Wahrscheinlichkeitstheorie und Statistik

Josef Teichmann
unter Verwendung von Skripten von Hans Föllmer, Hansruedi Künsch,
Martin Schweizer und Sara van der Geer

Frühling 2021

Grundbegriffe

Die Axiome von Kolmogorov

Sei $\Omega \neq \emptyset$ irgendeine nichtleere Menge, $\mathcal{A}$ eine Kollektion von Teilmengen $A \subseteq \Omega$ und $\mathbb{P} : \mathcal{A} \to [0, 1]$ eine Abbildung von $\mathcal{A}$ in das Einheitsintervall.

Die Elemente $\omega \in \Omega$ interpretieren wir als die (im Modell in Betracht gezogenen) **möglichen Fälle**, die Teilmengen $A \in \mathcal{A}$ als die (im Modell zugelassenen) **Ereignisse**, und für $A \in \mathcal{A}$ interpretieren wir die Zahl $\mathbb{P}(A)$ als die (im Modell angenommene) **Wahrscheinlichkeit des Ereignisses** A.

Die folgenden **Axiome von Kolmogorov** (1933) verlangen nun, dass die Kollektion $\mathcal{A}$ der Ereignisse abgeschlossen ist unter abzählbaren Mengenoperationen, und dass die Zuordnung $A \to \mathbb{P}(A)$ "konsistent" ist im Sinne der Rechenregeln für Wahrscheinlichkeiten, die uns von den diskreten Modellen her schon vertraut sind.

Axiome von Kolmogorov

Das Tripel $(\Omega, \mathcal{A}, \mathbb{P})$ heisst ein **Wahrscheinlichkeitsraum**, wenn gilt:

1) $\mathcal{A}$ ist eine σ-**Algebra**, d.h.

$$\Omega \in \mathcal{A} \tag{1}$$

$$A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A} \tag{2}$$

$$A_1, A_2, \cdots \in \mathcal{A} \Rightarrow \bigcup_i A_i \in \mathcal{A} \tag{3}$$

2) $\mathbb{P}$ ist ein **Wahrscheinlichkeitsmass**, d.h.

$$\mathbb{P}(\Omega) = 1 \tag{4}$$

$$A_1, A_2, \cdots \in \mathcal{A},\ A_i \cap A_j = \emptyset\ (i \neq j) \Rightarrow \mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i) \tag{5}$$

Bemerkung

In der Sprache der Masstheorie besagt 2), dass $\mathbb{P}$ ein **normiertes Mass** ist, d.h. eine Mengenfunktion $\mathbb{P} : \mathcal{A} \to \mathbb{R}^+$, die im Sinne obiger Definition σ-**additiv** ist und auf 1 normiert.

Diskrete Wahrscheinlichkeitsräume mit $\mathcal{A} = 2^{\Omega}$

Gewichtsfunktionen

Der Grundraum Ω sei nicht leer, endlich oder abzählbar unendlich.

Sei $\mathcal{A}$ die Potenzmenge von Ω, dh die grösstmögliche σ-Algebra.

Wir stellen uns auf den axiomatischen Standpunkt und nehmen an, dass für jedes Ergebnis $\omega \in \Omega$ die Wahrscheinlichkeit gegeben ist, dass gerade dieses Ergebnis eintritt. Diese Wahrscheinlichkeit bezeichnen wir mit

$$p(\omega) := \mathbb{P}(\{\omega\}) \in [0, 1]. \tag{6}$$

Wir setzen ferner voraus, dass diese Wahrscheinlichkeiten $p(\omega)$, die wir auch als Gewichte auffassen können, normiert sind:

$$\sum_{\omega \in \Omega} p(\omega) = 1. \tag{7}$$

Gewichtsfunktionen

Das Wahrscheinlichkeitsmass auf der Potenzmenge $\mathcal{A}$ von Ω wird dann festgelegt durch

$$\mathbb{P}(A) = \sum_{\omega \in A} p(\omega) \ , \ A \in \mathcal{A}. \tag{8}$$

Dann gilt offensichtlich:

$$\mathbb{P}(\Omega) = 1, \tag{9}$$

und für paarweise disjunkte Ereignisse $A_1, A_2, \ldots$ (d.h. $A_i \cap A_j = \emptyset$ für $i \neq j$) ist

$$\mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i), \tag{10}$$

weil

$$\sum_{\omega \in \cup_i A_i} p(\omega) = \sum_i \sum_{\omega \in A_i} p(\omega).$$

Gewichtsfunktionen

Auf einer abzählbaren Menge Ω ist jede Abbildung $\mathbb{P} : 2^{\Omega} \to [0, 1]$, die obige Eigenschaften erfüllt, von der Form

$$\mathbb{P}(A) = \sum_{\omega \in A} p(\omega) \ , \ A \in \mathcal{A}, \tag{11}$$

und zwar mit

$$p(\omega) \equiv \mathbb{P}(\{\omega\}) . \tag{12}$$

Weitere Rechenregeln

Man sieht sofort, dass die folgenden Regeln gelten

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A) \tag{13}$$

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B), \tag{14}$$

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \le i_1 < \cdots < i_k \le n} \mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}) \tag{15}$$

$$A \subseteq B \Rightarrow \mathbb{P}(A) \le \mathbb{P}(B). \tag{16}$$

Beispiel

Wir betrachten das "Experiment" Anzahl Anrufe während einer festen Zeit bei einer Telefonzentrale. Das bedeutet, dass $\Omega = \{0, 1, 2, \dots\}$. Folgende Gewichte sind sinnvoll bei unabhängigem Auftreten von Anrufen

$$p(\omega) = e^{-\lambda} \frac{\lambda^{\omega}}{\omega!} \qquad (\omega \in \Omega = \{0, 1, 2, \dots\}), \tag{17}$$

wobei $\lambda > 0$ ein Parameter ist. Dieses Modell heisst die Poisson-Verteilung. Wenn wir das Ereignis $A =$ "mindestens 1 Anruf" $= \{1, 2, \dots\}$, betrachten, dann ist

$$\mathbb{P}(A) = 1 - \mathbb{P}(A^c) = 1 - \mathbb{P}(\{0\}) = 1 - e^{-\lambda}.$$

Der Erwartungswert einer Zufallsvariablen

Eine reellwertige Funktion auf Ω heisst auch **Zufallsvariable**. Je nachdem, welches Elementarereignis ω realisiert wird, ändert sich auch der realisierte Wert $X(\omega)$. Für eine Zufallsvariable

$$X : \Omega \to \mathbb{R}^1 \tag{18}$$

ist der Wertebereich $X(\Omega)$ auch wieder abzählbar, und durch die Gewichtung

$$x \in X(\Omega) \to \mathbb{P}(X = x) \equiv \mathbb{P}(\{\omega \in \Omega | X(\omega) = x\}) \tag{19}$$

ist ein Wahrscheinlichkeitsmass auf $X(\Omega)$ gegeben, die sogenannte **Verteilung der Zufallsvariablen** X. Manchmal ist es bequem, auch die Werte $\pm\infty$ für X zuzulassen.

Der Erwartungswert einer Zufallsvariablen

Jeder Zufallsvariablen X ordnen wir den **Erwartungswert**

$$\mathbb{E}(X) = \sum_{\omega \in \Omega} X(\omega) p(\omega) \tag{20}$$

zu, wobei wir sicherstellen müssen, dass die rechte Seite sinnvoll ist. Dies ist z.B. der Fall, wenn $X \geq 0$ (den Wert $+\infty$ lassen wir durchaus zu). Wenn X positive und negative Werte annimmt, benutzen wir die Zerlegung $X = X^+ - X^-$ von X in den Positivteil $X^+ = \max(X, 0)$ und den Negativteil $X^- = (-X)^+$ und setzen

$$\mathbb{E}(X) = \mathbb{E}(X^+) - \mathbb{E}(X^-) = \sum_{X(\omega)>0} X(\omega) p(\omega) - \sum_{X(\omega)<0} (-X(\omega)) p(\omega), \tag{21}$$

sofern nicht beide Summen rechts unendlich sind, d.h. $\min(\mathbb{E}(X^+), \mathbb{E}(X^-)) < \infty$.

Der Erwartungswert einer Zufallsvariablen

Der Erwartungswert lässt sich auch *mit Hilfe der Verteilung* von X ausdrücken:

$$\begin{aligned} \mathbb{E}(X) &= \sum_{x \in X(\Omega)} \sum_{\omega : X(\omega) = x} X(\omega) p(\omega) \\ &= \sum_{x \in X(\Omega)} x \cdot \mathbb{P}(X = x). \end{aligned} \tag{22}$$

Beispiel

Sei X die "Anzahl der Anrufe", also $X(\omega) = \omega$. Dann ist

$$\mathbb{E}(X) = \sum_{k=0}^{\infty} k\mathbb{P}(X = k) = \sum_{k=0}^{\infty} ke^{-\lambda} \quad \frac{\lambda^k}{k!} = \lambda, \tag{23}$$

d.h. der Parameter λ ist gerade die *erwartete* Anzahl Anrufe.

Beispiel

In einem einfachen Versicherungskontrakt ist die Leistung der Versicherung

$$X = \begin{cases} c & \text{falls das Ereignis } A \text{ eintritt,} \\ 0 & \text{sonst,} \end{cases}$$

also zufallsabhängig. Die (deterministische!) Gegenleistung des Versicherungsnehmers ist die Prämie, und ein erster Ansatz für eine faire Prämie ist gerade der Erwartungswert

$$\mathbb{E}(X) = c \cdot \mathbb{P}(A) + 0 \cdot \mathbb{P}(A^c) = c \cdot \mathbb{P}(A).$$

Linearität des Erwartungswertes

Aus der Definition ergibt sich sofort die **Linearität des Erwartungswertes:**

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y) \quad (a, b \in \mathbb{R}^1) \tag{24}$$

(sofern beide Seiten sinnvoll sind).

Zur Berechnung des Erwartungswerts ist folgendes Lemma oft nützlich

Lemma

Wenn X nur die Werte $0, 1, 2, \ldots$ annimmt, so ist

$$\mathbb{E}(X) = \sum_{n=0}^{\infty} \mathbb{P}(X > n). \tag{25}$$

Beweis.

$$\begin{aligned}\sum_{n=0}^{\infty} \mathbb{P}(X > n) &= \sum_{n=0}^{\infty} \sum_{k=n+1}^{\infty} \mathbb{P}(X = k) = \sum_{k=1}^{\infty} \sum_{n=0}^{k-1} \mathbb{P}(X = k) \\ &= \sum_{k=1}^{\infty} k\mathbb{P}(X = k) = \mathbb{E}(X).\end{aligned}$$

□

Laplace-Modelle

Sei Ω *endlich*. In vielen Situationen ist es sinnvoll, den **Laplace-Ansatz** $p(\omega) = c$ zu machen (Indifferenzprinzip, Prinzip vom unzureichenden Grunde, Symmetrie eines Würfels etc.). Es folgt

$$p(\omega) = \frac{1}{|\Omega|}, \tag{26}$$

und es ergibt sich

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{\text{Anzahl der günstigen Fälle}}{\text{Anzahl der möglichen Fälle}} \qquad (A \subseteq \Omega) \tag{27}$$

für jedes Ereignis $A \subseteq \Omega$. Man nennt $\mathbb{P}$ auch die **Gleichverteilung** auf Ω.

Garderobenproblem (Montmort, 1708)

n Mäntel werden zufällig an n Personen verteilt. Was ist die Wahrscheinlichkeit, dass keine Person ihren eigenen Mantel bekommt? Als Modell wählen wir hier die Menge aller Permutationen Ω von $\{1, \ldots, n\}$ und die Gleichverteilung $\mathbb{P}$ auf Ω. Für $i = 1, \ldots, n$ sei $A_i = \{\omega \in \Omega | \omega(i) = i\}$ das Ereignis, dass die i-te Person ihren eigenen Mantel bekommt. Dann interessieren wir uns also für das Ereignis $A = (\cup_{i=1}^{n} A_i)^c$.

Garderobenproblem (Montmort, 1708)

Zunächst berechnen wir

$$\begin{aligned}\mathbb{P}(A^c) &= \mathbb{P}(\cup_{i=1}^n A_i) = \sum_{k=1}^n (-1)^{k+1} \sum_{1\le i_1<\cdots<i_k\le n} \underbrace{\mathbb{P}(A_{i_1}\cap\cdots\cap A_{i_k})}_{=\frac{(n-k)!}{n!}} \\ &= \sum_{k=1}^n (-1)^{k+1}\binom{n}{k}\frac{(n-k)!}{n!} = -\sum_{k=1}^n \frac{(-1)^k}{k!}. \qquad (28)\end{aligned}$$

Daraus folgt

$$\mathbb{P}(A) = 1 + \sum_{k=1}^n \frac{(-1)^k}{k!} \longrightarrow e^{-1} \quad (n\to\infty). \qquad (29)$$

Meinungsumfragen, Zufallsauswahl, Urnenexperimente

In einer Urne befinden sich N durchnummerierte Kugeln, K rote und $N - K$ weisse. Es werde eine Stichprobe von n Kugeln mit, bzw. ohne Zurücklegen gezogen. Wenn ω_i die Nummer der beim i-ten Mal gezogenen Kugel bezeichnet, dann ist der Grundraum

$$\begin{aligned} &\text{Mit Zurücklegen:} && \Omega_1 = && \{(\omega_1, \ldots, \omega_n) | 1 \leq \omega_i \leq N\}, \\ &\text{Ohne Zurücklegen:} && \Omega_2 = && \{(\omega_1, \ldots, \omega_n) | 1 \leq \omega_i \leq N, \omega_i \neq \omega_j\}. \end{aligned}$$

Meinungsumfragen, Zufallsauswahl, Urnenexperimente

Wir nehmen als $\mathbb{P}_i$ die Gleichverteilung auf Ω_i $(i = 1, 2)$ und berechnen die Verteilung der Zufallsvariablen $X =$ Anzahl roter Kugeln in der Stichprobe. Sei $A_i = \{\omega \in \Omega_i | 1 \leq \omega_j \leq K$ für genau k Indizes $j\}$. Dann gilt $\mathbb{P}_i(X = k) = |A_i|/|\Omega_i|$.

Meinungsumfragen, Zufallsauswahl, Urnenexperimente

Mit Zurücklegen ist $|\Omega_1| = N^n$ und $|A_1| = K^k(N-K)^{n-k}\binom{n}{k}$, also folgt

$$\mathbb{P}_1(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Dies ist die sogenannte **Binomialverteilung** mit Parameter $p = K/N =$ Anteil roter Kugeln.

Meinungsumfragen, Zufallsauswahl, Urnenexperimente

Ohne Zurücklegen ist

$$\begin{aligned} |\Omega_2| &= N(N-1)\dots(N-n+1) = \binom{N}{n} n!, \\ |A_2| &= K\dots(K-k+1)(N-K)\dots(N-K-(n-k)+1)\binom{n}{k} \\ &= \binom{K}{k}\binom{N-K}{n-k} n!. \end{aligned}$$

Daraus folgt

$$\mathbb{P}_2\big(X=k\big) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \quad (k = 0, 1, \dots \min(n, K)).$$

Dies ist die sogenannte **hypergeometrische Verteilung**.

Meinungsumfragen, Zufallsauswahl, Urnenexperimente

Man sieht leicht, dass für $N \to \infty,\ K \to \infty,\ K/N \to p$ die hypergeometrische Verteilung gegen die Binomialverteilung konvergiert, dh. die von N abhängigen Wahrscheinlichkeiten $\mathbb{P}_2(X = k)$ konvergieren gegen die Werte $\mathbb{P}_1\big(X = k\big)$.

Meinungsumfragen, Zufallsauswahl, Urnenexperimente

Als Grundraum haben wir hier die Menge aller Ziehungen mit Berücksichtigung der Reihenfolge genommen. Wenn man an der Reihenfolge, in der die Kugeln gezogen werden, nicht interessiert ist, kann man auch die Grundräume

$$\Omega_3 = \{(\omega_1, \ldots, \omega_n) | 1 \leq \omega_1 \leq \omega_2 \leq \ldots \leq \omega_n \leq N\},$$

bzw. beim Ziehen ohne Zurücklegen

$$\Omega_4 = \{(\omega_1, \ldots, \omega_n) | 1 \leq \omega_1 < \omega_2 < \ldots < \omega_n \leq N\},$$

betrachten.

Meinungsumfragen, Zufallsauswahl, Urnenexperimente

Ob man die Gleichverteilung auf Ω_2 oder Ω_4 wählt, spielt keine Rolle. Man erhält die gleichen Wahrscheinlichkeiten für Ereignisse, die sich nicht auf die Reihenfolge beziehen, denn bei der Reduktion von Ω_2 auf Ω_4 werden jeweils $n!$ Elementarereignisse zu einem neuen Elementarereignis zusammengefasst. Beim Ziehen mit Zurücklegen ist das nicht der Fall: Dem Element $(1, 1, \ldots, 1) \in \Omega_3$ entspricht nur ein Element in Ω_1, dem Element $(1, 2, \ldots, n) \in \Omega_3$ entsprechen aber $n!$ verschiedene Elemente in Ω_1.

Diskrete Wahrscheinlichkeitsräume mit allgemeinem $\mathcal{A}$

Struktur von σ-Algebren auf höchstens abzählbaren Mengen

Der Grundraum Ω sei nicht leer, endlich oder abzählbar unendlich.

Sei $\mathcal{A}$ eine allgemeine σ-Algebra.

Eine nichtleere Menge $A \in \mathcal{A}$ heisst atomare Menge von $\mathcal{A}$ falls für alle $B \in \mathcal{A}$ gilt: falls $B \subset A$, dann $B = \emptyset$ oder $B = A$. In anderen Worten A ist minimal nicht leer bezueglich der Inklusion in $\mathcal{A}$.

Struktursatz für endliche Mengen

Sei Ω nicht leer und endlich sowie $\mathcal{A}$ eine σ-Algebra, dann ist die Menge der Atome $\text{Atom}(\mathcal{A})$ nicht leer und jedes Element von $\mathcal{A}$ lässt sich eindeutig als Vereinigung der in ihr enthaltenen Atome schreiben.

Der Beweis erfolgt aufgrund der Beobachtung dass jedes Ereignis $A \in \mathcal{A}$ zumindest ein Atom enthält, nämlich ein Ereignis mit minimaler positiver Kardinalität, und dass weiters zwei verschiedene Atome einen leeren Durchschnitt haben müssen.

Rechenregeln

Man sieht sofort, dass die selben Rechenregeln fuer Ereignisse $A, A_i \in \mathcal{A}$ gelten:

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A) \tag{30}$$

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B), \tag{31}$$

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \le i_1 < \cdots < i_k \le n} \mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}) \tag{32}$$

$$A \subseteq B \Rightarrow \mathbb{P}(A) \le \mathbb{P}(B). \tag{33}$$

Der Erwartungswert einer Zufallsvariablen

Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein allgemeiner Wahrscheinlichkeitsraum mit höchstens abzählbarem Ω.

Eine reellwertige Funktion auf Ω heisst auch $\mathcal{A}$(-messbare) **Zufallsvariable** falls $X^{-1}(B) \in \mathcal{A}$ für alle abgeschlossenen Mengen $B \subset \mathbb{R}$. Auf Atomen $A \in \text{Atom}(\mathcal{A})$ ist eine $\mathcal{A}$ messbare Zufallsvariable konstant, der Wert wird mit $X(A)$ bezeichnet.

Für eine Zufallsvariable

$$X : \Omega \to \mathbb{R}^1 \tag{34}$$

ist der Wertebereich $X(\Omega)$ auch wieder abzählbar, und durch die Gewichtsfunktion

$$x \in X(\Omega) \to \mathbb{P}(X = x) \equiv \mathbb{P}(\{\omega \in \Omega | X(\omega) = x\}) \tag{35}$$

ist die **Verteilung der Zufallsvariablen** X gegeben, ein Wahrscheinlichkeitsmass auf $X(\Omega)$ bezüglich der grösstmöglichen σ-Algebra.

Der Erwartungswert einer Zufallsvariablen

Jeder Zufallsvariablen X ordnen wir den **Erwartungswert**

$$\begin{aligned}\mathbb{E}(X) &= \sum_{x \in X(\Omega)} \sum_{\omega : X(\omega) = x} X(\omega) p(\omega) \\ &= \sum_{x \in X(\Omega)} x \cdot \mathbb{P}(X = x) . \end{aligned} \tag{36}$$

Im Falle von endlichen Wahrscheinlichkeitsräumen lässt sich das auch mit Atomen ausdrücken

$$\mathbb{E}(X) = \sum_{A \in \mathrm{Atom}(\mathcal{A})} X(A) \mathbb{P}(A) . \tag{37}$$

Wenn X positive und negative Werte annimmt, benutzen wir die Zerlegung $X = X^+ - X^-$ von X in den Positivteil $X^+ = \max(X, 0)$ und den Negativteil $X^- = (-X)^+$ und setzen

$$\mathbb{E}(X) = \mathbb{E}\left(X^+\right) - \mathbb{E}\left(X^-\right) \tag{38}$$

sofern nicht beide Summen rechts unendlich sind, d.h.

Linearität des Erwartungswertes

Aus der Definition ergibt sich sofort die **Linearität des Erwartungswertes:**

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y) \quad (a, b \in \mathbb{R}^1) \tag{39}$$

(sofern beide Seiten sinnvoll sind).

Zur Berechnung des Erwartungswerts ist folgendes Lemma oft nützlich

Lemma

Wenn X nur die Werte $0, 1, 2, \ldots$ annimmt, so ist

$$\mathbb{E}(X) = \sum_{n=0}^{\infty} \mathbb{P}(X > n). \tag{40}$$

Beweis.

$$\begin{aligned}\sum_{n=0}^{\infty} \mathbb{P}(X > n) &= \sum_{n=0}^{\infty} \sum_{k=n+1}^{\infty} \mathbb{P}(X = k) = \sum_{k=1}^{\infty} \sum_{n=0}^{k-1} \mathbb{P}(X = k) \\ &= \sum_{k=1}^{\infty} k\mathbb{P}(X = k) = \mathbb{E}(X).\end{aligned}$$

Bedingte Wahrscheinlichkeiten

Die bedingte Wahrscheinlichkeit

Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein (diskreter) Wahrscheinlichkeitsraum, und sei $B \in \mathcal{A}$ ein Ereignis mit $\mathbb{P}(B) > 0$. Es gibt oft Fälle, wo wir zwar nicht das genaue Ergebnis des Versuchs erfahren, aber wenigstens, dass B eingetreten ist.

Dann wird häufig eine Modifikation der Wahrscheinlichkeiten gemäss

Definition

Die **bedingte Wahrscheinlichkeit von A gegeben B** ist

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{41}$$

vorgeschlagen. Wir werden dieses Vorgehen in diesem Abschnitt eingehend begründen.

Ein frequentistische Interpretation

In der frequentistische Interpretation ist ja $\mathbb{P}(C) \approx n_C/n =$ relative Häufigkeit von C. Die bedingte Wahrscheinlichkeit von A gegeben B ist dann analog ungefähr gleich der relativen Häufigkeit von A unter den Versuchen, wo B eingetreten ist. In Formeln

$$\mathbb{P}(A|B) \approx \frac{n_{A\cap B}}{n_B} = \frac{n_{A\cap B}/n}{n_B/n}.$$

Damit sollte obige Definition einleuchten.

Die bedingte Wahrscheinlichkeit als neues Wahrscheinlichkeitsmass

$\mathbb{P}(\cdot|B)$ ist eine neue Wahrscheinlichkeitsverteilung auf $(\Omega, \mathcal{A})$. Falls $\mathcal{A} = 2^{\Omega}$, dann ist die Gewichtsfunktion

$$p_B(\omega) \propto p(\omega) \quad \text{for all } \omega \in B, \quad p_B(\omega) = 0 \quad \text{for all } \omega \notin B.$$

Insbesondere, wenn $\mathbb{P}$ die Gleichverteilung auf Ω ist, dann ist $\mathbb{P}(\cdot|B)$ die Gleichverteilung auf B.

Beispiel

Wurf zweier Würfel. Man wählt $\Omega = \{(i,j) \mid 1 \leq i,j \leq 6\}$ und $\mathbb{P}$ die Gleichverteilung. Sei B_k das Ereignis "Augensumme $=k$" und $A_i =$ "Erster Würfel zeigt i". Dann $\mathbb{P}(A_i|B_7) = \frac{1}{36} : \frac{1}{6} = \frac{1}{6}$ für $i = 1, 2, \ldots, 6$; d.h. die Information, dass B_7 eingetreten ist, nützt nichts für die Prognose von A_i. Aber $\mathbb{P}(A_i|B_{11}) = \frac{1}{2}$ für $i = 5, 6$ und $\mathbb{P}(A_i|B_{11}) = 0$ für $i \leq 4$.

Satz von der totalen Wahrscheinlichkeit

Sei $(B_i)_{i\in I}$ eine disjunkte Zerlegung von Ω durch Ereignisse B_i (d.h. $\Omega = \bigcup_{i\in I} B_i, \quad B_i \cap B_j = \emptyset$ für $i \neq j$) . Dann gilt für beliebiges $A \in \mathcal{A}$:

$$\mathbb{P}(A) = \sum_{i:\mathbb{P}(B_i)>0} \mathbb{P}(A|B_i)\,\mathbb{P}(B_i)\,. \tag{42}$$

Beweis

Weil $A = \bigcup_{i \in I} (A \cap B_i)$, gilt

$$\mathbb{P}(A) = \sum_i \underbrace{\mathbb{P}(A \cap B_i)}_{= \mathbb{P}(A|B_i)\mathbb{P}(B_i)} .$$

Beispiel

Man wählt zuerst eine von zwei Urnen zufällig und zieht dann aus der gewählten Urne zufällig eine Kugel. Urne 1 enthält k weisse und ℓ rote Kugeln, Urne 2 sind $n-k$ weisse und $n-\ell$ rote. Dann

$$\mathbb{P}(\text{ Kugel weiss }) = \frac{k}{k+\ell} \cdot \frac{1}{2} + \frac{n-k}{2n-(k+\ell)} \cdot \frac{1}{2}$$

Für welche Werte von k und ℓ wird dies – bei festem n – maximal? Nach einiger Rechnung (zuerst $k+\ell = m$ festhalten) erhält man $k=1, \ell=0$, d.h. man verteilt also die Risiken am besten sehr ungleich. In diesem Fall ist

$$\mathbb{P}(\text{ Kugel weiss }) = \frac{1}{2}\ \frac{3n-2}{2n-1}\ \to\ \frac{3}{4}.$$

Ein weiteres nützliches Resultat

Für beliebige Ereignisse $A_1, \ldots, A_n$ gilt

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\,\mathbb{P}(A_2|A_1)\ldots\mathbb{P}(A_n|A_1 \cap \cdots \cap A_{n-1}), \quad (43)$$

sofern die linke Seite > 0 ist.

Beispiel

Wie gross ist die Wahrscheinlichkeit, dass n Personen alle an verschiedenen Tagen Geburtstag haben? Sei $G_i =$ Geburtstag der i-ten Person und $A_i = \{G_i \neq G_j \text{ für alle } j < i\}$. Dann

$$\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n) = 1 \cdot \frac{364}{365} \cdot \frac{363}{365} \cdot \cdots \cdot \frac{365 - n + 1}{365}.$$

Man erhält die numerischen Werte $= 0.49$ für $n = 23$, 0.11 für $n = 40$ und 0.03 für $n = 50$. Die Wahrscheinlichkeiten sind also viel kleiner als man intuitiv vermuten würde.

Satz von Bayes

Aus der Definition der bedingten Wahrscheinlichkeit folgt sofort die Bayessche Formel, welche den Zusammenhang zwischen $\mathbb{P}(A|B)$ und $\mathbb{P}(B|A)$ beschreibt:

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

Die bedingte Wahrscheinlichkeit ist also nicht symmetrisch. Die Grössen auf der rechten Seite hängen aber im allgemeinen voneinander ab. Mit dem Satz von der totalen Wahrscheinlichkeit folgt die Version

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\,\mathbb{P}(B)}{\mathbb{P}(A|B)\,\mathbb{P}(B) + \mathbb{P}(A|B^c)\,(1-\mathbb{P}(B))}. \tag{44}$$

Die Wahrscheinlichkeiten $\mathbb{P}(B)$, $\mathbb{P}(A|B)$ und $\mathbb{P}(A|B^c)$ auf der rechten Seite können beliebige Werte annehmen.

Beispiel

Von 145 Ihres Alters habe einer die Krankheit K (Pävalenz). Für das Ereignis $B =$ "Sie haben K" gilt also a priori $\mathbb{P}(B) = \frac{1}{145}$. Sie machen nun einen Test, und es tritt das Ereignis $A=$ "Ergebnis positiv" ein. Nun ist aber kein Test völlig fehlerfrei. Nehmen wir zum Beispiel an:

$$\mathbb{P}(A|B) = 0.96 \text{ (Sensitivität)}, \quad \mathbb{P}(A^c|B^c) = 0.94 \text{ (Spezifität)}.$$

Dann folgt mit obiger Formel

$$\mathbb{P}(B|A) = \frac{\frac{96}{100} \cdot \frac{1}{145}}{\frac{96}{100} \cdot \frac{1}{145} + \frac{6}{100} \cdot \frac{144}{145}} = \frac{1}{10}$$

(also noch kein Grund zur Panik, oder doch!).

Beispiel

Die Sensitivität und Spezifität von Test sind Ergebnisse von Studien, wobei hier zwischen realer und Laborsituation unterschieden werden muss.

Im konkreten Fall des neuen Coronavirus sind folgende Zahlen für den bekannten PCR Test in einer Studie angegeben: Sensitivität von 0.8 und Spezifität von 0.99. Für den Antigentest gibt es hier deutlich geringere Resultate abhängig vom klinischen Verlauf: Sensitivität von 0.41 bis 0.8 und eine Spezifität von 0.98.

Satz von Bayes

Betrachten wir an Stelle von (B, B^c) eine beliebige disjunkte Zerlegung von Ω, so erhalten wir den allgemeinen Satz von Bayes:

Ist $(B_i)_{i \in I}$ eine Zerlegung von Ω in disjunkte Ereignisse und $\mathbb{P}(A) \neq 0$, so ist

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(A|B_i) \cdot \mathbb{P}(B_i)}{\sum_j \mathbb{P}(A|B_j) \cdot \mathbb{P}(B_j)} \,. \tag{45}$$

Beweis und Bemerkung

Die Aussage folgt mit

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(A|B_i) \cdot \mathbb{P}(B_i)}{\mathbb{P}(A)}$$

und dem Satz von der totalen Wahrscheinlichkeit.

Dieses Resultat lässt sich wie folgt interpretieren: Wir haben verschiedene Hypothesen B_i mit a priori Wahrscheinlichkeiten $\mathbb{P}(B_i)$; unter der Hypothese B_i hat A die Wahrscheinlichkeit $\mathbb{P}(A|B_i)$. Wenn nun A eintritt, kann man aus den Grössen $\mathbb{P}(B_i)$ und $\mathbb{P}(A|B_i)$ die a posteriori Wahrscheinlichkeiten $\mathbb{P}(B_i|A)$ der verschiedenen Hypothesen berechnen.

Filtern

Wir betrachten einen **Nachrichtenkanal** mit Eingangsalphabet I und Ausgangsalphabet J. Sei A_j das Ereignis "Signal j wird empfangen" $(j \in J)$ und B_i das Ereignis "Signal i wird gesendet" $(i \in I)$.

Die Übertragung ist jedoch nicht fehlerfrei (Rauschen!). Die *Übertragungseigenschaften des Kanals* sind durch die bedingten Wahrscheinlichkeiten $\mathbb{P}(A_j|B_i) \quad (i \in I, j \in J)$ beschrieben, die *Struktur der Nachrichtenquelle* durch die Wahrscheinlichkeiten $\mathbb{P}(B_i) \quad (i \in I)$. Daraus kann der Empfänger die bedingten Wahrscheinlichkeiten $\mathbb{P}(B_i|A_j)$ gemäss dem Satz von Bayes berechnen.

Filtern

Gesucht ist nun eine **Dekodierung** $\varphi : J \to I$, für die das Ereignis

$$C_\varphi = \text{“ richtig dekodiert”} = \bigcup_j (A_j \cap B_{\varphi(j)})$$

maximale Wahrscheinlichkeit hat. Es gilt

$$\mathbb{P}(C_\varphi) = \sum_j \mathbb{P}(A_j) \cdot \mathbb{P}\left(B_{\varphi(j)} \middle| A_j\right), \tag{46}$$

und, offensichtlich können wir jeden Summanden einzeln maximieren. Wir erhalten also die folgende Lösung: Wähle $\varphi(j)$ für jedes $j \in J$ so, dass

$$\mathbb{P}\left(B_{\varphi(j)} \middle| A_j\right) = \max_i \mathbb{P}(B_i | A_j).$$

Filtern: Illustration im binären Fall $I = J = \{0, 1\}$

Sei

$$p_1 \equiv \mathbb{P}(A_1|B_1), \quad p_0 \equiv \mathbb{P}(A_0|B_0), \quad \alpha \equiv \mathbb{P}(B_1)$$

Es gibt vier mögliche Dekodierungen gemäss folgender Tabelle:

Dekodierung φ	C_φ	$\mathbb{P}(C_\varphi)$
$\varphi_1 \equiv 1$	B_1	α
$\varphi_2 \equiv 0$	B_0	$1-\alpha$
$\varphi_3(1) = 1, \quad \varphi_3(0) = 0$	$(A_1 \cap B_1) \cup (A_0 \cap B_0)$	$\alpha p_1 + (1-\alpha)p_0$
$\varphi_4(1) = 0, \quad \varphi_4(0) = 1$	$(A_1 \cap B_0) \cup (A_0 \cap B_1)$	$\alpha(1-p_1) + (1-\alpha)(1-p_0)$

Filtern: Illustration im binären Fall $I = J = \{0, 1\}$

Wenn sowohl p_0 als auch p_1 grösser als 0.5 sind, dann ist die optimale Dekodierung gegeben durch

$$\varphi = \begin{cases} \varphi_2 & \text{falls } 0 \leq \alpha \leq \frac{1-p_0}{1-p_0+p_1} \\ \varphi_1 & \text{falls } \frac{p_0}{p_0+1-p_1} \leq \alpha \leq 1 \\ \varphi_3 & \text{sonst} \end{cases}$$

Bedingter Erwartungswert auf ein Ereignis B

Sei $(\Omega, \mathcal{A}, P)$ ein diskreter Wahrscheinlichkeitsraum mit $\mathcal{A} = 2^{\Omega}$ (das geht natürlich allgemein aber wir nehmen es hier der Einfachheit halber an). Für ein Ereignis $B \in \mathcal{A}$ mit $\mathbb{P}(B) > 0$, gibt die bedingte Wahrscheinlichkeit $\mathbb{P}(A|B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$ an, wie wahrscheinlich das Ereignis A ist, wenn B eingetreten ist. Entsprechend gibt der bedingte Erwartungswert

$$\mathbb{E}(X|B) = \frac{\mathbb{E}(1_B X)}{\mathbb{P}(B)} = \sum_{x \in X(\Omega)} x\mathbb{P}(X = x|B) = \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\{\omega\}|B)$$

an, welchen Wert man für die Zufallsvariable X im Mittel erwartet, wenn man Information über das Eintreten von B erhalten hat.

Bedingter Erwartungswert als Zufallsvariable

Dieser elementare Begriff von bedingten Erwartungswerten ist jedoch oft nicht ausreichend. Für ein allgemeineres Konzept betrachten wir eine Partition $\mathcal{B} = (B_i)_{i \in I}$ von Ω, d.h. eine Zerlegung von Ω in disjunkte, nicht leere Teilmengen, wobei I eine abzählbare Indexmenge bezeichnet. Dann definieren wir die **Zufallsvariable**

$$\mathbb{E}(X|\mathcal{B})(\omega) = \sum_{i \in I, \mathbb{P}(B_i) > 0} \mathbb{E}(X|B_i)\ 1_{B_i}(\omega).$$

Ihr Wert ist festgelegt, sobald man weiss, welches B_i realisiert wurde (man braucht also den genauen Wert von ω nicht), und sie gibt an, welchen Wert man dann für X erwartet. Diese Zufallsvariable wird daher als bedingte Erwartung von X gegeben $\mathcal{B}$ bezeichnet.

Bedingter Erwartungswert als bester Schätzer

Sei X eine Zufallsvariable auf $(\Omega, \mathcal{A}, \mathbb{P})$ mit $\mathbb{E}\left(X^2\right) < \infty$ und sei $\mathcal{B} = (B_i)_{i \in I}$ eine Partition von Ω. Dann wird

$$\mathbb{E}\left(\left(X - \sum_{i \in I, \mathbb{P}(B_i) > 0} c_i 1_{B_i}\right)^2\right)$$

minimal für $c_i = \mathbb{E}\left(X 1_{B_i}\right) / \mathbb{P}\left(B_i\right)$, d.h. die bedingte Erwartung ergibt die beste Prognose von X auf Grund der Partition $\mathcal{B}$ ("beste" im Sinne des mittleren quadratischen Fehlers).

Beweis

Wir schreiben kurz $\sum_i$ für $\sum_{i\in I,\mathbb{P}(B_i)>0}$. Dann gilt für beliebige Wahl der c_i:

$$\begin{aligned}\mathbb{E}\left(X\sum_i c_i 1_{B_i}\right) &= \sum_i c_i \mathbb{E}\left(X1_{B_i}\right) = \mathbb{E}\left(\sum_i c_i \frac{\mathbb{E}\left(X1_{B_i}\right)}{\mathbb{P}\left(B_i\right)} 1_{B_i}\right) \\ &= \mathbb{E}\left(\sum_{i,j} c_i \frac{\mathbb{E}\left(X1_{B_j}\right)}{\mathbb{P}\left(B_j\right)} 1_{B_i} 1_{B_j}\right) = \mathbb{E}\left(\mathbb{E}\left(X|\mathcal{B}\right)\sum_i c_i 1_{B_i}\right)\end{aligned}$$

weil $1_{B_i}1_{B_j} = 0$ $(i \neq j)$ und $1_{B_i}1_{B_i} = 1_{B_i}$. Also folgt

$$\mathbb{E}\left(\left(X - \mathbb{E}\left(X|\mathcal{B}\right)\right)\sum_i c_i 1_{B_i}\right) = 0\,. \tag{47}$$

Beweis

Insbesondere gilt auch $\mathbb{E}\left((X - \mathbb{E}(X|\mathcal{B}))(\mathbb{E}(X|\mathcal{B}) - \sum_i c_i 1_{B_i})\right) = 0$, und damit

$$\mathbb{E}\left((X - \sum_i c_i 1_{B_i})^2\right) = \mathbb{E}\left((X - \mathbb{E}(X|\mathcal{B}))^2\right) + \mathbb{E}\left((\mathbb{E}(X|\mathcal{B}) - \sum_i c_i 1_{B_i})^2\right).$$

Der zweite Term rechts ist immer positiv ausser wenn $c_i = \mathbb{E}(X 1_{B_i}) / \mathbb{P}(B_i)$.

Bemerkung

Die Formel (47) zeigt, dass $\mathbb{E}(X|\mathcal{B})$ die orthogonale Projektion von X auf den Unterraum aller Funktionen der Form $\sum_i c_i 1_{B_i}$ ist bezüglich des Skalarprodukts

$$\langle X, Y \rangle := \sum_{\omega} X(\omega) Y(\omega) p(\omega)$$

auf der Menge aller Zufallsvariablen.

Unabhängigkeit

Unabhängigkeit

Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein (diskreter) Wahrscheinlichkeitsraum.

Eine Kollektion von Ereignissen $(A_i; i \in I)$ heisst **(stochastisch) unabhängig** wenn gilt:

$$J \subseteq I \text{ endlich } \Rightarrow \mathbb{P}\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} \mathbb{P}(A_i)\,. \tag{48}$$

Bemerkungen

- Unabhängigkeit ist keine Eigenschaft der Ereignisse per se, sondern eine Eigenschaft in Bezug auf die Wahrscheinlichkeitsverteilung $\mathbb{P}$.
- Für zwei Ereignisse A, B mit positiver Wahrscheinlichkeit gilt:

$$A, B \text{ sind unabhängig} \iff \mathbb{P}(A|B) = \mathbb{P}(A) \iff \mathbb{P}(B|A) = \mathbb{P}(B)\,. \tag{49}$$

Bemerkungen

- Die paarweise Unabhängigkeit impliziert noch nicht (49). Zum Beispiel sind beim Wurf zweier Münzen die Ereignisse

  $$\begin{aligned} A &= \text{“Erster Wurf Kopf”} \\ B &= \text{“Zweiter Wurf Kopf”} \\ C &= \text{“Die Ergebnisse der zwei Würfe sind verschieden”} \end{aligned}$$

  paarweise unabhängig, aber $\mathbb{P}(A \cap B \cap C) = 0 \neq \mathbb{P}(A) \cdot \mathbb{P}(B) \cdot \mathbb{P}(C)$.

- Bei endlich vielen Ereignissen $A_1, \ldots, A_n$ genügt es nicht, die Produktformel in (48) für $J = \{1, \ldots, n\}$ zu fordern. Dies sieht man sofort ein, wenn man $A_1 = \emptyset$ und $A_2 = A_3$ wählt mit $0 < \mathbb{P}(A_2) < 1$.

- Wie bei der bedingten Wahrscheinlichkeiten, *postuliert* man oft die Unabhängigkeit gewisser Ereignisse, um Wahrscheinlichkeiten festzulegen.

Lemma

Lemma

Die Ereignisse $A_i \quad (i \in I)$ seien unabhängig. Sei $B_i = A_i$ oder $= A_i^c$. Dann sind auch die Ereignisse $B_i \quad (i \in I)$ unabhängig.

Beweis

Für disjunkte endliche Mengen $J, K \subseteq I$ ist

$$\mathbb{P}\left(\bigcap_{i\in J} A_i \cap \bigcap_{i\in K} A_i^c\right) = \prod_{i\in J} \mathbb{P}(A_i) \prod_{i\in K} \mathbb{P}(A_i^c) \tag{50}$$

zu zeigen. Wir benützen Induktion nach $k = |K|$. Für $k = 0$ ist (50) richtig, und aus der Gültigkeit für $|K| = k$ folgt

$$\begin{aligned}\mathbb{P}\left(\bigcap_{i\in J} A_i \cap \bigcap_{i\in K} A_i^c \cap A_j^c\right) &= \mathbb{P}\left(\bigcap_{i\in J} A_i \cap \bigcap_{i\in K} A_i^c\right) - \mathbb{P}\left(\bigcap_{i\in J} A_i \cap A_j \cap \bigcap_{i\in K} A_i^c\right) \\ &= \prod_{i\in J} \mathbb{P}(A_i) \prod_{i\in K} \mathbb{P}(A_i^c) \cdot (1 - \mathbb{P}(A_j)),\end{aligned}$$

d.h. (50) gilt auch für $\tilde{K} = K \cup \{j\}$.

Bemerkung

Für endliches I ist die Unabhängigkeit äquivalent zur Gültigkeit von (50) für alle $J \subseteq I$, $K = J^c$.

Unabhängigkeit von Zufallsvariablen

Schliesslich definieren wir noch die Unabhängigkeit von Zufallsvariablen.

Definition

Eine Kollektion von diskreten Zufallsvariablen $(X_i; i \in I)$ heisst **unabhängig**, falls die Ereignisse $(\{X_i = y\}; i \in I, y \in \mathbb{R})$ unabhängig sind für jede Wahl von y aus dem Wertebereich von X_i (sonst haben wir ja die leere Menge).

Erwartungswerte unabhängiger Zufallsvariablen

Lemma

Wenn die diskreten Zufallsvariablen $X_1, \ldots, X_n$ unabhängig sind, dann gilt

$$\mathbb{E}\left(\prod_{i=1}^{n} g_i(X_i)\right) = \prod_{i=1}^{n} \mathbb{E}\left(g_i(X_i)\right)$$

für beliebige Funktionen g_i (sofern die Erwartungswerte existieren).

Beweis.

Es gilt wegen (22), bzw. gemäss der Definition der Unabhängigkeit

$$\begin{aligned}
\mathbb{E}\left(\prod_{i=1}^{n} g_i(X_i)\right) &= \sum_{x_1,\ldots,x_n} \prod_{i=1}^{n} g_i(x_i)\mathbb{P}\left(X_1 = x_1, \ldots, X_n = x_n\right) \\
&= \sum_{x_1,\ldots,x_n} \prod_{i=1}^{n} g_i(x_i)\mathbb{P}\left(X_i = x_i\right) \\
&= \prod_{i=1}^{n} \sum_{x_i} g_i(x_i)\mathbb{P}\left(X_i = x_i\right) = \prod_{i=1}^{n} \mathbb{E}\left(g_i(X_i)\right).
\end{aligned}$$

□

Die Irrfahrt

Definition

Die Irrfahrt (random walk, marche aléatoire) ist ein Modell für die zufällige Bewegung eines Teilchens auf dem eindimensionalen Gitter $\mathbb{Z} = \{\dots, -1, 0, 1, \dots\}$, das in 0 startet und in jeder Periode einen Schritt $+1$ oder -1 macht, jeweils mit Wahrscheinlichkeit $\frac{1}{2}$. Man kann das auch anders interpretieren, z.B. als die Bilanzentwicklung eines Spielers bei einem fairen Glücksspiel, als die Kursentwicklung einer Aktie, usw.

Modell für N Perioden

Sei Ω die Menge aller binären Folgen der Länge N, also $\Omega = \{\omega = (x_1, \ldots, x_N) \mid x_i \in \{+1, -1\}\}$. Ferner betrachten wir die Zufallsvariablen $X_k(\omega) = k$-te Komponente von $\omega = (x_1, \ldots, x_N) \in \Omega$ und $S_n(\omega) = \sum_{k=1}^{n} X_k(\omega)$. X_k ist also der Schritt bzw. Ertrag in der k-ten Periode und S_n die Position bzw. Bilanz nach n Perioden. Wir starten stets im Ursprung, d.h. $S_0(\omega) = 0$.

Modell für N Perioden

Für jedes $\omega \in \Omega$ erhält man eine (diskrete) **Trajektorie** (Pfad, Bilanzentwicklung) $(n, S_n(\omega)), (n = 0, \ldots, N)$. Sei nun $\mathbb{P}$ die **Gleichverteilung auf** Ω, also

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = 2^{-N} \cdot |A| \qquad (A \subseteq \Omega). \tag{51}$$

Definition

Die Folge der Zufallsvariablen $S_n (n = 0, \ldots, N)$ auf dem Wahrscheinlichkeitsraum $(\Omega, 2^{\Omega}, \mathbb{P})$ heisst **Irrfahrt** (mit Start in 0).

Wir nehmen also an, dass die binären Folgen $\omega \in \Omega$ bzw. die entsprechenden Trajektorien alle gleichwahrscheinlich sind.

Erste Schlussfolgerung

Aus (51) folgt

$$\mathbb{P}(X_k = +1) = \frac{2^{N-1}}{2^N} = \frac{1}{2} \quad (k = 1, \ldots, N). \tag{52}$$

Ebenso gilt für beliebige Indizes $1 \leq k_1 < \cdots < k_\ell \leq N$ und für jede Wahl von $x_k = \pm 1$

$$\mathbb{P}(X_{k_1} = x_{k_1}, \ldots, X_{k_\ell} = x_{k_\ell}) = \frac{2^{N-\ell}}{2^N} = 2^{-\ell}. \tag{53}$$

Insbesondere bilden die ersten $n < N$ Schritte einer Irrfahrt eine Irrfahrt mit n Perioden. Aus (52) folgt

$$\mathbb{E}(X_k) = (+1) \cdot \mathbb{P}(X_k = +1) + (-1) \cdot \mathbb{P}(X_k = -1) = 0 \tag{54}$$

und daraus folgt, wegen der Linearität des Erwartungswerts

$$\mathbb{E}(S_n) = \sum_{k=1}^{n} \mathbb{E}(X_k) = 0 \qquad (n = 0, \ldots, N). \tag{55}$$

Verteilung von S_n

Für festes n nimmt die Zufallsvariable S_n Werte $x \in \{-n, -n+2, \ldots, n-2, n\}$ an, und zwar mit den folgenden Wahrscheinlichkeiten:

$$\mathbb{P}(S_n = 2k - n) = \binom{n}{k} 2^{-n} \quad (k = 0, 1, \ldots, n). \tag{56}$$

Für alle anderen x ist $\mathbb{P}(S_n = x) = 0$.

Beweis

Sei U_n die Anzahl der "Schritte nach oben" bis zum Zeitpunkt n, d.h.

$$U_n = \sum_{k=1}^{n} 1_{\{X_k=+1\}}.$$

Dann ist $S_n = U_n - (n - U_n) = 2U_n - n$.

$$|\{U_n = k\}| = \binom{n}{k} 2^{N-n} \Rightarrow \mathbb{P}(U_n = k) = \binom{n}{k} 2^{N-n} 2^{-N} = \binom{n}{k} 2^{-n}.$$

Die Verteilung von S_n ist also eine "linear transformierte Binomialverteilung" mit $p = \frac{1}{2}$.

Asymptotik

Die Formel von Stirling besagt

$$n! \sim (\frac{n}{e})^n \sqrt{2\pi n}$$

($a_n \sim b_n$ heisst $\lim_{n\to\infty} \frac{a_n}{b_n} = 1$). Daraus folgt, dass

$$\mathbb{P}(S_{2n} = 0) = \mathbb{P}(S_{2n-1} = 1) = \binom{2n}{n} 2^{-2n} \sim \frac{1}{\sqrt{\pi n}}. \tag{57}$$

Ferner gilt

$$\binom{n}{k} = \binom{n}{k-1} \frac{n-k+1}{k} \geq \binom{n}{k-1} \iff k \leq \frac{n+1}{2} \tag{58}$$

Also ist für n gerade $\mathbb{P}(S_n = x)$ maximal für $x = 0$, und für n ungerade ist $\mathbb{P}(S_n = x)$ maximal für $x = \pm 1$.

Reflexionsprinzip

Wir untersuchen hier die Verteilung der Zufallsvariable

$$T_a(\omega) = \min\{n > 0 \mid S_n(\omega) = a\} \tag{59}$$

(Erstes Erreichen des Niveaus $a \neq 0$, bzw. für $a = 0$ erste Rückkehr nach 0). Dabei setzen wir hier $\min \emptyset = N + 1$.

Die entscheidende Idee ist das folgende Lemma:

Lemma (Reflexionsprinzip)

Für $a > 0$ und $b \geq -a$ ist

$$\mathbb{P}(T_{-a} \leq n, S_n = b) = \mathbb{P}(S_n = -2a - b). \tag{60}$$

Beweis.

Durch Spiegelung des Pfades für Zeiten $\geq T_{-a}$ am Niveau $-a$ erhält man eine Bijektion zwischen Pfaden mit $T_{-a} \leq n, S_n = b$ und solchen mit $S_n = -2a - b$. □

Folgerungen

Aus dem Reflexionsprinzip ergibt sich insbesondere die Verteilung von T_{-a}:

Satz

Für $a \neq 0$ gilt

$$\mathbb{P}(T_{-a} \leq n) = 2\mathbb{P}(S_n < -a) + \mathbb{P}(S_n = -a) = \mathbb{P}(S_n \notin (-a, a]). \quad (61)$$

Beweis

Mit dem Additionssatz folgt

$$\begin{aligned}
\mathbb{P}(T_{-a} \leq n) &= \sum_{b=-\infty}^{\infty} \mathbb{P}(T_{-a} \leq n, S_n = b) \\
&= \sum_{b=-\infty}^{-a} \mathbb{P}(S_n = b) + \sum_{b=-a+1}^{\infty} \mathbb{P}(S_n = -2a - b) \\
&= \mathbb{P}(S_n \leq -a) + \mathbb{P}(S_n \leq -a - 1)\,.
\end{aligned}$$

Die letzte Gleichung in (61) folgt aus Symmetrie.

Rueckkehr zu $a \neq 0$

Folgerung

Für jedes $a \neq 0$ gilt

$$1.\ \mathbb{P}(T_a > N) \quad \to \quad 0, \quad 2.\ \mathbb{E}(T_a) = \sum_{k=1}^{N+1} k\mathbb{P}(T_a = k) \to \infty.$$

Bemerkung

Anschaulich bedeutet die erste Aussage, dass die Irrfahrt mit Wahrscheinlichkeit 1 jedes Niveau erreicht, und die zweite Aussage, dass man *sehr lange* darauf warten muss. Exakt lässt sich das jedoch erst in einem Modell mit unendlich vielen Perioden formulieren. Dort gilt dann

$$\mathbb{P}(T_a < \infty) = \lim_{N\to\infty} \mathbb{P}(T_a \leq N) = 1, \tag{62}$$

und

$$\mathbb{E}(T_a) = \sum_{k=1}^{\infty} k\mathbb{P}(T_a = k) = +\infty. \tag{63}$$

In einem Modell mit unendlich vielen Perioden ist Ω jedoch überabzählbar, und die Konstruktion von $\mathbb{P}$ erfordert Masstheorie.

Beweis

Sei $a > 0$. Dann gilt

$$\mathbb{P}(T_{-a} > N) = \mathbb{P}(S_N \in (-a, a]) \leq \frac{c}{\sqrt{\pi N}} \longrightarrow 0.$$

Ferner ist

$$\begin{aligned} \sum_{k=1}^{N+1} k\mathbb{P}(T_{-a} = k) &= \sum_{k=0}^{N} \mathbb{P}(T_{-a} > k) = \sum_{k=0}^{N} \mathbb{P}(S_k \in (-a, a]) \\ &\geq \sum_{k=1}^{N} \mathbb{P}(S_k \in \{0, 1\}) \to +\infty. \end{aligned}$$

Für T_a mit $a > 0$ folgt die Behauptung aus Symmetrie.

Rueckkehr zu $a = 0$

Es gilt folgende Aussage für T_0, die erste Rückkehrzeit nach Null:

Satz

$$\mathbb{P}(T_0 > 2n) = \mathbb{P}(S_{2n} = 0). \tag{64}$$

Beweis

Es gibt gleichviele Pfade der Länge $2n$, die stets oberhalb der x-Achse verlaufen, wie es Pfade der Länge $2n-1$ gibt, die nie -1 erreichen. Damit und mit einer Symmetrieüberlegung erhält man

$$\begin{aligned}\mathbb{P}(T_0 > 2n) &= \frac{1}{2}\mathbb{P}(T_{-1} > 2n-1) + \frac{1}{2}\mathbb{P}(T_1 > 2n-1) \\ &= \mathbb{P}(T_{-1} > 2n-1) \\ &= \mathbb{P}(S_{2n-1} \in (-1,1]) \;=\; \mathbb{P}(S_{2n-1} = 1) \;=\; \mathbb{P}(S_{2n} = 0)\,.\end{aligned}$$

Die Aussage $\mathbb{P}(T_0 > 2n) \to 0$ heisst auch "Die Irrfahrt ist rekurrent".

Das Arkussinus-Gesetz für den letzten Besuch in Null

Sei

$$L(\omega) = \max\{0 \leq n \leq 2N \mid S_n(\omega) = 0\}$$

der Zeitpunkt des *letzten Besuches in Null* vor $2N$. Vom Zeitpunkt $L(\omega)$ an gibt also einer der beiden Spieler die Führung nicht mehr ab.

Man könnte meinen, dass $L(\omega)$ – zumindest für grosses N – meistens nahe bei $2N$ sein wird, insbesondere also $\mathbb{P}(L \leq N) \to 0$ für $N \to \infty$.
Stattdessen gilt aber:

Arkusssinusgesetz

Die Verteilung von L ist die sogenannte diskrete Arkussinus-Verteilung:

$$\mathbb{P}(L = 2n) = \mathbb{P}(S_{2n} = 0) \cdot \mathbb{P}(S_{2N-2n} = 0) = 2^{-2N} \binom{2n}{n} \binom{2N-2n}{N-n} \quad (65)$$

Die Verteilung ist symmetrisch um N (insbesondere ist $\mathbb{P}(L \leq N) \approx \frac{1}{2}$), und U-förmig. Man muss also damit rechnen, dass der Gewinner die endgültige Führung entweder recht früh oder recht spät übernimmt, und zwar passiert beides mit derselben Wahrscheinlichkeit!

Beweis

Die Anzahl Pfade bestehend aus $2N$ Schritten mit $L = 2n$ ist das Produkt der Anzahl Pfade bestehend aus $2n$ Schritten mit $S_{2n} = 0$ und der Anzahl Pfade bestehend aus $2N - 2n$ Schritten mit $T_0 > 2N - 2n$. Also ist wegen der Gleichverteilung aller Pfade
$\mathbb{P}(L = 2n) = \mathbb{P}(S_{2n} = 0) \cdot \mathbb{P}(T_0 > 2N - 2n)$. Aufgrund von (64) ist der zweite Faktor gleich $\mathbb{P}(S_{2N-2n} = 0)$.

Warum "Arkussinus"?

$$\mathbb{P}(S_{2k} = 0) \approx \frac{1}{\sqrt{\pi k}} \Rightarrow \mathbb{P}(L = 2k) \approx \frac{1}{\pi\sqrt{k(N-k)}} = \frac{1}{N} f(\frac{k}{N})$$

$$\text{mit } f(x) = \frac{1}{\pi\sqrt{x(1-x)}}.$$

Daraus folgt

$$\mathbb{P}\left(\frac{L}{2N} \leq z\right) \approx \sum_{k:\frac{k}{N} \leq z} \frac{1}{N} f(\frac{k}{N}) \approx \int_0^z f(x)dx = \frac{2}{\pi} \arcsin \sqrt{z}. \qquad (66)$$

Spielsysteme

Für festes n

$$\mathbb{E}(S_n) = 0,$$

d.h. der "Ertrag" S_n nach n Perioden ist im Schnitt gleich 0.

Kann man mehr erreichen, wenn man die Bilanzentwicklung $S_n(\omega) \quad (n = 0, \ldots, N)$ in einem zufälligen, aber günstigen Moment $T(\omega)$ stoppt? Kann man also durch geschickte Wahl einer Strategie zum vorzeitigen Abbruch des Spiels im Schnitt einen positiven Gewinn erzielen?

Die Entscheidung, im Zeitpunkt n zu stoppen oder nicht, darf sich natürlich nur auf die Entwicklung der Trajektorie bis zu diesem Zeitpunkt stützen. Es ist also keine "Insider"–Information über die zukünftige Entwicklung zugelassen! Um das mathematisch zu präzisieren, definieren wir zunächst die Klasse derjenigen Ereignisse, die nur vom Verhalten der Trajektorie bis zum Zeitpunkt n abhängen:

Beobachtbare Ereignisse bis n

Definition

Ein Ereignis $A \subseteq \Omega$ heisst **beobachtbar bis zum Zeitpunkt** n, wenn es von der Form $\{\omega \mid (X_1(\omega), \ldots, X_n(\omega)) \in C\}$ ist für ein $C \subseteq \{-1, 1\}^n$. Die Menge aller bis zum Zeitpunkt n beobachtbaren Ereignisse bezeichnen wir mit $\mathcal{A}_n$. Für $n = 0$ definieren wir $\mathcal{A}_0 = \{\emptyset, \Omega\}$.

Bemerkung

Jedes $A \in \mathcal{A}_n$ ist offensichtlich eine Vereinigung von Ereignissen der Form $\{\omega \mid X_1(\omega) = x_1, \ldots, X_n(\omega) = x_n\}$. Diese Ereignisse bilden eine Partition von Ω. Ferner ist $A \in \mathcal{A}_n$ genau dann, wenn es von der Form $\{\omega \mid (S_1(\omega), \ldots, S_n(\omega)) \in D\}$ ist, wobei D jetzt eine Teilmenge aller möglichen Verläufe einer Irrfahrt mit n Perioden ist. Jedes $\mathcal{A}_n$ ist abgeschlossen gegen Komplementbildung und gegen (endliche) Vereinigungen und Durchschnitte. Es gilt

$$\mathcal{A}_0 \subset \mathcal{A}_1 \subset \cdots \subset \mathcal{A}_N = \text{ Potenzmenge von } \Omega. \tag{67}$$

Später wird für aufsteigende Mengensysteme, welche wie $(\mathcal{A}_n)$ abgeschlossen gegen Komplementbildung und abzählbare Vereinigungen sind, der Begriff *Filtration* eingeführt werden.

Stoppzeit

Definition

Eine Abbildung $T : \Omega \to \{0, \ldots, N\}$ heisst **Stoppzeit** wenn gilt:

$$\{\omega \mid T(\omega) = n\} \in \mathcal{A}_n \quad (n = 0, \ldots, N). \tag{68}$$

Bemerkung

(68) ist äquivalent zu $\{T \leq n\} \in \mathcal{A}_n$ für $n = 0, \ldots, N$, weil $\{T \leq n\} = \cup_{k=0}^{n} \{T = k\}$. Es ist auch äquivalent zu $\{T \geq n\} \in \mathcal{A}_{n-1}$ für $n = 1, \ldots, N$, weil $\{T \geq n\} = \{T \leq n-1\}^c$.

Beispiel

Sei T_c der erste Zeitpunkt > 0, in dem die Irrfahrt den Level $c \in \mathbb{Z}$ erreicht, also

$$T_c(\omega) = \min\{n > 0 \mid S_n(\omega) = c\} \qquad (\min \emptyset := +\infty).$$

Dann ist $\min(T_c, N)$ eine Stoppzeit, denn für $c \geq 0$ und $n \leq N$ gilt

$$\{T_c = n\} = \{S_1 < c, S_2 < c, \ldots, S_n = c\} \in \mathcal{A}_n,$$

Für $c < 0$ argumentiert man analog.

Beispiel

Betrachte den (ersten) Zeitpunkt, bei dem die Bilanz S_n maximal ist:

$$T(\omega) \equiv \min\{n \mid \ S_n(\omega) = \max\{S_k(\omega) \mid k = 0, 1, \ldots N\}\}.$$

Es ist z.B.
$\{\omega \mid T(\omega) = 0\} = \{\omega \mid S_1(\omega) \leq 0, S_2(\omega) \leq 0, \ldots, S_n(\omega) \leq 0\} \notin \mathcal{A}_0$, also ist dieses T keine Stoppzeit.

Beispiel

Eine beliebige Stoppzeit erhalten wir, indem wir eine absteigende Folge von Ereignissen $\Omega \supseteq A_1 \supseteq \ldots \supseteq A_N$ mit $A_n \in \mathcal{A}_{n-1}$ wählen und A_n als das Ereignis $\{T \geq n\}$ interpretieren. Das heisst wir setzen $T(\omega) = 0$ auf A_1^c, $T(\omega) = N$ auf A_N und $T(\omega) = n$ auf $A_n \cap A_{n+1}^c$ für $0 < n < N$. Weil $A_n \in \mathcal{A}_{n-1}$, muss A_n die Form $\{\omega \mid (X_1(\omega), \ldots, X_{n-1}(\omega)) \in C_{n-1}\}$ haben für eine Folge $C_{n-1} \subseteq \{-1, +1\}^{n-1}$. Damit $A_n \supseteq A_{n+1}$ gilt, muss die Projektion von C_n auf $\{-1, +1\}^{n-1}$ gleich C_{n-1} sein, d.h. aus $(x_1, \ldots, x_n) \in C_n$, folgt $(x_1, \ldots, x_{n-1}) \in C_{n-1}$. Man kann also für jedes n und jedes $(x_1, \ldots, x_{n-1}) \in C_{n-1}$ entscheiden, ob beide, einer oder keiner der zwei Punkte $(x_1, \ldots, x_{n-1}, +1)$ und $(x_1, \ldots, x_{n-1}, -1)$ zu C_n gehören sollen. In anderen Worten, wenn man beim n-ten Durchgang gespielt hat, dann kann die Entscheidung, auch im $(n+1)$-ten Durchgang zu spielen, von den Ausgängen $x_1, \ldots, x_n$ abhängen.

Optional Sampling

Der folgende Satz gibt die (ernüchternde und vom Experiment bestätigte) Antwort auf die eingangs gestellte Frage.

Satz

Für jede Stoppzeit T ist

$$\mathbb{E}(S_T) = 0$$

wobei $S_T(\omega) \equiv S_{T(\omega)}(\omega)$ den bei Benutzung der Stoppzeit T erzielten Ertrag bezeichnet.

Obiger Satz ist ein Spezialfall eines allgemeinen Satzes über die **Unmöglichkeit von (lohnenden!) Spielsystemen**, den wir jetzt formulieren und beweisen werden.

Systeme

Ein Spielsystem legt für jede Periode k fest, welcher Betrag $V_k(\omega)$ auf "+1" gesetzt wird, und zwar in Abhängigkeit von der bisherigen Entwicklung. Dieser Betrag kann auch gleich Null oder negativ sein (dann setzt man $-V_k$ auf "-1"). Der resultierende Ertrag in Periode k ist dann

$$V_k(\omega) \cdot X_k(\omega),$$

und der resultierende Gesamtertrag

$$(V \cdot S)_N(\omega) \equiv \sum_{k=1}^{N} V_k(\omega) \cdot X_k(\omega).$$

Systeme

Die Bedingung, dass der Einsatz für Periode k sich nicht schon auf Informationen über die weitere Entwicklung stützen darf, wird präzisiert durch die folgende

Definition

Ein **Spielsystem** ist eine Folge $V = (V_k)_{k=1,\dots,N}$ von Zufallsvariablen $V_k : \Omega \to \mathbb{R}^1$ derart, dass $V_1 = \text{const.}$ und für $k = 2, 3, \dots N$ existieren Funktionen $\varphi_k : \{-1, +1\}^{k-1} \to \mathbb{R}$ mit

$$V_k(\omega) = \varphi_k(X_1(\omega), \dots X_{k-1}(\omega)). \tag{69}$$

Systeme

Offensichtlich gilt für jedes Spielsystem

$$\{V_k = c\} \in \mathcal{A}_{k-1} \qquad (c \in \mathbb{R},\ k = 1, \ldots, N). \tag{70}$$

Umgekehrt folgt aus (70), dass V ein Spielsystem ist: Weil V_k nur endlich viele Werte c_j annehmen kann und weil $\{V_k = c_j\} \in \mathcal{A}_{k-1}$, erhalten wir

$$V_k = \sum_j c_j 1_{[V_k = c_j]} = \sum_j c_j 1_{C_j}((X_1, \ldots, X_{k-1}))$$

für Mengen $C_j \subseteq \{-1, +1\}^{k-1}$.

Stoppzeiten als Spielsystem

Jede **Stoppzeit** T lässt sich als Spielsystem auffassen, wenn wir setzen

$$V_k = 1_{\{T \geq k\}} = 1 - 1_{\{T \leq k-1\}}.$$

Die Bedingung (70) ist erfüllt, denn $\{T \leq k-1\} \in \mathcal{A}_{k-1}$ (T ist eine Stoppzeit!). Der resultierende Ertrag für dieses Spielsystem ist (wegen $T \leq N$)

$$(V \cdot S)_N = \sum_{k=1}^{N} 1_{\{T \geq k\}} X_k = \sum_{k=1}^{T} X_k = S_T.$$

"Sukzessives Verdoppeln des Einsatzes bis zur ersten $+1$**"** ist offensichtlich ein Spielsystem im Sinne der obigen Definition.

Verdoppelungsstrategie

Sei T eine Stoppzeit, und sei

$$V_k \equiv S_{k-1} I_{\{T \geq k\}} \qquad (k = 1, \ldots, N). \tag{71}$$

Dieses Spielsystem setzt liefert in Periode k den Ertrag

$$V_k X_k = \frac{1}{2} \left(S_k^2 - S_{k-1}^2 - 1 \right) I_{\{T \geq k\}}$$

(denn: $S_k^2 = (S_{k-1} + X_k)^2 = S_{k-1}^2 + 1 + 2 S_{k-1} \cdot X_k$) und den Gesamtertrag

$$(V \cdot S)_N = \frac{1}{2} \left(S_T^2 - T \right) \tag{72}$$

Optional Sampling

Satz

Für jedes Spielsystem $V = (V_k)_{k=1,\dots,N}$ *ist der erwartete Ertrag*

$$\mathbb{E}\left((V \cdot S)_N\right) = 0$$

Beweis

Wegen

$$\mathbb{E}\left((V \cdot S)_N\right) = \sum_{k=1}^{N} \mathbb{E}\left(X_k V_k\right)$$

(Linearität des Erwartungswertes) genügt es, $\mathbb{E}\left(X_k V_k\right) = 0$ zu zeigen. Gemäss der Definition eines Spielsystems ist aber

$$\begin{aligned}
\mathbb{E}\left(X_k V_k\right) &= \sum_{x_1,\ldots x_k} x_k \varphi_k(x_1,\ldots,x_{k-1}) \mathbb{P}\left(X_1 = x_1,\ldots,X_k = x_k\right) \\
&= \sum_{x_1,\ldots x_{k-1}} \varphi_k(x_1,\ldots,x_{k-1})\big(1 \cdot \mathbb{P}\left(X_1 = x_1,\ldots,X_{k-1} = x_{k-1}, X_k = +1\right) \\
&\quad (-1) \cdot \mathbb{P}\left(X_1 = x_1,\ldots,X_{k-1} = x_{k-1}, X_k = -1\right)\big),
\end{aligned}$$

und wegen (53) ist jeder Term in Klammern auf der rechten Seiten gleich $1 \cdot 2^{-k} - 1 \cdot 2^{-k} = 0$.

Waldsche Identitäten

Als Korollar erhalten wir den Stoppsatz und darüber hinaus die sogenannte

Folgerung

Waldsche Identität: *Für jede Stoppzeit T ist*

$$\mathbb{E}(S_T) = 0 \tag{73}$$

und

$$\mathbb{E}\left(S_T^2\right) = \mathbb{E}(T) \tag{74}$$

Beweis.

Wende Optinal Sampling auf die Spielsysteme an, die obige Erträge erzeugen. □

Bemerkung

In der stochastischen Analysis erkennt man die Irrfahrt als stochastischen Prozess mit Martingaleigenschaft. Der Begriff des Spielsystemes geht im allgemeineren Begriff der vorhersehbaren Strategie (des vorhersehbaren Prozesses) auf.

Unabhängige Bernoulli Experimente mit Erfolgsparameter p

Modell

Wir betrachten das folgende Modell für n 0-1-Experimente: Der Grundraum ist die Menge aller 0-1 Folgen der Länge n, d.h

$$\Omega = \{\omega = (x_1, \ldots, x_n) \mid x_i \in \{0, 1\}\}.$$

Die Zufallsvariablen X_i geben das Ergebnis im i-ten Experiment an, d.h.

$$X_i = \text{i-te Komponente von } \omega.$$

Modell

Im **Laplace-Modell** ist $\mathbb{P}$ die Gleichverteilung auf Ω, und das impliziert nach (52) und (53)

$$\mathbb{P}(X_i = +1) = \frac{1}{2} \quad (i = 1, \ldots, n) \tag{75}$$

$$\text{Die Ereignisse } \{X_i = +1\} \quad (i = 1, \ldots, n) \text{ sind unabhängig.} \tag{76}$$

Beliebiger Erfolgsparameter p

Wir geben uns jetzt einen beliebigen **Erfolgsparameter**

$$0 \leq p \leq 1$$

vor und suchen ein Wahrscheinlichkeitsmass $\mathbb{P}$ auf Ω, für das (76) gilt und ausserdem

$$\mathbb{P}(X_i = 1) = p \quad (i = 1, \ldots, n). \tag{77}$$

Verteilung der X_k

Durch diese Bedingungen ist $\mathbb{P}$ eindeutig festgelegt, denn für ein beliebiges $\omega = (x_1, \ldots, x_n)$ gilt wegen Unabhängigkeit, (48),

$$\mathbb{P}(\{\omega\}) = \mathbb{P}\left(\bigcap_{i=1}^{n}\{X_i = x_i\}\right) = \prod_{i=1}^{n}\mathbb{P}(X_i = x_i) = p^k(1-p)^{n-k},$$

also

$$\mathbb{P}(\{\omega\}) = p^k(1-p)^{n-k} \quad , \text{ falls } \sum_{i=1}^{n} x_i = k\,. \tag{78}$$

Umgekehrt ergeben sich darauas, aufgefasst als **Definition** von $\mathbb{P}$, die gewünschten Eigenschaften (76) und (77).

Verteilung S_n

Wenn wir 1 als "Erfolg" interpretieren, dann ist

$$S_n(\omega) = X_1(\omega) + \cdots + X_n(\omega) \tag{79}$$

die **Anzahl der Erfolge**. Wegen $\mathbb{E}(X_i) = p$ folgt

$$\mathbb{E}(S_n) = n \cdot p. \tag{80}$$

Die Verteilung von S_n ist gegeben durch

$$\begin{aligned} \mathbb{P}(S_n = k) &= \sum_{\omega=(x_1,\ldots,x_n):\ x_1+\cdots+x_n=k} \mathbb{P}(\{\omega\}) \\ &= \binom{n}{k} p^k (1-p)^{n-k} \quad (k = 0, \ldots, n). \end{aligned} \tag{81}$$

Dies ist die **Binomialverteilung mit Parametern p und n**.

Verteilung von S_n

Man kann dies auch mit folgender analytischer Method sehen: für jede reelle Zahl λ gilt

$$\begin{aligned}
\mathbb{E}\left(\exp(\mathrm{i}\lambda S_n)\right) &= \mathbb{E}\left(\exp\left(\mathrm{i}\lambda(X_1+\cdots+X_n)\right)\right) && (82)\\
&= \mathbb{E}\left(\exp(\mathrm{i}\lambda X_1)\right)\ldots\mathbb{E}\left(\exp(\mathrm{i}\lambda X_n)\right) && (83)\\
&= \left((1-p)+\exp(\mathrm{i}\lambda)p\right)^n && (84)\\
&= \sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}\exp(\mathrm{i}\lambda k) && (85)
\end{aligned}$$

mit Unabhängigkeit und der Eigenschaft, dass die Verteilungen der X_i identisch sind. Vergleicht man die letzte Summe mit dem Ausdruck

$$\sum_{k=0}^{n}\mathbb{P}\left(S_n=k\right)\exp(\mathrm{i}\lambda k)$$

für $\mathbb{E}\left(\exp(\mathrm{i}\lambda S_n)\right)$, dann erhält man durch Koeffizientvergleich (das gilt ja für alle λ!) die obige Formel.

Binomialverteilung

Die Binomialwahrscheinlichkeiten lassen sich leicht rekursiv berechnen. Sei $p_n(k) = \binom{n}{k} p^k q^{n-k}$ mit $q = 1 - p$. Dann gilt

$$p_n(k+1) = \frac{n-k}{k+1} \frac{p}{q} p_n(k),$$

insbesondere ist $p_n(k)$ ist maximal für $k \approx np$. In der Nähe von $k = np$ können wir $p_n(k)$ mit Hilfe der Stirling Formel approximieren. Mit einer Taylorentwicklung folgt daraus eine Approximation der Binomialverteilung für grosses n.

De Moivre Laplace

Satz (de Moivre-Laplace)

Es gilt

$$p_n(k) = \frac{1}{\sqrt{2\pi npq}} \exp\left(-\frac{(k-np)^2}{2npq}\right)(1 + r_n(k)) \tag{86}$$

mit

$$\sup\{|r_n(k)| : |k - np| \leq A\sqrt{n}\} \to 0 \quad \textit{für alle } A > 0 \quad (n \to \infty)$$

Beweis

Aus der Stirling-Formel $k! \sim \sqrt{2\pi k}(\frac{k}{e})^k$ folgt

$$p_n(k) \sim \frac{\sqrt{2\pi n} \quad n^n p^k q^{n-k}}{\sqrt{(2\pi)^2 k(n-k)}\ k^k (n-k)^{n-k}} = \frac{1}{\sqrt{2\pi n \frac{k}{n}(1-\frac{k}{n})}} \exp(n g_p(k/n))$$

wobei

$$g_p(x) = x(\log(p) - \log(x)) + (1-x)(\log(1-p) - \log(1-x)).$$

Beweis

Wir entwickeln als nächstes g_p in eine Taylorreihe an der Stelle $x = p$. Man erhält $g_p(p) = g_p'(p) = 0$ und $g_p''(p) = -1/(p(1-p))$. Daraus folgt, dass für $|k/n - p| \leq A/\sqrt{n}$

$$\exp(ng_p(k/n)) = \exp\left(-\frac{1}{2npq}(k-np)^2 + n\,\mathcal{O}(n^{-3/2})\right) .$$

Weiters kann natürlich der erste Ausdruck mit Taylor entwickelt werden, wir erhalten

$$\frac{1}{\sqrt{2\pi n\frac{k}{n}(1-\frac{k}{n})}} = \frac{1}{\sqrt{2\pi np(1-p)}} + \mathcal{O}(\frac{k}{n} - p)$$

Daraus folgt schliesslich die Behauptung.

Bemerkung

Durch Approximation des Integrals durch eine Riemannsumme erhält man ferner

$$\begin{aligned}&\mathbb{P}\left(np - A\sqrt{npq} \leq S_n \leq np + B\sqrt{npq}\right)\\ &\sim \int_{-A\sqrt{pq}}^{B\sqrt{pq}} \frac{1}{\sqrt{2\pi pq}} e^{-x^2/(2pq)} dx = \int_{-A}^{B} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = \Phi(B) - \Phi(-A),\end{aligned}$$

wobei

$$\Phi(B) = \int_{-\infty}^{B} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

tabelliert ist. Dies ist ein Spezialfall des **zentralen Grenzwertsatzes**, den wir später ausführlicher behandeln werden.

Mendels Versuche zur Genetik

Die Keimblätter von Gartenerbsen sind gelb oder grün. Gemäss Mendels Theorie enthält jede Pflanze ein Genpaar für die Farbe, entweder ge/ge oder ge/gr oder gr/gr oder gr/ge.

Das Gen ge ist dominant, so dass nur Pflanzen mit Gen gr/gr grüne Keimblätter aufweisen. Kreuzt man zwei reine Stämme, so erhält in der zweiten Generation mit Wahrscheinlichkeit $\frac{1}{4}$ grüne Keimblätter (alle Genpaare sind gleich wahrscheinlich).

Gregor Mendel erhielt unter 8023 Pflanzen der zweiten Generation 2001 Pflanzen mit grünen Keimblättern. Da
$np = \frac{1}{4}8023 = 2005.75$, $\sqrt{npq} = 38.8$, ist

$$\frac{2001 - np}{\sqrt{npq}} = -0.12.$$

Mendels Versuche zur Genetik

Also ist die Wahrscheinlichkeit für eine mindestens so grosse Abweichung $|S_n - np|$ wie Mendel sie beobachtete, $\approx 1 - (\Phi(0.12) - \Phi(-0.12)) \approx 90\%$, d.h. die Übereinstimmung zwischen Theorie und Experiment ist sehr gut.

Beachte dass man dies auch anders interpretieren kann: die Wahrscheinlichkeit für eine so kleine Abweichung vom Wert np ist cirka 10%. In der Tag ergab eine statistische Analyse aller Datensätze Gregor Mendels, dass mit hoher Wahrscheinlichkeit die schon sehr schönen Ergebnisse noch etwas "geschönt" wurden. Man geht allerdings davon aus, dass damit keine unlautere Absicht verbunden war, sondern nur eine unpräzise Durchführung des Experimentes sichtbar werden.

Es wird aber auch ein weiteres Mal klar, dass jede willentliche Veränderung der Handschrift des Zufalls starke Spuren hinterlässt.

Poissonapproximation der Binomialverteilung

Die Approximation der Binomialverteilung im Satz von de Moivre-Laplace gilt für p fest und $n \to \infty$. Je näher p bei 0 oder 1 ist, desto schlechter ist sie. Hier besprechen wir die Approximation durch eine Poisson(λ)-Verteilung, welche gut ist für n gross und p klein, wobei $np = \lambda$. Das heisst, man hat viele 0-1-Experimente mit kleiner Erfolgswahrscheinlichkeit (Bsp. Radioaktiver Zerfall: viele Teilchen, kleine Zerfallswahrscheinlichkeit für jedes Teilchen; oder Schadenfälle bei einer Versicherung: viele Policen, kleine Schadenwahrscheinlichkeit für jede Police).

Poissonapproximation der Binomialverteilung

Satz

Für k fest, $n \to \infty, p \to 0$ mit $np \to \lambda$ gilt:

$$\binom{n}{k} p^k (1-p)^{n-k} \longrightarrow \frac{\lambda^k}{k!} e^{-\lambda}. \tag{87}$$

Beweis

Der Beweis ergibt sich aus folgendem Limes:

$$\begin{aligned}\binom{n}{k} p^k (1-p)^{n-k} &= \frac{1}{k!}(np)^k \left(1 - \frac{np}{n}\right)^n \frac{n(n-1)\dots(n-k+1)}{n^k}(1-p)^{-k} \\ &\longrightarrow \frac{1}{k!}\lambda^k e^{-\lambda} \cdot 1 \cdot 1.\end{aligned}$$

Da der Erwartungswert der Poissonverteilung gleich λ ist, d.h. $np \to \lambda$ bedeutet, dass der Erwartungswert der Binomialverteilung gegen den Erwartungswert der Poissonverteilung konvergiert.

Beispiel

Rutherford (1910) beobachtete die Zerfälle eines radioaktiven Präparats in $n = 2608$ Zeitintervallen von 7.5 Sekunden. n_k ist die Anzahl Intervalle mit genau k Zerfällen. Wenn die Anzahl Zerfälle in 7.5 Sekunden Poisson-verteilt ist, dann sollten nach der frequentistischen Interpretation der Wahrscheinlichkeit $n_k \approx np_k = ne^{-\lambda}\lambda^k/k!$ sein. Um dies zu überprüfen, müssen wir λ wählen.

Beispiel

Wir ersetzen dazu den Erwartungswert durch das arithmetische Mittel

$$\frac{\text{Gesamtzahl Zerfälle}}{\text{Anzahl Intervalle}} = \frac{\sum_{k=0}^{\infty} k n_k}{n} = \frac{10097}{2608} = 3.87$$

Die p_k's in der dritten Zeile der folgenden Tabelle sind mit $\lambda = 3.87$ berechnet.

k	$=$	0	1	2	3	4	5	6	7	8	9	10	11	≥ 12
n_k	$=$	57	203	383	525	532	408	273	139	45	27	10	4	2
np_k	$=$	54	210	407	525	508	394	254	141	68	29	11	4	1

Die Übereinstimmung scheint sehr gut zu sein.

Unabhängiges Ankommen

Eine andere Begründung der Poissonverteilung beruht auf folgendem Satz.

Satz

Seien X_1 und X_2 zwei unabhängige Zufallsvariable mit

$$\mathbb{P}(X_1 = k | X_1 + X_2 = n) = \binom{n}{k} 2^{-n}$$

für alle k, n mit $0 \leq k \leq n$. Dann sind X_1 und X_2 Poissonverteilt mit dem gleichen λ.

Um die Bedeutung dieses Satzes zu verstehen, seien X_1 die Anzahl Zerfälle in $[0, T]$ und X_2 die Anzahl Zerfälle in $[T, 2T]$. Die Voraussetzung in obigem Satz bedeutet dann, dass jeder der n Zerfälle in $[0, 2T]$ mit Wahrscheinlichkeit $\frac{1}{2}$ in $[0, T]$ geschah und mit Wahrscheinlichkeit $\frac{1}{2}$ in $[T, 2T]$, unabhängig von den andern Zerfällen.

Beweis

$$\begin{aligned}\frac{1}{n} &= \frac{\mathbb{P}(X_1 = n | X_1 + X_2 = n)}{\mathbb{P}(X_1 = n-1 | X_1 + X_2 = n)} = \frac{\mathbb{P}(X_1 = n, X_2 = 0)}{\mathbb{P}(X_1 = n-1, X_2 = 1)} \\ &= \frac{\mathbb{P}(X_1 = n) \cdot \mathbb{P}(X_2 = 0)}{\mathbb{P}(X_1 = n-1) \cdot \mathbb{P}(X_2 = 1)}\end{aligned}$$

Also ist mit $\lambda = \mathbb{P}(X_2 = 1) / \mathbb{P}(X_2 = 0)$:

$$\mathbb{P}(X_1 = n) = \frac{\lambda}{n} \mathbb{P}(X_1 = n-1) = \ldots = \frac{\lambda^n}{n!} \mathbb{P}(X_1 = 0).$$

Da sich die Wahrscheinlichkeiten zu eins addieren müssen, folgt $\mathbb{P}(X_1 = 0) = e^{-\lambda}$. Für X_2 geht es analog.

Summen unabhängiger Poissonverteilungen

Schliesslich beweisen wir noch, dass die Summe zweier unabhängiger poissonverteilter Zufallsvariable wieder poissonverteilt ist. In den obigen Beispielen ist die Wahl der Intervallänge also irrelevant.

Satz

Wenn X_1 und X_2 unabhängig und Poisson(λ_i)-verteilt sind, dann ist $X = X_1 + X_2$ Poisson($\lambda_1 + \lambda_2$)-verteilt.

Beweis

Entweder man approximiert X_i durch 2 unabhängige binomialverteilte Zufallsvariable mit Erfolgsparameter $p = \frac{1}{n}$ und $n_i = [\lambda_i n]$, für die die Additionseigenschaft offensichtlich ist (aber man bei der Approximation ein bisschen arbeiten muss), oder man rechnet direkt nach (später werden wir für diese Operation den Begriff der Faltung einführen):

$$\begin{aligned}
\mathbb{P}(X = k) &= \sum_{j=0}^{k} \mathbb{P}(X_1 = j, X_2 = k - j) = \sum_{j=0}^{k} \mathbb{P}(X_1 = j) \cdot \mathbb{P}(X_2 = k - j) \\
&= \sum_{j=0}^{k} e^{-\lambda_1} \frac{\lambda_1^j}{j!} e^{-\lambda_2} \frac{\lambda_2^{k-j}}{(k-j)!} \\
&= e^{-(\lambda_1+\lambda_2)} \frac{1}{k!} \sum_{j=0}^{k} \binom{k}{j} \lambda_1^j \lambda_2^{k-j} \\
&= e^{-(\lambda_1+\lambda_2)} \frac{(\lambda_1 + \lambda_2)^k}{k!} .
\end{aligned}$$

Stirlings Formel via Poissonapproximation

Wir können die Eigenschaft dass Summen von unabhängigen poissonverteilten Zufallsvariablen wieder poissonverteilt sind auch nützen, um die Stirlingsche Formel zu beweisen:

- Mit obigem Satz ist klar, dass

$$\mathbb{P}(S_n = n) = \frac{\exp(-n)n^n}{n!}$$

  gilt, falls $S_n = X_1 + \ldots + X_n$ and X_i seien unabhängig poissonverteilt.

- Weiters können wir die charakteristische Funktion einer Poissonverteilung berechen, nämlich

$$\mathbb{E}\left(\exp(\mathrm{i}\lambda X_1)\right) = \sum_{k=0}^{\infty} \exp(\mathrm{i}\lambda k)\exp(-1)\frac{1}{k!} = \exp\left(\exp(\mathrm{i}\lambda) - 1\right)$$

Stirlings Formel via Poissonapproximation

- Mit der letzten Formel können wir aber auch die charakteristische Funktion von $\frac{S_n-n}{\sqrt{n}}$ berechnen, nämlich

$$\mathbb{E}\left(\exp\left(\mathrm{i}\lambda\frac{S_n-n}{\sqrt{n}}\right)\right)=\left(\exp\left(\exp(\mathrm{i}\lambda/\sqrt{n})-1-\mathrm{i}\lambda/\sqrt{n}\right)\right)^n,$$

  der zu $\exp(-\lambda^2/2)$ konvergiert mit $n\to\infty$.

- Der letzte Ausdruck ist aber auch Limes von

$$\sum_{l\in\mathbb{Z}}\exp(\mathrm{i}\lambda l/\sqrt{n})\exp(-l^2/2n)\frac{1}{\sqrt{2\pi n}}$$

  durch Riemannsche Approximation von $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp(\mathrm{i}\lambda x)\exp(-x^2/2)dx$.

Stirlings Formel via Poissonapproximation

- Koeffizientenvergleich zwischen der korrekt berechneten charakteristischen Funktion von $\frac{S_n - n}{\sqrt{n}}$ und obiger Formel für $l = 0$ liefert

$$\mathbb{P}(S_n = n) = \frac{\exp(-n)n^n}{n!} \sim \frac{1}{\sqrt{2\pi n}}$$

as $n \to \infty$.

Allgemeine Wahrscheinlichkeitsräume

Einführung

Der bisher benutzte Rahmen eines diskreten Wahrscheinlichkeitsraumes erweist sich für viele Situationen und Fragestellungen als zu eng. Viele Phänomene treten erst im Grenzübergang zu einem überabzählbarem Modell deutlich hervor (z.B. beim Arkussinus-Gesetz oder beim zentralen Grenzwertsatz), und andere Fragen lassen sich überhaupt erst in einem überabzählbaren Modell exakt formulieren. Ein Beispiel dafür ist die Frage nach

$$\mathbb{P}\left(\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}X_i \leq \frac{1}{2}\right)$$

bei sukzessiven unabhängigen 0-1-Experimenten $X_1, X_2, \ldots$.

Deshalb führen wir jetzt den allgemeinen Begriff des Wahrscheinlichkeitsraumes ein. Dazu braucht man Masstheorie. Wir werden diese nicht systematisch entwickeln, sondern begnügen uns mit Hinweisen, wo diese gebraucht wird.

Axiome von Kolmogorov

Sei $\Omega \neq \emptyset$ irgendeine nichtleere Menge, $\mathcal{A}$ eine Kollektion von Teilmengen $A \subseteq \Omega$ und $\mathbb{P} : \mathcal{A} \to [0, 1]$ eine Abbildung von $\mathcal{A}$ in das Einheitsintervall.

Die Elemente $\omega \in \Omega$ interpretieren wir als die (im Modell in Betracht gezogenen) **möglichen Fälle**, die Teilmengen $A \in \mathcal{A}$ als die (im Modell zugelassenen) **Ereignisse**, und für $A \in \mathcal{A}$ interpretieren wir die Zahl $\mathbb{P}(A)$ als die (im Modell angenommene) **Wahrscheinlichkeit des Ereignisses** A.

Die folgenden **Axiome von Kolmogorov** (1933) verlangen nun, dass die Kollektion $\mathcal{A}$ der Ereignisse abgeschlossen ist unter abzählbaren Mengenoperationen, und dass die Zuordnung $A \to \mathbb{P}(A)$ "konsistent" ist im Sinne der Rechenregeln für Wahrscheinlichkeiten, die uns von den diskreten Modellen her schon vertraut sind.

Axiome von Kolmogorov

Definition

Das Tripel $(\Omega, \mathcal{A}, \mathbb{P})$ heisst ein **Wahrscheinlichkeitsraum**, wenn gilt:

1) $\mathcal{A}$ ist eine σ-**Algebra**, d.h.

$$\Omega \in \mathcal{A} \tag{88}$$

$$A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A} \tag{89}$$

$$A_1, A_2, \cdots \in \mathcal{A} \Rightarrow \bigcup_i A_i \in \mathcal{A} \tag{90}$$

2) $\mathbb{P}$ ist eine **Wahrscheinlichkeitsverteilung**, d.h.

$$\mathbb{P}(\Omega) = 1 \tag{91}$$

$$A_1, A_2, \cdots \in \mathcal{A},\ A_i \cap A_j = \emptyset\ (i \neq j) \Rightarrow \mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i) \tag{92}$$

Bemerkungen

In der Sprache der Masstheorie besagt 2), dass $\mathbb{P}$ ein **normiertes Mass** ist, d.h. eine Mengenfunktion $\mathbb{P} : \mathcal{A} \to \mathbb{R}^+$, die im Sinne von (92) σ-**additiv** ist und im Sinne von (91) **normiert**.

Beispiel für Grenzen der Axiome: zufällige Wahl einer natürlichen Zahl

Auf die Frage, mit welcher Wahrscheinlichkeit man bei zufälliger Wahl einer natürlichen Zahl eine gerade Zahl erhält, ist man versucht, die Antwort $\frac{1}{2}$ zu geben. Allgemeiner würde man dann für eine Menge $A \subseteq \Omega = \{1, 2, \dots\}$ den Ansatz

$$\mathbb{P}(A) = \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} I_A(n)$$

(relative Häufigkeit von A in Ω) machen, sofern A zur Klasse $\mathcal{A}_0$ derjenigen Mengen gehört, für die dieser Limes existiert.

Beispiel für Grenzen der Axiome: zufällige Wahl einer natürlichen Zahl

Es gibt aber keine Wahrscheinlichkeitsverteilung im Sinne der obigen Definition, die mit diesem Ansatz verträglich wäre! Für jedes n ist nämlich $\{n\} \in \mathcal{A}_0$ mit $\mathbb{P}(\{n\}) = 0$, aus (92) würde also für jede Menge $A \subseteq \Omega$

$$\mathbb{P}(A) = \mathbb{P}\left(\bigcup_{n \in A} \{n\}\right) = \sum_{n \in A} \mathbb{P}(\{n\}) = 0$$

folgen. Es geht nur dann, wenn man bereit ist, die σ-Additivität durch die einfache Additivität zu ersetzen, aber eine solche "finite" Wahrscheinlichkeitstheorie ist in ihren technischen Möglichkeiten sehr viel begrenzter.

Standardbeispiel: Gleichverteilung auf dem Einheitsintervall

Sei

$$\Omega = [0, 1],$$

und sei $\mathcal{A}$ die kleinste σ-Algebra, welche alle Intervalle $[a, b] \subseteq [0, 1]$ enthält. Dann gibt es genau eine Wahrscheinlichkeitverteilung $\mathbb{P} : \mathcal{A} \to [0, 1]$, bei der die Wahrscheinlichkeit eines Intervalls dessen Länge ist:

$$\mathbb{P}([a, b]) = b - a, \tag{93}$$

nämlich das auf $[0, 1]$ eingeschränkte **Lebesguemass**. Man kann $\mathcal{A}$ zwar noch etwas erweitern zur Klasse der "Lebesgue-messbaren" Mengen, aber es ist nicht klar, ob man bis zur vollen Potenzmenge gehen kann. Mit dem Auswahlaxiom kann man relativ leicht zeigen, dass es keine Wahrscheinlichkeitsverteilung auf der σ-Algebra aller Teilmengen $A \subseteq [0, 1]$ gibt, die translationsinvariant ist (d.h. $\mathbb{P}(\tau_x(A)) = \mathbb{P}(A)$ für alle A und alle x, wobei $\tau_x(y) = y + x \pmod 1$).

Standardbeispiel: Gleichverteilung auf dem Einheitsintervall

Für jede feste Zahl $\omega \in [0,1]$ folgt aus (93) $\mathbb{P}(\{\omega\}) = 0$. Anders als im vorigen Beispiel kann man aber nicht für beliebige Teilmengen $A \in \mathcal{A}$

$$\mathbb{P}(A) = \mathbb{P}\left(\bigcup_{\omega \in A} \{\omega\}\right) = \sum_{\omega \in A} \mathbb{P}(\{\omega\}) = 0$$

schliessen: nach (92) ist das nur für $\mathrm{abzählbare}$ Mengen A zulässig. Zum Beispiel ist $\mathbb{P}(A) = 0$ für $A =$ "alle rationalen Zahlen". Bei der Simulation der Gleichverteilung auf einem Taschenrechner ("Zufallszahlen") sieht das natürlich anders aus!

Einfache Folgerungen

Eine σ-Algebra $\mathcal{A}$ ist abgeschlossen gegen *abzählbare* Mengenoperationen, folglich

$$A_1, A_2, \cdots \in \mathcal{A} \Rightarrow \bigcap_i A_i = \left(\bigcup_i A_i^c \right)^c \in \mathcal{A},$$

oder das Ereignis $A_\infty =$ "Unendlich viele der Ereignisse A_i treten ein" gehört zu $\mathcal{A}$, denn

$$A_\infty = \bigcap_n \bigcup_{k \geq n} A_k.$$

Bemerkung zur Wahl der σ-Algebra $\mathcal{A}$:

Wir haben gesehen, dass man als σ-Algebra $\mathcal{A}$ der zugelassenen Ereignisse nicht immer die Klasse *aller* Teilmengen von Ω wählen *kann*. Oft *will* man es auch gar nicht, weil man sich nur für eine gewisse Teilklasse von Ereignissen interessiert (vgl. die Diskussion der Spielsysteme).

In der Regel ist es so, dass man eine gewisse Klasse $\mathcal{A}_0$ von Teilmengen im Auge hat, die jedenfalls als Ereignisse zugelassen sein sollten, und dass man dann übergeht zu der **"von $\mathcal{A}_0$ erzeugten σ-Algebra"**:

$$\mathcal{A} = \sigma(\mathcal{A}_0) = \bigcap_{\mathcal{B} \supseteq \mathcal{A}_0, \mathcal{B} \text{ ist } \sigma-\text{Algebra}} \mathcal{B}\,. \tag{94}$$

Dies *ist* eine σ-Algebra, und zwar offensichtlich die **kleinste σ-Algebra, welche $\mathcal{A}_0$ enthält**.

Weitere einfache Folgerungen

Eine Mengenfunktion $\mathbb{P} : \mathcal{A} \to [0,1]$ mit

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i) \tag{95}$$

für je *endlich* viele paarweise disjunkte Ereignisse $A_1, \ldots, A_n \in \mathcal{A}$ nennt man auch **additiv**. Aus der Additivität und der Normierung (91) ergeben sich die üblichen elementaren Rechenregeln für Wahrscheinlichkeiten wie im diskreten Fall

$$\begin{aligned} \mathbb{P}(A^c) &= 1 - \mathbb{P}(A), \quad \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B), \\ \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) &= \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \le i_1 < \cdots < i_k \le n} \mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}), \text{ usw.} \end{aligned}$$

Durch die σ-**Additivität** (92), bei der auch **abzählbar** viele paarweise disjunkte Ereignisse $A_1, A_2, \ldots$ zugelassen sind, kommen zusätzliche **Stetigkeitseigenschaften** ins Spiel. Genauer gilt:

Stetigkeitseigenschaften

Satz

Ist $\mathbb{P} : \mathcal{A} \to [0,1]$ *additiv, so sind die folgenden Aussagen äquivalent:*

$$\mathbb{P} \text{ ist } \sigma\text{-additiv}, \tag{96}$$

$$A_1 \subseteq A_2 \subseteq \dots\ , \quad A_n \in \mathcal{A} \Rightarrow \mathbb{P}\left(\bigcup_n A_n\right) = \lim_n \mathbb{P}(A_n)\,, \tag{97}$$

$$A_1 \supseteq A_2 \supseteq \dots\ , \quad A_n \in \mathcal{A} \Rightarrow \mathbb{P}\left(\bigcap_n A_n\right) = \lim_n \mathbb{P}(A_n)\,. \tag{98}$$

Beweis

Für $A_1 \subseteq A_2 \subseteq \dots$ setzen wir $B_1 \equiv A_1, B_n \equiv A_n - A_{n-1}$ $(n \geq 2)$. Diese B_n $(n = 1, 2, \dots)$ sind paarweise disjunkt, es gilt $\bigcup_n B_n = \bigcup_n A_n$, und aus der Additivität folgt

$$\lim_n \mathbb{P}(A_n) = \lim_n \mathbb{P}\left(\bigcup_{i=1}^{n} B_i\right) = \lim_n \sum_{i=1}^{n} \mathbb{P}(B_i) = \sum_{i=1}^{\infty} \mathbb{P}(B_i)\,.$$

Daraus ergibt sich die erste Äquivalenz. Die zweite Äquivalenz folgt durch Komplementbildung.

Korollar

Folgerung

Für beliebige $A_1, A_2, \cdots \in \mathcal{A}$ gilt

$$\mathbb{P}\left(\bigcup_k A_k\right) \leq \sum_k \mathbb{P}(A_k)\,.$$

Beweis.

$$\mathbb{P}\left(\bigcup_k A_k\right) = \lim_n \mathbb{P}\left(\bigcup_{k=1}^{n} A_k\right) \leq \lim_n \sum_{k=1}^{n} \mathbb{P}(A_k) = \sum_k \mathbb{P}(A_k)\,.$$

The Borel-Cantelli Lemma

Die **Unabhängigkeit von Ereignissen** ist definiert wie in (48)

Lemma (Borel-Cantelli)

Sei $A_1, A_2, \cdots \in \mathcal{A}$ eine Folge von Ereignissen, und sei

$$A_\infty = \bigcap_n \left(\bigcup_{k \geq n} A_k \right) = \text{“unendlich viele der } A_k \text{ treten ein”}.$$

1) *Aus $\sum_k \mathbb{P}(A_k) < \infty$ folgt stets $\mathbb{P}(A_\infty) = 0$.*
2) *Sind die Ereignisse $A_1, A_2, \ldots$ unabhängig, so folgt aus $\sum_k \mathbb{P}(A_k) = \infty$ umgekehrt auch $\mathbb{P}(A_\infty) = 1$.*

Beweis

Die Folge $(\bigcup_{k\geq n} A_k)_n$ ist absteigend. Also folgt

$$\mathbb{P}(A_\infty) = \lim_n \mathbb{P}\left(\bigcup_{k\geq n} A_k\right) \leq \lim_n \sum_{k\geq n} \mathbb{P}(A_k) = 0.$$

Beweis

Für die zweite Aussage gehen wir aus von

$$\mathbb{P}\left(\bigcap_{k\geq n} A_k^c\right) = \lim_m \mathbb{P}\left(\bigcap_{k=n}^{m} A_k^c\right) = \prod_{k\geq n} \mathbb{P}\left(A_k^c\right) = \prod_{k\geq n} (1 - \mathbb{P}\left(A_k\right)).$$

Weil $1 - x \leq e^{-x}$ gilt, folgt also für jedes feste n

$$\mathbb{P}\left(\bigcap_{k\geq n} A_k^c\right) \leq \exp(-\sum_{k\geq n} \mathbb{P}\left(A_k\right)) = 0.$$

Da die Folge $(\bigcap_{k\geq n} A_k^c)_n$ aufsteigend ist, folgt

$$\mathbb{P}\left(A_\infty^c\right) = \mathbb{P}\left(\bigcup_n (\bigcap_{k\geq n} A_k^c)\right) = \lim_n \mathbb{P}\left(\bigcap_{k\geq n} A_k^c\right) = 0.$$

Sukzessive unabhängige 0-1-Experimente

Sei

$$\Omega = \{\omega = (x_1, x_2, \dots) | x_i \in \{0, 1\}\}$$

die (überabzählbare!) Menge aller 0-1-Folgen, X_i die durch $X_i(\omega) = x_i$ definierte Zufallsvariable und $\mathcal{A}$ die von den Ereignissen $\{\omega | X_i(\omega) = 1\}$ $(i = 1, 2, \dots)$ erzeugte σ-Algebra. Ferner sei $0 \leq p \leq 1$ ein **"Erfolgsparameter".**

Struktursatz für allgemeines p

Es gibt genau eine Wahrscheinlichkeitsverteilung $\mathbb{P}$ auf $(\Omega, \mathcal{A})$ derart, dass gilt:

$$\mathbb{P}(X_i = 1) = p \quad (i = 1, 2, \dots) \tag{99}$$

Die Ereignisse $\{X_i = 1\}$ $(i = 1, 2, \dots)$ sind unabhängig bezüglich $\mathbb{P}$. (100)

Beweis

Zunächst folgt für jedes $n \geq 1$ und für jede Wahl von $x_i \in \{0, 1\}$ aus (99) und (100), dass mit $k = \sum_{i=1}^{n} x_i$ gelten muss

$$\begin{aligned} \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) &= \prod_{i=1}^{n} \mathbb{P}(X_i = x_i) \\ &= p^k (1-p)^{n-k}. \end{aligned}$$

Damit ist $\mathbb{P}$ festgelegt auf der Kollektion aller Ereignisse, die sich als endliche Vereinigung von Ereignissen der Form $\{X_1 = x_1, \ldots, X_n = x_n\}$ darstellen lassen $(= \bigcup_{n \geq 0} \mathcal{A}_n)$. Für die Erweiterung von $\mathbb{P}$ auf die volle σ-Algebra $\mathcal{A}$ benötigt man nun einen **Fortsetzungssatz der Masstheorie**. □

Bemerkung

Für $p = \frac{1}{2}$ folgt die Existenz von $\mathbb{P}$ auch aus der Existenz des Lebesguemasses auf $[0, 1]$.

Für $p = 1$ (bzw. $= 0$) ist die Konstruktion von $\mathbb{P}$ natürlich sehr einfach:

$$\mathbb{P}(A) \equiv \begin{cases} 1 & \text{falls } (1, 1, 1, \dots) \in A \\ 0 & \text{sonst.} \end{cases}$$

Bemerkung

Sei nun $0 < p < 1$. Für jedes $\omega = (x_1, x_2, \dots) \in \Omega$ ist dann

$$\mathbb{P}(\{\omega\}) = 0 \tag{101}$$

denn mit $k(n) = \sum_{i=1}^{n} x_i$ gilt

$$\begin{aligned} \mathbb{P}(\{\omega\}) &= \mathbb{P}\left(\bigcap_{n=1}^{\infty} \{X_n = x_n\}\right) = \lim_{n \uparrow \infty} \mathbb{P}(X_1 = x_1, \dots, X_n = x_n) \\ &= \lim_{n \uparrow \infty} p^{k(n)} (1-p)^{n-k(n)} = 0. \end{aligned}$$

Binäre Texte

Satz

Sei $[x_1, \dots, x_N]$ ein "binärer Text" mit $x_i \in \{0, 1\}$. Dann ist die Wahrscheinlichkeit, dass irgendwann dieser Text erscheint (und dies nicht nur einmal, sondern sogar unendlich oft !) gleich eins.

Beweis.

Wir betrachten die Ereignisse

$$A_k \equiv \{X_{(k-1)N+1} = x_1, \dots, X_{kN} = x_N\} \quad (k = 1, 2, \dots).$$

Diese sind unabhängig und haben alle dieselbe Wahrscheinlichkeit $\mathbb{P}(A_k) > 0$. Mit dem zweiten Teil des Lemmas von Borel-Cantelli folgt also die Behauptung. □ □

Messbare Abbildungen

Definition

Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, $\widetilde{\Omega} \neq \emptyset$ und $\widetilde{\mathcal{A}}$ eine σ-Algebra von Teilmengen $A \subseteq \widetilde{\Omega}$.

Definition

Eine Abbildung $\varphi : \Omega \to \widetilde{\Omega}$ heisst **messbar** (bezüglich $\mathcal{A}$ und $\widetilde{\mathcal{A}}$), wenn gilt:

$$A \in \widetilde{\mathcal{A}} \Rightarrow \varphi^{-1}(A) = \{\omega | \varphi(\omega) \in A\} \in \mathcal{A}. \tag{102}$$

Bemerkung

Wird $\widetilde{\mathcal{A}}$ von irgendeinem Mengensystem $\widetilde{\mathcal{A}}_0$ erzeugt, gilt also $\widetilde{\mathcal{A}} = \sigma(\widetilde{\mathcal{A}}_0)$, so genügt es, statt (102) zunächst nur

$$A \in \widetilde{\mathcal{A}}_0 \Rightarrow \varphi^{-1}(A) \in \mathcal{A} \tag{103}$$

zu fordern. Denn

$$\{A \subseteq \widetilde{\Omega} | \varphi^{-1}(A) \in \mathcal{A}\}$$

ist eine σ-Algebra, die $\widetilde{\mathcal{A}}_0$ umfasst, damit aber auch $\widetilde{\mathcal{A}} = \sigma(\widetilde{\mathcal{A}}_0)$, und das ist gleichbedeutend mit (102).

Verteilung einer messbaren Abbildung

Satz

Ist $\varphi : \Omega \to \widetilde{\Omega}$ *messbar, so ist durch*

$$\widetilde{\mathbb{P}}(A) \equiv \mathbb{P}\left(\varphi^{-1}(A)\right) \quad (A \in \widetilde{\mathcal{A}}) \tag{104}$$

eine Wahrscheinlichkeitsverteilung $\widetilde{\mathbb{P}}$ *auf* $(\widetilde{\Omega}, \widetilde{\mathcal{A}})$ *definiert.* $\widetilde{\mathbb{P}}$ *heisst das* **Bild von** $\mathbb{P}$ *unter* φ *bzw. die* **Verteilung von** φ *unter* $\mathbb{P}$.

Beweis

Zunächst gilt offensichtlich

$$\widetilde{\mathbb{P}}(\widetilde{\Omega}) = \mathbb{P}\left(\varphi^{-1}(\widetilde{\Omega})\right) = \mathbb{P}(\Omega) = 1.$$

Wir müssen also noch die σ-Additivität beweisen. Seien also $A_1, A_2, \ldots$ paarweise disjunkte Mengen in $\widetilde{\mathcal{A}}$. Die Ereignisse $B_n \equiv \varphi^{-1}(A_n)$ $(n = 1, 2, \ldots)$ gehören dann nach Voraussetzung zu $\mathcal{A}$ und sind offensichtlich paarweise disjunkt. Also

$$\widetilde{\mathbb{P}}\left(\bigcup_i A_i\right) = \mathbb{P}\left(\varphi^{-1}\left(\bigcup_i A_i\right)\right) = \mathbb{P}\left(\bigcup_i \varphi^{-1}(A_i)\right) = \sum_i \mathbb{P}\left(\varphi^{-1}(A_i)\right) =$$

$$= \sum_i \widetilde{P}(A_i).$$

□

Beispiel: binäre Darstellung

Sei $\Omega = [0,1]$, $\mathcal{A} = \sigma(\{[a,b] \mid 0 \leq a \leq b \leq 1\})$ und $\mathbb{P}$ die *Gleichverteilung auf* $[0,1]$, also das auf $[0,1]$ eingeschränkte Lebesguemass. Ferner sei wie im vorigen Abschnitt $\widetilde{\Omega}$ die Menge aller 0-1 Folgen und $\widetilde{\mathcal{A}}$ die von den Ereignissen $\{X_i = 0\}$ $(i = 1, 2, \dots)$ erzeugte σ-Algebra.

Wir definieren nun die Abbildung $\varphi : \Omega \longrightarrow \widetilde{\Omega}$ durch die binäre Darstellung der Zahlen im Einheitsintervall:

$$\varphi(\omega) = (\varphi_1(\omega),\ \varphi_2(\omega), \dots) \quad (\omega \in [0,1]).$$

Das heisst, wir setzen für $n = 0, 1, \dots$ $\varphi_{n+1}(\omega) = 0$ genau dann, wenn ω in einem Intervall $[k2^{-n}, k2^{-n} + 2^{-(n+1)})$ liegt für ein $k \in \{0, \dots, 2^n - 1\}$. Wenn die binäre Darstellung nicht eindeutig ist, nehmen wir also die Version, die mit lauter Nullen endet.

Beispiel: binäre Darstellung

Diese Abbildung ist messbar, denn für $n = 0, 1, \ldots$ ist

$$\varphi^{-1}(\{X_{n+1} = 0\}) = \{\varphi_{n+1} = 0\} = \bigcup_{k=0}^{2^n-1} [k2^{-n}, k2^{-n} + 2^{-(n+1)}) \in \mathcal{A}\,.$$

Wir setzen $\widetilde{\mathbb{P}}$ gleich dem Bild der Gleichverteilung unter φ. Dann ist

$$\widetilde{\mathbb{P}}(X_1 = x_1, \ldots, X_n = x_n) = \mathbb{P}(\underbrace{\varphi_1 = x_1, \ldots, \varphi_n = x_n}_{\text{Intervall der Länge } 2^{-n}}) = 2^{-n},$$

und daraus folgt, dass $\widetilde{\mathbb{P}}$ die Eigenschaften (99) und (100) mit $p = \frac{1}{2}$ hat.

Aus der Existenz des Lebesguemasses folgt also die Existenz eines exakten Modells für unendlich viele Würfe einer fairen Münze.

Reelwertige Zufallsvariablen

Sei $\mathcal{B}$ die Borelsche σ-Algebra auf $\mathbb{R}$, das heisst die von allen Intervallen der Form $(-\infty, b]$ mit $b \in \mathbb{R}$ erzeugte σ-Algebra. Man kann zeigen, dass $\mathcal{B}$ alle Intervalle, alle offenen Mengen und alle abgeschlossenen Mengen enthält.

Reelwertige Zufallsvariablen

Definition

Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Eine **Zufallsvariable** ist eine messbare Abbildung

$$X : (\Omega, \mathcal{A}) \longrightarrow (\mathbb{R}, \mathcal{B}).$$

Die **Verteilung** μ von X ist das Bild von $\mathbb{P}$ unter X, d.h. für jedes $A \in \mathcal{B}$

$$\mu(A) = \mathbb{P}\left(X^{-1}(A)\right) = \mathbb{P}\left(\{\omega | X(\omega) \in A\}\right) = \mathbb{P}\left(X \in A\right).$$

Bemerkung

Wenn wir nur an Ereignissen interessiert sind, welche die Verteilung der Zufallsvariable X betreffen, dann können wir den Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ vergessen und mit $(\mathbb{R}, \mathcal{B}, \mu)$ weiterarbeiten.

Beispiel: geometrische Verteilung

Sei $(\Omega, \mathcal{A}, \mathbb{P}) =$ das Modell für unendlich viele $0-1$ Experimente mit Erfolgsparameter p. Sei X die Zufallsvariable "Wartezeit auf die erste Eins" d.h.

$$X(\omega) = \min\{k \geq 1 | x_k = 1\}$$

wobei wir $\min \emptyset := \infty$ setzen.

Dann ist für $k = 1, 2, \ldots$

$$\mu(\{k\}) = \mathbb{P}(X = k) = \mathbb{P}(\{X_1 = \cdots = X_{k-1} = 0, X_k = 1\}) = (1-p)^{k-1} p,$$

Beispiel: geometrische Verteilung

und daher für beliebiges $A \in \mathcal{B}$

$$\mu(A) = p \sum_{k \in A} (1-p)^{k-1}.$$

Die Verteilung von $Y = X - 1$ heisst die **geometrische Verteilung**.

In diesem Beispiel tritt ∞ als möglicher Wert auf, und es ist von Vorteil, die Theorie so zu verallgemeinern, dass das auch zugelassen ist.

Verteilungsfunktion

Definition

Die durch

$$F(b) = \mathbb{P}(X \leq b) = \mu\big((-\infty, b]\big) \quad (b \in \mathbb{R})$$

definierte Funktion heisst **Verteilungsfunktion** von X bzw. von μ.

Rechenregeln

Aus den bekannten Rechenregeln folgt

$$\begin{aligned}
\mu\big((a,b]\big) &= F(b) - F(a) \quad (a < b) \\
\mu(\{a\}) &= \mu\left(\bigcap_{n=1}^{\infty}(a - \frac{1}{n}, a]\right) = \lim_n \mu\left((a - \frac{1}{n}, a]\right) \\
&= F(a) - F(a-) \qquad (= \text{ Sprunghöhe in } a\,.
\end{aligned}$$

Es gilt sogar, dass man aus F die Verteilung μ, d.h. $\mu(A)$ für alle $A \in \mathcal{B}$, erhalten kann.

Struktursatz

Satz

Jede Verteilungsfunktion hat die folgenden Eigenschaften

i) *Monotonie:* $a \leq b \Rightarrow F(a) \leq F(b)$

ii) *Rechtsstetigkeit:* $F(a) = \lim_{h \downarrow 0} F(a+h)$

iii) *Normierung:* $\lim_{a \to -\infty} F(a) = 0, \lim_{a \to +\infty} F(a) = 1.$

Umgekehrt ist jede Funktion mit diesen 3 Eigenschaften Verteilungsfunktion einer Zufallsvariablen.

Beweis

Für $a \leq b$ ist $(-\infty, a] \subseteq (-\infty, b]$, also $\mu\big((-\infty, a]\big) \leq \mu\big((-\infty, b]\big)$ und damit $F(a) \leq F(b)$, d.h. i).

Für $h_n \downarrow 0$ ist $\bigcap_n (a, a + h_n] = \emptyset$, also

$$0 = \mu\left(\bigcap_n (a, a + h_n]\right) = \lim_{n \uparrow \infty} \mu\big((a, a + h_n]\big) = \lim_{n \uparrow \infty} (F(a + h_n) - F(a)),$$

d.h. ii). Der Beweis von iii) geht analog wie bei ii).

Beweis

Für die Umkehrung definieren wir für $0 < t < 1$

$$F^{-1}(t) = \inf\{x | F(x) \geq t\}. \tag{105}$$

Es gilt

$$F^{-1}(t) \leq x \iff t \leq F(x). \tag{106}$$

Wir wählen nun als $(\Omega, \mathcal{A}, \mathbb{P})$ die Gleichverteilung auf $[0, 1]$ und setzen

$$X(\omega) = F^{-1}(\omega).$$

Dann ist

$$\mathbb{P}(X \leq b) = \mathbb{P}(\{\omega | X(\omega) \leq b\}) = \mathbb{P}\left(\{\omega | F^{-1}(\omega) \leq b\}\right) = \mathbb{P}(\{\omega | \omega \leq F(b)\})$$

d.h. F ist die Verteilungsfunktion von X. □

Quantile

Lemma

Wenn F i) - iii) von oben erfüllt sind und F^{-1} wie in (105) definiert ist, dann ist F^{-1} monoton wachsend, linksstetig und es gilt

i) $F^{-1}(F(x)) \leq x \quad (-\infty < x < \infty)$

ii) $t \leq F(F^{-1}(t)) \quad (0 < t < 1).$

Beweis

Wegen der Voraussetzungen i) und ii) ist

$$\{x|F(x) \geq t\} = [F^{-1}(t), \infty).$$

Damit ist die Monotonie von F^{-1} und $F(F^{-1}(t)) \geq t$ klar. Für $h_n \downarrow 0$ ist

$$\bigcap_n \{x|F(x) \geq t - h_n\} = \{x|F(x) \geq t\},$$

also ist F^{-1} linksstetig. $F^{-1}(F(x)) \leq x$ folgt schliesslich aus $x \in \{x'|F(x') \geq F(x)\}$. □

Bemerkung

Die hier durchgeführte Konstruktion einer Zufallsvariable mit vorgegebener Verteilungsfunktion ist auch von praktischer Bedeutung, nämlich bei der Erzeugung von Zufallszahlen mit beliebiger Verteilung aus der Gleichverteilung.

Gemäss der Definition, bzw. dem Lemma gilt

$$\mathbb{P}\left(X < F^{-1}(t)\right) \leq t \leq \mathbb{P}\left(X \leq F^{-1}(t)\right).$$

Man nennt daher die Grösse $F^{-1}(t)$ auch das t-**Quantil** von X. Wichtig ist insbesondere das 50%-Quantil, der sogenannte **Median**.

Typen von Verteilungen

Eine Zufallsvariable X heisst **diskret**, wenn es eine abzählbare Menge $A \subset \mathbb{R}$ gibt sodass $\mathbb{P}(X \in A) = 1$. Dann ist die Verteilungsfunktion

$$F(b) = \sum_{x \in A; x \leq b} \mathbb{P}(X = x) = \sum_{x \in A; x \leq b} \mathbb{P}(\{\omega \in \Omega \mid X(\omega) = x\})$$

bzw

$$F(b) = \sum_{x \in X(\Omega); x \leq b} \mathbb{P}(X = x)$$

eine Treppenfunktion mit Sprungstellen in A und Sprunghöhen $\mathbb{P}(X = x)$.

Typen von Verteilungen

Eine Zufallsvariable (bzw. deren Verteilung) heisst **absolut stetig** (genauer gesagt: bezüglich des Lebesguemasses), falls eine messbare Funktion $f : (\mathbb{R}, \mathcal{B}) \to (\mathbb{R}, \mathcal{B})$ existiert mit $f(x) \geq 0$ und $\int_{-\infty}^{\infty} f(x)dx = 1$, so dass

$$F(b) = \int_{-\infty}^{b} f(x)dx. \tag{107}$$

Die Funktion f heisst die **Dichte** von X. Jede Funktion der Form (107) erfüllt offensichtlich i) und iii) von Satz 31 und ist stetig. Letzteres ist klar für stückweise stetiges f, allgemein folgt es aus dem Konvergenzsatz von Lebesgue. Falls f stetig an der Stelle x ist, dann $f(x) = F'(x)$. Der Satz von Lebesgue besagt, dass F ohne zusätzliche Voraussetzungen sogar fast überall differenzierbar ist mit Ableitung f.

Uniform auf $[a, b]$: Bezeichnung $\mathcal{U}(a, b)$

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{sonst.} \end{cases}$$

$$F(x) = \begin{cases} 0 & x \leq a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x \geq b. \end{cases}$$

Die uniforme Verteilung wird eigentlich immer verwendet.

Exponential mit Parameter $\alpha > 0$: Bezeichnung: exp(α)

$$f(x) = \begin{cases} \alpha e^{-\alpha x} & x \geq 0 \\ 0 & x < 0. \end{cases}$$

$$F(x) = \begin{cases} 1 - e^{-\alpha x} & x \geq 0 \\ 0 & x < 0. \end{cases}$$

Die Exponentialverteilung wird für Warte- oder Überlebenszeiten verwendet.

Normal mit Parametern μ, σ^2: Bezeichnung: $\mathcal{N}(\mu, \sigma^2)$

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

(Dass $\frac{1}{\sqrt{2\pi}}$ die richtige Normierung ist, folgt aus der Analysis.) Hier ist $F(x)$ nicht in geschlossener Form darstellbar. Es gilt aber

$$F_{\mu,\sigma^2}(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy = \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = F_{0,1}(\frac{x-\mu}{\sigma}).$$

Die Dichte $f_{0,1}$ heisst die Standardnormalverteilungsdichte und wird meist mit φ bezeichnet. Die Verteilungsfunktion $F_{0,1}$ wird meist mit Φ bezeichnet und ist tabelliert.

Bemerkung

Die Normalverteilungsdichte hatten wir schon im Satz von de Moivre-Laplace als Approximation für die Binomialverteilung kennengelernt. Die Normalverteilung wird verwendet für Messfehler und andere Grössen, die man als Überlagerung vieler kleiner Effekte betrachten kann.

Nicht absolut stetige Verteilungen

Sei $(\Omega, \mathcal{A}, \mathbb{P})$ der Wahrscheinlichkeitsraum für unendlich viele unabhängige $0-1$ Experimente). Für $\omega = (x_1, x_2, \dots,)$ definieren wir die Zufallsvariable

$$X(\omega) = \sum_{k=1}^{\infty} x_k 2^{-k},$$

die Umkehrung der Abbildung φ. Um die Verteilungsfunktion von X zu bestimmen, schreiben wir $b \in (0,1)$ in der Binärdarstellung $\sum_{k=1}^{\infty} b_k 2^{-k}$ mit $b_k \in \{0,1\}$ (bei Mehrdeutigkeit nehmen wir die Version mit mit unendlich vielen $b_k = 0$). Dann ist $X(\omega) \leq b$ genau dann, wenn ein n existiert mit $x_k = b_k$ für $k < n$ und $x_n = 0,\ b_n = 1$ oder wenn $x_k = b_k$ für alle k.

Nicht absolut stetige Verteilungen

Also gilt mit $S_n = \sum_{k=1}^{n} b_k$

$$F(b) = \mathbb{P}\left(\bigcup_{n \geq 1; b_n = 1} \{X_1 = b_1, \ldots, X_{n-1} = b_{n-1}, X_n = 0\} \right) = \sum_{n=1}^{\infty} b_n p^{S_{n-1}} q^{n - S_{n-1}}.$$

Für $p = \frac{1}{2}$ erhalten wir insbesondere

$$\mathbb{P}(X \leq b) = \sum_{n=1}^{\infty} b_n 2^{-n} = b,$$

d.h. die Verteilung von X ist das Lebesguemass (insbesondere ist sie also absolut stetig).

Nicht absolut stetige Verteilungen

Für $p \neq \frac{1}{2}$ zeigt die folgende Überlegung, dass die Verteilungsfunktion stetig, aber nicht absolut stetig ist: Wenn eine Dichte f_p existieren würde, hätte man für jedes $A \in \mathcal{B}$

$$\mathbb{P}_p(X \in A) = \int_A f_p(x)dx.$$

Insbesondere ist also $\mathbb{P}_p(X \in A) = 0$ für alle A mit $\mathbb{P}_{\frac{1}{2}}(X \in A) = 0$. Aus dem starken Gesetz der grossen Zahlen folgt aber, dass es ein A gibt mit

$$\mathbb{P}_p(X \in A) = 1, \quad \mathbb{P}_{\frac{1}{2}}(X \in A) = 0.$$

Die Stetigkeit von F (für jedes $p \in (0,1)$) ist leicht einzusehen, denn es gibt höchstens zwei ω mit $X(\omega) = b$, also $\mathbb{P}(X = b) = 0$ für alle b. In der Masstheorie wird gezeigt, dass F sogar fast überall (bezüglich des Lebesguemasses) differenzierbar ist, aber diese Ableitung ist fast überall gleich null, und damit ist F nicht gleich dem Integral der Ableitung.

Transformation von Zufallsvariablen

Sei X eine Zufallsvariable auf $(\Omega, \mathcal{A}, \mathbb{P})$ und $g : (\mathbb{R}, \mathcal{B}) \to (\mathbb{R}, \mathcal{B})$ messbar. Dann ist leicht einzusehen, dass

$$Y(\omega) = g(X(\omega))$$

wieder eine Zufallsvariable ist. Sie hat die Verteilungsfunktion

$$F_Y(b) = \mathbb{P}\left(g(X) \leq b\right) = \mathbb{P}\left(X \in g^{-1}((-\infty, b])\right).$$

Beispiel

Sei $g(x) = x^2$. Dann ist

$$F_Y(b) = \mathbb{P}\left(-\sqrt{b} \leq X \leq \sqrt{b}\right) = F_X(\sqrt{b}) - F_X(-\sqrt{b})$$

falls F_X stetig. Falls F_X absolut stetig ist, dann ist auch Y absolut stetig, und zwar ist die Dichte

$$f_Y(b) = \frac{1}{2\sqrt{b}}(f_X(\sqrt{b}) + f_X(-\sqrt{b})).$$

Beispiel

Wenn $g(x) = ax + b$ mit $a > 0$ dann ist

$$F_Y(x) = F_X\left(\frac{x-b}{a}\right).$$

Falls eine Dichte f_X existiert, dann existiert auch f_Y und

$$f_Y(x) = \frac{1}{a} f_X(\frac{x-b}{a}).$$

Durch lineare Transformationen lassen sich also insbesondere die Verteilungen $\exp \alpha$ auf $\exp 1$, $\mathcal{U}(a,b)$ auf $\mathcal{U}(0,1)$ und $\mathcal{N}(\mu, \sigma^2)$ auf $\mathcal{N}(0,1)$ zurückführen.

Beispiel

Wenn g monoton wachsend und differenzierbar ist mit $g'(x) > 0$ für alle x, dann ist $F_Y(b) = F_X(g^{-1}(b))$. Falls die Dichte f_X existiert, dann existiert auch f_Y, und zwar ist

$$f_Y(x) = \frac{1}{g'(g^{-1}(x))} f_X(g^{-1}(x)).$$

Erwartungswert

Sei X eine Zufallsvariable auf $(\Omega, \mathcal{A}, \mathbb{P})$ mit Verteilung μ. Der Erwartungswert ist eine Kennzahl für die Lage der Verteilung von X; er gibt an, welchen Wert man im Mittel bei vielen unabhängigen Realisierungen erhält (vgl. das Gesetz der grossen Zahlen).

Definition

Für $X \geq 0$ ist der **Erwartungswert**

$$\mathbb{E}(X) = \int_\Omega X(\omega) d\mathbb{P}(\omega) = \int_{\mathbb{R}} x \mu(dx) \in [0, \infty].$$

Erwartungswert

Diese Integrale sind im Sinne der Masstheorie zu verstehen. Die Definition geht schrittweise: Zunächst betrachtet man einfache Zufallsvariable $X = \sum_{i=1}^{n} c_i 1_{A_i}(\omega)$ mit $A_i \in \mathcal{A}$ und setzt $\mathbb{E}(X) = \sum c_i \mathbb{P}(A_i)$. Dann schreibt man ein beliebiges nichtnegatives X als aufsteigenden Limes von einfachen Zufallsvariablen X_n und setzt $\mathbb{E}(X) = \lim_n \mathbb{E}(X_n)$. (Man muss zeigen, dass dies nicht davon abhängt, welche Darstellung man wählt).

Für eine Zufallsvariable X, die positive und negative Werte annimmt, setzen wir

$$X^+(\omega) = \max(X(\omega), 0) \quad , \quad X^-(\omega) = \max(-X(\omega), 0)$$

und definieren

$$\mathbb{E}(X) = \mathbb{E}\left(X^+\right) - \mathbb{E}\left(X^-\right),$$

sofern nicht beide Terme rechts $= +\infty$ sind.

Wenn X diskret ist, dann ist $\mathbb{E}(X) = \sum_{x_i \in X(\Omega)} x_i \mathbb{P}(X = x_i)$.

Erwartungswert

Wenn die Verteilung von X absolut stetig ist mit stückweise stetiger Dichte, dann ist

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

wobei das Integral im Riemann-Sinn genommen werden kann.

Eigenschaften des Erwartungswertes:

Linearität:

$$\mathbb{E}\left(\alpha_1 X_1 + \alpha_2 X_2\right) = \alpha_1 \mathbb{E}\left(X_1\right) + \alpha_2 \mathbb{E}\left(X_2\right). \tag{108}$$

Monotonie:

$$X \leq Y \Rightarrow \mathbb{E}\left(X\right) \leq \mathbb{E}\left(Y\right). \tag{109}$$

Monotone Stetigkeit:

$$0 \leq X_1 \leq X_2 \leq \cdots \Rightarrow \mathbb{E}\left(\lim_n X_n\right) = \lim_n \mathbb{E}\left(X_n\right) \tag{110}$$

Konvergenzsatz von Lebesgue: Sei $X_1, X_2, \ldots$ eine f.s. konvergente Folge von Zufallsvariablen. Wenn $|X_n(\omega)| \leq X(\omega)$ für alle n und $\mathbb{E}\left(X\right) < \infty$, dann

$$\mathbb{E}\left(\lim_n X_n\right) = \lim_n \mathbb{E}\left(X_n\right) \tag{111}$$

Transformation von Zufallsvariablen

Sei $g : (\mathbb{R}, \mathcal{B}) \to (\mathbb{R}, \mathcal{B})$ messbar und $Y(\omega) = g(X(\omega))$. Zur Berechnung von $\mathbb{E}(Y)$ könnte man zuerst die Verteilung von Y und dann $\mathbb{E}(Y) = \int x \mu_Y(dx)$ berechnen. Es geht aber einfacher:

$$\mathbb{E}(Y) = \int g(x) \mu_X(dx) = \begin{cases} \sum_{x_i \in X(\Omega)} g(x_i) \mathbb{P}(X = x_i) & (X \text{ diskret}) \\ \int g(x) f(x) dx & (X \text{ absolut stetig}). \end{cases}$$

Transformation von Zufallsvariablen

Insbesondere können wir also

$$\mathbb{E}(X^p) \qquad p = 1, 2, 3 \ldots \ (p - \text{tes Moment}) \tag{112}$$

$$\mathbb{E}(|X|^p) \qquad p > 0 \quad (p - \text{tes absolutes Moment}) \tag{113}$$

$$\mathbb{E}((X - \mathbb{E}(X))^p) \qquad p = 1, 2, 3 \ldots \ (p - \text{tes zentriertes Moment}) \tag{114}$$

direkt aus der Verteilung von X berechnen.

Varianz

Das zweite zentrierte Moment heisst die **Varianz** von X:

$$\operatorname{var}(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right) = \mathbb{E}\left(X^2\right) - (\mathbb{E}(X))^2 . \tag{115}$$

Die Wurzel der Varianz heisst die **Standardabweichung**:

$$\sigma(X) = \sqrt{\operatorname{var}(X)}. \tag{116}$$

Die Standardabweichung misst die Streuung von X um $\mathbb{E}(X)$ und ist damit eine wichtige Kennzahl der Verteilung von X. Wegen (115) gilt

$$\operatorname{var}(aX + b) = a^2 \operatorname{var}(X), \quad \sigma(aX + b) = |a|\sigma(X). \tag{117}$$

$\mathcal{L}^p$-Räume

Sei $\mathcal{L}^p = \mathcal{L}^p(\Omega, \mathcal{A}, \mathbb{P})$ die Menge der Zufallsvariablen X auf $(\Omega, \mathcal{A}, \mathbb{P})$ mit $\mathbb{E}(|X|^p) < \infty\}$. Durch

$$||X||_p = \mathbb{E}(|X|^p)^{1/p}$$

ist für $p \geq 1$ eine Halbnorm auf $\mathcal{L}^p$ definiert. Wenn wir Zufallsvariable X und Y mit $X = Y$ $\mathbb{P}-$ f.s. identifizieren, wird $\mathcal{L}^p$ zu einem **Banachraum** und $\mathcal{L}^2$ zu einem **Hilbertraum**.

Erwartungswert und Varianz der wichtigsten Verteilungen

Verteilung	$\mathbb{E}(X)$	$\mathrm{var}(X)$
Binomial (n, p)	np	$np(1-p)$
Hypergeometrisch (n, N, K)	$n\frac{K}{N}$	$n\frac{K}{N}(1-\frac{K}{N})\frac{N-n}{N-1}$
Poisson (λ)	λ	λ
Geometrisch (p)	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Uniform (a, b)	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential(α)	$\frac{1}{\alpha}$	$\frac{1}{\alpha^2}$
Normal (μ, σ^2)	μ	σ^2

Beweisskizze

Für $X \sim \text{Poisson}\,(\lambda)$ ist

$$\begin{aligned}\mathbb{E}\left(X^2\right) &= \sum_{k=0}^{\infty} k^2 \mathbb{P}\left(X=k\right) = \sum_{k=0}^{\infty} (k(k-1)+k)\mathbb{P}\left(X=k\right) \\ &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{k(k-1)}{k!}\lambda^k + \lambda = \lambda^2 + \lambda\end{aligned}$$

Also folgt $\text{var}(X) = \mathbb{E}\left(X^2\right) - \mathbb{E}\left(X\right)^2 = \lambda$.

Für $X \sim \mathcal{U}(0,1)$ ist $\mathbb{E}\left(X\right) = \frac{1}{2}$ aus Symmetrie. Ferner

$$\mathbb{E}\left(X^2\right) = \int_0^1 x^2 dx = \frac{1}{3},$$

also $\text{var}(X) = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$. Die Resultate für eine beliebige uniforme Verteilung folgen durch affine Transformation.

Beweisskizze

Für $X \sim \exp(1)$ ist mit partieller Integration

$$\mathbb{E}\left(X^k\right) = \int_0^\infty x^k e^{-x} dx = k!,$$

also $\mathbb{E}(X) = 1$ und $\operatorname{var}(X) = 2 - 1^2 = 1$.

Ungleichungen

Für eine nichtlineare Funktion g ist im Allgemeinen

$$\mathbb{E}\,(g(X)) \neq g(\mathbb{E}\,(X)).$$

Für ein konvexes oder konkaves g hat man wenigstens eine Ungleichung. Eine Funktion $g : \mathbb{R} \to \mathbb{R}$ heisst konvex, falls es zu jedem x_0 eine Stützgerade $\ell(x) = ax + b$ gibt mit

$$\ell(x) \leq g(x) \quad \text{für alle } x, \quad \ell(x_0) = g(x_0).$$

Jensen Ungleichung

Satz

Für eine Zufallsvariable X mit endlichem Erwartungswert und $g : \mathbb{R} \to \mathbb{R}$ konvex gilt

$$\mathbb{E}\left(g(X)\right) \geq g(\mathbb{E}\left(X\right)).$$

Beweis.

Sei ℓ die Stützgerade für $x_0 = \mathbb{E}\left(X\right)$. Dann gilt

$$g(\mathbb{E}\left(X\right)) = \ell(\mathbb{E}\left(X\right)) = \mathbb{E}\left(\ell(X)\right) \leq \mathbb{E}\left(g(X)\right).$$

Chebychev Ungleichung

Die folgende Ungleichung ist äusserst nützlich für Grenzwertsätze.

Satz (verallgemeinerte Chebyshev-Ungleichung)

Sei g eine nichtnegative, monoton wachsende Funktion auf $\mathbb{R}$. Dann gilt für jedes c mit $g(c) > 0$

$$\mathbb{P}(X \geq c) \leq \frac{\mathbb{E}(g(X))}{g(c)}.$$

Beweis.

Offensichtlich ist

$$1_{[X \geq c]} \leq \frac{g(X)}{g(c)}.$$

Also folgt die Behauptung. □ □

Beispiel

Wenn wir die Chebyshev Ungleichung auf $Y = |X|$ und $g(x) = \max(x, 0)$ anwenden, dann folgt

$$\mathbb{P}(|X| > c) \leq \frac{\mathbb{E}(|X|)}{c}.$$

Insbesondere impliziert $\mathbb{E}(|X|) = 0$, dass $\mathbb{P}(X = 0) = 1$, da

$$\mathbb{P}(X = 0) = \mathbb{P}\left(\bigcap_{n=1}^{\infty} \{|X| \leq \frac{1}{n}\}\right) = \lim_n \mathbb{P}\left(|X| \leq \frac{1}{n}\right).$$

Wenn wir die Chebyshev Ungleichung auf $Y = |X - \mathbb{E}(X)|$ und $g(x) = (\max(x, 0))^2$, dann folgt

$$\mathbb{P}(|X - \mathbb{E}(X)| > c) \leq \frac{\mathrm{var}(X)}{c^2}.$$

Insbesondere impliziert $\mathrm{var}(X) = 0$, dass X f.s. konstant ist.

Diese Ungleichungen sind zwar sehr einfach und universell gültig, dafür aber in spezifischen Fällen oft recht grob.

Multivariate Zufallsvariablen

Seien $X_1, X_2, \ldots, X_n$ Zufallsvariablen auf einem gemeinsamen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$. Wir betrachten den Zufallsvektor

$$\mathsf{X} = (X_1, X_2, \ldots, X_n). \tag{118}$$

Auf $\mathbb{R}^n$ sei $\mathcal{B}^n = \sigma(\{A_1 \times A_2 \times \cdots \times A_n | A_i \in \mathcal{B}\})$ die Borel-σ-Algebra. Mengen der Form $A_1 \times \cdots \times A_n$ nennen wir auch (verallgemeinerte) *Rechtecke*. Dann ist die Abbildung

$$\mathsf{X} : (\Omega, \mathcal{A}) \longrightarrow (\mathbb{R}^n, \mathcal{B}^n)$$

automatisch messbar. Das Bild μ von $\mathbb{P}$ unter X heisst die **gemeinsame Verteilung** von $X_1, \ldots, X_n$:

$$\mu(A) = \mathbb{P}\left(\mathsf{X}^{-1}(A)\right) = \mathbb{P}\left(\{\omega | \mathsf{X}(\omega) \in A\}\right) = \mathbb{P}[\mathsf{X} \in A] \quad (A \in \mathcal{B}^n). \tag{119}$$

Multivariate Zufallsvariablen

Wenn jedes X_i **diskret** ist, dann ist eine höchstens abzählbare Menge $B \subset \mathbb{R}^n$ gibt sodass $\mathbb{P}(X = B) = 1$ und

$$\mu(A) = \sum_{x \in B \cap A} \mathbb{P}(X = x) = \sum_{(x_1,\ldots,x_n) \in B \cap A} \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n). \quad (120)$$

Wenn die gemeinsame Verteilung **absolut stetig** ist, d.h. wenn es eine messbare Funktion $f : \mathbb{R}^n \to \mathbb{R}$ mit $f \geq 0$ und $\int_{\mathbb{R}^n} f(x)dx = 1$ derart gibt, dass

$$\mu(A) = \int_A f(x)dx. \quad (121)$$

Die Funktion f heisst wie im eindimensionalen Fall die Dichte.

Beispiel

Sei $A \in \mathcal{B}^n$ mit $0 < \lambda(A) < \infty$, wobei λ das Lebesguemass bezeichnet. Dann ist die **Gleichverteilung** auf A:

$$\mu_A(B) = \frac{\lambda(B \cap A)}{\lambda(A)} \tag{122}$$

eine absolut stetige Verteilung. Die Dichte ist konstant gleich $1/\lambda(A)$ auf A und null ausserhalb.

Neben diesen beiden Typen gibt es aber noch andere Verteilungen, z.B. die Gleichverteilung auf $\{x \in \mathbb{R}^n \mid ||x|| = 1\}$.

Marginalverteilungen

Aus der gemeinsamen Verteilung erhält man insbesondere auch die Verteilung von jedem X_i allein, die sogenannte (eindimensionale) **Rand-** oder **Marginalverteilung:**

$$\mu_i(B) = \mathbb{P}(X_i \in B) = \mu(\mathbb{R} \times \mathbb{R} \times \cdots \times \underbrace{B}_{i-te\ Stelle} \times \ldots \mathbb{R}) \ \ (B \in \mathcal{B}). \ \ (123)$$

Marginalverteilungen

Speziell für μ diskret

$$\mathbb{P}(X_i = x_i) = \sum_{(x_1,\ldots,x_i,\ldots x_n)\in B} \mathbb{P}(X_1 = x_1, \ldots, X_i = x_i, \ldots, X_n = x_n), \quad (124)$$

und für μ absolut stetig

$$\mu_i(B) = \int_{\mathbb{R}} \cdots \int_B \cdots \int_{\mathbb{R}} f(\mathrm{x}) d\mathrm{x}.$$

Das heisst, μ_i ist ebenfalls absolut stetig mit Dichte

$$f_i(x_i) = \int_{\mathbb{R}^{n-1}} f(\mathrm{x}) \quad dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_n. \quad (125)$$

Umgekehrt kann man aber aus den Randverteilungen nicht die gemeinsame Verteilung bestimmen.

Beispiel

Sei $n = 2$, $X_i(\Omega) = \{0, 1\}$ und α ein Parameter mit $-\frac{1}{4} \leq \alpha \leq \frac{1}{4}$. Wir setzen

$$\begin{array}{rcrcl} \mathbb{P}(X_1 = 0, X_2 = 0) & = & \mathbb{P}(X_1 = 1, X_2 = 1) & = & \frac{1}{4} + \alpha \\ \mathbb{P}(X_1 = 1, X_2 = 0) & = & \mathbb{P}(X_1 = 0, X_2 = 1) & = & \frac{1}{4} - \alpha \end{array}$$

Dann ist $\mathbb{P}(X_i = 0) = \mathbb{P}(X_i = 1) = \frac{1}{2}$ $(i = 1, 2)$ für jedes α.

Die gemeinsame Verteilung enthält eben noch zusätzliche Information, nämlich solche über die Abhängigkeiten zwischen den Variablen.

Bedingungen

Seien $X_1, X_2, \ldots, X_n$ Zufallsvariablen auf einem gemeinsamen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$. Wir betrachten den Zufallsvektor

$$\mathrm{X} = (X_1, X_2, \ldots, X_n). \tag{126}$$

Seien $g : \mathbb{R}^{n-1} \to \mathbb{R}$, $h : \mathbb{R} \to \mathbb{R}$ messbare Funktionen, dann gilt unter Verwendung der bedingten Dichte

$$f_{X|X_n}(x_1, \ldots, x_{n-1}, x_n) := \frac{f(x_1, \ldots, x_n)}{\int f(x_1, \ldots, x_{n-1}, x_n) dx_1 \ldots dx_{n-1}}$$

im absolut stetigen Fall (im diskreten Fall tritt sie natürlich analog auf, aber wird zur Schärfung der Intuition nicht verwendet)

$$\mathbb{E}\left(g(X_1, \ldots, X_{n-1})h(X_n)\right) = \begin{cases} \sum g(x_1, \ldots, x_{n-1}) \frac{\mathbb{P}\left(X_1 = x_1, \ldots, X_{n-1} = x_{n-1}, X_n = x_n\right)}{\mathbb{P}(X_n = x_n)} h(x_n) \mathbb{P}\left(X_n = x_n\right) & (X \text{ diskret}) \\ \int g(x_1, \ldots, x_{n-1}) f_{X|X_n}(x_1, \ldots, x_{n-1}, x_n) dx_1 \ldots dx_{n-1} h(x_n) \mu_n(dx_n) & (X \text{ absolut stetig}) \end{cases}$$

Unabhängigkeit

Definition

$X_1, \ldots, X_n$ heissen (stochastisch) **unabhängig** falls für alle $A_1, \ldots, A_n \in \mathcal{B}$ gilt:

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \cdot \cdots \cdot \mathbb{P}(X_n \in A_n)$$

bzw.

$$\mu(A_1 \times \cdots \times A_n) = \mu_1(A_1) \cdot \cdots \cdot \mu_n(A_n).$$

(Man sagt auch, μ ist das Produkt von $\mu_1, \ldots, \mu_n$).

Im Fall der Unabhängigkeit ist die gemeinsame Verteilung durch die Randverteilungen festgelegt (denn μ ist festgelegt, durch die Werte auf den verallgemeinerten Rechtecken).

Unabhängigkeit absolut stetiger Zufallsvariablen

Satz

Seien $X_1, \ldots, X_n$ unabhängig. Dann ist μ absolut stetig genau dann, wenn jedes μ_i absolut stetig ist. Ferner gilt $f(\mathrm{x}) = \prod_{i=1}^{n} f_i(x_i)$.

Der Beweis erfolgt mit dem Satz von Fubini.

Beispiel

Standard-Normalverteilung im $\mathbb{R}^n$. Seien $X_1, \ldots, X_n$ unabhängig und $\mathcal{N}(0,1)$-verteilt. Dann hat die gemeinsame Verteilung die Dichte

$$f(x_1, \ldots, x_n) = (2\pi)^{-n/2} \exp(-\frac{1}{2} \sum_{i=1}^{n} x_i^2).$$

Die Dichte ist also sphärisch symmetrisch, und man kann zeigen, dass es die einzige sphärisch symmetrische Verteilung ist, bei der die Komponenten unabhängig sind.

Daraus hat Maxwell geschlossen, dass der Geschwindigkeitsvektor eines Gasmoleküls diese Verteilung haben muss.

Transformationen

Sei X ein $n-$dimensionaler Zufallsvektor und $g : (\mathbb{R}^n, \mathcal{B}^n) \to (\mathbb{R}^m, \mathcal{B}^m)$ eine messbare Abbildung. Dann ist

$$\mathrm{Y}(\omega) = g(\mathrm{X}(\omega)) \tag{127}$$

ein $m-$dimensionaler Zufallsvektor. Ferner gilt

$$\mu_{\mathrm{Y}}(A) = \mu_{\mathrm{X}}(g^{-1}(A)).$$

Je nach der Art der Funktion g, kann man μ_{Y} mehr oder weniger explizit angeben.

Transformationssatz

Satz

Sei $g : \mathrm{R}^n \to \mathrm{R}^n$ *linear und umkehrbar, d.h.* $g(\mathrm{x}) = \mathrm{m} + B\mathrm{x}$ *mit* $\det(B) \neq 0$. *Wenn* μ_{X} *absolut stetig ist, dann ist auch* μ_{Y} *absolut stetig und es gilt:*

$$f_{\mathrm{Y}}(\mathrm{x}) = \frac{1}{|\det(B)|} f_{\mathrm{X}}(B^{-1}(\mathrm{x} - \mathrm{m})). \tag{128}$$

Beweis.

Mithilfe einer Substitution erhält man

$$\mu_{\mathrm{Y}}(A) = \mu_{\mathrm{X}}[g^{-1}(A)] = \int_{g^{-1}(A)} f_{\mathrm{X}}(\mathrm{x})d\mathrm{x} = \int_A f_{\mathrm{X}}(g^{-1}(\mathrm{x})) \frac{1}{|\det(B)|} d\mathrm{x}.$$

□ □

Beispiel

Wenn X standard-normalverteilt ist, dann

$$f_Y(x) = (2\pi)^{-n/2} \frac{1}{\sqrt{|\det \Sigma|}} \exp\left(-\frac{1}{2}(x-m)^\top \Sigma^{-1}(x-m)\right) \tag{129}$$

wobei $\Sigma = BB^\top$ und wir x als Spaltenvektor auffassen. Dies ist die allgemeine nicht degenerierte (da $\det \Sigma \neq 0$) ***n*-dimensionale Normalverteilung** $\mathcal{N}_n(m, \Sigma)$.

Beispiel

Verteilung von Summen: Sei $\mathsf{X} = (X_1, X_2)$ eine zweidimensionale absolut stetige Zufallsvariable. Wir setzen $\mathsf{Z} = (X_1, Y)$ mit $Y = X_1 + X_2$.

Folglich $f_{\mathsf{Z}}(x_1, y) = f_{\mathsf{X}}(x_1, y - x_1)$. Also erhalten wir, dass die Verteilung der Summe $Y = X_1 + X_2$ absolut stetig ist mit Dichte

$$f_Y(y) = \int_{\mathbb{R}} f_{\mathsf{Z}}(x_1, y) dx_1 = \int_{\mathbb{R}} f_{\mathsf{X}}(x_1, y - x_1) dx_1.$$

Wenn X_1 und X_2 unabhängig sind, hat man

$$f_Y(y) = \int_{-\infty}^{\infty} f_1(x_1) f_2(y - x_1) dx_1 \tag{130}$$

d.h. f_Y ist die **Faltung** von f_1 und f_2.

Beispiel für Bedingungen

Sei X eine n dimensionale Zufallsvariable mit Dichte

$$f_{\mathsf{X}}(\mathsf{x}) = (2\pi)^{-n/2} \frac{1}{\sqrt{|\det \Sigma|}} \exp\left(-\frac{1}{2}(\mathsf{x}-\mathsf{m})^{\top}\Sigma^{-1}(\mathsf{x}-\mathsf{m})\right) \tag{131}$$

wobei Σ invertierbar ist und wir x als Spaltenvektor auffassen, dann ist die bedingte Dichte $f_{X|X_n}$ wieder normalverteilt
$\mathcal{N}(\mathsf{m}_{|x_n}, \Sigma_{|x_n} := ((\Sigma^{-1})_{i,j\leq n-1})^{-1})$ (man sammle einfach die Koeffizienten der Quadrate und des linearen Teils), wobei für $1 \leq i \leq n-1$

$$(\mathsf{m}_{|x_n})_i = \sum_{j=1}^{n-1} (\Sigma_{|x_n})_{ij} \Big(\sum_{k=1}^{n} (\Sigma^{-1})_{kj} m_k - (\Sigma^{-1})_{nj}(x_n - m_n) \Big)$$

gilt. Auch die Marginaldichte von X_n is normalverteilt $\mathcal{N}(m_{|X_n}, \Sigma_{|X_n})$.

Kovarianz und Korrelation

Sei $g : (\mathbb{R}^n, \mathcal{B}^n) \to (\mathbb{R}, \mathcal{B})$ messbar. Zur Berechnung von $\mathbb{E}(g(X))$ muss man die Verteilung von $Y = g(X)$ nicht bestimmen, sondern es gilt wie im eindimensionalen Fall

$$\mathbb{E}(g(X)) = \int_{\mathbb{R}^n} g(x)\mu(dx). \tag{132}$$

Im diskreten, bzw. absolut stetigen Fall, heisst das

$$= \sum_{(x_1,\ldots,x_n)\in B} g(x_1,\ldots,x_n)\, \mathbb{P}(X_1 = x_1,\ldots,X_n = x_n), \text{ bzw.} \tag{133}$$

$$= \int_{\mathbb{R}^n} g(x)f(x)dx. \tag{134}$$

Kovarianz und Korrelation

Insbesondere können wir das benutzen, um die Kovarianz von zwei Zufallsvariablen zu berechnen:

Definition

Die **Kovarianz** von X_1 und X_2 ist definiert als

$$\operatorname{cov}(X_1, X_2) = \mathbb{E}\left(\Big(X_1 - \mathbb{E}(X_1)\Big)\Big(X_2 - \mathbb{E}(X_2)\Big)\right)$$

Eigenschaften

Satz

Die Kovarianz hat die folgenden Eigenschaften

i) $\operatorname{cov}(X, X) = \operatorname{var}(X)$.

ii) $\operatorname{cov}(X_1, X_2) = \operatorname{cov}(X_2, X_1)$.

iii) $\operatorname{cov}(X_1, X_2) = \mathbb{E}(X_1 X_2) - \mathbb{E}(X_1)\,\mathbb{E}(X_2)$.

iv) $\operatorname{cov}(X_1, aX_2 + b) = a \operatorname{cov}(X_1, X_2)$.

v) $\operatorname{cov}(X_1, X_2 + X_3) = \operatorname{cov}(X_1, X_2) + \operatorname{cov}(X_1, X_3)$.

vi) $\operatorname{var}(X_1 + X_2) = \operatorname{var}(X_1) + \operatorname{var}(X_2) + 2\operatorname{cov}(X_1, X_2)$.

vii) $|\operatorname{cov}(X_1, X_2)| \leq \sigma(X_1)\sigma(X_2)$.

viii) *Wenn* X_1, X_2 *unabhängig sind, dann ist* $\operatorname{cov}(X_1, X_2) = 0$ *und daher* $\operatorname{var} X_1 + X_2 = \operatorname{var} X_1 + \operatorname{var} X_2$.

Beweis

Die ersten zwei Behauptungen sind offensichtlich aufgrund der Definition. Die nächsten vier Behauptungen folgen durch Ausrechnen und Anwenden der Regeln für den Erwartungswert. Die siebte Behauptung ist nichts anderes als die Cauchy-Schwarz-Ungleichung. Im diskreten Fall folgt die letzte Behauptung direkt. Analog erhalten wir im absolut stetigen Fall

$$\mathbb{E}(X_1 X_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 x_2 f_1(x_1) f_2(x_2) dx_1 dx_2 = \mathbb{E}(X_1)\,\mathbb{E}(X_2)\,.$$

Für den allgemeinen Fall braucht man Masstheorie. □

Beispiel

Eine binomial(n, p)-verteilte Zufallsvariable X lässt sich schreiben als $\sum_{i=1}^{n} X_i$ mit $X_1, \ldots, X_n$ unabhängig und binär. Also folgt
$\operatorname{var}(X) = \sum_{i=1}^{n} \operatorname{var}(X_i) = np(1-p)$, denn
$\operatorname{var}(X_i) = \mathbb{E}\left(X_i^2\right) - \mathbb{E}\left(X_i\right)^2 = p - p^2 = p(1-p)$.

Bemerkung

Die Umkehrung von (viii) ist falsch. Wenn z.B. $X_1 \sim \mathcal{N}(0,1)$ und $X_2 = X_1^2$, dann ist aus Symmetrie $\operatorname{cov}(X_1, X_2) = \mathbb{E}(X_1 X_2) = \mathbb{E}(X_1^3) = 0$. Hier sind aber X_1 und X_2 sehr stark abhängig.

Beispiel

Bei der n-dimensionalen Normalverteilung können wir die Kovarianzen $\operatorname{cov}(Y_i, Y_j)$ mit Hilfe der Regeln iv) und v) sofort berechnen, weil $\operatorname{cov}(X_i, X_j) = 0$ für $i \neq j$ und $\operatorname{cov}(X_i, X_i) = \operatorname{var} X_i = 1$. Man erhält $\operatorname{cov}(Y_i, Y_j) = (B^\top B)_{ij} = \Sigma_{ij}$.

Wenn Y eine n-dimensionale Normalverteilung hat und $\operatorname{cov}(Y_i, Y_j) = 0$ für $i \neq j$, dann zerfällt die gemeinsame Dichte in ein Produkt und damit sind die $Y_1, Y_2, \ldots, Y_n$ unabhängig. Die Umkehrung von Satz (viii) gilt daher bei gemeinsamer Normalverteilung.

Korrelation

Die Kovarianz verschiedener Paare von Zufallsvariablen lässt sich nicht direkt vergleichen, da sie auch von der Streuung der Variablen abhängt. Eine anschaulichere Kennzahl ist die **Korrelation**.

Definition

Die **Korrelation von** X_1 und X_2 ist

$$\rho(X_1, X_2) = \frac{\operatorname{cov}(X_1, X_2)}{\sigma(X_1)\sigma(X_2)}.$$

Korrelation

Wenn $\rho(X_1, X_2) = 0$ ($\Leftrightarrow \operatorname{cov}(X_1, X_2) = 0$) dann heissen X_1 und X_2 **unkorreliert**. Die Korrelation misst Stärke und Richtung des linearen Zusammenhangs zwischen den beiden Variablen. Es folgen sofort die Eigenschaften

$$\rho(aX_1 + b, cX_2 + d) = \rho(X_1, X_2) \text{ für } a > 0, \quad c > 0. \tag{135}$$

$$-1 \leq \rho(X_1, X_2) \leq +1. \tag{136}$$

Grenzwertsätze

Sei $X_1, X_2, \ldots$ eine Folge von Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$. Wir betrachten die Summen

$$S_n = X_1 + \cdots + X_n$$

und interessieren uns für das asymptotische Verhalten von S_n für $n \to \infty$.

Die **Gesetze der grossen Zahlen** beschreiben die Konvergenz der arithmetischen Mittel $\frac{1}{n} S_n$, während der **zentrale Grenzwertsatz** die Form der Verteilung der arithmetischen Mittel vom Grenzwert angibt. Der wichtigste Fall ist der, wo die X_i i.i.d. (unabhängig und identisch verteilt = **i**ndependent and **i**dentically **d**istributed) sind. Wir werden aber auch kurz diskutieren, inwieweit man auf die identische Verteilung verzichten kann. Was passiert, wenn die X_i's abhängig sind, ist ebenfalls untersucht worden; wir gehen aber darauf nicht ein.

Schwaches Gesetz der grossen Zahlen

Wir nehmen an, dass alle X_i einen gemeinsamen Erwartungswert $\mathbb{E}(X_i) = m$ haben. Wir sagen, dass das **schwache Gesetz der grossen Zahlen** gilt, falls für alle $\varepsilon > 0$

$$\mathbb{P}\left(\left|\frac{S_n}{n} - m\right| > \varepsilon\right) \longrightarrow 0 \quad \text{für } n \to \infty. \tag{137}$$

Mit der Chebyshev-Ungleichung folgt:

$$\mathbb{P}\left(\left|\frac{S_n}{n} - m\right| > \varepsilon\right) \leq \frac{\operatorname{var} S_n/n}{\varepsilon^2} = \frac{\operatorname{var} S_n}{n^2\varepsilon^2} \tag{138}$$

Wenn die X_i i.i.d. sind, dann $\operatorname{var} S_n = n \operatorname{var} X_1$ also gilt das schwache Gesetz der grossen Zahlen für X_i i.i.d. , $\mathbb{E}\left(X_i^2\right) < \infty$. Wenn die X_i unkorreliert sind, gilt $\operatorname{var} S_n = \sum_{i=1}^n \operatorname{var} X_i$ so dass $\sum_{i=1}^n \operatorname{var} X_i = o(n^2)$ hinreichend ist für das schwache Gesetz der grossen Zahlen.

Gegenbeispiel zum Gesetz der grossen Zahlen: die Cauchyverteilung

Sei (X_i) i.i.d. mit Dichte $f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$ (sog. **Cauchy-Verteilung**). Dann ist $\mathbb{E}(|X_i|) = \infty$, und mit Hilfe der Faltungsformel kann man zeigen, dass $\frac{S_n}{n}$ für alle $n \in \mathbb{N}$ wieder Cauchy-verteilt ist. Das heisst, dass $\frac{S_n}{n}$ immer gleich stark streut, die ausgleichende Wirkung des Zufalls spielt hier also nicht.

Eine Anwendung: Bernsteinpolynome

Wir verwenden das Gesetz der grossen Zahlen, um den **Satz von Weierstrass** zu beweisen, dass die Menge der Polynome dicht ist in $C[0,1]$, versehen mit der Supremums-Norm. Die **Bernsteinpolynome** vom Grad n auf $[0,1]$ sind definiert als

$$B_{n,k}(x) = \binom{n}{k} x^k (1-x)^{n-k} \quad (k = 0, 1, \ldots, n). \tag{139}$$

Eine stetige Funktion f auf $[0,1]$ kann durch die folgende Linearkombination der Bernsteinpolynome approximiert werden

$$B_n^f(x) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right) B_{n,k}(x).$$

Eine Anwendung: Bernsteinpolynome

Weshalb ist $B_n^f(x) \approx f(x)$? Eine probabilistische Begründung beruht darauf, dass $B_{n,k} = \mathbb{P}(S_n = k)$, wobei S_n die Anzahl Erfolge bei n Würfen mit Erfolgsparameter x bezeichnet. Also ist $B_n^f(x) = \mathbb{E}\left(f(\frac{S_n}{n})\right)$. Nach dem Gesetz der grossen Zahlen ist aber $\frac{S_n}{n} \approx x$ und f ist stetig.

Eine Anwendung: Bernsteinpolynome

Genauer gilt:

$$\begin{aligned} |B_n^f(x) - f(x)| &= \left|\mathbb{E}\left(f\left(\frac{S_n}{n}\right) - f(x)\right)\right| \leq \mathbb{E}\left(\left|f\left(\frac{S_n}{n}\right) - f(x)\right|\right) \\ &\leq 2\sup_u |f(u)|\ \mathbb{P}\left(\left|\frac{S_n}{n} - x\right| > \delta\right) \\ &+ \sup_{|u-v|\leq\delta} |f(u) - f(v)|\ \mathbb{P}\left(\left|\frac{S_n}{n} - x\right| \leq \delta\right). \end{aligned}$$

Wegen der gleichmässigen Stetigkeit von f ist der zweite Term rechts $\leq \varepsilon$ wenn δ klein genug ist. Der erste Term rechts ist wegen der Chebyshev-Ungleichung beschränkt durch

$$2\sup_u |f(u)| \frac{x(1-x)}{n\delta^2} \leq \frac{1}{2n\delta^2} \sup_u |f(u)| \leq \varepsilon,$$

wenn n gross genug ist. Damit haben wir gezeigt, dass

$$\sup_x |B_n^f(x) - f(x)| \longrightarrow 0 \text{ für } n \to \infty.$$

Starkes Gesetz der grossen Zahlen

Die stärkere Aussage

$$\lim_n \mathbb{P}\left(\bigcap_{k\geq n}\left\{\left|\frac{S_k}{k}-m\right|\leq\varepsilon\right\}\right)=1 \tag{140}$$

für alle $\varepsilon>0$, d.h. "das arithmetische Mittel *bleibt* von einem Zeitpunkt n an immer in der Nähe vom Erwartungswert m", ist wegen der Stetigkeit von $\mathbb{P}$ äquivalent zu:

$$\forall\varepsilon>0:\ \mathbb{P}\left(\bigcup_n\bigcap_{k\geq n}\left\{\left|\frac{S_k}{k}-m\right|\leq\varepsilon\right\}\right)=1 \quad \text{bzw.} \tag{141}$$

$$\mathbb{P}\left(\bigcap_{\varepsilon>0}\bigcup_n\bigcap_{k\geq n}\left\{\left|\frac{S_k}{k}-m\right|\leq\varepsilon\right\}\right)=\mathbb{P}\left(\lim_{n\to\infty}\frac{S_n}{n}=m\right)=1. \tag{142}$$

Wenn $\mathbb{P}\left(\frac{S_n}{n}\to m\right)=1$, dann sagen wir, dass das **starke Gesetz der grossen Zahlen** gilt.

Konvergenzbegriffe

Allgemein definieren wir für Zufallsvariablen $Z, Z_1, Z_2, \ldots$ auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$

a) **Stochastische Konvergenz** oder Konvergenz in Wahrscheinlichkeit von Z_n gegen Z:

$$\forall \varepsilon > 0 \quad \lim_{n \to \infty} \mathbb{P}\left(|Z_n - Z| > \varepsilon\right) = 0. \tag{143}$$

b) **Fast-sichere Konvergenz** von Z_n gegen Z:

$$\mathbb{P}\left(\{\omega | \lim Z_n(\omega) = Z(\omega)\}\right) = 1. \tag{144}$$

Struktursatz

Dann gilt

Satz

i) *Fast-sichere Konvergenz impliziert stochastische Konvergenz.*
ii) *Wenn* $\sum_n \mathbb{P}(|Z_n - Z| > \varepsilon) < \infty$ *für jedes* $\varepsilon > 0$, *dann konvergiert* Z_n *f.s. gegen* Z.

Fast sicher konvergente Teilfolgen

Wenn (Z_n) stochastisch gegen Z konvergiert, dann existiert eine Teilfolge (Z_{n_j}), welche f.s. gegen Z konvergiert.

Wähle n_j so, dass $\mathbb{P}\left(|Z_{n_j} - Z| > \frac{1}{j}\right) \leq \frac{1}{j^2}$. Dann gilt für alle $\varepsilon > 0$

$$\sum_j \mathbb{P}\left(|Z_{n_j} - Z| > \varepsilon\right) < \infty\,.$$

Beweis des Struktursatzes

i) Mit den gleichen Überlegungen wie oben ergibt sich:

$$\mathbb{P}\left(\lim_n Z_n = Z\right) = 1 \Leftrightarrow \lim_n \mathbb{P}\left(\bigcap_{k\geq n}\{|Z_k - Z| \leq \varepsilon\}\right) = 1\,.$$

für alle $\varepsilon > 0$.

ii) Mit dem Lemma von Borel-Cantelli folgt für jedes $\varepsilon > 0$

$$\sum_{n=1}^{\infty}\mathbb{P}\left(|Z_n - Z| > \varepsilon\right) < \infty \Rightarrow \mathbb{P}\left(\bigcap_n\bigcup_{k\geq n}\{|Z_k - Z| > \varepsilon\}\right) = 0.$$

Indem wir zum Komplement übergehen, erhalten wir

$$\mathbb{P}\left(\bigcup_n\bigcap_{k\geq n}\{|Z_k - Z| \leq \varepsilon\}\right) = 1$$

woraus die Behauptung folgt.

Gegenbeispiel

Sei (Y_i) i.i.d. mit Werten in $\mathbb{N}$, $\mathbb{P}(Y_i \geq k) = \frac{1}{k}$, und $Z_n = n^{-1} Y_n$. Dann konvergiert Z_n stochastisch gegen 0, aber $\mathbb{P}(Z_n \to 0) = 0$, denn $\sum_n \mathbb{P}(Z_n \geq 1) = \sum_n \mathbb{P}(Y_n \geq n) = \infty$, also sind wegen Borel-Cantelli f.s. unendlich viele Z_n grösser oder gleich 1.

Starkes Gesetz der grossen Zahlen

Satz

Sei (X_i) *i.i.d. mit* $\mathbb{E}\left(X_i^2\right) < \infty$. *Dann konvergiert* $\frac{S_n}{n}$ *fast sicher gegen* $m = \mathbb{E}\left(X_i\right)$.

Beweis

Wir dürfen annehmen, dass $X_i \geq 0$ (zerlege X_i in $X_i^+ = \max(X_i, 0)$ und $X_i^- = \max(-X_i, 0)$). Es konvergiert die Teilfolge S_{n^2}/n^2 f.s. gegen m. Wir müssen also nur noch S_k/k für $n^2 \leq k \leq (n+1)^2$ untersuchen. Da $S_{k+1} \geq S_k$ ist, folgt:

$$\frac{n^2}{(n+1)^2}\,\frac{S_{n^2}}{n^2} = \frac{S_{n^2}}{(n+1)^2} \leq \frac{S_k}{k} \leq \frac{S_{(n+1)^2}}{n^2} = \frac{(n+1)^2}{n^2}\,\frac{S_{(n+1)^2}}{(n+1)^2}.$$

Die beiden Schranken links und rechts konvergieren f.s. gegen m, also auch $\frac{S_k}{k} \to m$ f.s.

Bemerkung

Notwendige und hinreichende Bedingungen für das starke Gesetz der grossen Zahlen wurden von Kolmogorov gefunden (ohne Beweis): Für (X_i) i.i.d. gilt $\frac{S_n}{n} \to m \in \mathbb{R}$ fast sicher genau dann, wenn $\mathbb{E}(|X_i|) < \infty$ und $m = \mathbb{E}(X_i)$.

Zentraler Grenzwertsatz

Seien X_i unabhängige Zufallsvariablen mit $\mathbb{E}(X_i) = m_i$ und $\operatorname{var} X_i = \sigma_i^2 < \infty$. Wir hatten bereits gesehen, dass für binäre X_i's die Form der Verteilung von S_n durch die Normalverteilung approximiert wird. Dieses Resultat gilt sehr viel allgemeiner. Da die Form der Verteilung unabhängig von Lage und Streuung ist, standardisieren wir S_n, so dass der Erwartungswert $= 0$ und die Varianz $= 1$ ist:

$$S_n^* = \frac{S_n - \mathbb{E}(S_n)}{\sqrt{\operatorname{var} S_n}} = \frac{S_n - \sum_{i=1}^n m_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}} \,. \tag{145}$$

Zentraler Grenzwertsatz

Wir wollen nun zeigen, dass die Verteilung von S_n^* unter gewissen Bedingungen gegen die **Standard-Normalverteilung** $\mathcal{N}(0,1)$ konvergiert. Wir verwenden dazu den folgenden Konvergenzbegriff:

Definition

Seien μ und μ_n Wahrscheinlichkeitsverteilungen auf $(\mathbb{R}, \mathcal{B})$. Wir sagen, dass μ_n **schwach** gegen μ **konvergiert**, falls

$$\int f d\mu_n \longrightarrow \int f d\mu \tag{146}$$

für alle f, welche stetig und beschränkt sind.

Bemerkung

Sei Z_n eine Zufallsvariable mit der Verteilung μ_n, d.h. $\mathbb{P}(Z_n \in B) = \mu_n(B)$ für $B \in \mathcal{B}$. Dann ist $\int f d\mu_n = \mathbb{E}(f(Z_n))$. Also bedeutet schwache Konvergenz von μ_n gegen μ, dass $\mathbb{E}(f(Z_n)) \to \mathbb{E}(f(Z))$ für alle stetigen und beschränkten Funktionen f, wobei $Z_n \sim \mu_n$ und $Z \sim \mu$. Auf welchem Raum Ω diese Zufallsvariablen definiert sind, spielt dabei keine Rolle.

Beispiel

$\mu_n = \mathcal{N}(c, \frac{1}{n})$ konvergiert schwach gegen die Verteilung μ, die in c konzentriert ist (Dirac-Mass). Mit einer Substitution erhält man

$$\int f d\mu_n - \int f d\mu = \frac{1}{\sqrt{2\pi}} \int (f(c + n^{-1/2}x) - f(c)) \exp(-\frac{1}{2}x^2) dx,$$

und die Behauptung folgt daher mit dem Konvergenzsatz von Lebesgue.

Verteilungskonvergenz via Verteilungsfunktionen

Seien μ und μ_n Wahrscheinlichkeitsverteilungen auf $(\mathbb{R}, \mathcal{B})$ mit Verteilungsfunktionen F und F_n. Dann sind die folgenden Aussagen äquivalent:

i) $\mu_n \longrightarrow \mu$ schwach.

ii) $F_n(x) \longrightarrow F(x)$ für jede Stetigkeitsstelle x von F.

iii) $\int f d\mu_n \longrightarrow \int f d\mu$ für alle $f \in C_b^3(\mathbb{R})$, wobei $C_b^3(\mathbb{R})$ die Menge aller dreimal stetig differenzierbaren Funktionen auf $\mathbb{R}$ bezeichnet, für die f, f', f'', f''' alle beschränkt sind.

Beispiel (Fortsetzung)

Die Verteilungsfunktion F_n von $\mathcal{N}(c, \frac{1}{n})$ ist $\Phi(\sqrt{n}(x-c))$, konvergiert also gegen 0 ($x < c$), bzw. gegen $\frac{1}{2}$ ($x = c$), bzw. gegen 1 ($x > c$). Die Verteilungsfunktion F des Dirac-Masses ist hingegen gleich 0 ($x < c$), bzw. gleich 1 ($x \geq c$), d.h. $F_n(c)$ konvergiert nicht gegen $F(c)$.

Beweis

Die Implikation "i) $\Rightarrow$ iii)" ist klar.

Für die Implikation "iii) $\Rightarrow$ ii)", nehmen wir an, es gelte $\int f d\mu_n \to \int f d\mu$ für alle $f \in C_b^3(\mathbb{R})$. Seien $x \in \mathbb{R}$ und $\delta > 0$ fest. Wir wählen ein $f \in C_b^3(\mathbb{R})$ mit

$$I_{(-\infty,x]} \leq f \leq I_{(-\infty,x+\delta]} \cdot$$

Beweis

Dann gilt

$$F_n(x) = \int I_{(-\infty,x]} d\mu_n \leq \int f d\mu_n$$

und

$$\int f d\mu \leq \int I_{(-\infty,x+\delta]} d\mu = F(x+\delta)\,.$$

Daraus folgt

$$\limsup F_n(x) \leq \lim \int f d\mu_n = \int f d\mu \leq F(x+\delta).$$

Analog folgt

$$\liminf F_n(x) \geq F(x-\delta).$$

Beweis

Wenn jetzt F stetig ist an der Stelle x, dann folgt mit $\delta \to 0$

$$F(x) \leq \liminf F_n(x) \leq \limsup F_n(x) \leq F(x),$$

also gilt überall "$=$" statt "$\leq$".

Es bleibt noch die Implikation "ii) $\Rightarrow$ i)" zu zeigen. Wir fixieren dazu ein stetiges, beschränktes f und ein $\varepsilon > 0$.

Zunächst bemerken wir, dass die Menge der Stellen, wo F unstetig ist, höchstens abzählbar ist (für jedes k gibt es nur endlich viele Stellen, wo F einen Sprung mit einer Höhe in $(2^{-k}, 2^{-k-1}]$ hat).

Beweis

Wegen ii) gibt es also Stetigkeitsstellen a und b von F, so dass

$$\inf_n \mu_n\big([a,b]\big) > 1-\varepsilon, \quad \mu\big([a,b]\big) > 1-\varepsilon.$$

Ferner ist f gleichmässig stetig auf $[a,b]$, d.h. es gibt ein δ, so dass $|f(x)-f(y)| \leq \varepsilon$ falls $a \leq x,y \leq b$ und $|x-y| \leq \delta$. Wir wählen als nächstes ein m und Stetigkeitsstellen x_i mit $a = x_0 < x_1 \ldots < x_m = b$ und $x_{i+1} - x_i \leq \delta$ und setzen

$$f_m = \sum_{i=1}^{m} f(x_{i-1}) I_{(x_{i-1},x_i]}.$$

Beweis

Dann gilt

$$\begin{aligned} |f(x) - f_m(x)| &\leq \varepsilon \text{ falls } a < x \leq b \\ |f(x) - f_m(x)| &\leq \sup_x |f(x)| \text{ sonst,} \end{aligned}$$

also ist

$$|\int (f - f_m) d\mu_n(x)| \leq \varepsilon(1 + \sup_x |f(x)|)$$

(und analog mit μ anstelle von μ_n). Ferner ist

$$\int f_m d\mu_n = \sum_{i=1}^{m} f(x_{i-1})(F_n(x_i) - F_n(x_{i-1})),$$

also konvergiert wegen ii) $\int f_m d\mu_n$ gegen $\int f_m d\mu$.

Beweis

Damit ist für n gross genug

$$\begin{aligned} |\int f d\mu_n - \int f d\mu| &\leq |\int (f - f_m) d\mu_n| + |\int f_m d\mu_n - \int f_m d\mu| \\ &\quad + |\int (f - f_m) d\mu| \\ &\leq \varepsilon(3 + 2 \sup_x |f(x)|). \end{aligned}$$

□

Bemerkung

Die Summe zweier unabhängiger normalverteilter Zufallsvariablen ist wieder normalverteilt:

Lemma

Wenn X_1, X_2 *unabhängig sind und* $X_i \sim \mathcal{N}(m_i, \sigma_i^2)$ $(i = 1, 2)$, *dann ist* $X_1 + X_2 \sim \mathcal{N}(m_1 + m_2, \sigma_1^2 + \sigma_2^2)$.

Beweis.

Nachrechnen mit Hilfe der Faltung und quadratischem Ergänzen. □

Bemerkung

Die Normalverteilung ist also ein Fixpunkt bei der Summation von i.i.d. Zufallsvariablen. Der **Zentrale Grenzwertsatz** besagt nun, dass man bei Summation von i.i.d. Zufallsvariablen mit endlichem zweiten Moment stets gegen diesen Fixpunkt konvergiert:

Satz

Sei (X_i) *i.i.d. mit* $\mathbb{E}(X_i) = m$ *und* $\operatorname{var} X_i = \sigma^2 < \infty$. *Dann konvergiert die Verteilung von* S_n^* *schwach gegen* $\mathcal{N}(0,1)$, *d.h. (gemäss ii) oben)*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - nm}{\sigma\sqrt{n}} \leq x\right) = \Phi(x) \quad \textit{for all } x \in \mathbb{R}.$$

Satz von Lindeberg

Wir leiten diesen Satz aus einem viel allgemeineren Resultat ab, bei dem die Summanden nicht i.i.d. sein müssen. Die Verteilung der Summanden darf sogar noch von n abhängen (daher die zwei Indizes n und i).

Satz (Lindeberg)

Seien $X_{n,i}$ $(1 \leq i \leq n, n \in \mathbb{N})$ *Zufallsvariablen mit*

a) $X_{n,1}, \ldots, X_{n,n}$ *sind unabhängig für alle* n;

b) $\mathbb{E}(X_{n,i}) = 0, \quad \mathbb{E}(X_{n,i}^2) = \sigma_{n,i}^2 < \infty, \quad \sum_{i=1}^{n} \sigma_{n,i}^2 = 1;$

c) $\lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{E}\left(X_{n,i}^2 1_{[|X_{n,i}|>\varepsilon]}\right) = 0$ *für alle* $\varepsilon > 0$.

Dann konvergiert die Verteilung von $S_n = X_{n,1} + \cdots + X_{n,n}$ *schwach gegen* $\mathcal{N}(0,1)$.

Beweis des CLT aus dem Satz von Lindeberg

Setze $X_{n,i} = (X_i - m)/(\sigma\sqrt{n})$. Dann sind a) und b) offensichtlich erfüllt. Ferner folgt mit dem Konvergenzsatz von Lebesgue, dass für alle $\varepsilon > 0$

$$\sum_{i=1}^{n} \mathbb{E}\left(X_{n,i}^2 1_{[|X_{n,i}|>\varepsilon]}\right) = \frac{1}{\sigma^2}\mathbb{E}\left((X_1 - m)^2 1_{[|X_1-m|>\varepsilon\sigma\sqrt{n}]}\right) \longrightarrow 0 \quad (n \to \infty).$$

□

Bemerkung

Die Bedingung c) besagt, dass die $X_{n,i}$ klein sein müssen. Es gilt nämlich

$$\max_i \sigma_{n,i}^2 \leq \max_i \left(\varepsilon^2 + \mathbb{E}\left(X_{n,i}^2 1_{[|X_{n,i}|>\varepsilon]} \right) \right) \leq \varepsilon^2 + \sum_{i=1}^{n} \mathbb{E}\left(X_{n,i}^2 1_{[|X_{n,i}|>\varepsilon]} \right),$$

d.h. die Bedingung c) impliziert

$$\max_{1\leq i\leq n} \sigma_{n,i}^2 \longrightarrow 0 \text{ für } n \to \infty. \tag{147}$$

Wenn wir also $\varepsilon_n = \sqrt{\max_i \sigma_{n,i}}$ setzen, dann geht ε_n gegen Null und mit der Chebyshev-Ungleichung folgt $\mathbb{P}(|X_{n,i}| \leq \varepsilon_n) \to 1$.

Beweis des Satzes von Lindeberg

Wir prüfen die Bedingung iii) oben nach. Sei also f dreimal stetig differenzierbar mit f, f', f'', f''' beschränkt. Wir wählen Zufallsvariablen $Y_{n,i} \sim \mathcal{N}(0, \sigma_{n,i}^2)$ derart, dass für jedes n

$$Y_{n,1}, \ldots, Y_{n,n}, X_{n,1}, \ldots, X_{n,n} \text{ unabhängig sind.}$$

Dann ist $\sum_{i=1}^n Y_{n,i} \sim \mathcal{N}(0,1)$, also muss man $\mathbb{E}\left(f(\sum_{i=1}^n X_{n,i}) - f(\sum_{i=1}^n Y_{n,i})\right)$ abschätzen Wir tauschen dazu sukzessive ein $X_{n,i}$ gegen ein $Y_{n,i}$ aus.

Beweis des Satzes von Lindeberg

Sei

$$Z_{n,i} = X_{n,1} + \cdots + X_{n,i-1} + Y_{n,i+1} + \cdots + Y_{n,n}.$$

Mit einer Taylor-Entwicklung folgt dann

$$\begin{aligned} f(Z_{n,i} + X_{n,i}) - f(Z_{n,i} + Y_{n,i}) &= f'(Z_{n,i})(X_{n,i} - Y_{n,i}) \\ &\quad + \frac{1}{2} f''(Z_{n,i})(X_{n,i}^2 - Y_{n,i}^2) \\ &\quad + R_3(Z_{n,i}, X_{n,i}) + R_1(Z_{n,i}, Y_{n,i}), \end{aligned}$$

wobei

$$\begin{aligned} |R_1(z,x)| &= |x^3 f'''(z + \theta x)\frac{1}{6}| \leq C\ |x|^3, \quad \text{bzw} \\ |R_2(z,x)| &= |x^2 \frac{1}{2}(f''(z + \theta x) - f''(z))| \leq C\ x^2 \end{aligned}$$

und

Beweis des Satzes von Lindeberg

$$|R_3(z,x)| = |R_1(z,x) + R_2(z,x)| \leq C\ (\varepsilon x^2 + x^2 1_{[|x|>\varepsilon]}).$$

gelten, und die Gleichungen für alle x, z mit von f abhängigen Konstanten gültig sind.

Weil nach Konstruktion $Z_{n,i}$ sowohl von $X_{n,i}$ als auch von $Y_{n,i}$ unabhängig ist, folgt daraus

$$\begin{aligned}
& \left|\mathbb{E}\left(f(Z_{n,i}+X_{n,i}) - f(Z_{n,i}+Y_{n,i})\right)\right| \leq \left|\mathbb{E}\left(f'(Z_{n,i})\right)\left(\mathbb{E}\left(X_{n,i}\right) - \mathbb{E}\left(Y_{n,i}\right)\right)\right| \\
+ \quad & \left|\frac{1}{2}\mathbb{E}\left(f''(Z_{n,i})\right)\left(\mathbb{E}\left(X_{n,i}^2\right) - \mathbb{E}\left(Y_{n,i}^2\right)\right)\right| + \mathbb{E}\left(|R(Z_{n,i}, X_{n,i})|\right) + \mathbb{E}\left(|R(Z_{n,i}, Y_{n,i})|\right) \\
\leq \quad & C\Big(\varepsilon\sigma_{n,i}^2 + \mathbb{E}\left(X_{n,i}^2 1_{[|X_{n,i}|>\varepsilon]}\right) + \mathbb{E}\left(|Y_{n,i}|^3\right)\Big).
\end{aligned}$$

Beweis des Satzes von Lindeberg

Damit ist für alle $\varepsilon > 0$

$$\left|\mathbb{E}\left(f(\sum_i X_{n,i}) - f(\sum_i Y_{n,i})\right)\right| = \left|\sum_{i=1}^n \mathbb{E}\left(f(Z_{n,i} + X_{n,i}) - f(Z_{n,i} + Y_{n,i})\right)\right|$$
$$\leq C\left(\varepsilon + \sum_{i=1}^n \mathbb{E}\left(X_{n,i}^2 1_{[|X_{n,i}|>\varepsilon]}\right) + \sum_{i=1}^n \mathbb{E}\left(|Y_{n,i}|^3\right)\right).$$

Die rechte Seite ist $\leq 3C\varepsilon$ für n gross genug wegen Voraussetzung c), bzw. weil gilt

$$\sum_{i=1}^n \mathbb{E}\left(|Y_{n,i}|^3\right) = \sum_{i=1}^n \sigma_{n,i}^3 \sqrt{\frac{8}{\pi}} \leq \max(\sigma_{n,i})\sqrt{\frac{8}{\pi}},$$

was gegen null konvergiert. □

Korollar

Sei (X_i) eine i.i.d. Folge von d-dimensionalen Zufallsvektoren mit Verteilung μ und sei $f \in L^2(\mathbb{R}^d, \mu)$. Dann gilt

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(X_i) = \int_{\mathbb{R}^d} f(x)\mu(dx) = \mathbb{E}\left(f(X_i)\right) =: m, \quad \mathbb{P}\text{-f.s.}$$

und

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\frac{1}{n}\sum_{i=1}^{n} f(X_i) - m}{\sigma_f/\sqrt{n}} \leq x \right) = \Phi(x) \quad \text{for all } x \in \mathbb{R},$$

wobei $\sigma_f^2 := \mathrm{Var}[f(X_i)] = \int_{\mathbb{R}^d}(f(x) - m)^2\mu(dx)$. Das heisst der Fehler

$$\frac{1}{n} \sum_{i=1}^{n} f(X_i) - m$$

ist approximativ normalverteilt mit Mittel 0 und Standardabweichung $\sigma_f/\sqrt{n}$.

Beweis

Die erste Aussage folgt aus dem starken Gesetz der grossen Zahlen und die zweite aus dem Zentralen Grenzwertsatz. □

Bemerkung

Man kann also Integrale approximieren durch die Erzeugung von Zufallszahlen (Monte-Carlo Verfahren). Beachtenswert ist, dass die Genauigkeit nur von der Anzahl Replikate n und nicht von der Dimension d abhängt. In hohen Dimensionen ist Monte Carlo deterministischen Verfahren zur Approximation von Integralen überlegen!

Rundungsfehler

Seien $X_1, X_2, \ldots, X_n$ i.i.d. $\mathcal{U}[-\frac{1}{2}, \frac{1}{2}]$. Dann ist $\mathbb{E}(X_i) = 0$ und $\operatorname{var} X_i = \frac{1}{12}$, also gilt:

$$\mathbb{P}(a \leq S_n \leq b) = \mathbb{P}\left(a\sqrt{\frac{12}{n}} \leq S_n^* \leq b\sqrt{\frac{12}{n}}\right) \approx \Phi\left(b\sqrt{\frac{12}{n}}\right) - \Phi\left(a\sqrt{\frac{12}{n}}\right).$$

Wenn Rundungsfehler als unabhängig und gleichverteilt angenommen werden können, dann ist die Wahrscheinlichkeit, dass bei der Addition von $n = 100$ Zahlen höchstens eine Stelle verloren geht, gleich

$$\mathbb{P}(-5 < S_{100} < 5) \approx \Phi(\sqrt{3}) - \Phi(-\sqrt{3}) = 0.917$$

(im schlimmsten Fall sind es zwei Stellen).

Beispiel: Asymptotik des Medians

Dieses Beispiel zeigt die Flexibilität des Satzes von Lindeberg. Der **Median** m einer Verteilung F wurde definiert durch $m = F^{-1}(\frac{1}{2})$. Seien $X_1, X_2, \ldots$ i.i.d. mit Verteilungsfunktion F und Median $m = 0$. Ferner soll $F'(0)$ existieren und > 0 sein. Das heisst, dass Beobachtungen in der Nähe des Medians auftreten werden. Ferner sei Z_n der sogenannte **Stichprobenmedian** von $X_1, \ldots, X_n$, d.h. Z_n ist die mittlere Beobachtung, oder formelmässig $Z_n = X_{(k)}$ mit $k = [\frac{n}{2} + 1]$, wobei $X_{(1)} \leq \cdots \leq X_{(n)}$ die der Grösse nach geordneten Zufallsvariablen $X_1, \ldots, X_n$ bezeichnen und $[x]$ den ganzzahligen Teil von x.

Wir behaupten, dass die Verteilung von $\sqrt{n}Z_n$ schwach gegen $\mathcal{N}(0, \frac{1}{4F'(0)^2})$ konvergiert, d.h.

$$\mathbb{P}\left(\sqrt{n}\ Z_n \leq x\right) \to \Phi(2F'(0)x).$$

Beispiel: Asymptotik des Medians

Wir setzen $Y_{n,i} = 1_{[X_i > x/\sqrt{n}]}$. Dann ist

$$\sqrt{n}\, Z_n \leq x \Leftrightarrow X_{(k)} \leq \frac{x}{\sqrt{n}} \Leftrightarrow \sum_{i=1}^{n} Y_{n,i} \leq n - k.$$

Es gilt $\mathbb{E}(Y_{n,i}) = p_n$, $\operatorname{var} Y_{n,i} = p_n(1-p_n)$ mit $p_n = 1 - F(\frac{x}{\sqrt{n}})$. Wir standardisieren die $Y_{n,i}$, um nachher den Satz von Lindeberg anwenden zu können:

$$X_{n,i} = \frac{Y_{n,i} - p_n}{\sqrt{np_n(1-p_n)}}\,.$$

Beispiel: Asymptotik des Medians

Da

$$\sqrt{n}\, Z_n \leq x \Leftrightarrow S_n = \sum_{i=1}^{n} X_{n,i} \leq a_n = \frac{n-k-np_n}{\sqrt{np_n(1-p_n)}} \longrightarrow 2F'(0)x,$$

gilt für $\delta > 0$ fest und n gross genug

$$\begin{aligned} \mathbb{P}\left(S_n \leq 2F'(0)x - \delta\right) &\leq \mathbb{P}\left(S_n \leq a_n\right) = \mathbb{P}\left(\sqrt{n}Z_n \leq x\right) \\ &\leq \mathbb{P}\left(S_n \leq 2F'(0)x + \delta\right). \end{aligned}$$

Andrerseits $\mathbb{P}\left(S_n \leq 2F'(0)x \pm \delta\right) \longrightarrow \Phi(2F'(0)x \pm \delta)$, und daraus folgt die Behauptung. □

Bemerkung

Man sieht aus dem Beweis, dass das Resultat gültig bleibt, solange $k = \frac{n}{2} + o(\sqrt{n})$, insbesondere also für leicht andere Definitionen des Stichprobenmedians.

Einführung in die mathematische Statistik

Was ist Statistik?

Deskriptive Statistik fasst Datensätze zusammen und macht deren Besonderheiten sichtbar, mit Hilfe von Kennzahlen und Grafiken. Wir behandeln hier diesen Teil nicht.

Schliessende Statistik betrachtet die vorliegenden Beobachtungen als Realisierungen von Zufallsvariablen und zieht Rückschlüsse auf die zugrunde liegende Verteilung (und damit auch auf zukünftige Beobachtungen). Das Ziel ist es, Zufallsfehler und echte Effekte unter Angabe der verbleibenden Unsicherheit zu trennen.

Wahrscheinlichkeitstheorie ist deduktiver Natur, Statistik induktiver Natur. Statistik versucht zu erfassen, wie wir aus beschränkter Erfahrung lernen und verallgemeinern können. Wir geben zunächst drei Beispiele, welche die Fragestellungen der schliessenden Statistik illustrieren.

Beispiel: Aussersinnliche Wahrnehmung

An der University of California in Davis wurde 1973 folgendes Experiment durchgeführt:

Ein Computer wählte zufällig eines von 4 Symbolen, und ein Medium versuchte, durch aussersinnliche Wahrnehmung das gewählte Symbol herauszufinden. 15 Medien machten je 500 Versuche und waren insgesamt 2006 mal erfolgreich.

Die Grundfrage: "War hier aussersinnliche Wahrnehmung im Spiel?" lässt sich leider auch nicht mit Statistik beantworten, wohl aber die einfachere Frage: "Kann das Resultat Zufall sein?".

Beispiel: Aussersinnliche Wahrnehmung

Falls nur der Zufall spielt, dann ist die Anzahl Erfolge X binomial$(7500, \frac{1}{4})$-verteilt, und die Wahrscheinlichkeit, dass wir ein Ergebnis erhalten, welches mindestens so gut ist wie das tatsächlich vorliegende, lässt sich mit dem Zentralen Grenzwertsatz sehr genau approximieren:

$$\mathbb{P}(X \geq 2006) \approx 1 - \Phi\left(\frac{2006 - 1875}{\sqrt{7500 \cdot 0.25 \cdot 0.75}}\right) = 1 - \Phi(3.49) = 0.0002.$$

Entweder ist also etwas extrem Unwahrscheinliches eingetreten, oder die Hypothese "reiner Zufall" ist falsch. Der Zufall ist also keine zufriedenstellende Erklärung. Andere mögliche Erklärungen sind "Schlechte Durchführung des Experiments" (z. B. der Zufallsgenerator war nicht in Ordnung), oder "Aussersinnliche Wahrnehmung hat stattgefunden".

Beispiel: Aussersinnliche Wahrnehmung

Es ist wesentlich, dass man die Wahrscheinlichkeit eines *mindestens* so guten Resultates nimmt. Die Wahrscheinlichkeit, genau das vorliegende Resultat zu erhalten, ist nämlich in jedem Fall klein; aus dem Satz von de Moivre-Laplace folgt z.B.

$$\mathbb{P}(X = 1875) \approx \frac{1}{\sqrt{2\pi \cdot 7500 \cdot 0.25 \cdot 0.75}} = 0.0106.$$

Beispiel: Aussersinnliche Wahrnehmung

Wenn wir einmal annehmen, dass aussersinnliche Wahrnehmung mit Wahrscheinlichkeit γ zustande kommt und dass in den andern Fällen einfach geraten wurde, erhalten wir folgendes Modell: $X =$ Anzahl Erfolge $\sim$ Binomial(7500, p) mit $p = \gamma + (1-\gamma)\frac{1}{4} = \frac{3}{4}\gamma + \frac{1}{4}$. Man möchte natürlich wissen, wie gross γ ist. Der naheliegende Schätzwert ist

$$\hat{p} = 2006/7500 = 0.267 \Rightarrow \hat{\gamma} = 0.023.$$

Vermutlich ist dies aber nicht das wahre γ.

Beispiel: Aussersinnliche Wahrnehmung

Wir sind eher an einem Intervall von $\gamma-$Werten interessiert, welches das wahre γ mit grosser Wahrscheinlichkeit $1-\alpha$ einfängt. Es wird mit späterer Theorie folgen, dass $[0.006, 0.040]$ ein solches Intervall ist für $1-\alpha = 0.99$. Auch wenn also aussersinnliche Wahrnehmung im Spiel war, ist sie höchst unzuverlässig.

Für eine detailiertere Analyse verschiedener parapsychologischer Experimente, siehe J. Utts, Replication and Meta-Analysis in Parapsychology, Statistical Science **6** (1991), 363-403.

Beispiel: Vergleich zweier Lehrmethoden

Die beiden Methoden wurden an je 10 Testpersonen mit ähnlicher Vorbildung ausprobiert. Bei den Abschlussprüfungen ergaben sich die folgenden Resultate (bereits der Grösse nach geordnet):

Methode	1:	3	6	18	25	37	48	49	51	81	89,
Methode	2:	9	34	40	57	61	64	75	91	93	98.

Die wesentlichen Fragestellungen sind: Besteht ein echter Unterschied zwischen beiden Methoden, der nicht nur auf diesen einmaligen Test beschränkt bleibt? Falls ja, wie viele Punkte macht dieser echte Unterschied aus?

Beispiel: Vergleich zweier Lehrmethoden

Modellierung: Wir nehmen an, dass die beobachteten Werte Realisierungen von unabhängigen Zufallsvariablen sind, und zwar
Methode 1: $X_1, \ldots, X_n$ i.i.d $\sim F$,
Methode 2: $Y_1, \ldots, Y_m$ i.i.d.$\sim G$.
Falls beide Methoden gleich gut sind, ist $F = G$ (die sogenannte Nullhypothese), während der einfachste Fall eines echten Unterschieds lautet $G(x) = F(x - \Delta)$.

Beispiel: Vergleich zweier Lehrmethoden

Eine erste Methode zur Feststellung eines Unerschiedes ist:

Berechne W =Anzahl Paare (i,j) mit $X_i < Y_j$. Wenn die Nullhypothese stimmt, ist $\mathbb{P}(X_i < Y_j) = \frac{1}{2}$ (Achtung! Das gilt nicht immer), also $W \approx \frac{n \cdot m}{2}$. Somit verwerfen wir die Nullhypothese, wenn W zu stark von $\frac{n \cdot m}{2}$ abweicht (sogenannter Wilcoxon-Test). Im Beispiel nimmt W den Wert 73 an.

Beispiel: Vergleich zweier Lehrmethoden

Eine zweite Methode zur Feststellung eines Unerschiedes ist:

Berechne

$$T = \frac{(\overline{X} - \overline{Y})}{\sqrt{\frac{1}{n} + \frac{1}{m}} \sqrt{\frac{1}{n+m-2} \left(\sum (X_i - \overline{X})^2 + \sum (Y_i - \overline{Y})^2 \right)}}$$

Wenn die Nullhypothese stimmt, ist $\mathbb{E}(X_i) = \mathbb{E}(Y_i)$, also $T \approx 0$. Somit verwerfen wir die Nullhypothese, wenn $|T|$ bzw. T zu gross ist. Die Definition von T wird verständlicher, wenn man beachtet, dass $\overline{X} - \overline{Y}$ die Varianz $\sigma^2(1/n + 1/m)$ hat, wenn alle Zufallsvariablen unabhängig sind und die Varianz σ^2 haben. Die zweite Wurzel im Nenner ist einfach eine Schätzung von σ. Dieser Test heisst $t-$Test. Im Beispiel nimmt T den Wert 1.58 an.

Beispiel: Vergleich zweier Lehrmethoden

Um zu einer Entscheidung zu gelangen, müssen wir wissen, wie gross unter der Nullhypothese die Wahrscheinlichkeit einer Abweichung ≥ 23, bzw. ≥ 1.58 ist. Wir schauen also wieder alle Abweichungen an, die mindestens so extrem sind wie die tatsächlich vorliegende. Rechnungen, die später begründet werden, ergeben

$$\begin{array}{llll} \mathbb{P}(|W-50| \geq 23) & \approx & 0.05, & \mathbb{P}(W \geq 73) \approx 0.025. \\ \mathbb{P}(|T| > 1.58) & \approx & 0.12, & \mathbb{P}(T > 1.58) \approx 0.06. \end{array}$$

falls $F = G$. Man ist also an der Grenze, von dem, was noch als zufällige Abweichung plausibel erscheint. (Konventionell setzt man die Grenze bei 5% an). Wir sehen auch, dass es zwei mögliche Interpretationen von "mindestens so extrem" gibt (mit, bzw. ohne Absolutbetrag), und dass der Wilcoxon-Test hier empfindlicher ist.

Beispiel: Vergleich zweier Lehrmethoden

Für die zweite Fragestellung nach der Grösse des Unterschieds verschieben wir die Y_i um soviel, dass W bzw. T den Wert $\frac{nm}{2}$ bzw. 0 annimmt. Diese Verschiebung nehmen wir als Schätzung von Δ. Man erhält: $\widehat{\Delta} = 23$ (im Fall von W) bzw. $\widehat{\Delta} = 21.5$ (im Fall von T). Wie zuvor wäre ein Intervall von Δ-Werten, welches das wahre Δ mit vorgegebener Wahrscheinlichkeit einfängt, informativer.

Beispiel: Blutdruck in Abhängigkeit von Alter und anderen Faktoren

Ziele solcher Studien sind die Identifikation derjenigen Faktoren, die den Blutdruck beeinflussen, die Erkennung atypischer Fälle, die Prüfung blutdrucksenkender Medikamente etc.. Wir haben die folgenden Grössen

$$\begin{aligned} Y_i &= \text{Blutdruck der } i\text{−ten Versuchsperson} \\ x_{ij} &= \text{Wert des } j\text{−ten Faktors der } i\text{−ten Person,} \end{aligned}$$

wobei zum Beispiel $x_{i1} =$ Alter, $x_{i2} =$ Gewicht, $x_{i3} =$ biologisches Geschlecht kodiert als $0/1$ etc.

Beispiel: Blutdruck in Abhängigkeit von Alter und anderen Faktoren

Als Modell verwenden wir

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, \tag{148}$$

d.h. eine lineare Beziehung mit zufälligen Fehlern. Wir nehmen an, dass wir alle wesentlichen Faktoren erfasst haben, so dass wir annehmen können $\mathbb{E}(\varepsilon_i) = 0$. Die Annahme der Linearität der Beziehung ist nicht ganz so einschränkend, wie man zunächst denken könnte, weil wir die x_{ij} transformieren oder durch Kombination neue Faktoren bilden können, z.B. $x_{i,p+1} = \text{Alter}^2$ oder $x_{i,p+1} = \text{Alter mal biologisches Geschlecht}$ etc. Zur Überprüfung der Linearität sollte man aber mindestens die sogenannten **Streudiagramme** $((Y_i, x_{ij});\ i = 1, \ldots, n)$ anschauen für jedes j.

Beispiel: Blutdruck in Abhängigkeit von Alter und anderen Faktoren

Die folgenden statistischen Fragen möchte man beantworten:

- Wie kann man die unbekannten Parameter $\beta_0, \ldots, \beta_p$ schätzen? Welche von mehreren möglichen Schätzverfahren sind gut?
- Welche Parameter β_j können gleich null sein? (d.h. der entsprechende Faktor hat keinen Einfluss).
- In welchem Bereich wird der Blutdruck einer neuen Versuchsperson mit bestimmten Werten von $x_1, \ldots, x_p$ liegen?

Beispiel: Blutdruck in Abhängigkeit von Alter und anderen Faktoren

Das bekannteste Schätzverfahren besteht darin, die Summe der Fehlerquadrate

$$\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p}\beta_j x_{ij} - \beta_0)^2 \tag{149}$$

zu minimieren (Kleinste Quadrate, Gauss). Die Lösung $\widehat{\beta} = (\widehat{\beta}_0, \ldots, \widehat{\beta}_p)^T$ lässt sich formelmässig angeben als

$$\widehat{\beta} = (X^T X)^{-1} X^T Y$$

wobei Y der $n \times 1$ Vektor $(y_1, \ldots, y_n)^T$ (Beobachtungen von Y_i) ist und X die $n \times (p+1)$-Matrix mit erster Spalte $(1, \ldots, 1)'$ und $(j+1)$-ter Spalte $(x_{1j}, \ldots, x_{nj})'$ (sofern X maximalen Rang hat). Die numerische Berechnung von $\widehat{\beta}$ ist ebenfalls bestens untersucht.

Grundstrukturen

Zusammenfassend halten wir fest, dass die mathematische Statistik mit Klassen von möglichen Verteilungen arbeitet. Diese enthalten einen **strukturellen Parameter**, der üblicherweise p reelle Komponenten hat und direkt mit der ursprünglichen Fragestellung zusammenhängt, sowie meist noch **Störparameter**, welche oft unendlichdimensional sind und nur mittelbar von Interesse. Im ersten Beispiel ist der struktureller Parameter p bzw. γ, und es gibt keinen Störparameter. Im zweiten Beispiel ist der strukturelle Parameter Δ und der Störparameter F. Im dritten Beispiel ist der strukturelle Parameter $\beta_0, \ldots, \beta_p$ und der Störparameter die Verteilung der ε_i's, bzw. deren Varianzen und Kovarianzen, falls wir Normalverteilung annehmen können.

Formalisierung der Grundstrukturen

Wenn wir die zeitlichen Abläufe ausser Acht lassen, dann sehen wir folgende wahrscheinlichkeitstheoretische Situation:

$(\Omega, \mathcal{F}, \mathbb{P}_\theta)$ ist eine Familie von Wahrscheinlichkeitsräumen indiziert nach einem Parameter $\theta \in \Theta$. Weiters gibt es eine Zufallsvariable $X : \Omega \to E$, wobei $(E, \mathcal{G})$ ein Messraum ist, die den datenerzeugenden Prozess modelliert.

Es geht in der mathematischen Statistik darum in Abhängigkeit von Beobachtungen $x \in E$ Aussagen über $\theta \in \Theta$ zu machen.

Der wahrscheinlichkeitstheoretische Rahmen kann ein zufälliges Phänomen modellieren, Fehler, oder beides. Man kann auch Ω bzw $\mathcal{F}$ von $\theta \in \Theta$ abhängig machen, was in der Gebiet der Knightean Uncertainty füehrt. Man betrachtet gerne auch auf Θ eine a priori Verteilung für den Parameter $\theta \in \Theta$, was in das Gebiet der Bayesianischen Statistik führt.

Problemstellungen

Man unterscheidet ferner die folgenden 3 wichtigsten Problemstellungen:

- **Punktschätzung** eines unbekannten Parameters.
- Prüfverfahren, ob vorgegebene Parameterwerte mit den Daten verträglich sind (**statistischer Test**).
- Angabe von Schranken, die einen unbekannten Parameter mit vorgegebener Wahrscheinlichkeit eingrenzen (**Vertrauens- oder Konfidenzintervall**).

Daneben gibt es z.B. noch Tests für eine bestimmte Verteilung oder einen bestimmten Verteilungstyp (z.B. "Sind die Daten Poisson-verteilt mit beliebigem λ"?).

Punktschätzungen

Formalismus

Wir verwenden den folgenden Formalismus: Die Beobachtungen seien eine Realisierung eines $\mathbb{R}^n-$wertigen Zufallsvektors $\mathsf{X} = (X_1, X_2, \ldots, X_n)$. Als Modellverteilungen für X haben wir eine Klasse von Verteilungen $(\mu_\theta)_{\theta\in\Theta}$ auf $(\mathbb{R}^n, \mathcal{B}^n)$ zur Verfügung. Eine Punktschätzung von θ ist dann eine im Prinzip beliebige Abbildung

$$T : \mathbb{R}^n \to \Theta.$$

Wir legen also fest, wie unsere Schätzung für beliebige mögliche Daten, nicht nur für die konkret vorliegenden Werte, aussehen würde.

Formalismus

Weil X ein Zufallsvektor ist, ist $T(\mathrm{X})$ ein zufälliges Element von Θ. Um von der Verteilung von $T(\mathrm{X})$ sprechen zu können, brauchen wir eine σ-Algebra auf Θ, und wir müssen T als messbar voraussetzen. In allen praktisch relevanten Fällen ist das nie ein Problem, und wir nehmen an, dass die Verteilung $\nu_\theta[B] = \mathbb{P}_\theta\left(T \in B\right) = \mu_\theta(T^{-1}(B))$ $(B \subset \Theta)$ definiert ist.

Gerne schreiben wir μ_θ oder P_θ wenn wir Verteilungen am Raum der Beobachtungen selbst haben, dh der datengenerierende Prozess X die Identität ist. Sonst zumeist $\mathbb{P}_\theta$ auf Ω mit allgemeinen X. Es tritt natuerlich auch der Fall auf, wo X von θ abhängt, und damit auch μ_θ und $\mathbb{P}$ konstant ist (z.B. die Regression). Das passt aber auch gut in den Formalismus.

Formalismus

Oft ist man aber gar nicht am ganzen Parameter θ interessiert, sondern nur an gewissen Komponenten, dem sogenannten strukturellen Parameter, während man zusätzliche Störparameter gar nicht zu schätzen braucht (vergleiche die vorangegangene Diskussion). Dies berücksichtigen wir, indem wir den Parameter von Interesse $\eta = g(\theta)$ einführen, wobei g meist eine Projektion auf einen niedrig-dimensionalen Raum darstellt. Wenn wir einen Schätzer T von θ haben, dann schätzen wir $\eta = g(\theta)$ meist durch $U := g(T(X)) = g(T)$ (der letzte Ausdruck ist eine Kurzschreibweise – beachte, dass wir an dieser Stelle immer von Zufallsvariablen sprechen). Es gibt aber wie gesagt Situationen, wo es einfacher ist, direkt einen guten Schätzer U von η zu konstruieren.

Falls nun ein Wert von X angenommen wird, das heisst U an einem $\omega \in \Omega$ ausgewertet wird, schreiben wir gerne $\hat{\eta}$. Manchmal wird aber auch für U selbst $\hat{\eta}$ geschrieben, was natürlich zu Verwirrungen führen kann.

Punktschätzer

Wenn wir an $\eta = g(\theta)$ interessiert sind, dann sollte für einen guten Schätzer U von η die Verteilung von U möglichst um $\eta = g(\theta)$ herum konzentriert sein, und zwar nicht nur für ein festes θ, sondern für alle $\theta \in \Theta$.

Mean Squared Error

Wir nehmen hier an, dass $g : \theta \to \mathbb{R}$, d.h. wir sind an einer Komponente von θ interessiert. Das übliche Kriterium ist dann der **Mittlere Quadratische Fehler** (MSE=**m**ean **s**quared **e**rror)

$$\mathbb{E}_\theta \left(|U - g(\theta)|^2 \right),$$

doch gäbe es auch andere Kriterien wie z. B. $\mathbb{P}_\theta \left(|U - g(\theta)| > a \right)$ für ein festes a oder $\mathbb{E}_\theta \left(|U - g(\theta)| \right)$.

Bias

Wir führen den **Systematischen Schätzfehler/Bias**

$$b_U(\theta) = \mathbb{E}_\theta(U) - g(\theta)$$

ein und den sogenannten **Standardfehler**

$$\sigma_U(\theta) = \sqrt{\mathsf{var}[_\theta]U},$$

der die Grösse der zufälligen Schwankungen von U angibt. Dann können wir den mittleren quadratischen Fehler wie folgt zerlegen:

$$\mathbb{E}_\theta\left(|U - g(\theta)|^2\right) = \sigma_U^2(\theta) + b_U^2(\theta).$$

Man möchte, dass $\sigma_U(\theta)$ und $|b_U(\theta)|$ beide klein sind, doch typischerweise kann man den einen Fehler nur auf Kosten des andern verkleinern. Am einfachsten wird es, wenn wir verlangen, dass der systematischen Fehler null sein soll. Dann gilt es, den Standardfehler, oder äquivalent dazu, die Varianz zu minimieren.

Erwartungstreue

U heisst **erwartungstreu** für $g(\theta)$ falls

$$\mathbb{E}_\theta(U) = g(\theta) \text{ für alle } \theta \in \Theta, \quad \text{d.h. } b_U(\theta) \equiv 0.$$

Beispiel: Normalverteilung

Seien $X_1, \ldots, X_n$ *i.i.d.* $\sim \mathcal{N}(\mu, \sigma^2)$. Dann ist der unbekannte Parameter $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$, und ein Schätzer ist $T = (\overline{X}, S_n^2)$, wobei

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i, \quad S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

das arithmetische Mittel und die sogenannte Stichprobenvarianz bezeichnen.

Das statistische Modell ist klar: $\Omega = \mathbb{R}^n$, $X = \mathrm{id}$ and $\Theta = \mathbb{R} \times \mathbb{R}_+$.

Beispiel: Normalverteilung

Man sieht sofort, dass $\overline{X}$ erwartungstreu ist für $g_1(\theta) = \mu$. Um zu sehen, dass S_n^2 erwartungstreu ist für $g_2(\theta) = \sigma^2$, braucht man eine kleine Rechnung:

$$\begin{aligned}\mathbb{E}_\theta\left(\sum_{i=1}^n (X_i - \overline{X})^2\right) &= \mathbb{E}_\theta\left(\sum_{i=1}^n (X_i - \mu)^2\right) - n\mathbb{E}_\theta\left((\overline{X} - \mu)^2\right) \\ &= n\sigma^2 - n\sigma^2/n = (n-1)\sigma^2.\end{aligned}$$

(dies erklärt den am Anfang geheimnisvollen Nenner bei S_n^2!). Man beachte jedoch, dass S_n nicht erwartungstreu ist für $g_3(\theta) = \sigma$, denn wegen der Jensenschen Ungleichung ist $\mathbb{E}_\theta\left(\sqrt{S_n^2}\right) < \sqrt{\mathbb{E}_\theta\left(S_n^2\right)} = \sigma$.

Beispiel: Normalverteilung

In der mathematischen Statistik wird bewiesen, dass $\overline{X}$ und S_n^2 minimalen Standardfehler haben unter allen erwartungstreuen Schätzern für μ bzw. σ^2, und zwar simultan für alle θ's.

Beispiel: Regression

Wir betrachten das Regressionsmodell mit fest gegebenen erklärenden Variablen (x_{ij}) und zufälligen Fehlern ε_i mit $\mathbb{E}(\varepsilon_i) = 0$, $i = 1, \ldots, n$. Der unbekannte Parameter θ besteht dann aus dem strukturellen Parameter β und Störparametern, die sich auf die Verteilung der Fehler beziehen. Wir betrachten den Kleinste-Quadrate-Schätzer $\widehat{\beta}$. Wenn wir den Erwartungswert eines Zufallsvektors komponentenweise definieren, dann folgt aus der Formel (149) und der Linearität des Erwartungswertes, dass

$$\mathbb{E}_\theta\left(\widehat{\beta}\right) = (X'X)^{-1}X'\mathbb{E}_\theta(Y) = (X'X)^{-1}X'X\beta = \beta$$

d.h. $\widehat{\beta}_j$ ist erwartungstreu für β_j $(j = 0, 1, \ldots p)$.

Das statistische Modell ist ein allgemeiner Wahrscheinlichkeitsraum Ω der die ε_i trägt, $\Theta = \mathbb{R}^{p+1}$ und Y ist ein datengenerierender Prozess (beachte hier hängt Y von θ ab).

Asymptotik vieler Beobachtungen

Oft wird die Theorie einfacher im Grenzwert von unendlich vielen Beobachtungen, d.h. wir geben die Anzahl Beobachtungen durch einen zusätzlichen Index an: $(X_1, \ldots, X_n) \sim \mu_{n,\theta}$ auf $(\mathbb{R}^n, \mathcal{B}^n)$ und wir haben eine Folge von Schätzern (U_n) von $g(\theta)$, wobei $U_n : \mathbb{R}^n \to \mathbb{R}$

Konsistenz und asymptotisch normalverteilt

Dann untersuchen wir, ob und wie rasch sich die Folge der Verteilungen von U_n auf $g(\theta)$ konzentriert:

Definition

(U_n) heisst **konsistent** für $g(\theta)$ falls

$$\mathbb{P}_\theta\left(|U_n - g(\theta)| > \varepsilon\right) \to 0 \text{ für alle } \varepsilon > 0 \text{ und für alle } \theta.$$

(U_n) heisst **asymptotisch normalverteilt** mit asymptotischer Varianz $\tau^2(\theta)$, und wir schreiben $U_n \approx\sim \mathcal{N}(g(\theta), \frac{1}{n}\tau^2(\theta))$, falls für alle θ

$$\mathbb{P}_\theta\left(\sqrt{n}(U_n - g(\theta)) \leq x\right) \longrightarrow \Phi(\frac{x}{\tau(\theta)}) \text{ für alle } x.$$

Bei einem asymptotisch normalverteilten Schätzer bestimmt die asymptotische Varianz die Genauigkeit: je kleiner $\tau^2(\theta)$, desto genauer der Schätzer.

Beispiel: Normalverteilung

Es seien wie im vorigen Beispiel $X_1, \ldots, X_n, \ldots \quad i.i.d. \sim \mathcal{N}(\mu, \sigma^2)$. Das arithmetische Mittel $\overline{X}$ ist konsistent und exakt $\mathcal{N}(\mu, \frac{1}{n}\sigma^2)$- verteilt ist. Dies zeigt insbesondere, dass die asymptotische Varianz im Allgemeinen vom ganzen Parameter θ abhängt, und nicht nur von $g(\theta)$. Die Stichproben-Varianz S_n^2 zerlegen wir wie oben:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \frac{n}{n-1} - (\overline{X} - \mu)^2 \frac{n}{n-1} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 + \text{ Rest.}$$

Beispiel: Normalverteilung

Der erste Term ist wieder ein arithmetisches Mittel, auf das wir das Gesetz der grossen Zahlen und den Zentralen Grenzwertsatz anwenden können, und man kann zeigen, dass der Rest asymptotisch vernachlässigbar ist bei beiden Grenzwertsätzen. Also ist S_n^2 konsistent und asymptotisch $\mathcal{N}(\sigma^2, \frac{1}{n}2\sigma^4)$-verteilt, denn $\mathbb{E}\left((X_i - \mu)^4\right) = 3\sigma^4$.

Das arithmetische Mittel ist nicht der einzig mögliche Schätzer von μ. Als Alternative können wir zum Beispiel den Stichprobenmedian U_n betrachten. Mit einem Symmetrieargument folgt, dass die Verteilung von U_n symmetrisch bezüglich μ ist. Daher ist U_n auch erwartungstreu. Wir wissen ferner bereits, dass der Stichprobenmedian U_n asymptotisch normalverteilt ist mit asymptotischer Varianz $\sigma^2\frac{\pi}{2}$). Der Stichprobenmedian ist bei Normalverteilung also weniger genau als das arithmetische Mittel: er streut etwa $\sqrt{\pi/2} \approx 1.25$ mal so viel.

Beispiel: Normalverteilung

Bei breiteren Verteilungen sieht die Sache aber anders aus: Bei der Cauchy-Verteilung mit unbekanntem Lageparameter μ, d.h. X_i i.i.d. mit

$$X_i \sim \frac{1}{\pi} \frac{1}{1 + (x - \mu)^2} dx$$

hat z.B. das arithmetische Mittel die gleiche Verteilung für jedes n, ist also nicht konsistent, der Stichprobenmedian hingegen schon.

Beispiel: Ausreisser

Die Genauigkeit eines Schätzers ist aber nicht das einzige Kriterium. Eine Rolle spielt auch die Empfindlichkeit eines Schätzers auf vereinzelte Ausreisser (grobe Fehler, Beobachtungen mit anderer Verteilung). Die einfachste Formalisierung dieser Empfindlichkeit ist der sogenannte **Bruchpunkt**, der definiert ist als

$$\varepsilon^*(x_1, \ldots, x_n) = \frac{1}{n} \max\{k \in \mathbb{N}_0 \,;\, \sup\{|U(y_1, \ldots, y_n)| \,;\, \#\{y_i \neq x_i\} = k\} < \infty$$

Das heisst also, dass U $\varepsilon^* n$ Ausreisser $x_1, \ldots, x_n$ verkraften kann, aber nicht $\varepsilon^* n + 1$.

Den Bruchpunkt sollte man als Schätzer zu einem gegebenen Schätzer verstehen.

Beispiel: Ausreisser

Das arithmetische Mittel hat offensichtlich Bruchpunkt null. Das α-gestutzte Mittel ($0 < \alpha \leq \frac{1}{2}$), das definiert ist durch Weglassen der $k = [\alpha n]$ kleinsten und k grössten Beobachtungen, hat Bruchpunkt $\varepsilon^* = k/n \approx \alpha$. Für $\alpha = \frac{1}{2}$ ist das gestutzte Mittel der Stichprobenmedian, welcher maximalen Bruchpunkt hat.

Konstruktion von Schätzern

Maximum-Likelihood-Methode (MLE)

Die wichtigste allgemein anwendbare Methode zur Konstruktion von Schätzern ist die **Maximum-Likelihood-Methode**.

Zunächst sei die Verteilung der Beobachtungen μ_θ diskret. Dann definieren wir die Likelihoodfunktion $L : \Theta \longrightarrow \mathbb{R}$

$$L(\theta) = \mu_\theta\big(x_1, \ldots, x_n\big)$$

für feste Beobachtungen $(x_1, \ldots, x_n)$. $L(\theta)$ gibt an, wie wahrscheinlich die gemachten Beobachtungen sind, wenn die zugrunde liegende Verteilung μ_θ ist.

Maximum-Likelihood-Methode

Wenn man θ nicht kennt, ist es plausibel anzunehmen, dass man einen typischen Wert beobachtet hat, d.h. man wird θ schätzen als

$$T(x_1, \ldots, x_n) \in \arg\max_{\theta} L(\theta).$$

Die Bezeichnung arg max bedeutet, dass wir dasjenige Argument suchen, bei dem die Funktion ihr Maximum annimmt.

Falls das Maximum an mehreren Stellen angenommen wird, wählt man willkürlich eine davon. Wenn das Maximum nicht angenommen wird, ist arg max nicht definiert.

Maximum-Likelihood-Methode

Im absolut stetigen Fall definieren wir analog

$$L(\theta) = f_\theta(x_1, \ldots, x_n),$$

wobei f_θ die Dichte bezeichnet.

Statt $L(\theta)$ ist es oft einfacher $\log L(\theta)$ zu maximieren. Oft, aber nicht immer, findet man den Maximum-Likelihood-Schätzer durch Ableiten und null setzen von $\log L(\theta)$.

Lokalisationsmodelle

Seien $X_1, \ldots, X_n$ i.i.d. mit Dichte $f(x - \theta)$. Dann ist

$$L(\theta) = f(x_1 - \theta) \ldots f(x_n - \theta).$$

Je nach Wahl von f erhält man dann klarerweise andere Schätzer, z.B.

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) \Rightarrow \log L(\theta) = -\frac{1}{2} \sum_{i=1}^{n} (x_i - \theta)^2 + C$$

impliziert $T(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i$.

Lokalisationsmodelle

Für

$$f(x) = \frac{1}{2}\exp(-|x|) \Rightarrow \log L(\theta) = -\sum_{i=1}^{n} |x_i - \theta| + C$$

erhalten wir den Stichprobenmedian

$$\begin{array}{lcll} T & \in & [x_{(k)}, x_{(k+1)}] & \text{falls} \quad n = 2k \\ T & = & x_{(k+1)} & \text{falls} \quad n = 2k+1, \end{array}$$

wobei $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ die geordnete Stichprobe bezeichnet. Dies kann man wie folgt einsehen.

Lokalisationsmodelle

Die Funktion $\log L$ ist stetig und linear auf jedem Intervall, das keine Beobachtungen enthält. Für $\theta < x_{(1)}$ ist die Steigung gleich n, und an einer Stelle $x_{(i)}$, die m mal vorkommt in $(x_1, x_2, \ldots x_n)$, nimmt die Steigung um $2m$ ab. Für $n = 2k + 1$ hat $\log L$ also ein Maximum bei $x_{(k+1)}$. Für $n = 2k$ und $x_{(k)} < x_{(k+1)}$ ist das arg max von $\log L$ das ganze Intervall $[x_{(k)}, x_{(k+1)}]$ (die Funktion ist auf diesem Intervall konstant).

Beispiel: das Regressionsmodell

Genau gleich argumentiert man im Regressionsmodell: Bei unabhängigen normalverteilten Fehlern mit konstanter Varianz ist der Maximum-Likelihood-Schätzer für β gleich dem Kleinste-Quadrate-Schätzer, das heisst wir suchen β sodass

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j - \beta_0)^2 \to \min!\,,$$

was genau dem Minimierungsproblem für unabhängig normalverteilte Fehler entspricht.

Beispiel: das Regressionsmodell

Wenn die Normalverteilung durch die Verteilung mit Dichte $f(x) = \frac{1}{2}\exp(-|x|)$ ersetzt wird, erhält man stattdessen den sogenannten L_1-Schätzer

$$\arg\min_{\beta} \sum_{i=1}^{n} |y_i - \sum_{j=1}^{p} \beta_j x_{ij} - \beta_0|.$$

Er wurde ursprünglich von Laplace vorgeschlagen. Wenn die Fehler eine Verteilung haben, die mehr Gewicht in die Tails legt, als die Normalverteilung, oder wenn das Problem sehr hochdimensional ist ($p \gg 1$), dann ist der L_1-Schätzer meist besser.

Beispiel: Schätzung der Varianz einer Normalverteilung

Seien $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$ mit $\theta = (\mu, \sigma^2)$. Dann ist

$$\begin{aligned} \log L(\mu, \sigma^2) &= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 - \frac{n}{2} \log \sigma^2 + C \\ \Rightarrow \quad T(x_1, \ldots, x_n) &= \left(\frac{1}{n} \sum_{i=1}^{n} x_i, \quad \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \frac{1}{n} \sum_{j=1}^{n} x_j \right)^2 \right). \end{aligned}$$

Der Maximum-likelihood-Schätzer für die Varianz ist also nicht erwartungstreu.

Beispiel: Binomialverteilung

Seien die Daten $X \sim \text{Binomial}(n, \theta)$, wobei die Erfolgswahrscheinlichkeit $0 < \theta < 1$ unbekannt ist. Für $x \in \{0, 1, \ldots, n\}$ ist dann

$$p_\theta(x) = \binom{n}{x} \theta^x (1-\theta)^{n-x},$$

und damit

$$\log L(\theta) = \log p_\theta(x) = \log \binom{n}{x} + x \log \theta + (n-x) \log(1-\theta).$$

Daraus folgt

$$\frac{d}{d\theta} \log p_\theta(x) = \frac{x}{\theta} - \frac{n-x}{1-\theta}.$$

Setzen wir das Null, so erhalten wir

$$\frac{x}{\theta} - \frac{n-x}{1-\theta} = 0,$$

und daraus ergibt sich

$$\hat{\theta} = \frac{x}{n}.$$

Beispiel: Multinomialverteilung

Seien die Daten $X_1, \ldots, X_n$ i.i.d. Kopien von X mit Werten in $\{1, \ldots, q\}$. Zum Beispiel kann X eine Klassenbezeichnung sein. Die Wahrscheinlichkeit für eine bestimmte Bezeichnung ist unbekannt, mit

$$P_\theta(X = j) =: \theta_j, \qquad j = 1, \ldots, q,$$

wobei

$$\theta \in \Theta = \left\{ \theta \in \mathbb{R}^q : \theta_j \geq 0 \text{ für alle } j, \text{ und } \sum_{j=1}^{q} \theta_j = 1 \right\}.$$

Wir erhalten

$$\log p_\theta(x) = \sum_{j=1}^{q} 1_{\{x=j\}} \log \theta_j.$$

Beispiel: Multinomialverteilung

Die log-Likelihood für $X = (X_1, \ldots, X_n)$ ist also

$$L(\theta) = \sum_{i=1}^{n} \log p_\theta(x_i) = \sum_{i=1}^{n} \sum_{j=1}^{q} 1_{\{x_i = j\}} \log \theta_j = \sum_{j=1}^{q} N_j \log \theta_j,$$

wobei

$$N_j := \sum_{i=1}^{n} 1_{\{x_i = j\}} = \#\{x_i = j\}$$

die Anzahl der Beobachtungen mit der Bezeichnung j $(j = 1, \ldots, q)$ zählt.

Beispiel: Multinomialverteilung

Um das Maximum der log-Likelihood unter der Bedingung $\sum_{j=1}^{q} \theta_j = 1$ zu finden, benutzen wir einen Lagrange-Multiplikator λ: wir maximieren

$$\sum_{j=1}^{q} N_j \log \theta_j + \lambda \left(1 - \sum_{j=1}^{q} \theta_j \right).$$

Differenzieren und Null setzen liefert die Gleichung

$$\frac{\partial}{\partial \theta_j} \left(\sum_{j=1}^{q} N_j \log \theta_j + \lambda \left(1 - \sum_{j=1}^{q} \theta_j \right) \right) = \frac{N_j}{\theta_j} - \lambda = 0.$$

Beispiel: Multinomialverteilung

Also erhalten wir

$$\hat{\theta}_j = \frac{N_j}{\lambda}, \qquad j = 1, \ldots, q.$$

Aus der Nebenbedingung folgt nun

$$1 = \sum_{j=1}^{q} \frac{N_j}{\lambda},$$

und wegen $\sum_{j=1}^{q} N_j = n$ erhalten wir $\lambda = n$. Der MLE ist also gegeben durch

$$\hat{\theta}_j = \frac{N_j}{n}, \qquad j = 1, \ldots, q.$$

Beispiel: Interval censoring

In diesem Beispiel betrachten wir einen Fall, wo der Parameterraum unendlich-dimensional ist. Wir präsentieren das, um zu illustrieren, dass Maximum Likelihood auch dann benutzt werden kann, wenn der Parameter nicht euklidisch ist.

Sei Z die Ankunftszeit der (normalen) Post. Die Ankunftszeit Z kann nie exakt beobachtet werden. Wir überprüfen den Briefkasten täglich zu einer zufälligen Zeit T. Dann ist entweder die Post schon angekommen, $Y = 1$, oder noch nicht, $Y = 0$. Das Ziel ist nun, eine Schätzung für die Verteilung von Z zu bestimmen. Dieses Problem heisst Intervall-zensiert.

Beispiel: Interval censoring

Sei F die Verteilungsfunktion von Z. Es gilt $P(Y = 1|T = t) = F(t)$ und $P(Y = 0|T = t) = 1 - F(t)$. Also ist die Dichte (wobei das dominierende Mass die Verteilung von T ist) gegeben durch

$$p_F(y, t) = F^y(t)(1 - F(t))^{1-y},$$

und damit gilt

$$\log p_F(y, t) = y \log F(t) + (1 - y) \log(1 - F(t)).$$

Wenn wir den Briefkasten während n Tagen überprüfen, so sind die Daten i.i.d. Kopien $X = \{Y_i, T_i\}_{i=1}^n$ von $X = (Y, T)$. Die log-Likelihood ist

$$L(F) = \sum_{i=1}^{n} \big(y_i \log F(t_i) + (1 - y_i) \log(1 - F(t_i))\big),$$

wobei der Parameter F über den Parameterraum Θ aller Verteilungsfunktionen läuft. Der (nichtparametrische) MLE ist

$$\hat{F} := \arg\max_{F \in \Theta} L(F).$$

Die Momentenmethode

Die Momentenmethode ist ein Vorgehen zur Konstruktion eines Schätzers für einen Parameter, welcher eine Verteilung beschreibt, wenn dieser Parameter endlich-dimensional ist, sagen wir mit Dimension d.

Sei X eine $\mathbb{R}$-wertige Zufallsvariable, und seien die Daten $X_1, \ldots, X_n$ i.i.d. Kopien von X.

k-tes Moment

Für $k \in \mathbb{N}$ ist das *k-te Moment* einer reelwertigen Zufallsvariable Y definiert als

$$\mu_k := E[Y^k]$$

(sofern der Erwartungswert existiert).

Das *k-te Stichprobenmoment* (oder *empirische Moment*) ist ein Schätzer und definiert als

$$\hat{\mu}_k := \frac{1}{n} \sum_{k=1}^{n} x_i^k, \qquad k \in \mathbb{N}$$

bei n Beobachtungen.

Die Momentenmethode

Nehmen wir an, dass X_1 eine Verteilung μ_θ habe, wobei $\theta \in \Theta \subseteq \mathbb{R}^d$.
Dann hängen auch die Momente von X von θ ab, d.h. wir haben

Ein *Momentenmethode-Schätzer* $\hat{\theta}$ ist eine Lösung der Gleichung

$$\mu_k(\theta)|_{\theta=\hat{\theta}} = \hat{\mu}_k, \qquad k = 1, \ldots, d$$

(unter der Annahme, dass eine Lösung existiert).

Die Momentenmethode

Mit der Momentenmethode kreiert man also d Gleichungen mit d Unbekannten und versucht, diese zu lösen. Diese d Gleichungen beruhen auf den Stichprobenmomenten. Der Parameter θ ist eine Lösung der d Gleichungen mit den Stichprobenmomenten ersetzt durch die theoretischen Momente. Weil die Stichprobenmoments nach dem Gesetz der grossen Zahlen nahe bei den theoretischen Momenten liegen, ist der Schätzer $\hat{\theta}$ "sinnvoll": falls die inverse Abbildung von $\theta \mapsto \{\mu_k(\theta)\}_{k=1}^d$ stetig ist, so liegt $\hat{\theta}$ nahe bei θ.

Beispiel: Normalverteilung

Seien die Daten $X_1, \ldots, X_n$ i.i.d. Kopien von $X \sim \mathcal{N}(\mu, \sigma^2)$, wobei sowohl $\mu \in \mathbb{R}$ als auch $\sigma^2 > 0$ unbekannt sind. Dann ist der Momentenmethode-Schätzer gegeben durch

$$\hat{\mu} = \bar{x},$$

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

Beispiel: Gammaverteilung

Sei $X \sim \mathrm{Gamma}(\alpha, \lambda)$ mit Dichte proportional zu $x^{\alpha-1} \exp(-\lambda x)$. Dann gilt

$$E_\theta X = \alpha/\lambda, \qquad \mathrm{Var}_\theta(X) = \alpha/\lambda^2.$$

Also ist $E_\theta X^2 = \alpha(\alpha+1)/\lambda^2$. Der Momentenmethode-Schätzer $(\hat{\alpha}, \hat{\lambda})$ muss also die zwei Gleichungen

$$\hat{\mu}_1 = \hat{\alpha}/\hat{\lambda}, \qquad \hat{\mu}_2 - \hat{\mu}_1^2 = \hat{\alpha}/\hat{\lambda}^2$$

erfüllen. Daraus folgt

$$\hat{\lambda} = \frac{\hat{\mu}_1}{\hat{\mu}_2 - \hat{\mu}_1^2},$$
$$\hat{\alpha} = \frac{\hat{\mu}_1^2}{\hat{\mu}_2 - \hat{\mu}_1^2}.$$

Beispiel: Mischung von Normalverteilungen

Nehmen wir an, die Zufallsvariable X habe die Dichte

$$\mu_\theta(x) := \pi_1 \frac{1}{\tau_1}\varphi\left(\frac{x-\nu_1}{\tau_1}\right) + (1-\pi_1)\frac{1}{\tau_2}\varphi\left(\frac{x-\nu_2}{\tau_2}\right),$$

wobei φ die Dichte der Standard-Normalverteilung ist. Zur Vereinfachung nehmen wir an, dass $\pi_1 = \frac{1}{2}$, $\nu_1 = 0$ und $\tau_1 = 1$ fixiert sind. Wir setzen $\nu := \nu_2$ und $\tau := \tau_2$. Der unbekannte Parameter ist $\theta = (\nu, \tau)$. Wir haben

$$E[X] = \frac{1}{2}\nu, \qquad E[X^2] = \frac{1}{2} + \frac{1}{2}(\nu^2 + \tau^2).$$

Also erfüllt der Momentenmethode-Schätzer $(\hat{\nu}, \hat{\tau})$ die Gleichungen

$$\frac{1}{2}\hat{\nu} = \hat{\mu}_1, \qquad \frac{1}{2} + \frac{1}{2}(\hat{\nu}^2 + \hat{\tau}^2) = \hat{\mu}_2.$$

Daraus erhalten wir

$$\hat{\nu} = 2\hat{\mu}_1,$$
$$\hat{\tau}^2 = 2\hat{\mu}_2 - 4\hat{\mu}_1^2 - 1.$$

Statistische Tests

Problemstellung

Die Beobachtungen seien wieder $X = (X_1, \ldots, X_n)$ und die möglichen Verteilungen $(\mu_\theta;\ \theta \in \Theta)$. Verschiedene Werte von θ entsprechen verschiedenen Hypothesen, und wir nehmen an, dass wir eine sogenannte Nullhypothese überprüfen wollen, die beschrieben ist durch eine Teilmenge $\theta \in \Theta_0 \subset \Theta$. Häufig bedeutet die Nullhypothese "kein Effekt" oder "reiner Zufall". Das Komplement Θ_0^c bezeichnet man üblicherweise als Alternative.

Problemstellung

Aufgrund des beobachteten X soll man entscheiden, ob die Nullhypothese zutrifft, d.h. ob die Verteilung von X gleich einem μ_θ mit $\theta \in \Theta_0$, sein kann. Es gibt nur zwei mögliche Entscheidungen: Entweder man behält die Nullhypothese bei, oder man lehnt sie ab. Offensichtlich gibt es dann zwei mögliche Fehlentscheidungen:

- Die Nullhypothese wird abgelehnt (verworfen), obwohl sie richtig ist (Fehler 1. Art),
- Die Nullhypothese wird akzeptiert (beibehalten), obwohl sie falsch ist (Fehler 2. Art).

Problemstellung

Ein statistischer Test ist eine Entscheidungsregel basierend auf der Beobachtung, d.h.

$$\varphi : (\mathbb{R}^n, \mathcal{B}^n) \longrightarrow \{0, 1\}$$

ist eine messbare Funktion, wobei $\varphi(x) = 0$ heisst "Die Nullhypothese wird akzeptiert" und $\varphi(x) = 1$ "Die Nullhypothese wird verworfen". Eine Entscheidungsregel φ definiert eine messbare Teilmenge $K \subset \mathbb{R}^n$ mit $\varphi = 1_K$. Diese Teilmenge heisst *Verwerfungsbereich oder kritischer Bereich des Tests*. Die Bestimmung eines Tests ist gleichbedeutend mit der Bestimmung des Verwerfungsbereiches.

Offensichtlich ist $\mathbb{E}_\theta(\varphi) = \int \varphi(x)\mu_\theta(dx) = \mathbb{P}_\theta(\varphi = 1)$ die Wahrscheinlichkeit, die Nullhypothese zu verwerfen. Ein guter Test sollte trennscharf sein im Sinne, dass

$$\mathbb{E}_\theta(\varphi) \quad \text{möglichst klein auf} \quad \Theta_0$$

und

$$\mathbb{E}_\theta(\varphi) \quad \text{möglichst gross auf} \quad \Theta_0^c.$$

Niveau und Macht

Wenn

$$\sup_{\theta \in \Theta_0} \mathbb{E}_\theta(\varphi) \leq \alpha$$

(d.h. die Wahrscheinlichkeit eines Fehlers 1. Art ist $\leq \alpha$), dann heisst φ ein Test zum **Niveau** α. Für $\theta \notin \Theta_0$ heisst $\mathbb{E}_\theta(\varphi)$ auch die **Macht des Tests** an der Stelle $\theta \notin \Theta_0$. Die Macht ist also Eins minus die Wahrscheinlichkeit eines Fehlers 2. Art.

Bemerkung

Üblicherweise wählt man ein Niveau α, z.B. 5% oder 1%, und sucht unter allen Tests mit diesem Niveau denjenigen, der $\mathbb{E}_\theta(\varphi)$ maximiert für ein festes $\theta \notin \Theta_0$, bzw. – sofern möglich – für alle $\theta \notin \Theta_0$ (sog. gleichmässig mächtigster Test). Die beiden Fehlerarten werden also nicht symmetrisch behandelt. Weil ein Test durchgeführt wird, um Kritiker und Skeptiker zu überzeugen, ist das Niveau wichtiger als die Macht.

Bemerkung

Es ist zu beachten, dass das Niveau α den Fehler erster Art begrenzt, also eine Ablehnung der Nullhypothese, obwohl sie richtig ist. Bei kleinem α ist die fälschliche Ablehnung der Nullhypothese daher ein seltenes Ereignis. Der Fehler zweiter Art hingegen ist deutlich unbestimmter: er hängt stark davon ab, welche Alternative man betrachtet, und seien Wahrscheinlichkeit konvergiert typischerweise gegen $1 - \alpha$, wenn $\theta \notin \Theta_0$ gegen Θ_0 konvergiert. Das heisst, gewisse Alternativen bleiben auch dann plausibel, wenn man die Nullhypothese akzeptiert. Deshalb ist das Beibehalten der Nullhypothese nicht ein Beweis, dass diese richtig ist.

Beispiel: Binomialverteilung

Wir betrachten 20 unabhängige $0-1$ Experimente mit unbekanntem Erfolgsparameter $p \in \Theta = [0,1]$. Die Beobachtung ist $X =$ Anzahl Erfolge $\sim$ Binomial $(20, p)$, und die Nullhypothese sei $\Theta_0 = [0, \frac{1}{2}]$. Als Beispiel kann man an den Vergleich zweier Behandlungen denken, wo wir die Anzahl Patienten zählen, bei denen die neue Behandlung eine bessere Wirkung hat als die alte. Die Nullhypothese wäre dann, dass die neue Behandlung höchstens gleich gut ist wie die alte.

Beispiel: Binomialverteilung

Ein grosses X spricht gegen $p \leq \frac{1}{2}$, daher setzen wir

$$\varphi(x) = \begin{cases} 1 & \text{falls} \quad x \geq c \\ 0 & \text{falls} \quad x < c \end{cases}$$

Dann ist $\mathbb{E}_p(\varphi) = \sum_{k=c}^{20} \binom{20}{k} p^k (1-p)^{20-k}$ monoton wachsend in p. Wir können c also als Funktion des Niveaus bestimmen, indem wir die Gleichung

$$2^{-20} \sum_{k=c}^{20} \binom{20}{k} \leq \alpha < 2^{-20} \sum_{k=c-1}^{20} \binom{20}{k}$$

lösen. Insbesondere ergibt ein $\alpha \in [0.021, 0.058]$ den Wert $c = 15$.

Beispiel: Binomialverteilung

Wie steht es mit der Macht? Man berechnet

p	0.6	0.7	0.8	0.9
$\mathbb{E}_p(\varphi)$	0.126	0.416	0.804	0.989

Wenn also $p = 0.7$ eine relevante Verbesserung ist, beträgt die Wahrscheinlichkeit, diese auch zu entdecken, nur etwas mehr als 40%, was sicher zu wenig ist. Die einzige Möglichkeit, die Macht zu vergrössern, ohne das Niveau zu vergrössern, ist hier eine grössere Stichprobe.

Bemerkung

Die Fehlerwahrscheinlichkeiten beim Testen beziehen sich auf die Unsicherheit vor der Durchführung. Nachdem der Test durchgeführt wurde und zur Verwerfung der Nullhypothese führte, dann kann man schliessen "Entweder ist die Nullhypothese falsch, oder ein seltenes Ereignis, dessen Wahrscheinlichkeit höchstens $= \alpha$ ist, ist eingetreten".

Hingegen ist es nicht korrekt im Falle der Verwerfung der Nullhypothese zu sagen "Die Nullhypothese ist höchstens mit Wahrscheinlichkeit α richtig". Erstens ist die Korrektheit der Nullhypothese im Allgemeinen kein zufälliges Ereignis, hat also auch keine Wahrscheinlichkeit.

Und selbst wenn man bereit ist, die Korrektheit der Nullhypothese als zufällig anzusehen, dann ist zweitens $\mathbb{P}(\varphi = 1 | \text{Nullhypothese richtig})$ nicht gleich $\mathbb{P}(\text{Nullhypothese richtig} | \varphi = 1)$: Die erste Wahrscheinlichkeit ist nach Konstruktion kleiner oder gleich α, die zweite Wahrscheinlichkeit muss mit der Bayesschen Regel berechnet werden und hängt davon ab, wie gross die sogenannte a priori Wahrscheinlichkeit der Nullhypothese ist.

P-Wert

In der Praxis wird oft der sogenannte **P-Wert** berechnet. Dazu muss man voraussetzen, dass man für jedes α einen Test φ_α festgelegt hat, und dass diese Tests kompatibel sind im Sinne, dass

$$\alpha' < \alpha \Rightarrow \varphi_{\alpha'} \leq \varphi_\alpha$$

(wenn eine Beobachtung auf einem bestimmten Niveau als verträglich mit der Nullhypothese angesehen wird, dann ist sie das erst recht, wenn man das Niveau verkleinert). Dann ist der P-Wert von x definiert als $\pi(\mathrm{x}) = \inf\{\alpha; \varphi_\alpha(\mathrm{x}) = 1\}$, d.h. der P-Wert ist "das kleinste Niveau, bei dem der Test die Nullhypothese gerade noch verwirft".

Bemerkung

Der P-Wert ist informativer als die Angabe der Testentscheidung auf einem festen Niveau, gewissermassen ein verfeinertes “Signifikanzmass”. Wenn der P-Wert kleiner oder gleich α ist, dann verwirft der Test auf dem Niveau α. Der P-Wert darf aber nicht als Wahrscheinlichkeit, dass die Nullhypothese richtig ist, interpretiert werden, vergleiche die Diskussion oben.

Bemerkung

Der P-Wert hängt ab von Daten ab und ist daher eine Zufallsvariable. Wie sieht seine Verteilung aus wenn X eingesetzt wird?

Lemma

Sei $\pi(X) = \inf\{\alpha; \varphi_\alpha(X) = 1\}$ *der P-Wert für die Tests* (φ_α) *mit Niveau* α $(0 \leq \alpha \leq 1)$. *Wenn* $\theta \in \Theta_0$, *dann gilt* $\mathbb{P}_\theta(\pi(X) \leq u) \leq u$. *Wenn* $\mathbb{P}_\theta(\varphi_\alpha(X) = 1) = \alpha$ *für alle* α, *dann gilt* $\mathbb{P}_\theta(\pi(X) \leq u) = u$, *d.h. der P-Wert ist uniform verteilt.*

Beweis

Weil $\alpha \mapsto \varphi_\alpha(x)$ monoton wachsend ist, impliziert $\pi(x) < u$, dass $\varphi_u(x) = 1$. Daraus folgt

$$\mathbb{P}_\theta\left(\pi(X) < u\right) \leq \mathbb{P}_\theta\left(\varphi_u(X) = 1\right) \leq u.$$

Weil $\mathbb{P}_\theta\left(\pi(X) \leq u\right)$ rechtsstetig und monoton wachsend ist, folgt die erste Behauptung. Ferner folgt aus $\varphi_u(x) = 1$, dass $\pi(x) \leq u$, also auf Grund der zusätzlichen Bedingung an $\mathbb{P}_\theta$ dass $u = \mathbb{P}_\theta\left(\varphi_u(X) = 1\right) \leq \mathbb{P}_\theta\left(\pi(X) \leq u\right)$. □

Optimale und randomisierte Tests

Im einfachen Fall, wo die Nullhypothese und die Alternative nur aus je einer Verteilung bestehen, kann man unter allen Tests zum Niveau α den mächtigsten bestimmen – zumindest, wenn man bereit ist, auch sogenannte randomisierte Tests zu betrachten. Für Anwendungen ist die Annahme, dass nur zwei Verteilungen möglich sind, natürlich zu einfach, aber es ist die Grundsituation, die man zuerst verstehen will, und die die Grundlage bildet für Optimalitätsaussagen in komplizierteren Situationen.

Optimale und randomisierte Tests

Wir bezeichnen die Nullhypothese mit μ_0 statt μ_{θ_0} und die Alternative mit μ_1 statt μ_{θ_1}. Wir beginnen mit einer heuristischen Überlegung und nehmen an, dass μ_0 und μ_1 diskret sind. Sei K der Verwerfungsbereich $\{x_i|\varphi(x_i) = 1\}$. Dann müssen wir $\sum_{x_i \in K} \mu_1(x_i)$ maximieren unter der Nebenbedingung $\sum_{x_i \in K} \mu_0(x_i) \leq \alpha$.

Optimale und randomisierte Tests

Wenn wir $\mu_0(x_i)$ als das Gewicht und $\mu_1(x_i)$ als den Wert des i-ten Gegenstandes interpretieren, dann ist dies das bekannte "Rucksackproblem" aus der diskreten Optimierung (K ist die Menge der Gegenstände, die in den Rucksack gepackt werden). Offensichtlich muss K aus denjenigen x_i bestehen, für die μ_1 gross und μ_0 klein ist. Die sogenannte "greedy" Strategie wählt diejenigen Gegenstände, bei denen das Verhältnis Wert:Gewicht, d.h. $\mu_1(x_i)/\mu_0(x_i)$, am grössten ist. Wegen der Diskretheit führt das jedoch nicht unbedingt zur optimalen Lösung. Der einfachste Ausweg lässt zu, dass wir vom i-ten Gegenstand einen Anteil $\varphi(x_i) \in [0, 1]$ einpacken können. Dann ist es offensichtlich optimal, die Gegenstände mit dem grössten Verhältnis $\mu_1(x_i)/\mu_0(x_i)$ auszuwählen.

Optimale und randomisierte Tests

Im Rahmen von Tests heisst dies, dass wir **randomisierte Tests** zulassen, d.h. wir betrachten messbare Funktionen

$$\varphi : (\mathbb{R}^n, \mathcal{B}^n) \longrightarrow [0, 1],$$

mit der Interpretation: "Wenn $\varphi(\mathrm{x}) = \gamma$, dann verwerfen wir die Nullhypothese gemäss einem unabhängigen Zufallsmechanismus mit Wahrscheinlichkeit γ" (also z.B. durch Werfen einer Münze mit Erfolgsparameter γ). Auch bei randomisierten Tests geben $\mathbb{E}_\theta(\varphi)$ $(\theta \in \Theta_0)$ und $1 - \mathbb{E}_\theta(\varphi)$ $(\theta \notin \Theta_0)$ die Fehlerwahrscheinlichkeiten 1. und 2. Art an. Wir müssen also $\mathbb{E}_1(\varphi)$ maximieren unter der Nebenbedingung $\mathbb{E}_0(\varphi) \leq \alpha$.

Optimale und randomisierte Tests

Randomisierte Tests kann man als gewöhnliche Tests basierend auf der erweiterten Beobachtung (X, U) auffassen, wobei U unabhängig von X und uniform auf $[0, 1]$ ist: Man setzt einfach $\Phi(x, u) = 1_{\{u \leq \varphi(x)\}}$. Den letzten Datenpunkt U besorgt man sich dann selber mit einem Zufallsmechanismus. Offensichtlich gilt (mit dem Satz von Fubini)

$$\mathbb{P}_\theta \left(\Phi(X, U) = 1\right) = \int \int_0^1 1_{\{u \leq \varphi(x)\}} du \mu_\theta(dx) = \int \varphi(x) \mu_\theta(dx),$$

d.h. Niveau und Macht des neuen Tests werden so berechnet wie oben angegeben.

Optimale und randomisierte Tests

Die Menge der randomisierten Tests ist eine *konvexe* Menge von Funktionen, weshalb die folgende Optimierungsaufgabe mit einem Dualitätsargument gelöst werden kann.

Um unnötige Annahmen über μ_i zu vermeiden, benutzen wir, dass μ_i $(i = 0, 1)$ bezüglich $\mu_0 + \mu_1$ immer eine Dichte p_i hat, d.h.

$$\mu_i(A) = \int_A p_i(x)(\mu_0(dx) + \mu_1(dx)) \quad \forall A \in \mathcal{B}^n.$$

Optimale und randomisierte Tests

Wenn beide μ_i diskret sind, ist

$$p_i(x) = \frac{\mu_i(x)}{\mu_0(x) + \mu_1(x)}.$$

Falls beide μ_i absolut stetig sind mit Dichten f_i, dann ist

$$p_i(x) = \frac{f_i(x)}{f_0(x) + f_1(x)}.$$

Damit können wir das Hauptresultat formulieren und beweisen:

Neyman-Pearson Lemma

Seien μ_0 und μ_1 zwei Wahrscheinlichkeitsmasse mit Dichten p_0 und p_1 bezüglich $\mu_0 + \mu_1$ und sei $\alpha \in [0, 1]$ gegeben. Dann

1. Es existiert ein randomisierter Test φ und ein $c \in [0, \infty]$ derart dass

$$\mathbb{E}_0(\varphi) = \alpha, \tag{150}$$

$$\varphi(x) = \begin{cases} 1 & \text{falls} \quad p_1(x) > c\ p_0(x) \\ 0 & \text{falls} \quad p_1(x) < c\ p_0(x). \end{cases} \tag{151}$$

   (wobei $\infty \cdot 0$ als 0 definiert ist).

2. Jeder Test, der (150) und (151) erfüllt, ist ein mächtigster Test zum Niveau α.

3. Jeder mächtigste Test zum Niveau α erfüllt (151) $(\mu_0 + \mu_1)$-fast überall. Er erfüllt auch (150), ausser wenn es einen Test φ' gibt mit $\mathbb{E}_1(\varphi') = 1$ und $\mathbb{E}_0(\varphi') < \alpha$.

Liklehoodquotient

$p_1(x)/p_0(x)$ heisst der **Likelihoodquotient** und der Test von 1. der Likelihoodquotiententest. Kurz gesagt ist also der Likelihoodquotiententest optimal.

Beweis

Die Lösung der Optimierungsaufgabe, einen Test zum Niveau α und maximaler Macht zu finden, ist die Suche nach einem optimalen $0 \leq \varphi \leq 1$, sodass

$$I(\varphi p_0) \leq \alpha, \quad I(\varphi p_1) \to \max$$

gilt. Hier bezeichne I das Integral bezüglich dem Mass $\mu_0 + \mu_1$. Diese Optimierungsaufgabe kann im Sinne Lagrangescher Multiplikatoren in folgende Aufgabe umgeformt werden: Suche einen Multiplikator $c \geq 0$ und ein $0 \leq \varphi \leq 1$ mit $I(\varphi p_0) = \alpha$, so dass φ die Zielfunktion

$$\varphi' \mapsto I(\varphi' p_1) - cI(\varphi' p_0) = I(\varphi'(p_1 - cp_0)) \quad (0 \leq \varphi' \leq 1)$$

maximiert. Diese Zielfunktion wird offensichtlich genau dann maximal, wenn (151) fast überall bez. $\mu_0 + \mu_1$ erfüllt ist.

Beweis

Wie kommt diese Umformung des Optimierungsproblemes zustande? Sei C eine kompakte, konvexe Menge in einem euklidischen Vektorraum, dann gilt offensichtlich: $\varphi \in C$ ist eine Lösung des Maximierungsproblemes genau dann wenn

$$I((\varphi' - \varphi)p_1) \leq 0$$

für alle $\varphi' \in C$ gilt. In anderen Worten: keine Bewegung weg von φ inerhalb von C kann zu einer Vergrösserung des Funktionals $\varphi' \mapsto I(\varphi' p_1)$ führen. Wenn es noch eine Nebenbedingung $I(\varphi' p_0) \leq \alpha$ gibt, dann unterscheidet man zwei Fälle:

Beweis

- Der Maximierer erfüllt die Nebenbedingung $I(\varphi p_0) = \alpha$. Dann ist $\varphi \in C$ mit $I(\varphi p_0) = \alpha$ eine Lösung des Maximierungsproblemes genau dann wenn es ein $c \geq 0$ gibt sodass

$$I((\varphi' - \varphi)(p_1 - cp_0)) \leq 0$$

für alle $\varphi' \in C$ gilt.

- Der Maximierer nimmt die Nebenbedingung nicht an, das heisst es gibt nur Maximierer mit $I(\varphi p_0) < \alpha$. Dann ist $\varphi \in C$ mit $I(\varphi p_0) \leq \alpha$ Lösung des Maximierungsproblems falls

$$I((\varphi' - \varphi)p_1) \leq 0$$

für alle $\varphi' \in C$ gilt. Das entspricht einem Lagrangeschen Multiplikator $c = 0$.

Beweis

Der Fall, dass die Menge der $\varphi \in C$ mit $I(\varphi p_0) \leq \alpha$ einpunktig ist, wird durch $c = +\infty$ codiert, weil dann das Maximierungsproblem sofort lösbar ist. Im Falle von Neyman-Pearson entspricht das dem Fall $\alpha = 0$.

Daraus folgt das Vorgehen: suche Lagrangeschen Multiplikator und $\varphi \in C$, sodass die Nebenbedingung angenommen wird, dh $I(\varphi p_0) = \alpha$. Dann hat man jedenfalls einen Maximierer gefunden, denn

$$I((\varphi' - \varphi)(p_1 - cp_0)) \leq 0$$

bedeutet $I(\varphi' p_1) \leq I(\varphi p_1)$ falls $I(\varphi' p_0) \leq I(\varphi p_0) = \alpha$.

Beweis

Man muss also den Vektor $p_1 - cp_0$ so wählen, dass er mit jedem Tangentialvektor $\varphi' - \varphi$ (an einem Randpunkt φ des Argumentebereiches) einen Winkel von mindestens $\pi/2$ bezüglich des durch I induzierten Skalarprodukts $\langle \varphi_1, \varphi_2 \rangle = I(\varphi_1 \varphi_2)$ einschliesst. In diesem Sinne hat das im folgenden konstruierte c die Bedeutung eines Lagrangeschen Multiplikators.

Beweis

Wir beweisen daher zunächst die erste Aussage, dh die Existenz eines Lagrangeschen Multiplikators $c \in [0, \infty]$. Dazu betrachten wir die kumulative Verteilungsfunktion F der Zufallsvariable $Q = p_1(X)/p_0(X)$ für $X \sim \mu_0$ (Q ist mit Wahrscheinlichkeit 1 definiert und endlich). Für $0 < \alpha < 1$, wählen wir $c = F^{-1}(1 - \alpha)$ mit der Quantilfunktion F^{-1}.

Wenn F an der Stelle c einen Sprung hat, dann ist die Sprunghöhe gleich $P_0(Q = c) = P_0(p_1(X) = cp_0(X)) > 0$, und wenn wir auf dieser Menge $\varphi = \gamma \in]0, 1[$ setzen mit geeignetem Wert von γ, dann gilt (150). Für $\alpha = 0$ setzen wir $c = +\infty$, d.h. $\varphi(x) = 0$ für alle x mit $p_0(x) > 0$ und $\varphi(x) = 1$ für alle x mit $p_0(x) = 0$. Für $\alpha = 1$ setzen wir $c = 0$ und $\varphi(x) \equiv 1$.

Beweis

Die Aussagen 2 und 3 folgen nun leicht. Sei φ der Test gemäss Aussage 1, der (150) und (151) erfüllt. Es gilt dann für jeden anderen randomisierten Test φ'

$$I(\varphi' p_1) \leq I(\varphi p_1) + c(I(\varphi' p_0) - I(\varphi p_0))$$

aufgrund der Konstruktion von φ: auf der Menge $p_1 - cp_0 > 0$ is φ maximal and auf $p_1 - cp_0 < 0$ verschwindet φ, also gilt die Ungleichung für jeden randomisierten Test. Wenn also $I(\varphi' p_0) \leq \alpha = I(\varphi p_0)$ gilt, dann ist $I(\varphi' p_1) \leq I(\varphi p_1)$, dh die zweite Aussage ist bewiesen. Damit ist bewiesen, dass an φ ein maximaler Wert der Zielfunktion angenommen wird. Es könnte aber auch sein, dass es weitere Maximierer gibt, allerdings mit geringerem Niveau.

Beweis

Für die dritte Aussage benutzt man, dass die obige Ungleichung strikt ist, ausser wenn (151) $(\mu_0 + \mu_1)$-fast überall auch für φ' gilt. Aus $I(\varphi' p_1) = I(\varphi p_1)$ folgt ausserdem weiters, dass entweder $I(\varphi' p_0) = I(\varphi p_0)$ oder $c = 0$. Wenn $c = 0$, dann ist $\varphi = 1$ für alle x mit $p_1(x) > 0$, dh die Macht von φ und damit auch von φ' ist 1. Der Wert von $I(\varphi' p_0)$ hängt noch davon ab, wie man φ' definiert auf $N = \{x; p_0(x) > 0, p_1(x) = 0\}$. Den minimalen Wert $1 - \mu_0(N)$ erreicht man, wenn man $\varphi' = 0$ setzt auf N, und dieser kann kleiner als α sein. □

Beispiel: Binomialverteilung

Binomialverteilung mit Parameter θ. Sei $\theta_0 < \theta_1$. Dann ist

$$\frac{p_1(x)}{p_0(x)} = \left(\frac{1-\theta_1}{1-\theta_0}\right)^n r(\theta_0, \theta_1)^x$$

wobei

$$r(\theta_0, \theta_1) = \frac{\theta_1}{1-\theta_1}\frac{1-\theta_0}{\theta_0} > 1,$$

denn $\theta/(1-\theta)$ ist strikt monoton wachsend. Also gilt

$$\frac{p_1(x)}{p_0(x)} \geq c \Leftrightarrow x \geq c'.$$

Beispiel: Binomialverteilung

Der Wert der kritischen Grenze c' ist bestimmt durch (150), d.h. er hängt nur von θ_0 und α ab. Also erhalten wir den gleichen optimalen Test für alle $\theta_1 > \theta_0$:

$$\varphi(x) = \begin{cases} 1 & \text{falls} \quad x > c' \\ \gamma & \text{falls} \quad x = c' \\ 0 & \text{falls} \quad x < c' \end{cases}$$

(sogenannter gleichmässig mächtigster Test für $\theta = \theta_0$ gegen $\theta \in (\theta_0, 1]$.)

Einige wichtige Tests

Wir besprechen hier einige der üblichen Tests für einfache, häufig auftretende Situationen. In allen praktischen Anwendungen ist der Verwerfungsbereich gegeben als Urbild einer sogenannten Teststatistik T: Die Nullhypothese wird verworfen, falls $T > c$, wobei die "kritische Grenze" c vom gewählten Niveau α abhängig ist. Wir beschränken uns darauf, die üblichen Teststatistiken anzugeben, ohne genauer zu untersuchen, ob und in welchem Sinne diese Teststatistiken optimal sind. Es sollte aber jeweils sofort einleuchten, dass grosse Werte von T gegen die Nullhypothese sprechen.

1-Stichproben-t-Test

Sei

$$X_1, \ldots, X_n \;\; i.i.d. \sim \mathcal{N}(\mu, \sigma^2),\; \Theta = \mathbb{R} \times \mathbb{R}_+ \;,\; \Theta_0 = \{\mu_0\} \times \mathbb{R}_+.$$

Diese Situation trifft man in der Qualitätskontrolle beim Test auf einen Sollwert μ_0 an, sowie bei verbundenen oder gepaarten Stichproben, wo jeweils 2 Behandlungen bei der gleichen Versuchseinheit durchgeführt werden. Im zweiten Fall ist X_i der Unterschied der Resultate für die beiden Behandlungen bei der i-ten Versuchseinheit, und meist hat man $\mu_0 = 0$, d.h. die Nullhypothese besagt, dass es keinen systematischen Unterschied zwischen den beiden Behandlungen gibt.

1-Stichproben-t-Test

Als Teststatistik wird in diesem Fall gewählt

$$T = \frac{\sqrt{n}(\overline{X} - \mu_0)}{S_n}, \quad S_n^2 = \frac{1}{n-1}\sum(X_i - \overline{X})^2, \quad \overline{X} = \frac{1}{n}\sum X_i,$$

und wir verwerfen die Nullhypothese, falls $|T|$ zu gross (sog. zweiseitiger t-Test), oder T gross ist.

Wir verwenden hier die leicht missbräuchliche Schreibweise, dass wir den datengenerierenden Prozess in den Schätzer bereits einsetzen.

1-Stichproben-t-Test

Die kritische Grenze wird bestimmt mit

Satz

Die Verteilung von T hat für $\mu = \mu_0$, $n \geq 2$ und alle $\sigma > 0$ die Dichte

$$f_{n-1}(t) = \frac{\Gamma(n/2)}{\sqrt{(n-1)\pi}\Gamma((n-1)/2)} \left(1 + \frac{t^2}{n-1}\right)^{-n/2}$$

(sog. $t-$Verteilung mit $n-1$ Freiheitsgraden).

1-Stichproben-t-Test

Damit lautet der Test

$$\varphi(x) = 1 \Longleftrightarrow |T| > c(n-1, \alpha)$$

wobei die kritische Grenze $c(n-1, \alpha)$ das $(1-\alpha/2)$-Quantil der t-Verteilung ist. Dieses ist tabelliert für die üblichen α's und nicht allzu grosse $n-1$'s. Im Grenzfall $n \to \infty$ erhält man die Quantile der Standard-Normalverteilung.

1-Stichproben-t-Test

Wenn man als Nullhypothese $\mu \leq \mu_0$ hat und als Alternative $\mu > \mu_0$ (d.h. man ist nur an Überschreitungen des Sollwertes, bzw. an Behandlungen, die zu einer Verbesserung führen, interessiert), nimmt man $\varphi(x) = 1 \Leftrightarrow T > c'(n-1, \alpha)$ (einseitiger t-Test). Analog geht es bei der Nullhypothese $\mu \geq \mu_0$.

Wenn man die Standardabweichung σ der Beobachtungen X_i kennt, verwenden wir σ statt S_n und entnehmen die kritischen Grenzen einer Tabelle der Normal- statt der t-Verteilung. Der Test heisst dann auch z-Test.

Vorzeichentest

Sei

$$X_1, \ldots, X_n \text{ i.i.d.} \sim F(x-\mu),\ \Theta = \mathbb{R} \times \{F | F(0) = \frac{1}{2}\},\ \Theta_0 : \mu = \mu_0.$$

Man testet also, ob der Median gleich μ_0 sein kann. Im Gegensatz zum t-Test nimmt man keine Normalverteilung mehr an. Aus technischen Gründen setzen wir voraus, dass F stetig ist in 0, so dass $\mathbb{P}(X_i = \mu) = 0$.

Die Teststatistik $T = \sum_{i=1}^{n} 1_{[X_i > \mu_0]}$ ist unter der Nullhypothese Binomial $(n, \frac{1}{2})$-verteilt, und wenn T zu stark abweicht von $\frac{n}{2}$, werden wir die Nullhypothese verwerfen. Wir verwenden also den Test

$$\varphi(x) = 1 \iff |T - n/2| > c,$$

wobei $c = c(n, \alpha)$ mit Hilfe der Binomialverteilung bestimmt wird.

Vergleich von t-Test und Vorzeichentest

Wenn man tatsächlich Normalverteilung hat, stimmt beim $t-$Test das Niveau exakt und in der Mathematischen Statistik wird gezeigt, dass er maximale Macht hat. Falls F symmetrisch ist, stimmt das Niveau beim $t-$Test genähert (wenn F nicht extrem langschwänzig ist) und beim Vorzeichentest exakt. Bei der Macht kommt es sehr auf F an: Je breiter die Verteilung von F ist, desto besser ist der Vorzeichentest.

2-Stichproben-t-Test

Seien

$$X_1, \ldots, X_n \sim \mathcal{N}(\mu_1, \sigma^2), \quad Y_1, \ldots, Y_m \sim \mathcal{N}(\mu_2, \sigma^2), \text{ alle unabhängig.}$$

Dann ist

$$\theta = (\mu_1, \mu_2, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+$$

Als Nullhypothese wählen wir $\Theta_0 = \{\mu_1 = \mu_2\}$, d.h. unter der Nullhypothese haben alle Variablen die gleiche Verteilung. Die Teststatistik ist

$$T = \frac{(\overline{X} - \overline{Y})/\sqrt{1/n + 1/m}}{\sqrt{\left(\sum (X_i - \overline{X})^2 + \sum (Y_i - \overline{Y})^2\right)/(n + m - 2)}}.$$

2-Stichproben-t-Test

Satz

Für alle $\theta \in \Theta_0$ hat T die $t-$Verteilung mit $n+m-2$ Freiheitsgraden.

Der Test lautet also

$$\varphi(x) = 1 \iff |T| > c,$$

wobei man $c = c(n+m-2, \alpha)$ aus Tabellen erhält. Analog geht es bei Nullhypothesen $\mu_1 \leq \mu_2$, bzw. $\mu_1 \geq \mu_2$.

2-Stichproben-Wilcoxon-Test (oder Mann-Whitney U-Test)

Seien

$$X_1, \ldots, X_n \sim F, \quad Y_1, \quad \ldots, Y_m \sim G, \text{ alle unabhängig.}$$

Wir wollen jetzt die Verteilungen F und G nicht näher spezifizieren, aber aus technischen Gründen nehmen wir an, dass beide stetig sind.

2-Stichproben-Wilcoxon-Test (oder Mann-Whitney U-Test)

Es ist also $\Theta = \mathcal{F} \times \mathcal{F}$, wobei $\mathcal{F}$ die Menge der stetigen Verteilungsfunktionen bezeichnet. Die Nullhypothese ist $\Theta_0 = \{F = G\}$, und als Teststatistik nehmen wir

$$W = \sum_{i=1}^{n} \sum_{j=1}^{m} 1_{[X_i < Y_j]} \,.$$

2-Stichproben-Wilcoxon-Test (oder Mann-Whitney U-Test)

Zur Durchführung des Test brauchen wir die Verteilung von W unter der Nullhypothese. Sei $(Z_i)_{1\leq i\leq n+m}$ die kombinierte Stichprobe, $Z_i = X_i \;\; (i \leq n),\; Z_i = Y_{i-n} \;\; (i > n)$. Unter der Nullhypothese sind die Z_i i.i.d., also gilt aus Symmetriegründen für alle Permutationen π

$$\mathbb{P}\left(Z_{\pi(1)} < Z_{\pi(2)} < \cdots < Z_{\pi(n+m)}\right) = \frac{1}{(n+m)!}.$$

Da W konstant ist auf $\{Z_{\pi(1)} < \cdots < Z_{\pi(n+m)}\}$, hängt die Verteilung von W nicht von F ab und kann durch Abzählen gefunden werden ($\rightarrow$ Tabellen für kleine n, m).

2-Stichproben-Wilcoxon-Test (oder Mann-Whitney U-Test)

Für grössere n, m verwendet man:

Lemma

Unter der Nullhypothese gilt

$$\mathbb{E}(W) = \frac{nm}{2}, \quad \operatorname{var} W = \frac{nm(n+m+1)}{12},$$

$$\mathbb{P}\left(\frac{W - \mathbb{E}(W)}{\sqrt{\operatorname{var} W}} \leq x\right) \longrightarrow \Phi(x) \quad (x \in \mathbb{R}; \quad n, m \to \infty).$$

2-Stichproben-Wilcoxon-Test (oder Mann-Whitney U-Test)

$\mathbb{E}(W) = nm \cdot \mathbb{P}(X_i < Y_j) = \frac{nm}{2}$ folgt aus Symmetriegründen. Ferner

$$\operatorname{var} W = \sum_i \sum_j \sum_k \sum_\ell \operatorname{cov}\left(1_{[X_i<Y_j]}, 1_{[X_k<Y_\ell]}\right)$$

und

$$\begin{aligned}
\operatorname{cov}\left(1_{[X_i<Y_j]}, 1_{[X_k<Y_\ell]}\right) &= \mathbb{E}\left(1_{[X_i<Y_j]}1_{[X_k<Y_\ell]}\right) - \left(\mathbb{E}\left(1_{[X_i<Y_j]}\right)\right)^2 \\
&= 0 \text{ falls alle Indizes verschieden} \\
&= \frac{1}{4} \text{ falls } i = k, j = \ell \\
&= \mathbb{P}(X_i < \min(Y_j, Y_\ell)) - \frac{1}{4} \text{ falls } i = k, j \neq \ell \\
&= \mathbb{P}(\max(X_i, X_k) < Y_j) - \frac{1}{4} \text{ falls } i \neq k, j = \ell.
\end{aligned}$$

2-Stichproben-Wilcoxon-Test (oder Mann-Whitney U-Test)

Aus Symmetrie $\mathbb{P}(X_i < \min(Y_i, Y_\ell)) = \mathbb{P}(\max(X_i, X_k) < Y_j) = \frac{1}{3}$. Also

$$\operatorname{var} W = nm\frac{1}{4} + nm(m-1)\frac{1}{12} + nm(n-1)\frac{1}{12} = \frac{nm}{12}(n+m+3-1-1).$$

Auf den Beweis der asymptotischen Normalität verzichten wir.

Vergleich von $t-$Test und Wilcoxon-Test

In der Mathematischen Statistik zeigt man, dass das Niveau beim Wilcoxon-Test immer exakt stimmt, während es beim $t-$Test meist nur genähert richtig ist. Entscheidend ist jedoch, dass der Wilcoxon-Test oft eine wesentlich grössere Macht hat als der $t-$Test, und dass auch im ungünstigsten Fall seine Macht nur wenig kleiner ist als beim t-Test. Deshalb sollte man eher den Wilcoxon-Test verwenden.

Chi-Quadrat-Anpassungstest

Wir betrachten n unabhängige Wiederholungen eines Experiments mit k möglichen Ausgängen. Die Wahrscheinlichkeit für den i-ten Ausgang sei θ_i. Dann ist

$$\Theta = \{(\theta_1, \ldots, \theta_k) |\ \theta_i \geq 0, \sum_{i=1}^{k} \theta_i = 1\}.$$

Wir betrachten zuerst eine einfache Nullhypothese, wo $\Theta_0 = \{\theta_0\} = \{(\theta_{10}, \ldots, \theta_{k0})\}$.

Chi-Quadrat-Anpassungstest

Uniforme Zufallszahlengeneratoren. Wir zerlegen $(0, 1]$ in die Intervalle $I_i = (\frac{i-1}{k}, \frac{i}{k}]$ und schauen, in welchem Intervall die Zufallszahl liegt. Wir haben also k Ausgänge, und wenn der Generator korrekt ist, dann sind alle Ausgänge gleich wahrscheinlich, d.h. $\theta_0 = (\frac{1}{k}, \ldots, \frac{1}{k})$.

Wir arbeiten nicht direkt mit den Resultaten der einzelnen Wiederholungen, sondern schauen nur darauf, wie häufig die verschiedenen Ausgänge vorkommen. Wir betrachten also die Zufallsvariablen $N_i =$ Anzahl Wiederholungen mit Ausgang i, $\quad i = 1, \ldots, k$.

Chi-Quadrat-Anpassungstest

Die Verteilung von $(N_1, \ldots, N_k)$ ist die sogenannte **Multinomial-Verteilung**:

$$\mathbb{P}_\theta\left(N_1 = n_1, \ldots, N_k = n_k\right) = \frac{n!}{n_1! \ldots n_k!} \quad \theta_1^{n_1} \ldots \theta_k^{n_k}$$

mit $\mathbb{E}\left(N_i\right) = n\theta_i$, $\operatorname{var} N_i = n\theta_i(1-\theta_i)$ und $\operatorname{cov}\left(N_i, N_j\right) = -n\theta_i\theta_j$ $(i \neq j)$.

Chi-Quadrat-Anpassungstest

Als Teststatistik wählen wir eine gewichtete Summe der quadrierten Abweichungen der N_i's von ihrem Erwartungswert:

$$\chi^2 = \sum_{i=1}^{k} \frac{(N_i - n\theta_{i0})^2}{n\theta_{i0}},$$

und wir verwerfen, falls $\chi^2 > c$ (hier hat man praktisch immer einseitige Tests). Um c als Funktion von k und α zu bestimmen, müssen wir die Verteilung von χ^2 unter der Nullhypothese kennen. Im Prinzip kann man diese aus der Multinomial-Verteilung berechnen. Meist benützt man aber das Resultat (ohne Beweis), dass χ^2 asymptotisch die gleiche Verteilung hat wie $\sum_{i=1}^{k-1} Y_i^2$, wobei die $Y_1, \ldots, Y_{k-1}$ i.i.d. $\mathcal{N}(0,1)$-verteilt sind.

Chi-Quadrat-Anpassungstest

Diese Verteilung heisst die **Chiquadrat-Verteilung** mit $k-1$ **Freiheitsgraden**. Sie hat die Dichte

$$f_{k-1}(x) = 2^{-\frac{k-1}{2}} \Gamma\left(\frac{k-1}{2}\right)^{-1} e^{-x/2} x^{(k-3)/2}.$$

Die Grenzen $c = c(k, \alpha)$ kann man daher genähert aus Tabellen der Chiquadrat-Verteilung bestimmen.

Chi-Quadrat-Anpassungstest

Für $k = 2$ folgt obiges Resultat aus dem Zentralen Grenzwertsatz. Es gilt nämlich

$$\chi^2 = \frac{(N_1 - n\theta_{10})^2}{n\theta_{10}} + \frac{(n - N_1 - n(1-\theta_{10}))^2}{n(1-\theta_{10})} = \frac{(N_1 - n\theta_{10})^2}{n\theta_{10}(1-\theta_{10})}.$$

Also hat χ^2 asymptotisch die gleiche Verteilung wie Y^2, wobei $Y \sim \mathcal{N}(0,1)$. Als Faustregel gilt, dass die Approximation brauchbar ist, sofern wenn etwa 80% der $n\theta_{i0} \geq 4$ und der Rest ≥ 1. Sonst muss man Klassen zusammenfassen.

Chi-Quadrat-Anpassungstest

Mit einer kleinen Modifikation können wir auch **zusammengesetzte Nullhypothesen** testen. Sei

$$\Theta_0 = \{(\theta_1(\eta), \ldots, \theta_k(\eta)) \big| \eta \in E\}, \quad E \subset \mathbb{R}^r \text{ offen } .$$

Ferner sei $\widehat{\eta}$ eine Schätzung von η. Dann verwenden wir die Teststatistik

$$\chi^2 = \sum_{i=1}^{k} \frac{(N_i - n\ \theta_i(\widehat{\eta}))^2}{n\theta_i(\widehat{\eta})}$$

und verwerfen, wenn $\chi^2 > c$.

Chi-Quadrat-Anpassungstest

Die Verteilung von χ^2 ist schwierig zu bestimmen und hängt von der verwendeten Schätzmethode ab. Falls $\widehat{\eta}$ der Maximum-Likelihood-Schätzer ist, d.h.

$$\widehat{\eta} = \arg\min_{\eta} \sum_{i=1}^{k} N_i \log \theta_i(\eta),$$

dann hat χ^2 asymptotisch eine Chiquadrat-Verteilung mit $k - r - 1$ Freiheitsgraden.

Kontingenztafeln

Wir haben zwei Merkmale mit p, bzw. q Ausprägungen, und die Nullhypothese lautet, dass die Merkmale unabhängig sind. Ein Beispiel sind die beiden Merkmale biologisches Geschlecht und Händigkeit (in diesem Fall ist die Anzahl der Merkmale jeweils durch zwei gegeben, $p = q = 2$, obwohl wir jeweils eine grössere Anzahl nehmen sollten).

Nimmt man jede Kombination von Merkmalen als einen Ausgang, so haben wir genau die Situation für den Chiquadrat-Test:

$$\begin{aligned} k &= pq \\ \Theta &= \{(\theta_{11}, \theta_{12}, \ldots, \theta_{pq})\}, \\ \theta_{ij} &= \mathbb{P}(1.\ \text{Merkmal } = i,\ \ 2.\ \text{Merkmal } = j) \\ E &= \left\{(\eta_1, \ldots, \eta_p), \sum \eta_i = 1\right\} \times \left\{(\xi_1, \ldots, \xi_q), \sum \xi_j = 1\right\}, \\ \theta_{ij}(\eta, \xi) &= \eta_i \xi_j, \quad r = p + q - 2. \end{aligned}$$

Kontingenztafeln

Man findet:

$$\begin{aligned} \widehat{\eta}_i &= N_{i\cdot}/n = \sum_{j=1}^{q} N_{ij}/n, \\ \widehat{\xi}_j &= N_{\cdot j}/n = \sum_{i=1}^{p} N_{ij}/n. \end{aligned}$$

Also ist

$$\chi^2 = \sum_{i,j} \frac{(N_{ij} - N_{i\cdot}N_{\cdot j}/n)^2}{N_{i\cdot}N_{\cdot j}/n}$$

und wir haben $k - r - 1 = (p-1)(q-1)$ Freiheitsgrade.

Kontingenztafeln

Als Zahlenbeispiel entnehmen wir aus einer amerikanischen Untersuchung

	Männer	Frauen	insgesamt	
rechtshändig	2′780	3′281	6′061 =	$N_{1\cdot}$
nicht rechtshändig	311	300	611 =	$N_{2\cdot}$
insgesamt	3′091	3′581	6′672 =	n
	$= N_{\cdot 1}$	$= N_{\cdot 2}$		

Kontingenztafeln

Das ergibt die unter der Nullhypothese erwarteten Häufigkeiten

	Männer	Frauen	insgesamt	
rechtshändig	2′808	3′253	6′061 =	N_1.
nicht rechtshändig	283	328	611 =	N_2.
insgesamt	3′091	3′581	6′672 =	n
	$= N_{\cdot 1}$	$= N_{\cdot 2}$		

und damit

$$\begin{aligned}\chi^2 &= \frac{(2'780-2'808)^2}{2808}+\frac{(3'281-3'253)^2}{3253}+\frac{(311-283)^2}{283}+\frac{(300-328)^2}{328}\\ &= 5.68,\end{aligned}$$

mit 1 Freiheitsgrad. Auf dem Niveau $\alpha = 5\%$ wird die Nullhypothese verworfen, auf dem Niveau $\alpha = 1\%$ wird sie hingegen beibehalten. Wir haben also eine gewisse, aber nicht sehr starke statistische Evidenz gegen die Unabhängigkeit von Linkshändigkeit und biologischem Geschlecht.

Kontingenztafeln

In unserer Diskussion wurde angenommen, dass die N_{ij} multinomialverteilt sind, d.h. im Beispiel oben muss man die Personen zufällig aus einer grossen Population auswählen und danach deren Geschlecht und Händigkeit feststellen. Oft führt man aber die Untersuchung so durch, dass die Spaltentotalen im Voraus fixiert werden. In unserem Beispiel würde man also die Anzahl Männer und Frauen im Voraus festlegen und diese separat zufällig auswählen. Dann hat man q Multinomialverteilungen mit p Ausprägungen, und die Nullhypothese lautet, dass die Wahrscheinlichkeiten für jede Ausprägung gleich sind. Man kann zeigen, dass auch in dieser Situation die gleiche Teststatistik angezeigt ist und dass diese unter der Nullhypothese ebenfalls eine genäherte Chiquadrat-Verteilung mit $(p-1)(q-1)$ Freiheitsgraden hat.

Vertrauensintervalle

Die Beobachtungen seien $X = (X_1, \ldots, X_n)$, die Modellverteilungen für X seien $(\mu_\theta)_{\theta \in \Theta}$ und der interessierende Parameter sei $g(\theta), g : \Theta \to \mathbb{R}$.

Definition

Seien $\underline{T}, \overline{T}$ $\mathbb{R}^n \to \mathbb{R}$ zwei messbare Funktionen mit $\underline{T} < \overline{T}$. Dann heisst $(\underline{T}(X), \overline{T}(X))$ ein **Vertrauensintervall** für $g(\theta)$ zum Niveau $1 - \alpha$, falls

$$\forall \theta \in \Theta : \mathbb{P}_\theta \left(\underline{T}(X) < g(\theta) < \overline{T}(X) \right) > 1 - \alpha \, .$$

Bemerkung

Ein Vertrauensintervall fängt also den interessierenden Parameter mit Wahrscheinlichkeit $1 - \alpha$ ein. Zufällig sind die Intervallgrenzen, nicht der Parameter. Wenn man N unabhängige Wiederholungen des Experiments machen würde, erhielte man N verschiedene Vertrauensintervalle, von denen ungefähr $N(1 - \alpha)$ das wahre $g(\theta)$ enthalten.

Beispiel: Normalverteilung

Seien $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$ mit $\theta = (\mu, \sigma^2)$ und $g(\theta) = \mu$.

Ferner sei $t(n-1, 1-\alpha/2)$ das $(1-\frac{\alpha}{2})$-Quantil der t-Verteilung mit $n-1$ Freiheitsgraden und

$$S_n^2 = \frac{1}{n-1} \sum (X_i - \overline{X})^2$$

die Schätzung der Varianz σ^2. Dann ergibt

$$\overline{X} \pm t(n-1, 1-\alpha/2) \frac{S_n}{\sqrt{n}}$$

ein Vertrauensintervall zum Niveau $1-\alpha$, denn

$$\overline{X} - t(n-1, 1-\alpha/2) \frac{S_n}{\sqrt{n}} \leq \mu \leq \overline{X} + t(n-1, 1-\alpha/2) \frac{S_n}{\sqrt{n}}$$

$$\Leftrightarrow \quad \frac{\sqrt{n}|\overline{X} - \mu|}{S_n} \leq t(n-1, 1-\alpha/2).$$

Beispiel: Normalverteilung

In diesem Beispiel besteht das Vertrauensintervall aus denjenigen Werten μ, für die die Nullhypothese $\mathbb{E}(X_i) = \mu$ akzeptiert wird. Ein solcher Zusammenhang besteht allgemein.

Dualitätssatz

Satz (Dualitätssatz)

Sei C eine messbare Teilmenge von $\mathbb{R}^n \times \mathbb{R}$ mit den messbaren Schnitten $A(\gamma) = \{x \in \mathbb{R}^n | (x, \gamma) \in C\}$ und $B(x) = \{\gamma \in \mathbb{R} | (x, \gamma) \in C\}$. Dann sind die beiden folgenden Aussagen äquivalent:

i) *Für jedes γ ist $\varphi(x) = 1_{A(\gamma)^c}(x)$ ein Test der Nullhypothese $g(\theta) = \gamma$ zum Niveau α mit Verwerfungsbereich $A(\gamma)^c$.*

ii) *$B(X)$ bildet einen Vertrauensbereich für $g(\theta)$ zum Niveau $1 - \alpha$.*

Beweis

i) besagt $\mathbb{P}_\theta(\mathsf{X} \notin A(\gamma)) \leq \alpha$ für alle θ mit $g(\theta) = \gamma$. ii) besagt $\mathbb{P}_\theta(g(\theta) \in B(\mathsf{X})) > 1 - \alpha$ für alle θ, für alle γ. Nach Definition von $A(\gamma)$ und $B(\mathsf{x})$ gilt aber

$$g(\theta) \in B(\mathsf{x}) \Leftrightarrow (\mathsf{x}, g(\theta)) \in C \Leftrightarrow \mathsf{x} \in A(g(\theta)).$$

Daraus folgt die Behauptung. □

Bemerkung

Familien von Tests für die Nullhypothesen $g(\theta) = \gamma$ sind also äquivalent zu Vertrauensbereichen. Diese sind häufig, aber nicht immer Intervalle.

Beispiel: Binomialverteilung

Sei $X \sim$ Binomial(n, p)-verteilt mit $\theta = p \in (0, 1)$. Zunächst bestimmen wir ein exaktes Vertrauensintervall. Wir beginnen mit einem Test für $p = p_0$. Wir fixieren α und bestimmen $\underline{k}(p_0)$ und $\overline{k}(p_0)$ derart dass,

$$\begin{aligned}
\sum_{j=0}^{\underline{k}-1} \binom{n}{j} p_0^j (1-p_0)^{n-j} &\leq \frac{\alpha}{2} < \sum_{j=0}^{\underline{k}} \binom{n}{j} p_0^j (1-p_0)^{n-j} \\
\sum_{j=\overline{k}}^{n} \binom{n}{j} p_0^j (1-p_0)^{n-j} &> \frac{\alpha}{2} \geq \sum_{j=\overline{k}+1}^{n} \binom{n}{j} p_0^j (1-p_0)^{n-j}.
\end{aligned}$$

Beispiel: Binomialverteilung

Der Test

$$\varphi = 0 \Leftrightarrow X \in \{\underline{k}(p_0), \ldots, \overline{k}(p_0)\}$$

hat dann Niveau α. Der entsprechende Vertrauensbereich nach Satz 54 ist

$$B(x) = \{p;\ \underline{k}(p) \leq x \leq \overline{k}(p)\}.$$

Da $\sum_{j=0}^{k} \binom{n}{j} p^i (1-p)^{n-j}$ für festes k monoton fallend in p ist, ergibt sich

$$B(x) = \left[\underline{p}(x), \overline{p}(x)\right],$$

wobei

$$\begin{aligned} \sum_{j=0}^{x} \binom{n}{j} \overline{p}(x)^j (1-\overline{p}(x))^{n-j} &= \frac{\alpha}{2} \quad (x \neq n), \\ \sum_{j=x}^{n} \binom{n}{j} \underline{p}(x)^j (1-\underline{p}(x))^{n-j} &= \frac{\alpha}{2} \quad (x \neq 0). \end{aligned}$$

Für $x = n$ ist $\overline{p} = 1$, für $x = 0$ ist $\underline{p} = 0$.

Beispiel: Binomialverteilung

Ein einfacheres, genähertes Vertrauensintervall erhalten wir aus dem Zentralen Grenzwertsatz. Asymptotisch ist

$$\frac{X - np}{\sqrt{np(1-p)}} \sim \mathcal{N}(0,1).$$

Mit $z_\alpha = \Phi^{-1}(1 - \frac{\alpha}{2})$ hat daher der Test

$$\varphi = 0 \Leftrightarrow |X - np_0| \leq z_\alpha \sqrt{np_0(1-p_0)}$$

das Niveau $\approx \alpha$ für die Nullhypothese $p = p_0$.

Beispiel: Binomialverteilung

Der entsprechende Vertrauensbereich ist

$$B(x) = \{p;\ |x - np| \leq z_\alpha \sqrt{np(1-p)}\}.$$

Durch Umformen erhält man

$$(x - np)^2 \leq z_\alpha^2 np(1-p) \Longleftrightarrow p^2(n + z_\alpha^2) - 2p(x + \frac{z_\alpha^2}{2}) + \frac{x^2}{n} \leq 0,$$

was äquivalent ist zu

$$B(x) = \left[\underline{p}(x), \overline{p}(x)\right]$$

wobei $\underline{p}, \overline{p}$ die Stellen sind, wo in obiger Ungleichung das Gleichheitszeichen gilt. Man erhält

$$\begin{aligned} \underline{p}(x) &= \frac{x + z_\alpha^2/2 - z_\alpha \sqrt{x(1-x/n) + z_\alpha^2/4}}{n + z_\alpha^2}, \\ \overline{p}(x) &= \frac{x + z_\alpha^2/2 + z_\alpha \sqrt{x(1-x/n) + z_\alpha^2/4}}{n + z_\alpha^2}. \end{aligned}$$

Bayes-Statistik

Grundgedanke

Zuvor nahmen wir an, dass X eine Verteilung P_θ hat auf dem Raum der Beobachtungen $\mathcal{X}$, wobei $\theta \in \Theta$ ein unbekannter Parameter ist, den wir *schätzen* möchten.

In der *frequentistischen Statistik* nimmt man das unbekannte θ als fixiert und nicht zufällig an. Dahinger steht der Gedanke dass bei genügend vielen unabhängigen Experimenten die Verteilung P_θ durch die Frequenz des Auftretens identifizierbar ist, da

$$P_\theta(A) = \lim_{n\to\infty} \frac{\sum_{i=1}^n 1_A(X^{(i)})}{n}\,.$$

In der *Bayes-Statistik* nimmt man hingegen an, dass θ selbst zufällig ist, das heisst einer Verteilung gehorcht. Häufig wird dieser Zufall als ein quantitativer Ausdruck von Information betrachtet, aber nicht notwendigerweise.

Grundbegriffe

Nehmen wir an, dass die Familie $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$ durch ein σ-endliches Mass ν dominiert wird. In früheren Abschnitten haben wir für $\theta \in \Theta$ die entsprechenden Gewichtsfunktion oder Dichten als

$$p_\theta(x) = \frac{dP_\theta}{d\nu}(x) \qquad \text{für } x \in \mathcal{X}$$

geschrieben. In Bayes-Notation ist

$$p_\theta(x) = p(x|\theta) \qquad \text{für } x \in \mathcal{X}$$

die Dichte von X, gegeben dass der Parameterwert θ ist. Damit das technisch funktioniert, nehmen wir an, dass Θ ein messbarer Raum ist.

Grundbegriffe

Sei Π eine gegebene Wahrscheinlichkeitsverteilung auf Θ, die *a-priori-Verteilung*. Für ein dominierendes Mass μ ist die *a-priori-Dichte* von θ gegeben durch

$$w(\theta) := \frac{d\Pi}{d\mu}(\theta) \qquad \text{für } \theta \in \Theta.$$

Ist Θ abzählbar, so nehmen wir als $w(\cdot)$ die Gewichtsfunktion von θ.

Grundbegriffe

Die *Randdichte* von X ist definiert durch

$$p(x) = \int p(x|\theta) w(\theta) d\mu(\theta)$$

$$= \begin{cases} \sum_\theta p(x|\theta) w(\theta) & \text{für } \theta \text{ diskret,} \\ \int_\theta p(x|\theta) w(\theta) d\theta & \text{für } \theta \text{ absolutstetig,} \end{cases} \qquad \text{für } x \in \mathcal{X}.$$

Grundbegriffe

Für $p(x) > 0$ ist die *a-posteriori-Dichte* von θ gegeben $X = x$ definiert durch

$$w(\theta|x) := \frac{p(x|\theta)w(\theta)}{p(x)}.$$

Die a-posteriori-Dichte ist also durch die Bayes-Regel gegeben.

Bemerkung

Mit dem Bayes-Ansatz führen die Daten X zu einer a-posteriori-Verteilung für θ.

Aber vielleicht will man auch einen Punktschätzer für θ, also einen einzelnen Wert, der repräsentativ für den Parameter ist. Das könnte der Erwartungswert oder der Median der a-posteriori-Verteilung sein (im Fall, wo $\Theta \subseteq \mathbb{R}$). Ein anderer repräsentativer Wert ist der wahrscheinlichste Wert für θ, gegeben die Daten X.

Der Vorteil der Bayes-Statistik ist dass man vielfältige Information in Gestalt einer Verteilung über die Lage des Parameters θ gegeben eine Beobachtung $x \in \mathcal{X}$ bekommt.

Maximum A posterior Schätzer

Der *Maximum-a-posteriori-Schätzer* (kurz *MAP-Schätzer*) ist

$$\hat{\theta}_{\mathrm{MAP}} := \hat{\theta}_{\mathrm{MAP}}(x) := \arg\max_{\theta \in \Theta} w(\theta|x),$$

sofern das Maximum existiert.

Bemerkung

Um $\hat{\theta}_{\mathrm{MAP}}$ zu bestimmen, braucht man nicht die Randverteilung $p(\cdot)$ zu berechnen; es gilt nämlich

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta \in \Theta} p(x|\theta) w(\theta).$$

Wir benutzen das Proportionalitäts-Symbol $\propto$: für reellwertige Funktionen f und g mit Definitionsbereich Θ schreiben wir $f(\theta) \propto g(\theta) \neq 0$, falls $f(\theta)/g(\theta)$ nicht von $\theta \in \Theta$ abhängt. Dann gilt $w(\theta|x) \propto p(x|\theta) w(\theta)$.

Wir können auch

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta \in \Theta} \big(\log p(x|\theta) + \log w(\theta) \big)$$

schreiben. Anders gesagt heisst das, dass der MAP-Schätzer die log-Likelihood $\log L_X(\theta)$ mit einem "Regularisierungsterm" $\log w(\theta)$ maximiert.

Beispiel: Normalverteilung

$X = (X_1, \ldots, X_n)$ eine i.i.d. Stichprobe aus einer $\mathcal{N}(\theta, 1)$-Verteilung. Nehmen wir an, die a-priori-Verteilung auf θ sei eine $\mathcal{N}(0, 1/\lambda^2)$-Verteilung, wobei $\lambda > 0$ gegeben ist. Dann ist

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta \in \mathbb{R}} \left(-\frac{1}{2} \sum_{i=1}^{n} (x_i - \theta)^2 - \frac{1}{2} \lambda^2 \theta^2 \right) = \frac{\bar{x}}{1 + \lambda^2 / n}.$$

Beispiel: Normalverteilung

Wir sehen also, dass der MAP-Schätzer eine geschrumpfte Version des ML-Schätzers $\bar{x}$ ist. Das ist auch plausibel, denn die $\mathcal{N}(0, 1/\lambda^2)$-a-priori-Verteilung hat eine Präferenz für Werte von θ in der Nähe von 0, und das widerspiegelt sich im MAP-Schätzer durch die Schrumpfung von $\bar{x}$ gegen 0.

Bayes-Entscheidungen in Klassifikationen

Betrachten wir zwei gegebene Dichten $p_0(x)$ und $p_1(x)$ für $x \in \mathcal{X}$. Wenn wir eine Beobachtung X haben, so wollen wir sie Klassifizieren, ob sie aus der Verteilung P_0 mit Dichte p_0 oder aus P_1 mit Dichte p_1 kommt. Sei die a-priori-Verteilung gegeben durch

$$w(\theta) = \begin{cases} w_0 & \text{für } \theta = 0, \\ w_1 & \text{für } \theta = 1, \end{cases}$$

für gegebene $0 < w_0 < 1$ und $w_1 = 1 - w_0$.

Bayes-Entscheidungen in Klassifikationen

Dann ist der MAP Schätzer gegeben durch

$$\hat{\theta}_{\mathrm{MAP}} = \begin{cases} 1, & \text{falls } \frac{p_1(x)}{p_0(x)} > \frac{w_0}{w_1}, \\ q, & \text{falls } \frac{p_1(x)}{p_0(x)} = \frac{w_0}{w_1}, \\ 0, & \text{falls } \frac{p_1(x)}{p_0(x)} < \frac{w_0}{w_1}, \end{cases}$$

wobei $q \in \{0, 1\}$ beliebig ist. Hier benutzen wir

$$w(\theta|x) = \begin{cases} p_0(x)w_0/p(x) & \text{für } \theta = 0, \\ p_1(x)w_1/p(x) & \text{für } \theta = 1. \end{cases}$$

Bayes-Entscheidungen in Klassifikationen

Man beachte, dass

$$p(x) = w_0 p_0(x) + w_1 p_1(x) \qquad \text{für } x \in \mathcal{X}$$

eine Mischung von p_0 und p_1 ist.

Der Schätzer $\hat{\theta}_{\mathrm{MAP}}$ heisst hier eine *Bayes-Entscheidung*, und wir schreiben dafür φ_{Bayes}. Man beachte, dass φ_{Bayes} die gleiche Form hat wie ein Neyman–Pearson-Test.

Bayes-Entscheidungen in Klassifikationen

Nun schreiben wir das Klassifikationsproblem um. Wir benutzen dafür die Notation Y statt θ. Sei Y mit Werten in $\{0, 1\}$ eine Klassifikation und bezeichne X mit Werten in $\mathcal{X}$ die Eigenschaften. Wir schreiben

$$\eta(x) = P(Y = 1|X = x) \qquad \text{für } x \in \mathcal{X}.$$

Dann ist die Bayes-Entscheidung, die wahrscheinlichste Klassifikation zu wählen: gegeben $X = x$ prognostizieren wir $Y = 1$, falls $\eta(x) > 1/2$, wir prognostizieren $Y = 0$, falls $\eta(x) < 1/2$, und wir randomisieren, falls $\eta(x) = 1/2$. Anders gesagt bedeutet das

$$\varphi_{\text{Bayes}}(X) = 1_{\{\eta(X)>1/2\}} + q1_{\{\eta(X)=1/2\}}.$$

Bayes-Entscheidungen in Klassifikationen

Man kann diese Situation auch mit Hilfe von *Entscheidungstheorie* beschreiben. Es gibt zwei mögliche Aktionen, $a = 0$ (wir klassifizieren als von p_0 stammend) und $a = 1$ (wir klassifizieren als von p_1 stammend). Der *Aktionenraum* ist also $\mathcal{A} := \{0, 1\}$. Wir definieren die Verlustfunktion als die Indikatorfunktion des Ereignisses, dass man einen Fehler macht, also

$$L(y, a) := 1_{\{y \neq a\}} \qquad \text{für } (y, a) \in \{0, 1\}^2.$$

Also hat man einen Verlust von 1, wenn man die falsche Aktion wählt. Wir nennen eine Funktion $\varphi : \mathcal{X} \to \{0, 1\}$ eine Entscheidung und definieren ihr *Risiko* als

$$R(y, \varphi) := E\left(L(y, \varphi(X)) \middle| Y = y\right).$$

Bayes-Entscheidungen in Klassifikationen

Also ist

$$R(y, \varphi) = \begin{cases} P_0(\varphi(X) = 1) & \text{für } y = 0, \\ P_1(\varphi(X) = 0) & \text{für } y = 1. \end{cases}$$

Wir definieren dann das *Bayes-Risiko* von φ als das erwartete (bezüglich Y) Risiko, wobei $P(Y = 1) = w_1$ und $P(Y = 0) = w_0$; also erhalten wir

$$r_w(\varphi) = w_0 P_0\big(\varphi(X) = 1\big) + w_1 P_1\big(\varphi(X) = 0\big) = P\big(\varphi(X) \neq Y\big).$$

Eine *Bayes-Entscheidung* ist dann ein Minimierer des Bayes-Risikos, d.h.

$$\varphi_{\text{Bayes}} = \arg \min_{\varphi: \mathcal{X} \to \{0,1\}} r_w(\varphi).$$

Bayes-Inferenz für die Binomialverteilung

Sei $X|\theta \sim \text{Binomial}(n, \theta)$ und $\theta \sim \text{Beta}(r, s)$. Der Erwartungswert der a-priori-Verteilung ist dann $E\theta = \frac{r}{r+s}$. Die a-posteriori-Dichte ist

$$w(\theta|x) \propto p(x|\theta)w(\theta) \propto \theta^x(1-\theta)^{n-x}\theta^{s-1}(1-\theta)^{r-1} = \theta^{x+s-1}(1-\theta)^{n-x+r-1}.$$

Bayes-Inferenz für die Binomialverteilung

Also ist $\theta|X = x \sim \mathrm{Beta}(x + r, n - x - s)$ und der a-posteriori-Erwartungwert ist

$$E(\theta|x) = \frac{x + r}{n + r + s}.$$

Der MAP-Schätzer ist

$$\hat{\theta}_{\mathrm{MAP}} = \frac{x + r - 1}{n + s + r - 2}.$$

Bayes-Inferenz für die Binomialverteilung

Beginnt man zum Beispiel mit einer Gleichverteilung als a-priori-Verteilung, so findet man als a-posteriori-Verteilung

$$w(\theta|x) = (n+1)\binom{n}{x}\theta^x(1-\theta)^{n-x},$$

und $\hat{\theta}_{\mathrm{MAP}}$ ist dasselbe wie der Maximum-Likelihood-Schätzer $\hat{\theta}_{\mathsf{MLE}} = x/n$.

Bayes-Inferenz für die Normalverteilung

Sei $X|\theta \sim \mathcal{N}(\theta, \sigma^2)$ mit $\theta \in \mathbb{R}$ und bekanntem σ^2 (Varianz des Rauschens). Ferner gelte $\theta \sim \mathcal{N}(m, \tau^2)$ für ein gegebenes $\tau^2 > 0$. Dann ist die a-posteriori-Verteilung gegeben durch

$$\theta|x \sim \mathcal{N}\left(m + \frac{\tau^2}{\tau^2 + \sigma^2}(x - m), \frac{\tau^2\sigma^2}{\tau^2 + \sigma^2}\right).$$

Wir sehen, dass der a-posteriori-Erwartungswert

$$E(\theta|x) = m + \frac{\tau^2}{\tau^2 + \sigma^2}(x - m)$$

ist. In diesem Fall ist das auch der MAP-Schätzer.

Bayes-Inferenz für die Normalverteilung

Die Herleitung der Formel ergibt sich aus der Betrachtung einer normalverteilten Zufallsvariable (θ, X) mit Erwarungswert (m, m) und Kovarianzmatrix Σ

$$\begin{pmatrix} \tau^2 & \tau^2 \\ \tau^2 & \tau^2 + \sigma^2 \end{pmatrix}.$$

Unter Zuhilfenahme der Formel auf Seite 223. Wir haben dass Σ^{-1} gleich

$$\begin{pmatrix} \frac{\sigma^2+\tau^2}{\sigma^2\tau^2} & -\frac{\tau^2}{\sigma^2\tau^2} \\ -\frac{\tau^2}{\sigma^2\tau^2} & \frac{\tau^2}{\sigma^2\tau^2} \end{pmatrix}.$$

Dann ist $\Sigma_{|X} = \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}$ und

$$m_{|X} = \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\left(\frac{\sigma^2+\tau^2}{\sigma^2\tau^2}m + \frac{\tau^2}{\sigma^2\tau^2}(x-m)\right) = m + \frac{\tau^2}{\sigma^2+\tau^2}(x-m).$$

Bemerkung

Im Gegensatz zur informationskodierenden Bedeutung der Verteilung von θ im allgemeinen Bayesianischen Kontext, kann man hier θ als Signal behandeln, das selbst eine stochastische Natur haben kann.

Das statistische Modell ist dann einfach durch Überlagern des Rauschens gegeben.

Die Rekonstruktion von θ aus Beobachtungen X nennt man Filtern. Es handelt sich um eine der zentralen Anwendungen der modernen Stochastik.

Rekursive Anwendung der Bayes-inferenz für Normalverteilung

Im Falle eines weissen Rauschens, welches mit der Zeit skaliert, wird in einem Intervall der Länge Δt ein Rauschen mit Varianz $\sigma^2/\Delta t$ addiert. Bezeichnen wir oben die Signalvarianz zum Zeitpunkt 0 mit τ_0, und den Erwartungswert mit m_0, dann erhält man folgende Rekursion für $k \geq 0$

$$m_k = m_{k-1} + \frac{\tau_k^2 (x_{k-1} - m_{k-1})\Delta t}{\sigma^2}$$

und

$$\tau_k^2 = \frac{\sigma^2 \tau_{k-1}^2}{\sigma^2 + \tau_{k-1}^2 \Delta t} \approx \tau_{k-1}^2 - \frac{\tau_{k-1}^4 \Delta t}{\sigma^2} \,.$$

Diese Gleichungen haben dann auch eine Interpretation als gewöhnliche Differentialgleichungen für $\Delta t \to 0$:

$$\frac{d}{dt} m_t = \frac{\tau_t^2}{\sigma^2}(x_t - m_t)\,, \quad \frac{d}{dt}\tau_t^2 = -\frac{\tau_t^4}{\sigma^2}\,.$$

Bemerkung

Hier kann noch zusätzlich eine Dynamik in θ eingefüegt werden (die die Normalverteilungsannahme nicht stört, dh eine affine Transformation $m \mapsto am + b$). Das führt dann zu allgemeinen Filtergleichungen der Bauart

$$\frac{d}{dt} m_t = a m_t + b + \frac{\tau_t^2}{\sigma^2}(x_t - m_t)\,, \quad \frac{d}{dt}\tau_t^2 = -\frac{\tau_t^4}{\sigma^2} + a^2\tau_t^2\,,$$

wodurch das Problem der verrauschten kontinuierlichen Beobachtung einer Trajektorie gelöst wird, welche aus der Weiterpropagation einer Zufallsvariable θ_0, die normalverteilt $\mathcal{N}(m_0, \tau_0^2)$ ist, unter einer Dynamik mit Vektorfeld $m \mapsto am + b$ entsteht.