) \cdots P_n(A_n)$$

mit

$$P_j(A_j) = \sum_{\omega_j \in A_j} p_j(\omega_j).$$

Wir nennen P ein *unabhängiges Produkt* und schreiben wir $P = P_1 \otimes \ldots \otimes P_n$.

Beispiel 4.2 (Hardy–Weinberg-Gesetz). Ein Allel komme in einer Population in den beiden Ausprägungen A und a vor. Die Genotypen AA, $Aa(= aA)$ und aa treten dann mit Wahrscheinlichkeiten $u, 2v, w$ auf, wobei $u + 2v + w = 1$ gelten muß. Die Paarungen erfolgen entsprechend den Wahrscheinlichkeiten der Genotypen in der Population. Es ergeben sich die folgenden Wahrscheinlichkeiten.

Genotyp der Eltern	dessen Wahrscheinlichkeit	bedingte Wahrscheinlichkeit der Genotypen in der nächsten Generation		
		AA	Aa oder aA	aa
$AA \times AA$	u^2	1	0	0
$AA \times Aa$	$2uv$	1/2	1/2	0
$Aa \times AA$	$2uv$	1/2	1/2	0
$AA \times aa$	uw	0	1	0
$Aa \times Aa$	$4v^2$	1/4	1/2	1/4
$aa \times AA$	uw	0	1	0
$Aa \times aa$	$2vw$	0	1/2	1/2
$aa \times Aa$	$2vw$	0	1/2	1/2
$aa \times aa$	w^2	0	0	1

Für die nächste Generation resultieren also für die Genotypen die Wahrscheinlichkeiten

$$u_1 = p_1(AA) = u^2 + \frac{2uv}{2} + \frac{2uv}{2} + \frac{4v^2}{4} = (u+v)^2,$$
$$w_1 = p_1(aa) = w^2 + \frac{2vw}{2} + \frac{2vw}{2} + \frac{4v^2}{4} = (v+w)^2,$$
$$2v_1 = 1 - (u+v)^2 - (v+w)^2 = 2(u+v)(v+w).$$

In der übernächsten Generation gilt dann

$$u_2 = (u_1+v_1)^2 = \big((u+v)^2+(u+v)(v+w)\big)^2 = (u+v)^2(u+v+v+w)^2 = (u+v)^2 = u_1.$$

Analog erhält man $w_2 = (v+w)^2 = w_1$ und $v_2 = v_1$. Nach der ersten Generation ändert sich also die Verteilung der Genotypen nicht mehr.

5 Bedingte Wahrscheinlichkeiten

Bei den gekoppelten Experimenten hatten wir mit bedingten Wahrscheinlichkeiten begonnen und daraus Wahrscheinlichkeiten für Pfade berechnet. Jetzt gehen wir umgekehrt vor.

Beispiel 5.1 (Roulette). Wir nehmen vereinfacht an, daß von den Zahlen $0, \dots, 36$ die ungeraden Zahlen rot und die geraden schwarz seien; 0 sei grün. Wir wissen:
a) Die Wahrscheinlichkeit, eine 7 zu ziehen, beträgt 1/37.
b) Wenn bekannt ist, daß eine rote Zahl gezogen wurde, dann ist es mit Wahrscheinlichkeit 1/18 die 7.
Der Ergebnisraum ist $\Omega = \{0, \dots, 36\}$ mit $p(j) = 1/37$. Es gilt $A = \{\text{rot}\} = \{1, 3, \dots, 37\}$ und $B = \{7\}$. Die (bedingte) Wahrscheinlichkeit, daß die gezogene Zahl die 7 ist, wenn bekannt ist, daß die gezogene Zahl rot ist, ergibt sich als

$$P(B|A) = \frac{1}{18} = \frac{1/37}{18/37} = \frac{P(A \cap B)}{P(A)}.$$

Das ist der Anteil von B an A.

Definition 5.1. Gegeben sei ein Grundraum Ω mit einem Wahrscheinlichkeitsmaß P und Ereignisse $A, B \subset \Omega$ mit $P(A) > 0$. Die *bedingte Wahrscheinlichkeit von B gegeben A* ist definiert als

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

Satz 5.1. *Für jede Menge A ist $P(\cdot|A)$ ist ein Wahrscheinlichkeitsmaß.*

Beweis. Es gilt $P(B|A) = P(A \cap B)/PA \geq 0$ und $P(\Omega|A) = P(A)/P(A) = 1$. Schließlich gilt noch die Additivität,

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

□

Beispiel 5.2 (Roulette als gekoppeltes Experiment). Wir können (künstlich) ein einzelnes Roulettespiel als gekoppeltes Experiment wie in Kapitel 4 auffassen, indem wir erst die Farbe (g, r oder s) und dann die gezogene Zahl angeben. Der Ergebnisraum ist dann

$$\Omega = \{g, r, s\} \times \{0, 1, \dots, 36\}.$$

Für die Farben gelten die Wahrscheinlichkeiten $p_1(g) = 1/37$ und $p_1(r) = p_1(s) = 18/37$, und die bedingten Wahrscheinlichkeiten der Zahlen sind dann $p_2(0|g) = 1$ und $p_2(i|g) = 0$ für $i = 1, \dots, 36$ und

$$p_2(i|r) = \begin{cases} 1/18, & i = 1, 3, \dots, 35, \\ 0, & i = 0, 2, 4, \dots, 36, \end{cases} \qquad p_2(i|s) = \begin{cases} 1/18, & i = 2, 4, \dots, 36, \\ 0, & i = 0, 1, 3, \dots, 35. \end{cases}$$

Die Wahrscheinlichkeit, die 7 zu ziehen, ist dann

$$p(7) = p(r, 7) = p_1(r)p_2(7|r) = \frac{18}{37}\frac{1}{18} = \frac{1}{37}.$$

Beispiel 5.3 (Würfeln). Wir werfen erst einen roten und dann einen schwarzen Würfel.
a) Die Wahrscheinlichkeit, daß der rote Würfel eine 6 zeigt, ist dann 1/6.
b) Die bedingte Wahrscheinlichkeit, daß der rote Würfel eine 6 zeigt, wenn die Augensumme beider Würfel 11 ist, beträgt

$$P(\text{rote 6}|\text{Augensumme 11}) = \frac{P(\text{rote 6 und Augensumme 11})}{P(\text{Augensumme 11})} = \frac{1/36}{2/36} = \frac{1}{2}.$$

Das sieht man auch direkt: Möglich für 11 sind $(5,6)$ und $(6,5)$; günstig für eine rote 6 ist $(6,5)$.

Der nächste Satz zeigt, wie sich ein Wahrscheinlichkeitsmaß aus bedingten Wahrscheinlichkeiten rekonstruieren läßt.

Satz 5.2 (Satz von der totalen Wahrscheinlichkeit). *Ist* $\Omega = \sum_{k=1}^{n} A_k$ *mit* $P(A_k) > 0$*, dann gilt*

$$P(B) = \sum_{k=1}^{n} P(B|A_k)P(A_k).$$

Beweis. Mit der Additivität des Wahrscheinlichkeitsmaßes und der Definition der bedingten Wahrscheinlichkeit folgt

$$P(B) = \sum_{k=1}^{n} P(B \cap A_k) = \sum_{k=1}^{n} \frac{P(B \cap A_k)}{P(A_k)} P(A_k) = \sum_{k=1}^{n} P(B|A_k)P(A_k).$$

□

Der nächste Satz zeigt, wie man bedingte Wahrscheinlichkeiten aus bedingten Wahrscheinlichkeiten in der umgekehrten Richtung rekonstruiert.

Satz 5.3 (Bayessche Regel). *Unter den Voraussetzungen von Satz 5.2 gilt*

$$P(A_m|B) = \frac{P(B|A_m)P(A_m)}{P(B)} = \frac{P(B|A_m)P(A_m)}{\sum_{k=1}^{n} P(B|A_k)P(A_k)}.$$

Beweis. Es gilt

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

Wende dann auf den Nenner Satz 5.2 an. □

Beispiel 5.4 (Diagnose). Ein Test stuft eine Person mit 5% Wahrscheinlichkeit fälschlich als krank ein, mit 10% Wahrscheinlichkeit fälschlich als gesund. Der Anteil der Kranken an den Getesteten beträgt 2%.
Bezeichne die *Zustände* "gesund" und "krank" mit G und K, und die *Diagnosen* "gesund" bzw. "krank"mit g bzw. k. Gegeben sind die bedingten Wahrscheinlichkeiten

$$p(k|G) = 0.05, \qquad p(g|K) = 0.1, \qquad p(K) = 0.02.$$

Die Gegenwahrscheinlichkeiten sind dann

$$p(g|G) = 0.95, \qquad p(k|K) = 0.9, \qquad p(G) = 0.98.$$

Uns interessieren die umgekehrten bedingten Wahrscheinlichkeiten.
a) Wie wahrscheinlich ist es, daß eine als gesund diagnostizierte Person tatsächlich krank ist?
Satz 5.3 liefert

$$p(K|g) = \frac{p(g|K)p(K)}{p(g|K)p(K) + p(g|G)p(G)} = \frac{0,1 \cdot 0.02}{0,1 \cdot 0.02 + 0.95 \cdot 0.98} \approx 0.002.$$

b) Wie wahrscheinlich ist es, daß eine als krank diagnostizierte Person tatsächlich gesund ist?
Mit Satz 5.3 erhalten wir

$$p(G|k) = \frac{p(k|G)p(G)}{p(k|G)p(G) + p(k|K)p(K)} \approx 0.73.$$

Der Unterschied erklärt sich aus dem geringen Anteil der Kranken.

6 Unabhängige Ereignisse

Beispiel 6.1 (Würfeln). Wir werfen erst einen roten und dann einen schwarzen Würfel. Es gilt $P(\text{rote 6}) = 1/6$, aber

$$P(\text{rote 6}|\text{Augensumme 11}) = \frac{1}{2},$$
$$P(\text{rote 6}|\text{Augensumme nicht 11}) = \frac{5/36}{34/36} = \frac{5}{34}.$$

Hier sind die bedingten Wahrscheinlichkeiten nicht gleich der unbedingten. Aber

$$P(\text{rote 6}|\text{Augensumme gerade}) = \frac{3/36}{18/36} = \frac{3}{18} = \frac{1}{6},$$
$$P(\text{rote 6}|\text{Augensumme ungerade}) = \frac{3/36}{18/36} = \frac{1}{6}.$$

Hier stimmen die bedingten Wahrscheinlichkeiten mit der unbedingten überein. Der nächste Satz erklärt, warum das so ist.

Satz 6.1. *Es gilt $P(A|B) = P(A|B^c)$ genau dann, wenn $P(A \cap B) = P(A)P(B)$, also $P(A|B) = P(A)$ und $P(B|A) = P(B)$.*

Beweis. $P(A|B) = P(A|B^c)$ bedeutet, daß

$$\frac{P(A \cap B)}{P(B)} = \frac{P(A \cap B^c)}{P(B^c)}.$$

Dann gilt

$$P(A) = P(A \cap B) + P(A \cap B^c) = P(A \cap B)\Big(1 + \frac{P(B^c)}{P(B)}\Big) = \frac{P(A \cap B)}{P(B)}.$$

Gilt umgekehrt $P(A \cap B) = P(A)P(B)$, so erhalten wir

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

Aus $P(A \cap B) = P(A)P(B)$ folgt auch

$$P(A \cap B^c) = P(A) - P(A \cap B) = P(A)(1 - P(B)) = P(A)P(B^c);$$

also gilt auch $P(A|B^c) = P(A)$. □

Definition 6.1. Sei P ein Wahrscheinlichkeitsmaß auf Ω und $A, B \subset \Omega$. Dann heißen A und B *unabhängig*, wenn

$$P(A \cap B) = P(A)P(B).$$

Definition 6.2. Sei P ein Wahrscheinlichkeitsmaß auf Ω. Die Ereignisse $A_1, \ldots, A_n$ in Ω heißen *unabhängig*, wenn für $k = 2, \ldots, n$ und alle Teilmengen $\{i_1, \ldots, i_k\}$ von $\{1, \ldots, n\}$ gilt:

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \cdots P(A_{i_k}).$$

Bemerkung 6.1. Um Unabhängigkeit nachzuweisen, reicht es *nicht* aus, die Produktformel nur für $k = n$ nachzuweisen. Seien zum Beispiel A und B abhängig und $C = \emptyset$. Dann gilt $P(A \cap B \cap C) = P(A)P(B)P(C)$, obwohl die Ereignisse nicht unabhängig sind.

Bemerkung 6.2. *Paarweise* Unabhängigkeit reicht ebenfalls *nicht* aus, um Unabhängigkeit nachzuweisen. Werfen wir zum Beispiel nacheinander zwei Münzen und betrachten die Ereignisse $A = \{\text{erste Münze Kopf}\}$, $B = \{\text{zweite Münze Kopf}\}$ und $C = \{\text{genau einmal Kopf}\}$. Man sieht sofort, daß diese Ereignisse *paarweise* unabhängig sind, aber $0 = P(\emptyset) = P(A \cap B \cap C) \neq P(A)P(B)P(C) = 1/8$.

7 Mittelwert und Erwartungswert

Definition 7.1. Sei $\Omega = \{a_1, \ldots, a_n\} \subset \mathbb{R}$. Das Wahrscheinlichkeitsmaß P auf Ω sei gegeben durch $P(\{a_i\}) = p_i$. Der *Mittelwert* von P ist

$$\mu(P) = \sum_{i=1}^{n} p_i a_i.$$

Beispiel 7.1. Bei 2 von 2000 Losen gewinnt man 1000 Euro, sonst gewinnt man nur 1 Euro. Der erwartete Gewinn ist der Mittelwert

$$\mu(P) = \frac{1998}{2000} 1 + \frac{2}{2000} 1000 \approx 1.999.$$

Ein Lospreis von 2 Euro wäre also schon (knapp) ungünstig.

Beispiel 7.2. Wir werfen einen Würfel. Dann ergibt sich die "mittlere" Augenzahl als $\frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3.5$.

Definition 7.2. Sei X eine Zufallsvariable auf (Ω, P). Der *Erwartungswert* von X ist der Mittelwert von P^X.

Die Zufallsvariable X auf Ω nehme die Werte $a_1, \ldots, a_n$ an. Dann ist die induzierte Verteilung P^X von X gegeben durch $p_i = P^X(\{a_i\}) = P(\{\omega \in \Omega : X(\omega) = a_i\}) = P(X = a_i)$. Der Erwartungswert von X ergibt sich also als

$$EX = \sum_{i=1}^{n} p_i a_i = \sum_{i=1}^{n} P(X = a_i) a_i.$$

Beispiel 7.3. Sei $\Omega = \{1, \ldots, 2000\}$ und $P(\{\omega_i\}) = 1/2000$. Setze $X(1) = X(2) = 1000$ und $X(j) = 1$ sonst. Diese Zufallsvariable beschreibt den möglichen Gewinn im obigen Beispiel mit den Losen. Der erwartete Gewinn ergibt sich jetzt als

$$EX = \mu(P^X) = P(X = 1) \cdot 1 + P(X = 1000) \cdot 1000 = \frac{1998}{2000} 1 + \frac{2}{2000} 1000.$$

8 Die Kolmogorov-Axiome und das Lebesgue-Borel-Maß

In diesem Kapitel führen wir allgemeine Experimente ein. Sie können auch abzählbar oder sogar überabzählbar viele Ergebnisse haben, zum Beispiel beliebige reellwertige Ergebnisse. Wir brauchen (für die Theorie) die Additivität von Wahrscheinlichkeitsmaßen auch für *abzählbar* viele disjunkte Ereignisse. Dann können wir nicht mehr für *alle* Ereignisse Wahrscheinlichkeiten einführen.

Beispiel 8.1. Betrachte das Intervall $[0, 1]$. Die Wahrscheinlichkeit eines Teilintervalls (a, b) sei durch das *Lebesgue-Maß* $P((a, b)) = b - a$ gegeben. Es ist nicht möglich, dieses Maß abzählbar additiv auf *alle* Teilmengen von $[0, 1]$ fortzusetzen. (Das soll hier nicht begründet werden.)

Wir umgehen das Problem im folgenden, indem wir uns auf "einfache" Ereignisse beschränken. Das reicht für alle Anwendungen. Zur Bequemlichkeit nehmen wir zusätzlich an, daß Komplemente und abzählbare Vereinigungen wieder zu diesen "einfachen" Ereignissen gehören.

Die Kolmogorovschen Axiome: Unter einem *Wahrscheinlichkeitsraum* (oder *(Zufalls-)Experiment*) verstehen wir einen Grundraum $\Omega \neq \emptyset$ mit einer *σ-Algebra* $\mathcal{F}$ über Ω (einen *meßbaren Raum*) und einem *Wahrscheinlichkeitsmaß* P auf $\mathcal{F}$.

Definition 8.1. Ein Mengensystem $\mathcal{F} \subset \mathcal{P}(\Omega)$ ist eine *σ-Algebra*, wenn es folgende Eigenschaften besitzt:

(i) $\Omega \in \mathcal{F}$;

(ii) Wenn $A \in \mathcal{F}$ dann auch $A^c \in \mathcal{F}$;

(iii) Wenn $A_i \in \mathcal{F}$ für $i \in \mathbb{N}$, dann auch $\bigcup_{i \in \mathbb{N}} A_i \in \mathcal{F}$.

Definition 8.2. Eine Mengenfunktion P auf $\mathcal{F}$ heißt *Wahrscheinlichkeitsmaß*, falls $P(\Omega) = 1$, $P(A) \geq 0$ für $A \in \mathcal{F}$ gilt und die Mengenfunktion abzählbar additiv (*σ-additiv*) ist:

$$P\Big(\sum_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} P(A_i) \quad \text{für paarweise disjunkte } A_i \in \mathcal{F}.$$

Gilt $P(\Omega) = 1$ *nicht*, dann heißt P *Maß*.

Lemma 8.1. *Ist $\mathcal{S} \subset \mathcal{P}(\Omega)$, dann existiert eine kleinste σ-Algebra $\sigma(\mathcal{S})$, die $\mathcal{S}$ enthält.*

Eine wichtige σ-Algebra ist die Borel-Algebra über $\mathbb{R}^m$. Sie wird von den Quadern erzeugt. Im folgenden bezeichne $\langle a, b \rangle$ ein offenes, halboffenes oder abgeschlossenes Intervall.

Definition 8.3. Die *Borel-(σ-)Algebra* $\mathcal{B}^m$ ist die kleinste σ-Algebra, die alle Quader $\langle a_1, b_1\rangle \times \cdots \times \langle a_m, b_m\rangle$ enthält.

$\mathcal{B}^m$ wird zu Beispiel auch schon von den abgeschlossenen "unteren Quadranten" $(-\infty, q_1] \times \cdots \times (-\infty, q_m]$, $q_i \in \mathbb{Q}$, erzeugt.

Ein wichtiges Maß auf der Borel-Algebra $\mathcal{B}^m$ ist das *Lebesgue–Borel-Maß* λ^m. Es mißt das "Volumen" einer Menge. Auf dem Einheitswürfel $[0,1]^m$ ist es ein Wahrscheinlichkeitsmaß. Für eine (m-dimensionalen) beschränkten Quader ist das Volumen das Produkt der Kantenlängen

$$\lambda^m\big(\langle a_1, b_1\rangle \times \cdots \times \langle a_m, b_m\rangle\big) = \prod_{i=1}^{m}(b_i - a_i).$$

Satz 8.2. *Das Lebesgue–Borel-Maß λ^m läßt sich von den beschränkten Quadern eindeutig zu einem Maß auf $\mathcal{B}^m$ fortsetzen.*

9 Lebesgue-Integral

A. Zufallsvariablen: Um das Lebesgue-Integral zu definieren, führen wir "Zufallsvariablen" allgemeiner als in Kapitel 1 ein. Seien $(\Omega, \mathcal{F})$ und $(\Omega', \mathcal{F}')$ meßbare Räume.

Definition 9.1. Eine Abbildung $X : \Omega \to \Omega'$ heißt *meßbar*, wenn Urbilder meßbarer Mengen meßbar sind; das heißt, für alle $A' \in \mathcal{F}'$ gilt

$$X^{-1}(A') = \{\omega : X(\omega) \in A'\} = (X \in A') \in \mathcal{F}.$$

Ist $\mathcal{F}' = \sigma(\mathcal{S})$, so genügt es zu zeigen, daß $X^{-1}(S) \in \mathcal{F}$ für alle $S \in \mathcal{S}$. Dazu muß man nur nachprüfen, daß $\mathcal{T} = \{A' \subset \Omega' : X^{-1}(A') \in \mathcal{F}\}$ eine σ-Algebra ist. Da sie $\mathcal{S}$ enthält, muß dann $\mathcal{F}' = \sigma(\mathcal{S}) \subset \mathcal{T}$ gelten.

Meßbare Funktionen mit $\Omega' = \mathbb{R}^m$ heißen *Zufallsvektoren*; mit $\Omega' = \mathbb{R}$ heißen sie *Zufallsvariablen*.

Proposition 9.1. *Die Menge der Zufallsvariablen ist abgeschlossen unter Linearkombinationen, abzählbaren Suprema, abzählbaren Infima und Limites.*

Beweis. Es reicht zu zeigen, daß Urbilder offener unterer Halbstrahlen meßbar sind. Für die Skalarmultiplikation mit $\alpha > 0$ gilt das wegen $(\alpha X < r) = (X < r/\alpha)$. Für die Addition schreibt man

$$(X + Y < r) = \bigcup_{u \in \mathbb{Q}} \big((X < u) \cap (Y < r - u)\big).$$

□

Zufallsvariablen lassen sich als Limites meßbarer Treppenfunktionen charakterisieren. Dazu führen wir zunächst die *Indikatorfunktion* einer Menge A ein,

$$1_A(\omega) = \begin{cases} 1, & \omega \in A. \\ 0, & \omega \notin A. \end{cases}$$

Definition 9.2. Eine Funktion $X : \Omega \to \mathbb{R}$ heißt *Treppenfunktion*, wenn sie nur endlich viele Werte annimmt.

Eine Treppenfunktion läßt sich in der Form $\sum_{i=1}^{n} a_i 1_{A_i}$ mit paarweise disjunkten A_i schreiben. Sie ist meßbar genau dann, wenn alle A_i meßbar sind.

Proposition 9.2. *Sei $X : \Omega \to [0, \infty)$ eine (nichtnegative) Zufallsvariable. Dann existieren meßbare Treppenfunktionen X_n mit $X_n \uparrow X$ punktweise, und umgekehrt.*

Beweis. Wähle

$$X_n(\omega) = \begin{cases} (k-1)2^{-n}, & (k-1)2^{-n} \le X(\omega) < k2^{-n}, k = 1, \dots, n2^n, \\ n, & X(\omega) \ge n. \end{cases}$$

□

B. Lebesgue-Integral: Das Lebesgue-Integral wird durch "algebraische Induktion" eingeführt: Man beginnt mit den einfachsten Integranden und setzt das Integral dann linear und stetig fort.

Sei $(\Omega, \mathcal{F})$ ein meßbarer Raum und μ ein Maß auf $\mathcal{F}$. Für meßbare Indikatorfunktionen $X = 1_A$ definieren wir das Integral als

$$\int X \, d\mu = \mu(A).$$

Für meßbare Treppenfunktionen $X = \sum_{i=1}^{n} a_i 1_{A_i}$ definieren wir das Integral als

$$\int X \, d\mu = \sum a_i \mu(A_i).$$

Für nichtnegative Zufallsvariablen X wählen wir meßbare Treppenfunktionen X_n mit $X_n \uparrow X$ und definieren das Integral als

$$\int X \, d\mu = \sup \int X_n \, d\mu.$$

Bezeichne $X^+ = X 1_{(X \ge 0)}$ und $X^- = -X 1_{(X \le 0)}$ den *Positiv-* bzw. *Negativteil* einer Zufallsvariablen X. Dann gilt $X = X^+ - X^-$. Wir nennen X *integrierbar*, wenn $\int X^+ \, d\mu$ oder $\int X^- \, d\mu$ endlich ist, und definieren das Integral als

$$\int X \, d\mu = \int X^+ \, d\mu - \int X^- \, d\mu.$$

Die folgenden nützlichen Konvergenzsätze beweisen wir nicht.

Satz 9.3 (Levi)**.** *Seien X_n nichtnegative Zufallsvariablen mit $X_n \uparrow X$ punktweise, dann ist X integrierbar, und*

$$\int X_n dP \uparrow \int X dP.$$

Satz 9.4 (Lebesgue). *Seien X_n und X Zufallsvariablen mit $X_n \to X$ punktweise, und gilt $|X_n| \leq Y$ für ein Y mit $\int Y dP < \infty$, so gilt*

$$\int X_n dP \to \int X dP.$$

C. Induzierte Maße: Seien $(\Omega, \mathcal{F})$ und $(\Omega', \mathcal{F}')$ meßbare Räume, $\mu|\mathcal{F}$ ein Maß und $X : \Omega \to \Omega'$ eine meßbare Abbildung. Das durch X und μ auf Ω' *induzierte* Maß μ^X ist definiert durch

$$\mu^X(A') = \mu(X^{-1}(A')) = \mu(X \in A'), \quad A' \in \mathcal{F}'.$$

Daß das tatsächlich ein Maß ist, rechnet man wie folgt nach. Offensichtlich gilt $\mu^X(A') = \mu(X \in A') \geq 0$. Sind $A'_n \in \mathcal{F}'$ paarweise disjunkt, so auch $X^{-1}(A'_n)$. Daher folgt die σ-Additivität von μ^X,

$$\begin{aligned}\mu^X\Big(\sum_n A'_n\Big) &= \mu\Big(X^{-1}\Big(\sum_n A'_n\Big)\Big) = \mu\Big(\sum_n X^{-1}(A'_n)\Big) \\ &= \sum_n \mu(X^{-1}(A'_n)) = \sum_n \mu^X(A'_n).\end{aligned}$$

Proposition 9.5. *Sei $(\Omega'', \mathcal{F}'')$ ein weiterer meßbarer Raum und $Y : \Omega' \to \Omega''$ eine weitere meßbare Abbildung. Dann gilt für alle $C'' \in \mathcal{F}''$:*

$$(\mu^X)^Y(C'') = \mu^{Y \circ X}(C'').$$

Beweis. Für $C'' \in \mathcal{F}''$ gilt

$$\begin{aligned}(\mu^X)^Y(C'') &= \mu^X(Y^{-1}(C'')) = \mu(X^{-1}(Y^{-1}(C''))) \\ &= \mu((Y \circ X)^{-1}(C'')) = \mu^{Y \circ X}(C'').\end{aligned}$$

□

D. Integrale bzgl. induzierter Maße: Gegeben seien ein Maßraum $(\Omega, \mathcal{F}, \mu)$, meßbare Räume $(\Omega', \mathcal{F}')$ und $(\mathbb{R}, \mathcal{B})$ sowie meßbare Abbildungen $X : \Omega \to \Omega'$ und $Y : \Omega' \to \mathbb{R}$.

Proposition 9.6. *Ist $Y \circ X$ μ-integrierbar, so gilt*

$$\int Y d\mu^X = \int Y \circ X d\mu.$$

Beweis. Wir verwenden wieder algebraische Induktion wie bei der Konstruktion des Lebesgue-Integrals. Für meßbare Indikatorfunktionen $Y = 1_B$ gilt

$$\int 1_B d\mu^X = \mu^X(B) = \mu(X^{-1}(B)) = \int 1_{X^{-1}(B)} d\mu = \int 1_B \circ X d\mu.$$

Für meßbare Treppenfunktionen $Y = \sum_{i=1}^n b_i 1_{B_i}$ gilt

$$\int Y d\mu^X = \sum_i b_i \int 1_{B_i} d\mu^X = \sum_i b_i \int 1_{B_i} \circ X d\mu = \int Y \circ X d\mu.$$

Für nichtnegative Zufallsvariablen $Y \geq 0$ wählen wir meßbare Treppenfunktionen Y_n mit $Y_n \uparrow Y$, also auch $Y_n \circ X \uparrow Y \circ X$. Wir haben gerade gezeigt, daß $\int Y_n d\mu^X = \int Y_n \circ X d\mu$. Jetzt wenden wir auf beiden Seiten den Satz von Levi an.
Für eine beliebige Zufallsvariable Y schreiben wir $Y = Y^+ - Y^-$. Die Behauptung folgt wegen $(Y \circ X)^+ = Y^+ \circ X$ und $(Y \circ X)^- = Y^- \circ X$. □

Definition 9.3. Sei $(\Omega, \mathcal{F}, P)$ ein Wahrscheinlichkeitsraum. Der *Mittelwert* von P ist $\mu(P) = \int xP(dx)$. Der *Erwartungswert* von X ist $\int XdP = \int xP^X(dx)$.

10 Produktmaße und unabhängige Zufallsvariablen

Die Unabhängigkeit von *Ereignissen* haben wir in Kapitel 6 beschrieben. Die Unabhängigkeit von *Experimenten* beschreibt man am besten mit Produktmaßen (auf Produkträumen).

Definition 10.1. Gegeben seien meßbare Räume $(\Omega_i, \mathcal{F}_i)$, $i \in \mathbb{N}$. Die von den meßbaren *Zylindern*

$$A_1 \times \cdots \times A_n \times \Omega_{n+1} \times \cdots, \quad A_i \in \mathcal{F}_i,$$

erzeugte σ-Algebra $\otimes_{i\in\mathbb{N}}\mathcal{F}_i$ auf $\times_{i\in\mathbb{N}}\Omega_i$ heißt *Produkt-σ-Algebra.*

Ist P ein Wahrscheinlichkeitsmaß auf $\otimes_i \mathcal{F}_i$, so heißt das durch

$$P_i A_i = P(\Omega_1 \times \cdots \times \Omega_{i-1} \times A_i \times \Omega_{i+1} \times \cdots)$$

definierte Wahrscheinlichkeitsmaß die i-te *Randverteilung.*

Satz 10.1. *Sind $(\Omega_i, \mathcal{F}_i, P_i)$ Wahrscheinlichkeitsräume, so läßt sich die Mengenfunktion*

$$P(A_1 \times \cdots \times A_n \times \Omega_{n+1} \times \cdots) = P_1(A_1) \cdots P_n(A_n)$$

eindeutig auf $\otimes_i \mathcal{F}_i$ zu einem Wahrscheinlichkeitsmaß fortsetzen. Es wird $\otimes_i P_i$ geschrieben und heißt Produktmaß.

Definition 10.2. Sei $(\Omega, \mathcal{F}, P)$ ein Wahrscheinlichkeitsraum. Zwei Mengen $A, B \in \mathcal{F}$ heißen *unabhängig*, wenn

$$P(A \cap B) = P(A)P(B).$$

Definition 10.3. Zwei Mengensysteme $\mathcal{G}, \mathcal{H} \subset \mathcal{F}$ heißen *unabhängig*, wenn A und B unabhängig sind für alle $A \in \mathcal{G}$ und $B \in \mathcal{H}$.

Definition 10.4. Seien $(\Omega, \mathcal{F}, P)$ ein Wahrscheinlichkeitsraum. Außerdem seien $(\Omega', \mathcal{F}')$ und $(\Omega'', \mathcal{F}'')$ meßbare Räume sowie $X : \Omega \to \Omega'$ und $Y : \Omega \to \Omega''$ meßbare Abbildungen. Dann heißen X, Y *unabhängig*, wenn $X^{-1}(\mathcal{F}'), Y^{-1}(\mathcal{F}'')$ unabhängig sind.

Bemerkung 10.1. Folgende Aussagen sind äquivalent:

1. X, Y sind unabhängig.
2. $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$ für $A \in \mathcal{F}', B \in \mathcal{F}''$.

3. $P^{(X,Y)}(A \times B) = P^X(A)P^Y(B) \quad$ für $A \in \mathcal{F}', B \in \mathcal{F}''$.

4. $P^{(X,Y)} = P^X \otimes P^Y$.

5. $P^{(X,Y)}$ ist ein Produktmaß.

Seien Ω_1, Ω_2 endliche Mengen. Dann ist die Gleichverteilung auf $\Omega_1 \times \Omega_2$ das unabhängige Produkt der Gleichverteilungen auf Ω_i. Denn für $A_1 \times A_2$ mit $A_1 \in \mathcal{F}_1$ und $A_2 \in \mathcal{F}_2$ gilt

$$\frac{|A_1 \times A_2|}{|\Omega_1 \times \Omega_2|} = \frac{|A_1|}{|\Omega_1|}\frac{|A_2|}{|\Omega_2|}.$$

Die Unabhängigkeit zweier Münzen, Würfel oder Roulettespiele hatten wir oben immer stillschweigend angenommen.

Definition 10.5. Mengen $A_i \in \mathcal{F}$, $i = 1, \ldots, n$, $n > 2$, heißen *unabhängig*, wenn für alle $k \geq 2$ und alle Indexmengen $\{i_1, \ldots, i_k\} \subset \{1, \ldots, n\}$ gilt

$$P\Big(\bigcap_{j=1}^{k} A_{i_j}\Big) = \prod_{j=1}^{k} P(A_{i_j}).$$

Entsprechend für Mengensysteme und Zufallsvariablen.

11 Einige diskrete Verteilungen

Wir definieren diskrete Experimente jetzt etwas allgemeiner als in Kapitel 1. Die σ-Additivität bekommen wir dabei geschenkt.

Proposition 11.1. *Sei $(\Omega, \mathcal{F})$ ein meßbarer Raum. $\mathcal{F}$ enthalte alle Einpunktmengen. Gegeben seien $\omega_i \in \Omega$, $i \in \mathbb{N}$, mit Wahrscheinlichkeiten $p_i \geq 0$ und $\sum_{i\in\mathbb{N}} p_i = 1$. Dann definiert*

$$P(A) = \sum_{\omega_i \in A} p_i, \quad A \in \mathcal{F},$$

ein Wahrscheinlichkeitsmaß auf $\mathcal{F}$. Es heißt diskret.

Beweis. Die Eigenschaften $P(A) \geq 0$ und $P(\Omega) = 1$ sind offensichtlich erfüllt. Seien nun $A_j \in \mathcal{F}$ für $j \in \mathbb{N}$ paarweise disjunkte Mengen. Dann gilt mit dem großen Umordnungssatz

$$P\Big(\sum_{j\in\mathbb{N}} A_j\Big) = \sum_{\omega_i \in \sum_{j\in\mathbb{N}} A_j} p_i = \sum_{j\in\mathbb{N}} \sum_{\omega_i \in A_j} p_i = \sum_{j\in\mathbb{N}} P(A_j).$$

□

Eine Zufallsvariable $X : \Omega \to \mathbb{R}$ heißt *diskret*, wenn das induzierte Wahrscheinlichkeitsmaß $P^X|\mathcal{B}$ diskret ist.

Beispiel 11.1. Eine Zufallsvariable X heißt *Laplace-verteilt* (oder *gleichverteilt*) auf $\{a_1, \ldots, a_n\}$, wenn $P(X = a_i) = 1/n$.

Beispiel 11.2. Eine Zufallsvariable X heißt *hypergeometrisch* verteilt mit den Parametern N, K und n, in Zeichen $X \sim H_{N,K,n}$, wenn

$$P(X=k) = H_{N,K,n}\{k\} = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}, \quad k = \max\{0, K+n-N\}, \ldots, \min\{n, K\}.$$

Definition 11.1. Seien $X_1, \ldots, X_n$ unabhängige Zufallsvariablen. Die Verteilung von $X_1 + \cdots + X_n$ heißt *Faltung*.

Satz 11.2. *Sind X, Y unabhängig mit Werten in $\mathbb{Z}$, so gilt*

$$P(X+Y=i) = \sum_{k\in\mathbb{Z}} P(X=i-k)P(Y=k).$$

Beweis. Schreibe das Ereignis $(X+Y=i)$ als disjunkte Vereinigung

$$(X+Y=i) = \sum_{k\in\mathbb{Z}} (X=i-k, Y=k).$$

□

Beispiel 11.3. Seien X und Y unabhängige und auf $\{1,\ldots,6\}$ Laplace-verteilte Zufallsvariablen. Dann ist die Verteilung von $X+Y$ gegeben durch

$$\begin{aligned} P(X+Y=i) &= \sum_k P(X=i-k)P(Y=k) \\ &= \frac{1}{36}\big|\{k : 1 \le i-k \le 6, 1 \le k \le 6\}\big| \\ &= \frac{1}{36}\big|\{k : \max\{1, i-6\} \le k \le \min\{6, i-1\}\}\big|. \end{aligned}$$

Definition 11.2. Eine Zufallsvariable X heißt *Bernoulli-verteilt* mit dem Parameter p, in Zeichen $X \sim B_{1,p}$, wenn

$$P(X=1) = B_{1,p}\{1\} = p = 1 - P(X=0) = B_{1,p}\{0\}.$$

Beispiel 11.4. Sei $(\Omega, \mathcal{F}, P)$ eine Wahrscheinlichkeitsraum und $A \in \mathcal{F}$. Dann ist die Indikatorfunktion $X = 1_A$ Bernoulli-verteilt mit dem Parameter $p = P(X=1) = P(A)$.

Definition 11.3. Eine Zufallsvariable X heißt *binomialverteilt* mit den Parametern n und p, in Zeichen $X \sim B_{n,p}$, wenn

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

Satz 11.3. *Sind $X_1, \ldots, X_n$ unabhängig und $B_{1,p}$-verteilt, dann ist $X_1 + \cdots + X_n$ verteilt nach $B_{n,p}$.*

Beweis. Seien $\delta_i \in \{0,1\}$ für $i = 1, \ldots, n$. Dann gilt

$$P(X_1 = \delta_1, \ldots, X_n = \delta_n) = \prod_{i=1}^n P(X_i = \delta_i) = p^{\sum_{i=1}^n \delta_i} (1-p)^{n - \sum_{i=1}^n \delta_i}.$$

Für $\sum_{i=1}^n \delta_i = k$ gibt es $\binom{n}{k}$ Möglichkeiten. Also folgt

$$P\Big(\sum_{i=1}^n X_i = k\Big) = \binom{n}{k} p^k (1-p)^{n-k} = B_{n,p}\{k\}.$$

□

Beispiel 11.5. Sie spielen n-mal ein Spiel, das Sie jeweils mit Wahrscheinlichkeit p gewinnen. Wie wahrscheinlich ist es dann, daß Sie genau einmal gewinnen?
Hier wurde stillschweigend angenommen, daß die Ergebnisse unabhängig sind. Wir beschreiben sie mit unabhängigen und $B_{1,p}$-verteilten Zufallsvariablen $X_1, \ldots, X_n$, bei denen der Wert 1 für ein gewonnenes Spiel steht. Die gesuchte Wahrscheinlichkeit ist also

$$P(X_1 + \cdots + X_n = 1) = \binom{n}{1} p^1 (1-p)^{n-1} = np(1-p)^{n-1}.$$

Für kleines p ist das ungefähr np.

Definition 11.4. Eine Zufallsvariable X heißt *geometrisch verteilt* mit dem Parameter p, in Zeichen $X \sim G_p$, wenn

$$P(X = k) = G_p\{k\} = p(1-p)^{k-1}, \quad k = 1, 2, \ldots.$$

Definition 11.5. Sind $X_1, X_2, \ldots$ Bernoulli-verteilt, so heißt die Zufallsvariable $W_1 = \inf\{k : X_k = 1\}$ die *Wartezeit bis zum ersten Erfolg*. Induktiv definiert man die *Wartezeit bis zum n-ten Erfolg* durch $W_n = \inf\{k > W_{n-1} : X_k = 1\}$.

Satz 11.4. *Sind $X_1, X_2, \ldots$ unabhängig und $B_{1,p}$-verteilt, dann ist $W_1 \sim G_p$.*

Beweis. Es gilt $W_1 = k$ genau dann, wenn die ersten $k-1$ Spiele verloren und das k-te gewonnen wurde. Daher gilt

$$P(W_1 = k) = P(X_1 = \cdots = X_{k-1} = 0, X_k = 1) = (1-p)^{k-1} p.$$

□

Definition 11.6. Eine Zufallsvariable X heißt *negativ binomialverteilt*, in Zeichen $X \sim B^-_{n,p}$, wenn

$$P(X = k) = B^-_{n,p}\{k\} = \binom{k-1}{n-1} p^n (1-p)^{k-n}, \quad k = n, n+1, \ldots.$$

Satz 11.5. *Sind $X_1, X_2, \ldots$ unabhängig $B_{1,p}$-verteilt, dann ist die Wartezeit W_n bis zum n-ten Erfolg $B^-_{n,p}$-verteilt.*

Beweis. Setze $Y_1 = W_1$ und $Y_i = W_i - W_{i-1}$ für $i \geq 2$. Die Y_i bezeichnen die Wartezeiten vom $(i-1)$-ten bis zum i-ten Erfolg. Es gilt $W_n = \sum_{i=1}^n Y_i$. Die Zufallsvariablen Y_i sind unabhängig und G_p-verteilt. Also ist die Verteilung von W_n die n-fache Faltung von G_p-verteilten Zufallsvariablen. Wir haben deshalb

$$P(Y_1 = k_1, \ldots, Y_n = k_n) = \prod_{i=1}^n P(Y_i = k_i) = p^n (1-p)^{\sum_{i=1}^n k_i - n}.$$

Es gibt $\binom{k-1}{n-1}$ Möglichkeiten für $\sum_{i=1}^n k_i = k$ mit $k_i \geq 1$. Also folgt

$$P(W_n = k) = P(Y_1 + \cdots + Y_n = k) = \binom{k-1}{n-1} p^n (1-p)^{k-n}.$$

□

Es gilt $B^-_{1,p} = G_p$.

Definition 11.7. Eine Zufallsvariable X heißt *Poisson-verteilt* mit dem Parameter λ, in Zeichen $X \sim P_\lambda$, wenn

$$P(X = k) = P_\lambda\{k\} = \frac{\lambda^k}{k!}\, e^{-\lambda}, \quad k = 0, 1, \dots.$$

Satz 11.6. *Für $k = 0, 1, \dots$ gilt $B_{n,\lambda/n}\{k\} \to P_\lambda\{k\}$, wenn $n \to \infty$.*

Beweis. Schreibe

$$B_{n,\lambda/n}\{k\} = \binom{n}{k}\Big(\frac{\lambda}{n}\Big)^k\Big(1 - \frac{\lambda}{n}\Big)^{n-k} = \frac{n!}{n^k(n-k)!}\frac{\lambda^k}{k!}\Big(1 - \frac{\lambda}{n}\Big)^n\Big(1 - \frac{\lambda}{n}\Big)^{-k}.$$

Der erste und letzte Faktor konvergieren gegen 1, und $(1 - \lambda/n)^n$ konvergiert gegen $e^{-\lambda}$. □

Beispiel 11.6. Ein Teig enthalte so viele Rosinen, daß im Schnitt 10 Rosinen auf ein Brötchen kommen. Wie wahrscheinlich ist es, daß ein bestimmtes Brötchen keine Rosine enthält?
Man braucht n Rosinen für $n/10$ Brötchen. Jede Rosine gerät mit Wahrscheinlichlichkeit $10/n$ und unabhängig von den anderen in das gegebene Brötchen. Die Anzahl der Rosinen in diesem Brötchen folgt also der Faltung von n jeweils $B_{1,10/n}$-verteilten Zufallsvariablen, also einer $B_{n,10/n}$-Verteilung. Die Wahrscheinlichkeit für keine Rosine ist also $B_{n,10/n}\{0\}$. Nach dem vorigen Satz ist das für große n ungefähr $P_{10}\{0\} = e^{-10}$.

Beispiel 11.7. Wenn eine Buchseite im Schnitt vier Druckfehler enthält, so ist die Wahrscheinlichkeit für eine Seite mit mehr als zwei Druckfehlern wie im vorigen Beispiel ungefähr

$$P_4\{3, 4, \dots\} = 1 - P_4\{0, 1, 2\} = 1 - e^{-4}(1 + 4 + 8) = 1 - 13e^{-4}.$$

12 Stetige Wahrscheinlichkeitsmaße

Neben den diskreten Wahrscheinlichkeitsmaßen sind die im folgenden definierten "stetigen" Wahrscheinlichkeitsmaße wichtig. Auch sie sind werden wieder automatisch σ-additiv sein.

Proposition 12.1. *Sei $p : \mathbb{R}^m \to [0, \infty)$ meßbar mit $\int p\, d\lambda^m = 1$. Dann definiert*

$$P(A) = \int_A p\, d\lambda^m$$

ein Wahrscheinlichkeitsmaß auf $\mathcal{B}^m$. Es heißt stetig*; die Funktion p heißt* Dichte.

Beweis. Die Eigenschaften $P(A) \geq 0$ und $P(\mathbb{R}^m) = 1$ sind offensichtlich. Seien nun A_i für $i \in \mathbb{N}$ paarweise disjunkte Mengen in $\mathcal{B}^m$. Dann gilt mit dem Satz über die

monotone Konvergenz

$$\begin{aligned} P\Big(\sum_{i=1}^{\infty} A_i\Big) &= \int_{\sum_{i\in\mathbb{N}} A_i} p\, d\lambda^m = \int 1_{\sum_{i\in\mathbb{N}} A_i}\, p\, d\lambda^m \\ &= \int \sum_{i=1}^{\infty} 1_{A_i}\, p\, d\lambda^m = \int \lim_{n\to\infty} \sum_{i=1}^{n} 1_{A_i}\, p\, d\lambda^m \\ &= \lim_{n\to\infty} \int \sum_{i=1}^{n} 1_{A_i}\, p\, d\lambda^m = \lim_{n\to\infty} \sum_{i=1}^{n} \int_{A_i} p\, d\lambda^m \\ &= \sum_{i=1}^{\infty} \int_{A_i} p\, d\lambda^m = \sum_{i=1}^{\infty} P(A_i). \end{aligned}$$

□

Eine Zufallsvektor $X : \Omega \to \mathbb{R}^m$ heißt *stetig*, wenn das induzierte Wahrscheinlichkeitsmaß $P^X|\mathcal{B}^m$ stetig ist.

Beispiel 12.1. Sei $A \in \mathcal{B}^m$. Die *Gleichverteilung* auf A besitzt die Dichte

$$p(x) = \frac{1}{\int_A d\lambda^m} 1_A(x) = \frac{1}{\lambda^m(A)} 1_A(x).$$

Bemerkung 12.1. Ist $P|\mathcal{B}^{m_1} \otimes \mathcal{B}^{m_2}$ stetig mit Dichte p, so ist P ein unabhängiges Produkt auf $\mathbb{R}^{m_1} \times \mathbb{R}^{m_2}$ (genau dann), wenn $p(x, y) = p_1(x)p_2(y)$. Dann ist $P = P_1 \otimes P_2$, wobei P_i die Dichte p_i hat.

Beweis. Gilt $p(x, y) = p_1(x)p_2(y)$, so folgt

$$\begin{aligned} P(A \times B) &= \int_{A\times B} p(x, y)\lambda^{m_1+m_2}(dx, dy) \\ &= \int_A p_1(x)\lambda^{m_1}(dx) \int_B p_2(y)\lambda^{m_2}(dy) = P_1(A)P_2(B). \end{aligned}$$

□

Definition 12.1. Die *Exponentialverteilung* E_a mit Parameter $a > 0$ hat die Dichte

$$p(t) = \begin{cases} \frac{1}{a}\, e^{-t/a}, & t > 0 \\ 0, & \text{sonst.} \end{cases}$$

13 Verteilungsfunktionen und Transformationssätze

Ein Wahrscheinlichkeitsmaß ist eine *Mengen*funktion; auf der Borel-Algebra $\mathcal{B}$ läßt sie sich einfach durch eine *Punkt*funktion beschreiben.

Definition 13.1. Sei $P|\mathcal{B}$ ein Wahrscheinlichkeitsmaß. Dann heißt die durch

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

definierte Funktion F auf $\mathbb{R}$ die *Verteilungsfunktion* von P.

Satz 13.1. *Für eine Verteilungsfunktion F gilt*

1. *F ist nichtfallend;*
2. *$\lim_{t\to\infty} F(t) = 1$ und $\lim_{t\to-\infty} F(t) = 0$;*
3. *F ist rechtsstetig.*

Der Beweis ist eine Übungsaufgabe. Die folgenden beiden Sätze beweisen wir nicht.

Satz 13.2. *Zu jeder Funktion F mit den Eigenschaften 1.–3. aus obigem Satz gibt es genau ein Wahrscheinlichkeitsmaß mit Verteilungsfunktion F.*

Satz 13.3 (Satz über Stammfunktionen)**.** *Ist eine Verteilungsfunktion F differenzierbar, dann ist F' die zugehörige Dichte.*

Proposition 13.4 (Transformationssatz für Verteilungsfunktionen)**.** *Sei X eine Zufallsvariable mit Verteilungsfunktion F, und sei $T : \mathbb{R} \to \mathbb{R}$ streng monoton wachsend und stetig mit $T(\mathbb{R}) = (a, b)$, wobei $-\infty \le a < b \le \infty$. Dann hat $T \circ X$ die Verteilungsfunktion*

$$F^T(y) = \begin{cases} 0, & \le a, \\ F(T^{-1}(y)), & y \in (a,b), \\ 1, & \ge b. \end{cases}$$

Beweis. Für $y \in (a, b)$ gilt

$$P(T \circ X \le y) = P^X(T \le y) = P^X((-\infty, T^{-1}(y)]).$$

□

Proposition 13.5 (Transformationssatz für Dichten)**.** *Sei X eine Zufallsvariable mit Dichte p, und sei $T : \mathbb{R} \to \mathbb{R}$ differenzierbar mit $T(\mathbb{R}) = (a, b)$ und $T' > 0$ oder $T' < 0$. Dann hat $T \circ X$ die Dichte*

$$p^T(y) = \begin{cases} \dfrac{p(T^{-1}(y))}{|T'(T^{-1}(y))|}, & y \in (a,b), \\ 0, & \text{sonst.} \end{cases}$$

Beweis. Nach dem Satz über Stammfunktionen erhalten wir die Dichte durch Ableiten der Verteilungsfunktion aus dem Transformationssatz für Verteilungsfunktionen. □

Beispiel 13.1. Hat die Zufallsvariable X die Dichte p^X, so hat $aX + b$ die Dichte

$$p^{aX+b}(y) = \frac{1}{|a|} p^X\Big(\frac{y-b}{a}\Big).$$

Beweis. Setze im Transformationssatz für Dichten $T(x) = ax+b$. Dann gilt $T^{-1}(y) = (y-b)/a$ und $T'(x) = a$, also $T'(T^{-1}(y)) = a$. □

Definition 13.2. Die eindimensionale *Normalverteilung* N_{μ,σ^2} mit den Parametern $\mu \in \mathbb{R}$ und $\sigma > 0$ hat die Dichte

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}.$$

Ist X verteilt nach $N_{0,1}$, so ist $\sigma X + \mu$ verteilt nach N_{μ,σ^2}. Das folgt aus dem vorigen Beispiel.

Definition 13.3. Die *Cauchy-Verteilung* $C_a|\mathcal{B}$ mit Parameter $a > 0$ hat die Dichte

$$p(x) = \frac{1}{\pi a} \frac{1}{1 + (x/a)^2}.$$

Ist X verteilt nach C_1, so ist aX verteilt nach C_a.

Bemerkung 13.1. Der Mittelwert der Cauchy-Verteilung existiert nicht. Es genügt, dies für C_1 nachzuweisen. Es gilt

$$\int_0^\infty xp(x)dx = \frac{1}{\pi} \int_0^\infty \frac{x}{1+x^2} dx = \frac{1}{2\pi} \log(1+x^2)\Big|_0^\infty = \infty$$

und analog

$$\int_{-\infty}^0 xp(x)dx = -\infty.$$

Beispiel 13.2. Ist X nichtnegativ mit Dichte p^X, so hat X^a für $a \neq 0$ die Dichte

$$p^{X^a}(y) = \frac{1}{|a|} y^{(1-a)/a} p^X\left(y^{1/a}\right), \quad y > 0.$$

Beweis. Setze im Transformationssatz für Dichten $T(x) = x^a$. Dann gilt $T^{-1}(y) = y^{1/a}$ und $T'(x) = ax^{a-1}$, also $T'(T^{-1}(y)) = ay^{(a-1)/a}$. □

Definition 13.4. Die *Gamma-Verteilung* $\Gamma_{a,b}|\mathcal{B}$ mit Parametern $a, b > 0$ hat die Dichte

$$p(x) = \frac{1}{a^b \Gamma(b)} x^{b-1} e^{-x/a}, \quad x > 0.$$

Die Gamma-Funktion ist dabei definiert durch

$$\Gamma(b) = \int_0^\infty x^{b-1} e^{-x} dx.$$

Sie genügt der Funktionalgleichung $\Gamma(b+1) = b\Gamma(b)$. Wegen $\Gamma(1) = 1$ folgt daraus $\Gamma(n) = (n-1)!$. Weiterhin gilt $\Gamma(1/2) = \sqrt{\pi}$.

Der Parameter a von $\Gamma_{a,b}$ ist ein Skalenparameter. Außerdem gilt $\Gamma_{a,1} = E_a$.

Beispiel 13.3. Ist X verteilt nach N_{0,σ^2}, so ist X^2 verteilt nach $\Gamma_{2\sigma^2,1/2}$.

Beweis. Die Funktion $\Gamma(x) = x^2$ ist zwar nicht injektiv, jedoch ist N_{0,σ^2} symmetrisch um 0. Ist X verteilt nach N_{0,σ^2}, so hat $|X|$ also die Dichte

$$p(x) = \frac{2}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2}, \quad x > 0.$$

Nach dem Transformationssatz für Dichten hat $X^2 = |X|^2$ also mit $\sqrt{\pi} = \Gamma(1/2)$ die Dichte

$$\frac{2}{|2|} y^{(1-2)/2} \frac{1}{\sigma\sqrt{2\pi}} e^{-y/2\sigma^2} = \frac{1}{(2\sigma^2)^{1/2}\Gamma(1/2)} y^{\frac{1}{2}-1} e^{-y/2\sigma^2}.$$

□

In Kapitel 11 hatten wir schon die Faltung zweier diskreter Zufallsvariablen ausgerechnet. Für stetige Zufallsvariablen gilt ein analoges Resultat.

Proposition 13.6 (Faltung von Dichten)**.** *Hat (X,Y) die Dichte p, dann hat $X+Y$ die Dichte*

$$p^{X+Y}(x) = \int p(x-y,y)dy.$$

Sind insbesondere X und Y unabhängig mit Dichten p_1 und p_2, so gilt $p(x,y) = p_1(x)p_2(y)$, also hat die Faltung $X+Y$ die Dichte

$$p^{X+Y}(x) = \int p_1(x-y)p_2(y)dy.$$

Beweis. Für die Verteilungsfunktion von $X+Y$ gilt

$$\begin{aligned} P(X+Y\le r) &= \int_{X+Y\le r} p(x,y)d(x,y) = \int \left(\int_{X\le r-y} p(x,y)dx\right)dy \\ &= \int \left(\int_{-\infty}^{r} p(x-y,y)dx\right)dy = \int_{-\infty}^{r}\left(\int p(x-y,y)dy\right)dx. \end{aligned}$$

Die Dichte ergibt sich aus dem Satz über Stammfunktionen. □

Bemerkung 13.2. Sind X und Y unabhängig und $\Gamma_{a,b}$- bzw. $\Gamma_{a,c}$-verteilt, so ist $X+Y$ $\Gamma_{a,b+c}$-verteilt.

Beweis. Wir bezeichnen die Dichte von $\Gamma_{a,b}$ mit $p_{a,b}$. Die Dichte von $X+Y$ ergibt sich aus der Proposition über die Faltung von Dichten und mit der Variablentransformation $y \mapsto xy$ als

$$\begin{aligned} p^{X+Y}(x) &= \int p_{a,b}(x-y)p_{a,c}(y)dy \\ &= \frac{1}{a^{b+c}\Gamma(b)\Gamma(c)}\int (x-y)^{b-1}e^{-(x-y)/a}1_{(0,\infty)}(x-y)y^{c-1}e^{-y/a}1_{(0,\infty)}(y)dy \\ &= \frac{1}{a^{b+c}\Gamma(b)\Gamma(c)}\int_0^x (x-y)^{b-1}e^{-(x-y)/a}y^{c-1}e^{-y/a}dy \\ &= \frac{1}{a^{b+c}\Gamma(b)\Gamma(c)}x^{b-1+c}e^{-x/a}\int_0^1 (1-y)^{b-1}y^{c-1}dy. \end{aligned}$$

Das verbleibende Integral wird auch als Beta-Funktion von (b,c) bezeichnet. Es hat den Wert $\Gamma(b)\Gamma(c)/\Gamma(b+c)$. □

Beispiel 13.4 (Maxwellsche Geschwindigkeitsverteilung). Der Geschwindigkeitsvektor eines Moleküls werde mit $X=(X_1,X_2,X_3)$ bezeichnet. Die drei Komponenten seien unabhängig und N_{0,σ^2}-verteilt. Gesucht ist die Verteilung der *Geschwindigkeit*, also der Länge $\|X\| = (X_1^2+X_2^2+X_3^2)^{1/2}$ des Geschwindigkeitsvektors. Wir haben oben gesehen, daß X_i^2 nach $\Gamma_{2\sigma^2,1/2}$ verteilt ist. Nach dem vorigen Beispiel ist $X_1^2+X_2^2+X_3^2$ verteilt nach $\Gamma_{2\sigma^2,3/2}$ mit der Dichte

$$p^{X_1^2+X_2^2+X_3^2}(x) = \frac{1}{(2\sigma^2)^{3/2}\Gamma(3/2)}x^{\frac{3}{2}-1}e^{-x/2\sigma^2} = \frac{1}{\sigma^3\sqrt{2\pi}}x^{1/2}e^{-x/2\sigma^2}, \quad x>0.$$

Durch Transformation mit $T(x)=x^{1/2}$ erhalten wir daraus die Dichte der Geschwindigkeit $(X_1^2+X_2^2+X_3^2)^{1/2}$. Sie ist

$$p^{\|X\|}(y) = \frac{\sqrt{2}}{\sigma^3\sqrt{\pi}}y^2e^{-y^2/2\sigma^2}, \quad y>0.$$

14 Gesetz der großen Zahl

Sei $(\Omega, \mathcal{F}, P)$ ein Wahrscheinlichkeitsraum. Für das Folgende bemerken wir, daß der Erwartungswert ein Integral, also *linear* im Integranden ist, d.h. es gilt $E(cX) = cEX$ und $E(X+Y) = EX + EY$.

Definition 14.1. Sei X eine Zufallsvariable mit endlichem Erwartungswert $\mu(X) = EX$. Die *Varianz* von X ist

$$\mathrm{Var}(X) = \sigma^2(X) = E[(X - EX)^2] = E[X^2] - 2\mu(X)^2 + \mu(X)^2 = E[X^2] - (EX)^2.$$

Es gilt $\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$, also $\sigma(cX) = |c|\sigma(X)$. Es gilt $\mathrm{Var}(X+c) = \mathrm{Var}(X)$, also auch $\sigma(X+c) = \sigma(X)$. Diese beiden Eigenschaften machen die *Standardabweichung* $\sigma(X)$ zu einem vernünftigen Maß für die *Streuung* von X.

Definition 14.2. Seien X und Y Zufallsvariablen mit endlicher Varianz. Dann heißt $\mathrm{Cov}(X,Y) = E[(X - EX)(Y - EY)]$ die *Kovarianz* von X und Y.

Bemerkung 14.1. Sind $X_1, \ldots, X_n$ Zufallsvariablen mit endlicher Varianz, so gilt

$$\mathrm{Var}\Big(\sum_{i=1}^n X_i\Big) = \sum_{i=1}^n \mathrm{Var}(X_i) + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j).$$

Beweis. Wir dürfen annehmen, daß $EX_i = 0$ gilt. (Sonst betrachten wir $Y_i = X_i - EX_i$.) Es gilt

$$E\Big[\Big(\sum_{i=1}^n X_i\Big)^2\Big] = E\Big[\sum_{i=1}^n X_i^2 + \sum_{i\neq j} X_i X_j\Big] = \sum_{i=1}^n E[X_i^2] + \sum_{i\neq j} E[X_i X_j].$$

□

Satz 14.1. *Sind X und Y unabhängig mit endlichen Varianzen, so gilt*

$$E[XY] = EXEY.$$

Beweis. Algebraische Induktion: Indikatorfunktionen 1_A und 1_B sind unabhängig genau dann, wenn A und B unabhängig sind. Dann gilt

$$E[1_A 1_B] = E1_{A\cap B} = P(A\cap B) = P(A)P(B) = E1_A E1_B.$$

Ein anderer Beweis benutzt den Satz von Fubini:

$$\begin{aligned} E[XY] &= \int xyP^{(X,Y)}(dx,dy) = \iint xyP^X(dx)P^Y(dy) \\ &= \int xP^X(dx) \int yP^Y(dy) = EXEY. \end{aligned}$$

□

Folgerung 14.2. *Sind $X_1, \ldots, X_n$ unabhängig mit endlichen Varianzen, dann gilt*

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

Beweis. Mit Satz 14.1 gilt

$$\mathrm{Cov}(X_i, X_j) = E[(X_i - EX_i)(X_j - EX_j)] = E[X_i - EX_i]E[X_j - EX_j] = 0.$$

□

Satz 14.3 (Chebyshev-Ungleichung)**.** *Ist X eine Zufallsvariable mit endlicher Varianz, dann gilt für alle $c > 0$,*

$$P(|X - EX| \geq c) \leq \frac{\mathrm{Var}(X)}{c^2}.$$

Beweis. Es gilt $P(|X - EX| \geq c) = E1_{\{|X-EX|\geq c\}}$ und

$$1_{\{|X-EX|\geq c\}} \leq \Big(\frac{X - EX}{c}\Big)^2.$$

Der Erwartungswert ist monoton. □

Satz 14.4 (Gesetz der großen Zahl)**.** *Sind $X_1, \ldots, X_n$ unabhängig und identisch verteilt mit endlicher Varianz, so gilt für alle $\varepsilon > 0$,*

$$P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} X_i - EX_1\Big| \geq \varepsilon\Big) \leq \frac{1}{n\varepsilon^2}\mathrm{Var}(X_1).$$

Beweis. Wir wenden die Chebyshev-Ungleichung für $c = \varepsilon$ und $X = (1/n)\sum_{i=1}^{n} X_i$ an. Dann gilt $EX = EX_1$ und

$$\mathrm{Var}(X) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n}\mathrm{Var}(X_1).$$

□

15 Zentraler Grenzwertsatz

Nach dem Gesetz der großen Zahl gilt für unabhängige und identisch verteilte Zufallsvariablen $X_1, X_2, \ldots$ mit endlicher Varianz in einem gewissen Sinne

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to EX_1.$$

Die Konvergenzgeschwindigkeit ist $n^{-1/2}$, denn die standardisierte Summe

$$n^{1/2}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i - EX_1\Big) = n^{-1/2}\sum_{i=1}^{n}(X_i - EX_i)$$

hat Erwartungswert 0 und eine endliche Varianz $\mathrm{Var}(X_1)$. Diese Varianz ist auch positiv, es sei denn, X_1 ist eine Konstante. Es gilt sogar, daß die standardisierte Summe für große n ungefähr normalverteilt ist, ebenfalls mit Mittelwert 0 und Varianz $\mathrm{Var}(X_1)$.

Satz 15.1 (Zentraler Grenzwertsatz). *Sind $X_1, X_2, \ldots$ unabhängige und identisch verteilte Zufallsvariablen mit endlicher Varianz $\mathrm{Var}(X_1) = \sigma^2$ und Mittelwert $EX_i = \mu$, so gilt für beliebige Zahlen a und b mit $a < b$, daß*

$$P\Big(a < n^{-1/2}\sum_{i=1}^{n}(X_i - \mu) < b\Big) \to N_{0,\sigma^2}(a,b) = \frac{1}{\sigma\sqrt{2\pi}}\int_a^b e^{-\frac{x^2}{2\sigma^2}}dx.$$

Beispiel 15.1 (Symmetrische Irrfahrt). Seien $X_1, X_2, \ldots$ unabhängige Zufallsvariablen mit $P(X_i = 1) = P(X_i = -1) = \frac{1}{2}$. Dann gilt $\mu = EX_1 = 0$ und $\sigma^2 = \mathrm{Var}(X_1) = 1$. Aus dem Zentralen Grenzwertsatz folgt also für die Irrfahrt $\sum_{i=1}^n X_i$, $n = 1, 2, \ldots$, daß

$$P\Big(-n^{1/2} < \sum_{i=1}^{n} X_i < n^{1/2}\Big) = P\Big(-1 < n^{-1/2}\sum_{i=1}^{n} X_i < 1\Big) \to N_{0,1}(-1,1) \approx 0,7.$$

Beispiel 15.2 (Meßfehler). Mit einem Instrument machen wir Messungen $X_1, \ldots, X_n$ einer physikalischen Konstante μ. Die Meßfehler $X_i - \mu$ seien unabhängig und identisch verteilt mit Mittelwert 0 und Varianz σ^2. Wir schätzen μ mit dem arithmetischen Mittel der Messungen,

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

Nachdem Gesetz der großen Zahl konvergiert der Schätzer $\bar{X}_n$ gegen μ. In welchem Bereich liegt der Schätzfehler $\bar{X}_n - \mu$ ungefähr mit Wahrscheinlichkeit 0.95?
Nach dem Zentralen Grenzwertsatz ist

$$n^{1/2}\frac{\bar{X}_n - \mu}{\sigma} = n^{-1/2}\sum_{i=1}^{n}\frac{X_i - \mu}{\sigma}$$

annähernd $N_{0,1}$-verteilt. Aus Tabellen entnehmen wir $N_{0,1}(-1.96, 1.96) \approx 0.95$. Also gilt

$$\begin{aligned} P\big(|\bar{X}_n - \mu| < 1,96n^{-1/2}\sigma\big) &= P\Big(-1.96 < n^{1/2}\frac{\bar{X}_n - \mu}{\sigma} < 1.96\Big) \\ &\to N_{0,1}(-1.96, 1.96) \approx 0.95. \end{aligned}$$

Der Schätzfehler von $\bar{X}_n$ liegt also etwa mit Wahrscheinlichkeit 0.95 im Bereich $(-1.96n^{-1/2}\sigma, 1.96n^{-1/2}\sigma)$. Für $\sigma = 1$ und $n = 100$ ergibt sich z.B. $1.96n^{-1/2}\sigma = 0.196$.

Beispiel 15.3 (Rundungsfehler). Die Zahlen $R_1, \ldots, R_n$ werden auf ganze Zahlen gerundet, also $R_i = Z_i + S_i$ mit ganzzahligem Anteil $Z_i \in \mathbb{Z}$ und Rundungsfehler $S_i \in [-1/2, 1, 2)$. Beim Summieren $\sum_{i=1}^n R_i$ entsteht dann der Rundungsfehler $\sum_{i=1}^n S_i$. Wir können annehmen, daß die Rundungsfehler S_i unabhängig und gleichverteilt auf dem Intervall $[-1/2, 1/2)$ sind. Dann gilt $ES_1 = 0$ und

$$\mathrm{Var}(S_1) = E[S_1^2] = \int_{-1/2}^{1/2} x^2 dx = \frac{1}{12}.$$

Der Zentrale Grenzwertsatz liefert

$$P\Big(\Big|\sum_{i=1}^{n} S_i\Big| < n^{1/2}t\Big) = P\Big(\Big|n^{-1/2}\sum_{i=1}^{n} S_i\Big| < t\Big) \to N_{0,1/12}(-t,t).$$

Für $n = 100$ wird der Rundungsfehler mit Wahrscheinlichkeit 0,95 kleiner als 5,7 sein. Dies zeigt man wie eben mit $\sigma^2 = 1/12$.

Beispiel 15.4 (Normalapproximation der Binomialverteilung). Ist $X \sim B_{n,p}$, dann ist $n^{-1/2}(X - np)$ ungefähr $N_{0,p(1-p)}$-verteilt.

Beweis. Nach Satz 11.3 ist X verteilt wie $X_1 + \cdots + X_n$, wenn die X_i unabhängig und $B_{1,p}$-verteilt sind. Es gilt $EX_1 = 0 \cdot (1-p) + 1 \cdot p = p$ und

$$\mathrm{Var}(X_1) = E[(X_1 - EX_1)^2] = (1-p)(0-p)^2 + p(1-p)^2 = p(1-p).$$

Also ist $n^{-1/2}(X - np)$ verteilt wie $n^{-1/2} \sum_{i=1}^n (X_i - p)$. Wende darauf den Zentralen Grenzwertsatz an. □

16 Schwache Konvergenz und Konvergenz in Wahrscheinlichkeit

Das Gesetz der großen Zahl verwendet einen Konvergenzbegriff (für $n \to \infty$), den wir allgemein wie folgt definieren. Seien $X, X_1, X_2, \ldots$ Zufallsvariablen. Die zugehörigen Verteilungsfunktionen bezeichnen wir mit $F, F_1, F_2, \ldots$.

Definition 16.1. Die Folge X_n konvergiert gegen X *in Wahrscheinlichkeit* (oder *stochastisch*; in Zeichen $X_n = X + o_p(1)$), wenn

$$P(|X_n - X| > \varepsilon) \to 0 \quad \text{für alle } \varepsilon > 0.$$

Der zentrale Grenzwertsatz verwendet folgenden Konvergenzbegriff.

Definition 16.2. Die Folge X_n konvergiert *schwach* gegen X (oder *in Verteilung*; in Zeichen $X_n \Rightarrow X$), wenn

$$F_n(t) \to F(t) \quad \text{für alle } t, \text{ in denen } F \text{ stetig ist.}$$

Definition 16.3. Die Folge X_n ist *beschränkt in Wahrscheinlichkeit* (in Zeichen $X_n = O_p(1)$), wenn für alle $\varepsilon > 0$ ein $c > 0$ existiert, so daß für alle (hinreichend großen) n gilt:

$$P(|X_n| > c) \le \varepsilon.$$

Bemerkung 16.1. Wenn $X_n \to X$ in Wahrscheinlichkeit, dann gilt auch $X_n = O_p(1)$.

Beweis. Es gilt $\{|X| > c\} \downarrow \emptyset$ für $c \uparrow \infty$. Für $\varepsilon > 0$ gibt es also ein $c > 0$, so daß $P(|X| > c) \le \varepsilon$. Andererseits gilt $P(|X_n - X| > \varepsilon) \le \varepsilon$ für alle hinreichend großen n. Wir dürfen $\varepsilon \le c$ annehmen. Dann erhalten wir

$$P(|X_n| > 2c) \le P(|X| > c) + P(|X_n - X| > c) \le 2\varepsilon.$$

□

Beispiel 16.1 (Summen unabhängiger Zufallsvariablen). Seien $X_1, X_2, \ldots$ unabhängig und identisch verteilt mit endlicher Varianz σ^2 und Erwartungswert μ. Dann gilt

$$\frac{1}{n} \sum_{i=1}^n X_i \to \mu \quad \text{in Wahrscheinlichkeit,}$$

$$n^{-1/2} \sum_{i=1}^n (X_i - \mu) \Rightarrow X \sim N_{0,\sigma^2},$$

$$n^{-1/2} \sum_{i=1}^n (X_i - \mu) = O_p(1).$$

Bemerkung 16.2. Wenn $X_n \to X$ in Wahrscheinlichkeit und $Y_n \to Y$ in Wahrscheinlichkeit, dann gilt auch $X_n + Y_n \to X + Y$ in Wahrscheinlichkeit und $X_n Y_n \to XY$ in Wahrscheinlichkeit.

Beweis. Für die Summe schreiben wir

$$P(|X_n + Y_n - (X+Y)| > \varepsilon) \le P(|X_n - X| > \varepsilon/2) + P(|Y_n - Y| > \varepsilon/2).$$

Für das Produkt schreiben wir $X_n Y_n - XY = X_n(Y_n - Y) + (X_n - X)Y$. Für $\varepsilon > 0$ gibt es ein $c > 0$, so daß $P(|X_n| > c) \le \varepsilon$ und $P(|Y| > c) \le \varepsilon$. Es ergibt sich

$$P(|X_n(Y_n - Y)| > \varepsilon) \le P(|X_n| > c) + P(|Y_n - Y| > \varepsilon/c) \le \varepsilon + o(1).$$

Ähnlich verfahre man für $(X_n - X)Y$. □

Bemerkung 16.3. Wenn $X_n \Rightarrow X$, dann gilt $X_n = O_p(1)$.

Beweis. Die Verteilungsfunktion von X sei stetig in c und $-c$. Dann gilt

$$P(|X_n| > c) = P(X_n > c) + P(X_n \le -c) \to P(X > c) + P(X \le -c).$$

□

Lemma 16.1 (Slutsky)**.** *Gilt $X_n \Rightarrow X$ und $Y_n \to 0$ in Wahrscheinlichkeit, so gilt $X_n + Y_n \Rightarrow X$.*

Beweis. Die Verteilungsfunktion von X sei stetig in t. Für jedes $\delta > 0$ gilt

$$\begin{aligned} P(X_n + Y_n \le t) &\le P(X_n + Y_n \le t, |Y_n| \le \delta) + P(|Y_n| > \delta) \\ &\le P(X_n \le t + \delta) + o(1) \to F(t + \delta). \end{aligned}$$

Analog gilt

$$P(X_n + Y_n \le t) \ge P(X_n \le t - \delta) + o(1) \to F(t - \delta).$$

Jetzt δ klein wählen. □

Die folgende (wichtige) Proposition ist eine Übungsaufgabe.

Proposition 16.2. *Wenn $X_n \to c$ in Wahrscheinlichkeit und f stetig in c ist, so gilt $f(X_n) \to f(c)$ in Wahrscheinlichkeit.*

Ebenso wichtig ist die folgende Proposition.

Proposition 16.3. *Wenn $a_n(X_n - c) \Rightarrow Z$ für $a_n \to \infty$ und f stetig differenzierbar in c ist, so gilt*

$$a_n\big(f(X_n) - f(c)\big) \Rightarrow f'(c)Z.$$

Beweis. Mit einer Taylor-Entwicklung ergibt sich

$$f(X_n) = f(c) + (X_n - c)f'(c) + (X_n - c)\int_0^1 \big(f'(X_n + t(X_n - c)) - f'(c)\big)dt.$$

Nach dem Satz von der dominierten Konvergenz gilt

$$\int_0^1 \big(f'(c + t(x - c)) - f'(c)\big)dt \to 0 \quad \text{für } x \to c.$$

Aus $a_n(X_n - c) \Rightarrow Z$ folgt $a_n(X_n - c) = O_p(1)$ und $X_n = c + o_p(1)$, also

$$\int_0^1 \big(f'(X_n + t(X_n - c)) - f'(c)\big)dt = o_p(1).$$

□

17 Empirische Schätzer

Jetzt wechseln wir von der Wahrscheinlichkeitstheorie zur Statistik. Von jetzt an schreiben wir den zugrundeliegenden meßbaren Raum immer in der "kanonischen" Darstellung: Wir setzen $\Omega = \mathbb{R}^{\mathbb{N}}$ und $\mathcal{F} = \mathcal{B}^{\mathbb{N}}$. Die Zufallsvariable X_i bezeichnet jetzt immer die Beobachtung des i-ten Experiments: Für $x = (x_1, x_2, \dots) \in \mathbb{R}^{\mathbb{N}}$ setzen wir $X_i(x) = x_i$, die i-te Koordinate von x. (Diese Notation ist eigentlich redundant, vereinfacht aber häufig die Schreibweise.) Außerdem werden wir immer annehmen, daß die Experimente unabhängig voneinander und identisch sind. Das bedeutet, daß die Verteilung von $X_1, \dots, X_n$ ein unabhängiges Produkt der Form $P^n = P \otimes \cdots \otimes P$ mit n identischen Faktoren ist.

Das Wahrscheinlichkeitsmaß P kennen wir nicht oder zumindest nicht vollständig. Wir nehmen also an, daß es in einer Familie $\mathcal{P}$ von Wahrscheinlichkeitsmaßen liegt, dem *Modell*. Wir beobachten $X_1, \dots, X_n$ und wollen den Wert $t(P)$ eines reellwertigen Funktionals $t : \mathcal{P} \to \mathbb{R}$ schätzen.

Ein *Schätzer* T_n ist eine messbare Funktion von $\mathbb{R}^n$ nach $\mathbb{R}$.

Definition 17.1. Ein Schätzer T_n heißt *erwartungstreu* für t, wenn $E_P T_n = t(P)$ für alle $P \in \mathcal{P}$. (Für den Erwartungswert schreiben wir jetzt manchmal E_P statt E, um die Abhängigkeit von P auszudrücken.)

Ein Schätzer T_n heißt *konsistent* für t, wenn $T_n \to t(P)$ in W. für alle $P \in \mathcal{P}$. (Solche *asymptotischen* Eigenschaften beziehen sich natürlich auf eine *Folge* von Schätzern.)

Ein Schätzer T_n heißt *asymptotisch normal* unter P mit Varianz $\sigma^2(P)$, wenn

$$n^{1/2}(T_n - t(P)) \Rightarrow X \sim N_{0,\sigma^2(P)}.$$

Beispiel 17.1. Sei $t(P) = \mu(P) = E_P X$. Der *empirische Schätzer* dafür ist $T_n = \frac{1}{n}\sum_{i=1}^n X_i$. Gilt $E_P[X^2] < \infty$ für alle $p \in \mathcal{P}$, so ist T_n erwartungstreu, konsistent und asymptotisch normal.

Beweis. Die Erwartungstreue folgt aus der Linearität des Erwartungswerts,

$$E_P T_n = \frac{1}{n}\sum_{i=1}^n E_P X_i = E_P X = t(P).$$

Die Konsistenz folgt aus dem Gesetz der großen Zahl,

$$\frac{1}{n}\sum_{i=1}^n X_i \to E_P X = t(P) \quad \text{in Wahrscheinlichkeit.}$$

Die asymptotische Normalität folgt aus dem Zentralen Grenzwertsatz,

$$n^{1/2}(T_n - t(P)) = n^{-1/2}\sum_{i=1}^n (X_i - \mu(P)) \Rightarrow X \sim N_{0,\sigma^2(P)}.$$

□

Entsprechendes gilt allgemein für das Schätzen von Erwartungswerten $Eh(X)$. Der *empirische Schätzer* dafür ist

$$\mathbb{E}_n h = \frac{1}{n}\sum_{i=1}^{n} h(X_i).$$

Wir setzen $L_2(P) = \{h : E_P[h^2(X)] < \infty\}$ und

$$L_{2,0}(P) = \{h \in L_2(P) : E_P[h(X)] = 0\}.$$

Bemerkung 17.1. Gilt $h \in L_2(P)$ für $P \in \mathcal{P}$, so ist der empirische Schätzer $\mathbb{E}_n h$ für $E_P[h(X)]$ erwartungstreu, konsistent, und asymptotisch normal mit Varianz $\sigma_P^2(h) = E_P[h^2(X)] - (E_P[h(X)])^2$. Das zeigt man genau wie eben für $(1/n)\sum_{i=1}^n X_i$.

Definition 17.2. Ein Schätzer T_n heißt *asymptotisch linear* mit *Einflußfunktion* $a \in L_{2,0}(P)$ für $t(P)$, wenn

$$n^{1/2}(T_n - t(P)) = n^{-1/2}\sum_{i=1}^{n} a(X_i) + o_p(1).$$

Ein solcher Schätzer ist konsistent und asymptotisch normal mit Varianz $E[a^2(X)]$. Der empirische Schätzer $\mathbb{E}_n h$ ist nicht nur asymptotisch linear, sondern sogar (exakt) linear.

Beispiel 17.2. Ist f stetig in $E_p h$, so ist $f(\mathbb{E}_n h)$ konsistent für $f(E_p h)$ (aber nicht (exakt) erwartungstreu). Ist f stetig differenzierbar in $E_p h$, so ist $f(\mathbb{E}_n h)$ asymptotisch linear mit Einflußfunktion $f'(\mu)(X - \mu)$, also insbesondere asymptotisch normal mit Varianz $(f'(\mu))^2\sigma^2$. Diese Aussagen ergeben sich direkt aus dem vorigen Kapitel.

Beispiel 17.3. Seien $X_1, X_2, \ldots$ unabhängig mit Verteilung P und endlicher Varianz σ_P^2. Der Mittelwert werde μ_P genannt. Wir wissen schon, daß der empirische Schätzer $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$ erwartungstreu, konsistent und (exakt) linear für μ_P ist. Seine Genauigkeit hängt von $\sigma_P^2 = E[(X - \mu_P)^2]$ ab. Ein plausibler Schätzer dafür ist

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

Es gilt

$$\begin{aligned} n^{1/2}(\hat{\sigma}^2 - \sigma_P^2) &= n^{1/2}\frac{1}{n}\sum_{i=1}^{n}\left((X_i - \bar{X}_n)^2 - \sigma_P^2\right) \\ &= n^{-1/2}\sum_{i=1}^{n}\left((X_i - \mu_P)^2 - \sigma_P^2\right) - n^{1/2}(\bar{X}_n - \mu_P)^2 \end{aligned}$$

und $n^{1/2}(\bar{X}_n - \mu_P)^2 = o_p(1)$. Wenn $E[X^4] < \infty$ gilt, dann ist also $\hat{\sigma}^2$ asymptotisch linear mit Einflußfunktion $a(X) = (X - \mu_P)^2 - \sigma_P^2$; insbesondere asymptotisch linear mit Varianz $E[(X - \mu_P)^4] - \sigma_P^4$.

Beispiel 17.4. Seien $X_1, X_2, \ldots$ unabhängig und $B_{k,p}$-verteilt. mit bekanntem k. Weil $B_{k,p}$ wie $B_{1,p}^k$ verteilt ist, ergibt sich sofort $EX = kp$. Also ist das arithmetische Mittel $(1/n)\sum_{i=1}^n X_i$ ein erwartungstreuer, konsistenter und (exakt) linearer Schätzer für kp mit asymptotischer Varianz $kp(1-p)$. Also ist $(1/nk)\sum_{i=1}^n X_i$ erwartungstreu, konsistent und (exakt) linear für p mit asymptotischer Varianz $p(1-p)/k$. Die Binomialverteilung ist also informativer für p als die Bernoulliverteilung.

Weiß man etwas über die zugrundeliegenden Verteilungen, so gibt es oft *bessere* Schätzer als die empirischen.

Beispiel 17.5. Die Taxis einer Stadt tragen die Nummern $1, \ldots, N$. Wie können wir N schätzen? Wir beobachten n verschiedene Taxinummern. Sie können als unabhängige, auf $\{1, \ldots, N\}$ Laplace-verteilte Zufallsvariablen modelliert werden. Hier sind drei Schätzer für N:

1. Es gilt $EX_1 = (N+1)/2$, also $N = 2EX_1 - 1$; also ist $T_n^{(1)} = 2\bar{X}_n - 1$ erwartungstreu, konsistent und asymptotisch normal für N. Der Schätzer hat insbesondere die Konvergenzrate $n^{-1/2}$. Ein Nachteil: Er kann kleiner sein als die größte beobachtete Taxinummer.

2. Die größte beobachtete Taxinummer $T_n^{(2)} = \max\{X_1, \ldots, X_n\}$. Er ist nicht erwartungstreu, aber konsistent, denn irgendwann hat man alle Taxinummern beobachtet,

$$P(T_n^{(2)} < N) = P(X_1 < N, \ldots, X_n < N) = P(X_1 < N)^n = (1 - 1/N)^n \to 0.$$

Dieser Schätzer ist viel besser als $T_n^{(1)}$, denn seine Konvergenzrate ist $1/n$. Heuristisch folgt das daraus, daß die Beobachtungen ziemlich gleichmäßig über $\{1, \ldots, N\}$ verteilt sind, also die Lücke von der größten beobachteten Taxinummer $T_n^{(2)}$ bis N ungefähr $N/(n+1)$ ist.

3. Wir können $T_n^{(2)}$ verbessern, indem wir einen Schätzer für diese Lücke addieren,

$$T_n^{(3)} = \frac{n+1}{n} T_n^{(2)}.$$

18 Lineare Regression

Die Wirkung Y einer medizinischen Behandlung hängt im allgemeinen von anderen Merkmalen des Patienten ab, die wir in einem Vektor X zusammenfassen (zum Beispiel dem Krankenblatt). Wir nennen Y die *Zielvariable*, X die *Kovariablen*. Manchmal läßt sich diese Abhängigkeit mit einem *linearen Regressionsmodell* beschreiben,

$$Y = a^\top X + b + \varepsilon,$$

in dem X und ε unabhängig sind und der *Fehler* ε zentriert ist, $E\varepsilon = 0$. Wir kennen die Parameter a und b nicht, und auch nicht die Verteilungen von ε und X.

Wir haben unabhängige und identisch verteilte Beobachtungen (X_i, Y_i), $i = 1, \ldots, n$, aus diesem Modell und wollen a und b schätzen.

Definition 18.1. Die *Kleinste-Quadrate-Schätzer* $\hat{a}$ und $\hat{b}$ sind Minimierer von

$$\sum_{i=1}^n (Y_i - a^\top X_i - b)^2.$$

Wir bestimmen $\hat{a}$ und $\hat{b}$, indem wir die partiellen Ableitungen des obigen Ausdrucks nach a und b gleich Null setzen. Das liefert die *Normalengleichungen*

$$\sum_{i=1}^{n}(Y_i - a^\top X_i - b) = 0,$$
$$\sum_{i=1}^{n} X_i(Y_i - a^\top X_i - b) = 0.$$

Mit den Abkürzungen

$$c = \begin{pmatrix} b \\ a \end{pmatrix}, \qquad Z = \begin{pmatrix} 1 \\ X \end{pmatrix}$$

kann man die Normalengleichungen zusammenfassen,

$$\sum_{i=1}^{n} Z_i(Y_i - c^\top Z_i) = 0.$$

Der Kleinste-Quadrate-Schätzer $\hat{c} = (\hat{b}, \hat{a})^\top$ ergibt sich also als

$$\hat{c} = \Big(\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^\top\Big)^{-1} \frac{1}{n}\sum_{i=1}^{n} Z_i Y_i.$$

Er ist konsistent, denn aus $Y = c^\top Z + \varepsilon = Z^\top c + \varepsilon$ folgt

$$E[ZY] = E[ZZ^\top]c + E[Z\varepsilon] = E[ZZ^\top]c.$$

Das letzte Gleichheitszeichen gilt, weil Z und ε unabhängig sind mit $E[\varepsilon] = 0$, also $E[Z\varepsilon] = E[Z]E[\varepsilon] = 0$ gilt. Es ergibt sich $c = (EZZ^\top)^{-1}EYZ$, und dagegen konvergiert $\hat{c}$ in Wahrscheinlichkeit.

19 Kernschätzer für Dichten

Seien $X_1, \ldots, X_n$ unabhängige reellwertige Beobachtungen mit Dichte f. Sei $x \in \mathbb{R}$. Wir wollen $f(x)$ schätzen. Wenn f stetig in x ist und b eine kleine positive Zahl, so gilt

$$P\big(x - b/2, x + b/2\big) = \int_{x-b/2}^{x+b/2} f(t)dt \approx f(x) \cdot b.$$

Der empirische Schätzer für $P\big(x - b/2, x + b/2\big)$ ist der Anteil der Beobachtungen in $(x - b/2, x + b/2\big)$, nämlich

$$\frac{1}{n}\sum_{i=1}^{n} 1_{(x-b/2,x+b/2)}(X_i) = \frac{1}{n}\#\{i : x - b/2 < X_i < x + b/2\}.$$

Ein einfacher Schätzer für $f(x)$ ist also

$$\hat{f}(x) = \frac{1}{b}\frac{1}{n}\sum_{i=1}^{n} 1_{(x-b/2,x+b/2)}(X_i) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{b}1_{(-b/2,b/2)}(x - X_i) = \frac{1}{n}\sum_{i=1}^{n} K_b(x - X_i)$$

mit den Abkürzungen $K_b(t) = K(x/b)/b$ und $K(t) = 1_{(-1/2,1/2)}(t)$. Es ist aber meist besser, statt der Indikatorfunktion K andere, zum Beispiel glatte, Funktionen zu nehmen. Das führt zu folgender Definition.

Definition 19.1. Sei K eine Funktion mit $\int K(t)dt = 1$. Setze $K_b(t) = K(t/b)/b$. Der *Kernschätzer* mit *Kern* K und Bandweite b_n ist definiert durch

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_{b_n}(x - X_i).$$

Um Konsistenz von $\hat{f}_n(x)$ zu zeigen, gehen wir wie folgt vor. Mit der Chebyshev-Ungleichung 14.3 gilt zunächst

$$P(|\hat{f}_n(x) - f(x)| > \varepsilon) \le \frac{1}{\varepsilon^2} E[(\hat{f}_n(x) - f(x))^2].$$

Der Erwartungswert $E[(\hat{f}_n(x) - f(x))^2]$ ist der *mittlere quadratische Fehler* von $\hat{f}_n(x)$. Wir können ihn in die Summe aus der *Varianz* und dem Quadrat des *Bias* von $\hat{f}_n(x)$ zerlegen,

$$\begin{aligned} E[(\hat{f}_n(x) - f(x))^2] &= E[(\hat{f}_n(x) - E[\hat{f}_n(x)])^2] + \big(E[\hat{f}_n(x)] - f(x)\big)^2 \\ &= \mathrm{Var}(\hat{f}_n(x)) + \big(\mathrm{Bias}(\hat{f}_n(x))\big)^2. \end{aligned}$$

Die Konsistenz von $\hat{f}_n(x)$ folgt also, wenn Bias und Varianz von $\hat{f}_n(x)$ für $n \to \infty$ gegen 0 gehen. Das zeigt die folgende Proposition.

Proposition 19.1. *Sei f stetig in x, sei K beschränkt mit beschränktem Träger, und gelte $b_n \to 0$ und $nb_n \to \infty$. Dann gilt* $\mathrm{Var}(\hat{f}_n(x)) \to 0$ *und* $\mathrm{Bias}(\hat{f}_n(x)) \to 0$, *also $\hat{f}_n(x) \to f(x)$ in Wahrscheinlichkeit.*

Beweis. Mit der Transformation $u = (x - y)/b_n$ und mit $\int K(t)dt = 1$, dem Satz von der dominierten Konvergenz und $b_n \to 0$ gilt

$$\begin{aligned} E[\hat{f}_n(x)] = E[K_{b_n}(x - X)] &= \int K_{b_n}(x - y) f(y) dy \\ &= \int K(u) f(x - b_n u) du = f(x) + \int K(u)\big(f(x - b_n u) - f(x)\big) du \to f(x), \end{aligned}$$

also $\mathrm{Bias}(\hat{f}_n(x)) \to 0$. Die Varianz von $\hat{f}_n(x)$ ist

$$\begin{aligned} \mathrm{Var}(\hat{f}_n(x)) &= \frac{1}{n}\mathrm{Var}(K_{b_n}(x - X_1)) \\ &= \frac{1}{n}\Big(\int K_{b_n}^2(x - y) f(y) dy - \Big(\int K_{b_n}(x - y) f(y) dy\Big)^2\Big). \end{aligned}$$

Wir brauchen nur noch zu zeigen, daß $\int K_{b_n}^2(x - y) f(y) dy = E[K_{b_n}(x - X_1)]$ langsamer als n wächst. Das folgt ähnlich wie eben bei $\int K_{b_n}(x - y) f(y) dy$,

$$\begin{aligned} \int K_{b_n}^2(x - y) f(y) dy &= \frac{1}{b_n}\int K^2(u) f(x - b_n u) du \\ &= \frac{1}{b_n} f(x) \int K^2(u) du + o\Big(\frac{1}{b_n}\Big) = O\Big(\frac{1}{b_n}\Big), \end{aligned}$$

also wegen $nb_n \to \infty$

$$\mathrm{Var}(\hat{f}_n(x)) = O\Big(\frac{1}{nb_n}\Big) \to 0.$$

□

20 Maximum-Likelihood-Schätzer

In Kapitel 12 hatten wir Dichten bezüglich des Lebesgue-Maßes λ^m eingeführt. Das brauchen wir jetzt allgemeiner. Ist $P|\mathcal{F}$ ein Wahrscheinlichkeitsmaß und $\mu|\mathcal{F}$ ein Maß, so heißt eine meßbare Funktion f eine *μ-Dichte* von P, wenn

$$P(A) = \int_A f(x)\mu(dx) \quad \text{für } A \in \mathcal{F}.$$

Die Verteilungen aus Kapitel 12 haben Lebesgue-Dichten. Für eine diskretes Wahrscheinlichkeitsmaß P auf $\omega_1, \omega_2, \ldots$ wählen wir das *Zählmaß* $\mu(\{\omega_i\}) = 1$ für $i = 1, 2, \ldots$. Dann hat P die μ-Dichte $f(\omega_i) = P(\{\omega_i\})$, denn

$$P(A) = \sum_{\omega_i \in A} P(\{\omega_i\}) = \int_A f(x)\mu(dx).$$

Die meisten diskreten Wahrscheinlichkeitsmaße aus Kapitel 11 leben auf $\mathbb{Z}$ oder einer Teilmenge davon.

Seien $X_1, \ldots, X_n$ unabhängige Beobachtungen mit einer μ-Dichte f_ϑ, die von einem (unbekannten) Parameter $\vartheta \in \Theta \subset \mathbb{R}^k$ abhängt. Wir wollen ϑ schätzen. Dazu wählen wir das ϑ, zu dem die Beobachtungen am besten passen ("maximum likelihood" haben).

Definition 20.1. Der *Maximum-Likelihood-Schätzer* $\hat{\vartheta}$ für ϑ ist derjenige Parameter ϑ, für den die Dichte $\prod_{i=1}^n f_\vartheta(X_i)$ maximiert wird.

Es ist manchmal bequem, statt der Dichte ihren Logarithmus zu maximieren, denn er verwandelt das Produkt in eine Summe. Falls $f_\vartheta(x)$ für jedes x in ϑ differenzierbar ist, verschwinden in $\vartheta = \hat{\vartheta}$ die partiellen Ableitungen des Logarithmus,

$$\sum_{i=1}^n \partial_{\vartheta=\hat{\vartheta}} \log f_\vartheta(X_i) = 0.$$

Beispiel 20.1. Die Bernoulli-Verteilung $B_{1,p}$ hat die Zähldichte

$$f_p(x) = p^x(1-p)^{1-x}, \quad x = 0, 1.$$

Sind $X_1, \ldots, X_n$ unabhängig und $B_{1,p}$-verteilt, so ist die Zähldichte von $(X_1, \ldots, X_n)$ das Produkt

$$\prod_{i=1}^n p^{X_i}(1-p)^{1-X_i} = p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i}.$$

Der Logarithmus ist

$$\sum_{i=1}^n X_i \log p + \Big(n - \sum_{i=1}^n X_i\Big) \log(1-p).$$

Die Ableitung davon verschwindet an der Stelle des Maximum-Likelihood-Schätzers $\hat{p}$ für p,

$$0 = \frac{\sum_{i=1}^n X_i}{p} - \frac{n - \sum_{i=1}^n X_i}{1-p} = \frac{\sum_{i=1}^n X_i - np}{p(1-p)}.$$

Also ist $\hat{p} = (1/n)\sum_{i=1}^n X_i$. Das ist gleichzeitig der empirische Schätzer für den Erwartungswert $p = EX_1$.

Beispiel 20.2. Seien $X_1, \dots, X_n$ unabhängig und N_{μ,σ^2}-verteilt. Als Parameter wählen wir $\vartheta = (\mu, \sigma^2)^\top$. Die (Lebesgue-)Dichte von N_{μ,σ^2} ist

$$f_\vartheta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(x-\mu)^2}{2\sigma^2}\Big).$$

Der Logarithmus der Dichte einer Beobachtung ist

$$\log f_\vartheta(x) = \log \frac{1}{\sqrt{2\pi}} - \frac{1}{2}\log\sigma^2 - \frac{(x-\mu)^2}{2\sigma^2}.$$

Der Logarithmus der Dichte von $(X_1, \dots, X_n)$ ist also

$$\sum_{i=1}^n \log f_\vartheta(X_i) = n \log \frac{1}{\sqrt{2\pi}} - \frac{n}{2}\log\sigma^2 - \sum_{i=1}^n \frac{(X_i-\mu)^2}{2\sigma^2}.$$

Wir setzen die partiellen Ableitungen nach μ und σ^2 gleich 0,

$$\frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \mu) = 0, \qquad \frac{1}{\sigma^4}\sum_{i=1}^n (X_i-\mu)^2 - \frac{n}{\sigma^2} = 0.$$

Wir lösen erst die erste, dann die zweite Gleichung und erhalten als Maximum-Likelihood-Schätzer $\hat\vartheta = (\hat\mu, \hat\sigma^2)^\top$ mit

$$\hat\mu = \frac{1}{n}\sum_{i=1}^n X_i, \qquad \hat\sigma^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \hat\mu)^2.$$

Das sind (wieder) die empirischen Schätzer für die Erwartungswerte $\mu = EX_1$ und $\sigma^2 = \mathrm{Var}(X_1)$.

Beispiel 20.3. In diesem Beispiel haben wir keine (nach ϑ) differenzierbare Dichte, und der Maximum-Likelihood-Schätzer ist auch kein empirischer Schätzer.

Für $\vartheta > 0$ hat die Gleichverteilung auf $[0, \vartheta]$ die (Lebesgue-)Dichte

$$f_\vartheta(x) = \frac{1}{\vartheta} 1_{[0,\vartheta]}(x).$$

Seien $X_1, \dots, X_n$ unabhängig gleichverteilt auf $[0, \vartheta]$. Die Dichte von $(X_1, \dots, X_n)$ ist dann

$$\frac{1}{\vartheta^n}\prod_{i=1}^n 1_{[0,\vartheta]}(X_i) = \frac{1}{\vartheta^n} 1_{[0,\vartheta]}(X_{(n)})$$

mit der Abkürzung $X_{(n)} = \max\{X_1, \dots, X_n\}$. Der Maximum-Likelihood-Schätzer ist also $\hat\vartheta = X_{(n)}$. Es gilt $E_\vartheta X = \frac{\vartheta}{2}$, also ist $\frac{2}{n}\sum_{i=1}^n X_i$ der *empirische* Schätzer für ϑ. Wir wissen schon aus dem Beispiel mit den Taxinummern, daß er wesentlich schlechter ist als der Maximum-Likelihood-Schätzer, weil er nur die Konvergenzrate $n^{-1/2}$ hat.

21 Tests

Seien $(\Omega, \mathcal{F})$ ein meßbarer Raum und $P_\vartheta|\mathcal{F}$ ein Wahrscheinlichkeitsmaß mit einem Parameter $\vartheta \in \Theta$. (In unseren Beispielen werden wir es wieder wie in den Kapiteln 17–20 mit $\Omega = \mathbb{R}^n$, $\mathcal{F} = \mathcal{B}^n$ und $P_\vartheta = P_\vartheta^n$ zu tun haben.) Wir wollen testen, ob der unbekannte Parameter ϑ in einer Menge $H \subset \Theta$, der *Hypothese*, enthalten ist. Das Komplement $K = \Theta \backslash H$ nennen wir *Alternative*.

Definition 21.1. Ein (randomisierter) *Test* (für H gegen K) ist eine meßbare Abbildung $\varphi : \Omega \to [0, 1]$. Ist $\varphi = 1_C$, so heißt der Test *nichtrandomisiert* und C der *kritische Bereich* Interpretation: Wird x beobachtet, entscheiden wir uns mit Wahrscheinlichkeit $\varphi(x)$ für K.

Testen ist im Prinzip leichter als Schätzen, da wir nicht ϑ, sondern nur H oder K wissen wollen. Häufig beruhen aber auch Tests auf Schätzern von ϑ, wie in den folgenden beiden Beispielen.

Beispiel 21.1. Wir wollen testen, ob eine Münze unverfälscht ist. Wir machen n Münzwürfe. Das sind Beobachtungen $X_1, \ldots, X_n$, die unabhängig und $B_{1,p}$-verteilt sind. Die Hypothese ist $H = \{1/2\}$; die Alternative ist $K = (0, 1)\backslash\{1/2\}$. Ein kritischer Bereich könnte dann die Form $C = \{|\bar{X}_n - 1/2| > c\}$ haben.

Beispiel 21.2. Ein neues fiebersenkendes Mittel soll daraufhin getestet werden, ob es besser ist als ein auf dem Markt schon eingeführtes. Wir messen Temperaturen $X_1, \ldots, X_n$ bei n Patienten und nehmen an, daß sie unabhängig und N_{μ,σ^2}-verteilt sind. Das schon eingeführte Mittel führe auch zu normalverteilten Temperaturen, aber möglicherweise mit anderem Mittelwert $\mu = \mu_0$ und anderer Varianz. Die Hypothese ist dann $H = \{(\mu, \sigma^2) : \mu \geq \mu_0\}$; die Alternative ist $K = \{(\mu, \sigma^2) : \mu < \mu_0\}$. Ein kritischer Bereich könnte dann die Form $C = \{\bar{X}_n \leq \mu_0 + c\}$ haben.

Wie sollen wir c wählen? Wir können zwei Arten von Fehlern machen. Entscheiden wir uns fälschlich für die Alternative, ist das ein Fehler *erster* Art. Im Mittel ist er

$$E_\vartheta \varphi, \quad \text{falls } \vartheta \in H.$$

Entscheiden wir uns fälschlich für die Hypothese, ist das ein Fehler *zweiter* Art. Im Mittel ist er

$$E_\vartheta(1 - \varphi) = 1 - E_\vartheta \varphi, \quad \text{falls } \vartheta \in K.$$

Wir können nicht beide gleichzeitig minimieren. Wir wählen deshalb eine Schranke für den mittleren Fehler erster Art und versuchen unter dieser Nebenbedingung an φ den mittleren Fehler zweiter Art zu minimieren, das heißt, die *Gütefunktion* $\vartheta \to E_\vartheta \varphi$ für $\vartheta \in K$ zu *maximieren*.

Wir behandeln Hypothese und Alternative nicht symmetrisch, weil gewöhnlich der Fehler erster Art größere Risiken birgt. (Zum Beispiel ein Medikament auf den Markt zu bringen, obwohl es nicht besser ist als die schon eingeführten.) Das ist auch der Grund für die Namen Fehler "erster" Art und "kritischer" Bereich.

Definition 21.2. Ein Test φ hat das *Niveau* α für H, wenn

$$E_\vartheta \varphi \leq \alpha \quad \text{für } \vartheta \in H.$$

Wir werden immer diese Nebenbedingung ganz ausschöpfen, weil wir die Gütefunktion $E_\vartheta\varphi$ für $\vartheta \in K$ möglichst groß machen wollen. Auch in den folgenden beiden Beispielen schöpfen wir das Niveau (asymptotisch) ganz aus.

Beispiel 21.3 (Fortsetzung des Beispiels 21.1). Beim Testen einer Münze auf Unverfälschtheit war die Hypothese $H = \{1/2\}$ und unser kritischer Bereich $C = \{|\bar{X}_n - 1/2| > c\}$. Um das Niveau α einzuhalten und auszuschöpfen, brauchen wir ein c mit

$$B^n_{1,1/2}(|\bar{X}_n - 1/2| > c) = \alpha.$$

Dieses c läßt sich exakt berechnen. Wir wollen es hier wenigstens asymptotisch bestimmen. Die Varianz von $B_{1,1/2}$ ist $1/4$. Bezeichnet Φ die Verteilungsfunktion der Standardnormalverteilung $N_{0,1}$ und setzen wir $c = n^{-1/2}b$, so erhalten wir mit dem Zentralen Grenzwertsatz

$$\begin{aligned} B^n_{1,1/2}(|\bar{X}_n - 1/2| > n^{-1/2}b) &= B^n_{1,1/2}\Big(\Big|n^{-1/2}\sum_{i=1}^n (X_i - 1/2)\Big| > b\Big) \\ &\to N_{0,1/4}(-b,b)^c = N_{0,1}(-2b,2b)^c \\ &= 1 - \Phi(2b) + \Phi(-2b) = 2 - 2\Phi(2b). \end{aligned}$$

Die rechte Seite ist α für $2b = \Phi^{-1}(1-\alpha/2)$. Also hat der Test bei n Beobachtungen ungefähr das Niveau α, wenn $c = n^{-1/2}b = n^{-1/2}\Phi^{-1}(1-\alpha/2)/2$. (Seine Genauigkeit ist also von der Ordnung $n^{-1/2}$, genau wie bei den empirischen *Schätzern*.)

Beispiel 21.4 (Fortsetzung des Beispiels 21.2). a) Nehmen wir zunächst an, daß wir die *Varianz* $\sigma^2 = \sigma_0^2$ bei den Temperaturen des neuen Medikaments kennen. Um das Niveau auszuschöpfen, brauchen wir ein c mit

$$N^n_{\mu_0,\sigma_0^2}(\bar{X}_n \le \mu_0 + c) = \alpha.$$

Für $c = n^{-1/2}b$ erhalten wir

$$\begin{aligned} N^n_{\mu_0,\sigma_0^2}(\bar{X}_n \le \mu_0 + n^{-1/2}b) &= N^n_{\mu_0,\sigma_0^2}\Big(n^{-1/2}\sum_{i=1}^n (X_i - \mu_0) \le b\Big) \\ &= N_{\mu_0,\sigma_0^2}(-\infty, b) = \Phi(b/\sigma_0). \end{aligned}$$

Das ist α für $b = \sigma_0\Phi^{-1}(\alpha)$.

b) Wenn die Varianz σ^2 beim neuen Medikament nicht bekannt ist, müssen wir σ_0^2 im kritischen Bereich, also in $c = n^{-1/2}b = n^{-1/2}\sigma_0\Phi^{-1}(\alpha)$, durch einen Schätzer ersetzen, zum Beispiel durch

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

Dann erhalten wir mit Teil a) dieses Beispiels und dem Lemma von Slutzky

$$\begin{aligned} &N^n_{\mu_0,\sigma_0^2}(\bar{X}_n \le \mu_0 + n^{-1/2}\hat{\sigma}\Phi^{-1}(\alpha)) \\ &= N^n_{\mu_0,\sigma_0^2}\Big(n^{-1/2}\sum_{i=1}^n (X_i - \mu_0) \le \hat{\sigma}\Phi^{-1}(\alpha)\Big) \\ &= N^n_{\mu_0,\sigma_0^2}\Big(n^{-1/2}\sum_{i=1}^n (X_i - \mu_0) - (\hat{\sigma} - \sigma)\Phi^{-1}(\alpha) \le \sigma\Phi^{-1}(\alpha)\Big) \to \alpha. \end{aligned}$$

Also wird das Niveau zumindest noch asymptotisch ausgeschöpft.

Sind H und K *einfach*, also einpunktig, so schreiben wir sie einfach als zwei Wahrscheinlichkeitsmaße P und Q. In diesem Fall gibt es immer einen "besten" Test zu vorgegebenem Niveau, den "Neyman–Pearson-Test".

Lemma 21.1 (Neyman–Pearson)**.** *Seien P und Q Wahrscheinlichkeitsmaße mit Dichten p und q. Sei $\alpha \in (0,1)$. Sei c das $(1-\alpha)$-*Quantil *von q/p unter P, das heißt*

$$c = \inf\{y : P(q > yp) \le \alpha\}.$$

Setze

$$a = \begin{cases} \frac{\alpha - P(q>cp)}{P(q=cp)}, & P(q=cp) > 0, \\ 0, & sonst \end{cases}$$

und

$$\psi = \begin{cases} 1, & > \\ a, & q = cp. \\ 0, & < \end{cases}$$

Für das Niveau gilt dann $E_P\psi = \alpha$; für die Güte gilt $E_Q\psi \ge E_Q\varphi$, falls $E_P\varphi \le \alpha$.

Beweis. Die erste Behauptung folgt aus

$$E_P\psi = P(q > cp) + aP(q = cp) = \alpha.$$

Wegen $(q - cp)(\psi - \varphi) \ge 0$ gilt

$$E_Q\psi - E_Q\varphi \ge c(E_P\psi - E_P\varphi) \ge 0,$$

also die zweite Behauptung. □

Der Test ψ heißt *Neyman–Pearson-Test* zum Niveau α für P gegen Q. Die Zahl c heißt *kritischer Wert*. Das Prinzip ähnelt dem des Maximum-Likelihood-Schätzers: Wir entscheiden uns für Q, wenn die Dichte von Q bezüglich P groß ist. Wir randomisieren so wenig wie möglich, nämlich nur, wenn $q = cp$, und dort immer gleich, mit a.

Beispiel 21.5. Seien $X_1, \dots, X_n$ unabhängig und $B_{1,p}$-verteilt. Dann haben die n Beobachtungen $(X_1, \dots, X_n)$ die Zähldichte

$$\prod_{i=1}^n p^{X_i}(1-p)^{1-X_i} = p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i} = \Big(\frac{p}{1-p}\Big)^{\sum_{i=1}^n X_i}(1-p)^n.$$

Der Neyman–Pearson-Test zum Niveau α für p gegen ein $q > p$ ist von der Form

$$\psi = \begin{cases} 1, & \\ a, & \Big(\frac{q}{1-q}\Big)^{\sum_{i=1}^n X_i}(1-q)^n \ \begin{matrix} > \\ = \\ < \end{matrix}\ c\Big(\frac{p}{1-p}\Big)^{\sum_{i=1}^n X_i}(1-p)^n. \\ 0, & \end{cases}$$

Der Quotient $p/(1-p)$ wächst mit p, denn die Ableitung

$$\partial_p \log \frac{p}{1-p} = \frac{1}{p} + \frac{1}{1-p}$$

ist positiv für $p \in (0,1)$. Also ist der Test von der Form

$$\psi = \begin{cases} 1, & \\ a, & \sum_{i=1}^n X_i = b, \\ 0, & \end{cases} \begin{matrix} > \\ \\ < \end{matrix}$$

wobei jetzt a und b so gewählt werden, daß $E_p\psi = \alpha$. Weil $\sum_{i=1}^n X_i$ wie $B_{n,p}$ verteilt ist, bedeutet das

$$B_{n,p}(b,\infty) + aB_{n,p}\{b\} = \alpha.$$

Der Test hängt *nicht* von $q > p$ ab. Er ist außerdem identisch mit dem Test, der auf dem empirischen Schätzer für p beruht.

Beispiel 21.6. Seien $X_1, \dots, X_n$ positiv, unabhängig und E_ϑ-verteilt, das heißt exponentialverteilt mit Parameter ϑ. Dann haben die n Beobachtungen $(X_1, \dots, X_n)$ die Lebesgue-Dichte

$$\prod_{i=1}^n \frac{1}{\vartheta} \exp(-X_i/\vartheta) = \frac{1}{\vartheta^n} \exp\Big(-\frac{1}{\vartheta}\sum_{i=1}^n X_i\Big).$$

Der Neyman-Pearson-Test zum Niveau α für ϑ gegen ein $\tau > \vartheta$ hat die Form

$$\psi = \begin{cases} 1, & \\ a, & \frac{1}{\tau^n} \exp\Big(-\frac{1}{\tau}\sum_{i=1}^n X_i\Big) \\ 0, & \end{cases} \begin{matrix} > \\ = \\ < \end{matrix} \frac{c}{\vartheta^n} \exp\Big(-\frac{1}{\vartheta}\sum_{i=1}^n X_i\Big),$$

anders geschrieben

$$\psi = \begin{cases} 1, & \\ a, & \frac{\vartheta^n}{\tau^n} \exp\Big(-\big(\frac{1}{\tau} - \frac{1}{\vartheta}\big)\sum_{i=1}^n X_i\Big) \\ 0, & \end{cases} \begin{matrix} > \\ = \\ < \end{matrix} c.$$

Wegen $1/\tau < 1/\vartheta$ ist der Test von der Form

$$\psi = \begin{cases} 1, & \\ a, & \sum_{i=1}^n X_i = b, \\ 0, & \end{cases} \begin{matrix} > \\ \\ < \end{matrix}$$

wobei wieder a und b so gewählt werden, daß $E_\vartheta\psi = \alpha$. (Hier bedeutet E_ϑ "Erwartungswert"). Wegen $E_\vartheta = \Gamma_{\vartheta,1}$ gilt nach der Bemerkung über die Faltung von Gammaverteilungen, daß $\sum_{i=1}^n X_i$ verteilt nach $\Gamma_{\vartheta,n}$ ist. Also ergibt sich b aus $\Gamma_{\vartheta,n}(b,\infty) = \alpha$. Eine Randomisierung ist unnötig, da $\Gamma_{\vartheta,n}(\{b\}) = 0$.

22 Exponentielle Familien, monotone Dichtequotienten, gleichmäßig beste Tests

Wir haben in den letzten beiden Beispielen gesehen, daß es Tests geben kann, die gegen alle Alternativen gleichzeitig optimal sind. Man nennt sie (etwas irreführend) "gleichmäßig" beste Tests. Solche Tests gibt es nur für bestimmte parametrische Familien von Dichten.

Definition 22.1. Eine Familie von Wahrscheinlichkeitsmaßen $P_\vartheta|\mathcal{F}$ mit $\vartheta \in \Theta$ heißt *exponentielle Familie* in T und $\eta(\vartheta)$, wenn sie μ-Dichten folgender Form hat,

$$f_\vartheta(x) = c(\vartheta)g(x)\exp\big(\eta(\vartheta)^\top T(x)\big).$$

Beispiel 22.1. Die Normalverteilungen N_{μ,σ_0^2} mit $\mu \in \mathbb{R}$ und bekanntem σ_0^2 haben die (Lebesgue-)Dichte

$$\begin{aligned} f_\mu(x) &= \frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\Big(-\frac{1}{2\sigma_0^2}(x-\mu)^2\Big) \\ &= \frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\Big(-\frac{\mu^2}{2\sigma_0^2}\Big)\exp\Big(-\frac{x^2}{2\sigma_0^2}\Big)\exp\Big(\frac{\mu x}{\sigma_0^2}\Big). \end{aligned}$$

Sie bilden also eine exponentielle Familie mit

$$c(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\Big(-\frac{\mu^2}{2\sigma_0^2}\Big), \qquad g(x) = \exp\Big(-\frac{x^2}{2\sigma_0^2}\Big)$$

und mit $T(x) = x$ und $\eta(\mu) = \mu/\sigma_0^2$.

Beispiel 22.2. Die $B_{n,p}$-Verteilungen mit $p \in (0,1)$ besitzen in den Punkten $k = 0,\dots,n$ die Zähldichte

$$f_p(k) = \binom{n}{k}p^k(1-p)^{n-k} = \binom{n}{k}(1-p)^n\exp\Big(k\log\frac{p}{1-p}\Big).$$

Sie bilden also eine exponentielle Familie mit

$$c(p) = (1-p)^n, \qquad g(k) = \binom{n}{k}$$

und mit $T(k) = k$ und $\eta(p) = \log\big(p/(1-p)\big)$.

Beispiel 22.3. Sind $X_1,\dots,X_n$ unabhängig mit μ-Dichte

$$f_\vartheta(x) = c(\vartheta)g(x)\exp\big(\eta(\vartheta)^\top T(x)\big),$$

so hat $(X_1,\dots,X_n)$ die μ^n-Dichte

$$\prod_{i=1}^n f_\vartheta(X_i) = c(\vartheta)^n\prod_{i=1}^n g(X_i)\exp\Big(\eta(\vartheta)^\top\sum_{i=1}^n T(X_i)\Big).$$

In den letzten Beispielen für Tests haben wir immer den Neyman–Pearson-Test umgeformt, indem wir die Dichte q unter der Alternative durch die Dichte p unter der Hypothese geteilt haben. Das schreiben wir jetzt etwas allgemeiner auf. Für $\vartheta \in \Theta$ seien f_ϑ μ-Dichten, die alle auf der gleichen Menge A positiv sind. Der *Dichteqotient* von P_τ nach P_ϑ ist definiert durch

$$L_{\vartheta\tau}(x) = \frac{f_\tau(x)}{f_\vartheta(x)}, \qquad x \in A.$$

Definition 22.2. Die Familie P_ϑ mit $\vartheta \in \Theta$ hat *monotone Dichtequotienten* in T, wenn T eine Zufallsvariable ist und für alle $\vartheta, \tau \in \Theta$ mit $\vartheta < \tau$ eine nichtfallende Funktion $H_{\vartheta\tau}$ existiert mit $L_{\vartheta\tau} = H_{\vartheta\tau} \circ T$.

Definition 22.3. Ein Test ψ zum Niveau α für H gegen K heißt *gleichmäßig bester Test zum Niveau* α, wenn

$$E_\vartheta \psi \geq E_\vartheta \varphi, \quad \vartheta \in K,$$

für alle Tests φ zum Niveau α für H.

Satz 22.1. *Hat P_ϑ, $\vartheta \in \Theta$, monotone Dichtequotienten in T und ist $\alpha \in (0,1)$, so existiert ein gleichmäßig bester Test zum Niveau α für $\tau \leq \vartheta$ gegen $\tau > \vartheta$, nämlich*

$$\psi = \begin{cases} 1, & > \\ a, & T = b. \\ 0, & < \end{cases}$$

Dabei sind a und b bestimmt durch $E_\vartheta \psi = P_\vartheta(T > b) + aP_\vartheta(T = b) = \alpha$. Für $\tau < \vartheta$ minimiert ψ das Niveau unter allen Tests φ mit $E_\vartheta \varphi \geq \alpha$.

Beweis. Wähle a und c wie in Lemma 21.1. Der Neyman–Pearson-Test zum Niveau α für ϑ gegen $\tau > \vartheta$ hat die Form

$$\psi = \begin{cases} 1, & > \\ a, & L_{\vartheta\tau} = c. \\ 0, & < \end{cases}$$

Wähle b mit $H_{\vartheta\tau}(b) = c$. Wegen $L_{\vartheta\tau} = H_{\vartheta\tau} \circ T$ kann man ψ wie in der Behauptung schreiben. Insbesondere hängt ψ nicht von τ ab. Also gilt $E_\tau \psi \geq E_\tau \varphi$ für $\tau > \vartheta$. Durch Umkehren der Ungleichungen folgt, daß $1 - \psi$ ein gleichmäßig bester Test zum Niveau $1 - \alpha$ für ϑ gegen $\tau < \vartheta$ ist. Insbesondere gilt $E_\tau(1 - \psi) \geq 1 - \alpha$, denn $1 - \psi$ ist besser als der Test $\varphi = 1 - \alpha$. Also hat ψ das Niveau α nicht nur in ϑ, sondern für alle $\tau \leq \vartheta$. □

Satz 22.2. *Sei P_ϑ mit $\vartheta \in \Theta \subset \mathbb{R}$ eine exponentielle Familie in $\eta(\vartheta)$ und $T(x)$. Ist η nichtfallend (oder nichtwachsend), so hat die Familie monotone Dichtequotienten in T (bzw. $-T$).*

Beweis. Der Dichtequotient hat die Form $L_{\vartheta\tau}(x) = H_{\vartheta\tau}(T(x))$ mit

$$H_{\vartheta\tau}(t) = \frac{c(\tau)}{c(\vartheta)} \exp\big((\eta(\tau) - \eta(\vartheta))t\big).$$

Ist η nichtfallend und $\vartheta < \tau$, so gilt $\eta(\tau) - \eta(\vartheta) \geq 0$. Also ist $H_{\vartheta\tau}(t)$ nichtfallend in t. Bei *nichtwachsendem* η ist $\eta(\tau) - \eta(\vartheta) \leq 0$, also $H_{\vartheta\tau}(t)$ *nichtwachsend* in t, also $H^-_{\vartheta\tau}(t) = H_{\vartheta\tau}(-t)$ nichtfallend in t. □

Beispiel 22.4. Seien $X_1, \ldots, X_n$ unabhängig und N_{μ,σ_0^2}-verteilt mit bekanntem σ_0^2. Diese Verteilungen bilden eine exponentielle Familie in $\eta(\mu) = \mu/\sigma_0^2$ und $T(x) = x$. Also folgt $(X_1, \ldots, X_n)$ einer exponentiellen Familie in $\eta(\mu) = \mu/\sigma_0^2$ und $T = \sum_{i=1}^n X_i$. Nach Satz 22.2 hat $(X_1, \ldots, X_n)$ also monotone Dichtequotienten in T.

Nach Satz 22.1 ist ein gleichmäßig bester Test zum Niveau α für $\mu \le \mu_0$ gegen $\mu > \mu_0$ gegeben durch

$$\psi = \begin{cases} 1, & T \begin{matrix} > \\ < \end{matrix} b. \\ 0, & \end{cases}$$

Randomisierung ist nicht nötig, denn T ist für $\mu = \mu_0$ verteilt wie $N_{n\mu_0, n\sigma_0^2}$, hat also eine Lebesgue-Dichte. Die Konstante b ergibt sich aus

$$N^n_{\mu_0,\sigma_0^2}\Big(n^{-1/2}\sum_{i=1}^n (X_i - \mu_0) > \sigma_0 \Phi^{-1}(1-\alpha)\Big) = \alpha,$$

also

$$N^n_{\mu_0,\sigma_0^2}\Big(T = \sum_{i=1}^n X_i > n\mu_0 + n^{1/2}\sigma_0\Phi^{-1}(1-\alpha)\Big) = \alpha.$$

Beispiel 22.5. Seien $X_1, \dots, X_n$ unabhängig und $B_{k,p}$-verteilt mit $p \in (0,1)$ und bekanntem k. Diese Verteilungen bilden eine exponentielle Familie in $\eta(p) = \log\big(p/(1-p)\big)$ und $T(x) = x$. Also folgt $(X_1, \dots, X_n)$ einer exponentiellen Familie in $\eta(p) = \log\big(p/(1-p)\big)$ und $T = \sum_{i=1}^n X_i$. Für die Ableitung von η gilt, wie schon einmal benutzt,

$$\eta'(p) = \frac{1}{p} + \frac{1}{1-p} > 0.$$

Also folgt $(X_1, \dots, X_n)$ einer Familie mit monotonen Dichtequotienten in T. Nach Satz 22.1 ist ein gleichmäßig bester Test zum Niveau α für $q \le p$ gegen $q > p$ gegeben durch

$$\psi = \begin{cases} 1, & \\ a, & T \begin{matrix} > \\ = \\ < \end{matrix} b. \\ 0, & \end{cases}$$

Für $q = p$ ist T verteilt wie $B_{nk,p}$. Die Konstanten a und b ergeben sich also aus

$$B_{nk,p}(b, \infty) + aB_{nk,p}(\{b\}) = \alpha.$$

Näherungsweise lassen a und b wie folgt berechnen. Die Verteilung $B_{nk,p}$ hat Mittelwert kp und Varianz $kp(1-p)$. Mit dem Zentralen Grenzwertsatz gilt also

$$B^n_{k,p}\Big(n^{-1/2}\sum_{i=1}^n (X_i - kp) > \big(kp(1-p)\big)^{1/2}\Phi^{-1}(1-\alpha)\Big) \to \alpha.$$

Der Test hat also für $q = p$ ungefähr das Niveau α, wenn

$$b = nkp + n^{1/2}\big(kp(1-p)\big)^{1/2}\Phi^{-1}(1-\alpha).$$

Beispiel 22.6. Seien $X_1, \dots, X_n$ unabhängig und P_λ-verteilt mit $\lambda > 0$. Die Zähldichte dieser Poisson-Verteilung in den Punkten $k = 0, 1, \dots$ ist

$$f_\lambda(k) = e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\frac{1}{k!}e^{k\log\lambda}.$$

Die Verteilungen bilden also eine exponentielle Familie in $\eta(\lambda) = \log\lambda$ und $T(k) = k$. Also folgt $(X_1, \dots, X_n)$ einer exponentiellen Familie in $\eta(\lambda) = \log\lambda$ und $T = \sum_{i=1}^n X_i$, und η ist streng monoton wachsend. Also folgt $(X_1, \dots, X_n)$ einer Familie

mit monotonen Dichtequotienten in T. Nach Satz 22.1 ist ein gleichmäßig bester Test zum Niveau α für $\mu \leq \lambda$ gegen $\mu > \lambda$ gegeben durch

$$\psi = \begin{cases} 1, & > \\ a, & T = b. \\ 0, & < \end{cases}$$

Für $\mu = \lambda$ ist T verteilt wie $P_{n\lambda}$. Die Konstanten a und b ergeben sich also aus

$$P_{n\lambda}(b, \infty) + aP_{n\lambda}(\{b\}) = \alpha.$$

Näherungsweise lassen a und b wie folgt berechnen. Die Verteilung P_λ hat Mittelwert λ und Varianz λ. Mit dem Zentralen Grenzwertsatz gilt also

$$P_\lambda^n\Big(n^{-1/2}\sum_{i=1}^n (X_i - \lambda) > \lambda^{1/2}\Phi^{-1}(1-\alpha)\Big) \to \alpha.$$

Der Test hat also für $\mu = \lambda$ ungefähr das Niveau α, wenn

$$b = n\lambda + n^{1/2}\lambda^{1/2}\Phi^{-1}(1-\alpha).$$

23 Konfidenzbereiche

Die Genauigkeit eines Schätzers kennt man ungefähr, wenn man seine asymptotische Varianz kennt, oder wenigstens einen Schätzer für diese Varianz hat. Damit läßt sich ein Intervall konstruieren, in dem der wahre Parameter näherungsweise mit vorgegebener Wahrscheinlichkeit liegt.

Manchmal geht das auch exakt. Seien zum Beispiel $X_1, \ldots, X_n$ unabhängig und N_{μ,σ_0^2}-verteilt mit bekanntem σ_0^2. Ein Schätzer für μ ist das Stichprobenmittel $\bar{X}_n$. Es gilt

$$n^{1/2}(\bar{X}_n - \mu) = n^{-1/2}\sum_{i=1}^n (X_i - \mu) \text{ ist verteilt wie } N_{0,\sigma_0^2}.$$

Insbesondere gilt

$$N_{\mu,\sigma_0^2}^n\Big(-\sigma_0\Phi^{-1}\Big(1-\frac{\alpha}{2}\Big) < n^{-1/2}\sum_{i=1}^n (X_i - \mu) < \sigma_0\Phi^{-1}\Big(1-\frac{\alpha}{2}\Big)\Big) = 1-\alpha.$$

Anders ausgedrückt gilt

$$N_{\mu,\sigma_0^2}^n(\mu \in B) = 1-\alpha$$

für

$$B = \Big\{\mu : \big|\bar{X}_n - \mu\big| < n^{-1/2}\sigma_0\Phi^{-1}\Big(1-\frac{\alpha}{2}\Big)\Big\}.$$

Wir nennen die zufällige (nämlich von den Beobachtungen abhängige) Teilmenge B des Parameterraums ein *Konfidenzintervall* für μ zum *Niveau* $1-\alpha$. Wenn μ wahr ist, liegt μ mit Wahrscheinlichkeit $1-\alpha$ in B.

Definition 23.1. Sei $P_\vartheta|\mathcal{F}$ mit $\vartheta \in \Theta \subset \mathbb{R}$ eine Familie von Wahrscheinlichkeitsmaßen. Eine Abbildung $B : \Omega \to \mathcal{P}(\mathbb{R})$ mit $\{\omega : \vartheta \in B(\omega)\} \in \mathcal{F}$ für jedes ϑ heißt *Konfidenzbereich* für ϑ.

Definition 23.2. Ein Konfidenzbereich B hat das *Niveau* $1-\alpha$, wenn

$$P_\vartheta\{\omega : \vartheta \in B(\omega)\} = 1-\alpha, \quad \vartheta \in \Theta.$$

Bemerkung 23.1. Für $\vartheta \in \Theta$ sei C_ϑ ein kritischer Bereich zum Niveau α für ϑ. Dann hat der Konfidenzbereich $B(\omega) = \{\vartheta : \omega \notin C_\vartheta\}$ das Niveau $1-\alpha$, denn $\alpha = P_\vartheta(C_\vartheta) = P_\vartheta(\omega : \vartheta \notin B(\omega))$.

Im obigen Beispiel mit N_{μ,σ_0^2}-verteilten Beobachtungen haben wir für jedes μ einen kritischen Bereich zum Niveau α für μ gegen $\nu \neq \mu$,

$$C_\mu = \Big\{(X_1,\dots,X_n) : \big|\bar{X}_n - \mu\big| \geq n^{-1/2}\sigma_0\Phi^{-1}\Big(1-\frac{\alpha}{2}\Big)\Big\}.$$

Daraus ergibt sich der obige Konfidenzintervall für μ zum Niveau $1-\alpha$.

Beispiel 23.1. Seien $X_1,\dots,X_n$ unabhängig und $B_{k,p}$-verteilt mit $p \in (0,1)$ und bekanntem k. Ähnlich wie in Beispiel 22.5 haben wir für jedes p einen kritischen Bereich zum asymptotischen Niveau α für p gegen $q \neq p$,

$$C_p = \Big\{(X_1,\dots,X_n) : \big|\bar{X}_n - kp\big| \geq n^{-1/2}(kp(1-p))^{1/2}\Phi^{-1}\Big(1-\frac{\alpha}{2}\Big)\Big\}.$$

Daraus ergibt sich ein Konfidenzintervall zum asymptotischen Niveau $1-\alpha$,

$$B = \Big\{p : \big|\bar{X}_n - kp\big| < n^{-1/2}(kp(1-p))^{1/2}\Phi^{-1}\Big(1-\frac{\alpha}{2}\Big)\Big\}.$$