Wahrscheinlichkeit & Statistik

Jens Eirik Saethre

22. Juni 2020

Inhaltsverzeichnis

Teil I
Wahrscheinlichkeitstheorie

1 Wahrscheinlichkeiten

1.1 Grundbegriffe

Def. 1.1 (**Ereignisraum**). *Ereignisraum* oder *Grundraum* $\mathbf{\Omega} \neq \emptyset$ ist Menge aller möglichen Ergebnisse des Zufallsexperiments. Seine Elemente $w \in \Omega$ heissen *Elementarereignisse*.

Def. 1.2 (**Potenzmenge, Ereignis**). Die *Potenzmenge* von Ω wird mit 2^Ω oder mit $\mathcal{P}(\Omega)$ bezeichnet und ist die Menge aller Teilmengen von Ω. Ein *Ereignis* ist ein solches Element der Potenzmenge, also $A \in \mathcal{P}(\Omega)$. Die Klasse aller beobachtbaren Ereignisse ist $\mathcal{F}$, eine Teilmenge der Potenzmenge.

Def. 1.3 (**σ-Algebra**). Ein Mengensystem $\mathcal{F}$ ist eine σ-Algebra, falls

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

(ii) für jedes $A \in \mathcal{F}$ ist auch Komplement $A^\complement \in \mathcal{F}$.

(iii) für jede Folge $(A_n)_{n\in\mathbb{N}}$ mit $A_n \in \mathcal{F}$ für alle $n \in \mathbb{N}$ ist auch $\bigcup_{n=1}^\infty A_n \in \mathcal{F}$.

Def. 1.4 (**Wahrscheinlichkeitsmass**). Ein *Wahrscheinlichkeitsmass* ist eine Abbildung $P : \mathcal{F} \to [0,1]$ mit folgenden Axiomen:

A0) $P[A] \geq 0 \quad \forall A \in \mathcal{F}$

A1) $P[\Omega] = 1$

A2) $P\left[\bigcup_{i=1}^\infty A_i\right] = \sum_{i=1}^\infty P[A_i]$ für disjunkte Ereignisse A_i.

Aus den Axiomen A1 und A2 lassen sich die folgenden Rechenregeln herleiten:

- $P[A^\complement] = 1 - P[A]$
- $P[\emptyset] = 0$ und $P[\Omega] = 1$
- $A \subseteq B \implies P[A] \leq P[B]$
- $P[A \cup B] = P[A] + P[B] - P[A \cap B]$

1.2 Diskrete Wahrscheinlichkeitsräume

Annahme: Ω ist **endlich** oder **abzählbar unendlich** und $\mathcal{F} = 2^\Omega$. Hier kann man das Wahrscheinlichkeitsmass definieren, in dem man die Wahrscheinlichkeiten der Elementarereignisse addiert.

Ist $\Omega = \{\omega_1, \ldots, \omega_N\}$ endlich mit $|\Omega| = N$ und sind alle ω_i gleich wahrscheinlich, also $p_i = 1/N$, so nennt man Ω einen **Laplace Raum** und P ist die *diskrete Gleichverteilung*. Die Wahrscheinlichkeit eines Ereignisses kann dann wie folgt berechnet werden:

$$P[A] = \frac{\text{Anz. Elementarereignisse in } A}{\text{Anz. Elementarereignisse in } \Omega} = \frac{|A|}{|\Omega|}$$

1.3 Bedingte Wahrscheinlichkeiten

Def. 1.5 (**Bedingte Wahrscheinlichkeit**). A, B Ereignisse und $P[A] > 0$. Die *bedingte Wahrscheinlichkeit* von B unter der Bedingung A ist definiert als

$$P[B \mid A] := \frac{P[B \cap A]}{P[A]}$$

Bei fixierter Bedingung A ist $P[\cdot \mid A]$ wieder ein Wahrscheinlichkeitsmass auf $(\Omega, \mathcal{F})$.

$\implies$ **Multiplikationsregel:** $P[A \cap B] = P[B \mid A] \cdot P[A]$ und *Additionsregel:* $P[A \cup B] = P[A] + P[B] - P[A \cap B]$

Satz 1.1 (**Satz der totalen Wahrscheinlichkeit**). Sei $A_1, \ldots, A_n$ eine Zerlegung von Ω in paarweise disjunkte Ereignisse, d.h. $\bigcup_{i=1}^{n} A_i = \Omega$ und $A_i \cap A_k = \emptyset \; \forall i \neq k$. Dann gilt:

$$P[B] = \sum_{i=1}^{n} P[B \mid A_i] \cdot P[A_i]$$

Beweis. Da $B \subseteq \Omega \implies B \cap \Omega = B = B \cap (\bigcup_{i=1}^{n} A_i) = \bigcup_{i=1}^{n} (B \cap A_i)$. Weiter sind alle Mengen der Art $(B \cap A_i)$ paarweise disjunkt, was bedeutet, dass $(B \cap A_i)$ eine disjunkte Zerlegung von B bilden. Damit folgt dann

$$P[B] = P\left[\bigcup_{i=1}^{n} (B \cap A_i)\right] = \sum_{i=1}^{n} P[B \cap A_i] = \sum_{i=1}^{n} P[B \mid A_i] \cdot P[A_i]$$

□

Bedingte Wahrscheinlichkeiten in mehrstufigen Experimenten können oft als Wahrscheinlichkeitsbäume dargestellt werden.

Satz 1.2 (**Satz von Bayes**). Sei $A_1, \ldots, A_n$ eine Zerlegung von Ω mit $P[A_i] > 0$ für $i = 1 \ldots n$ und B ein Ereignis mit $P[B] > 0$, dann gilt für jedes k

$$P[A_k \mid B] = \frac{P[B \mid A_k] \cdot P[A_k]}{\sum_{i=1}^{n} P[B \mid A_i] \cdot P[A_i]}$$

$$\text{einfacher: } P[A \mid B] = \frac{P[A \cap B]}{P[B]} = \frac{P[B \mid A] \cdot P[A]}{P[B \mid A] \cdot P[A] + P[B \mid \overline{A}] \cdot P[\overline{A}]}$$

Beweis. Verwende Definition der bedingten Wahrscheinlichkeit, wende im Zähler die Multiplikationsregel und im Nenner den Satz der totalen Wahrscheinlichkeit an. □

1.4 Unabhängigkeit

Def. 1.6 (**Unabhängigkeit von 2 Ereignissen**). Zwei Ereignisse A, B heissen *stochastisch unabhängig* falls $P[A \cap B] = P[A] \cdot P[B]$. Ist $P[A] = 0$ oder $P[B] = 0$, so sind zwei Ereignisse immer unabhängig. Ist $P[A] \neq 0$, dann gilt folgende Äquivalenz:

$$A, B \text{ sind unabhängig} \iff P[B \mid A] = P[B]$$

Analoges gilt falls $P[B] \neq 0$.

Def. 1.7 (**allgemeine Unabhängigkeit**). Ereignisse $A_1, \ldots, A_n$ heissen *stochastisch unabhängig*, falls für jede endliche Teilfamilie die Produktformel gilt. D.h. für ein $m \in \mathbb{N}$ und $\{k_1, \ldots, k_m\} \subseteq \{1, \ldots, n\}$ gilt immer

$$P\left[\bigcap_{i=1}^{m} A_{k_i}\right] = \prod_{i=1}^{m} P[A_{k_i}]$$

2 Diskrete Zufallsvariablen und Verteilungen

In diesem Kapitel ist $\Omega \neq \emptyset$ abzählbar oder endlich und $\mathcal{F} = 2^\Omega$ die Potenzmenge von Ω, und damit das Wahrscheinlichkeitsmass P gegeben durch seine Gewichte $p_i = P[\omega_i]$ für alle i.

2.1 Grundbegriffe

Def. 2.1 (**diskrete Zufallsvariable**). Eine *reellwertige diskrete Zufallsvariable* auf Ω ist eine Funktion $X : \Omega \to \mathbb{R}$ mit abzählbarem Wertebereich $\mathcal{W}(X) = \{x_1, \ldots, x_n\}$.

- die *Verteilungsfunktion* von X ist die Abbildung $F_X : \mathbb{R} \to [0,1]$ und ist definiert durch

$$t \mapsto F_X(t) := P[X \leq t] := P[\{\omega \mid X(\omega) \leq t\}]$$

- die *diskrete Dichte* von X ist die Funktion $p_X : \mathcal{W}(X) \to [0,1]$ und ist definiert durch

$$p_X(x_k) := P[X = x_k] = P[\{\omega \mid X(\omega) = x_k\}] \qquad \text{für } k = 1, 2$$

In unserem Fall mit Ω abzählbar und $\mathcal{F} = 2^\Omega$ ist jede Funktion $X : \Omega \to \mathbb{R}$ eine Zufallsvariable. Sind $\Omega, \mathcal{F}$ allgemeiner, dann muss die obige Definition der Verteilung so angepasst werden, dass die Menge $\{X \leq t\}$ ein beobachtbares Ereignis für jedes t ist, also in $\mathcal{F}$ ist. Das bedeutet, dass die Funktion X im allgemeinen Fall $\mathcal{F}$-messbar sein muss.

Def. 2.2 (**Indikatorfunktion**). Für jede Teilmenge $A \subseteq \Omega$ ist die *Indikatorfunktion* I_A von A definiert durch

$$I_A(\omega) := \begin{cases} 1 & \text{falls } \omega \in A \\ 0 & \text{falls } \omega \in A^\complement \end{cases}$$

In unserem Fall ist I_A für jedes $A \subseteq \Omega$ eine Zufallsvariable.

Eigenschaften der Dichte und Verteilungsfunktion

- die Verteilungsfunktion F_X ist vollständig durch die Dichte p_X festgelegt, nämlich:
  $F_X(t) = P[X \leq t] = \sum_{x_k \leq t} P[X = x_k] = \sum_{x_k \leq t} p_X(x_k)$
- für jedes $x_k \in \mathcal{W}(X)$ gilt $0 \leq p_X(x_k) \leq 1$ und $\sum_{x_k \in \mathcal{W}(X)} p_X(x_k) = 1$.
- ist $\mathcal{W} \subseteq \mathbb{R}$ nichtleer und abzählbar und $f : \mathcal{W} \to [0,1]$ eine Funktion mit $\sum_{w_k \in \mathcal{W}} f(w_k) = 1$, dann kann man einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, P)$ und darauf eine Zufallsvariable X konstruieren, deren Gewichtsfunktion gerade die Funktion f ist. Dazu genügt bspw. $\Omega := \mathcal{W}$, $\mathcal{F} := 2^\Omega, P[\{\omega\}] := f(\omega)$ und $X(\omega) = \omega$.
- Die Verteilung beschreibt das stochastische Verhalten einer Zufallsvariable. Das ist dasjenige Wahrscheinlichkeitsmass μ_X auf $\mathbb{R}$, das durch $\mu_X(B) := P[X \in B]$ definiert ist. Ist X diskrete Zufallsvariable $\implies$ μ_X heisst *diskrete Verteilung*. Damit kann man die Verteilung μ_X und die Gewichtsfunktion p_X direkt miteinander identifizieren: der einzige Unterschied besteht darin, dass μ_X als Argumente *Teilmengen* von $\mathcal{W}(\mathcal{X})$ hat, p_X hingegen *Elemente* von $\mathcal{W}(X)$. Folgende Formel beschreibt ihren Zusammenhang:

$$\mu_X(B) = P[X \in B] = \sum_{x_k \in B} p_X(x_k) \qquad \text{für } B \subseteq \mathcal{W}(X)$$

2.2 Erwartungswerte

Def. 2.3 (**Erwartungswert**). Sei X eine diskrete Zufallsvariable mit Gewichtsfunktion $p_X(x)$, dann ist der *Erwartungswert* definiert als

$$\mathbb{E}[X] := \sum_{x_k \in \mathcal{W}(X)} x_K \cdot p_X(x_k)$$

sofern diese Reihe absolut konvergiert. Ansonsten existiert der Erwartungswert nicht.

Man kann den Erwartungswert auch als Summe über Ω schreiben, falls er exisitert, denn dann gilt:

$$\mathbb{E}[X] = \sum_{\omega_i \in \Omega} X(\omega_i) P[\{\omega_i\}] = \sum_{\omega_i \in \Omega} p_i X(\omega_i)$$

(eine weitere Umformung existiert im Skript, Seite 43)

Satz 2.1 (**Erwartungswert von Funktionen von ZV**). Sei X eine diskrete Zufallsvariable mit Gewichtsfunktion $p_X(x)$ und $Y = g(X)$ für eine Funktion $g : \mathbb{R} \to \mathbb{R}$. Dann gilt

$$\mathbb{E}[Y] = \mathbb{E}[g(X)] = \sum_{x_k \in \mathcal{W}(X)} g(x_k) \cdot p_X(x_k)$$

sofern die Reihe absolut konvergiert.

Damit genügt es, die Verteilung von X zu kennen, man muss nicht extra die Verteilung von Y zuerst bestimmen, um den Erwartungswert von Y zu berechnen.

Satz 2.2 (**Eigenschaften des Erwartungswerts**). Seien X, Y Zufallsvariablen mit existentem Erwartungswert. Dann gilt:

(i) **Monotonie:** falls $X(\omega) \leq Y(\omega)$ für alle ω, so gilt $\mathbb{E}[X] \leq \mathbb{E}[Y]$

(ii) **Linearität:** für beliebige $a, b \in \mathbb{R}$ gilt: $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$

(iii) nimmt X nur Werte aus $\mathbb{N}_0 = \{0, 1, 2, \dots\}$ annimmt, dann gilt:

$$\mathbb{E}[X] = \sum_{j=1}^{\infty} P[X \geq j] = \sum_{l=0}^{\infty} P[X > l]$$

Def. 2.4 (**Varianz & Standardabweichung**). Sei X eine diskrete ZV mit $\mathbb{E}[X^2] < \infty$
dann definieren wir die *Varianz* von X als

$$\mathrm{Var}[X] := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$

und die *Standardabweichung* von X als

$$\sigma(X) = \mathrm{sd}(X) := \sqrt{\mathrm{Var}[X]}$$

Beides sind *Streuungsmasse* für die Verteilung von X

Schreiben wir $m_X := \mathbb{E}[X]$ und definieren die Funktion $g(x) := (x - m_X)^2$, dann erhalten wir

$$\mathrm{Var}[X] = \sum_{x_k \in \mathcal{W}(X)} (x_k - m_X)^2 \cdot p_X(x_K)$$

Lemma 2.1. Die Varianz von Zufallsvariablen hat folgende Eigenschaften:

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

(ii) $\mathrm{Var}[aX + b] = a^2 \cdot \mathrm{Var}[X]$

2.3 Gemeinsame Verteilungen & Unabhängige Zufallsvariablen

Def. 2.5 (**Gemeinsame Verteilung & Dichte**). Seien $X_1, \dots, X_n$ Zufallsvariablen. Die *gemeinsame Verteilungsfunktion* von $X_1, \dots, X_n$ ist die Abbildung $F : \mathbb{R}^n \to [0, 1]$ definiert durch

$$(x_1, \dots, x_n) \mapsto F(x_1, \dots, x_n) := P[X_1 \leq x_1, \dots, X_n \leq x_n]$$

Sind $X_1, \dots, X_n$ diskrete Zufallsvariablen, so definiert man ihre *gemeinsame Gewichtsfunktion* $p : \mathbb{R}^n \to [0, 1]$ durch

$$p(x_1, \dots, x_n) := P[X_1 = x_1, \dots, X_n = x_n]$$

. Es ist klar, dass $p(x_1, \dots, x_n) = 0$ falls das Ereignis $(x_1, \dots, x_n)$ nicht im gemeinsamen Wertebereich liegt.

Aus der gemeinsamen Gewichtsfunktion p erhält man die gemeinsame Verteilungsfunktion:

$$F(x_1, \dots, x_n) = \sum_{y_1 \leq x_1, \dots, y_n \leq x_n} p(y_1, \dots, y_n)$$

Def. 2.6 (**Randverteilung**). Sein X, Y Zufallsvariablen mit der gemeinsamen Verteilungsfunktion F. Dann ist die *Randverteilung* von X gegeben durch

$$F_X : \mathbb{R} \to [0, 1] \text{ mit } x \mapsto F_X(x) := P[X \leq x] = P[X \leq x, Y < \infty] = \lim_{y \to \infty} F(x, y)$$

Sind X, Y diskrete Zufallsvariablen mit $\mathcal{W}(Y) = \{y_1, y_2, \dots\}$ und gemeinsamer Gewichtsfunktion $p(x, y)$, so ist die *Gewichtsfunktion der Randverteilung* von X gegeben durch

$$\begin{aligned} p_X : \mathcal{W}(X) \to [0,1] \text{ mit } x \mapsto p_X(x) &= P[X = x] \\ &= \sum_{y_j \in \mathcal{W}(Y)} P[X = x, Y = y_j] \\ &= \sum_{y_j \in \mathcal{W}(Y)} p(x, y_j) \quad \text{für } x \in \mathcal{W}(X) \end{aligned}$$

Analoge Aussagen gelten natürlich für Y.

Für Vektoren von diskreten Zufallsvariablen $(X_1, \dots, X_n)$ definiert man die Randverteilungen für jeden möglichen *Teilvektor* von $(X_1, \dots, X_n)$. Es gibt also eindimensionale, aber auch multi-dimensionale Randverteilungen!

Bei zweidimensionalen diskreten Zufallsvariablen erhält man die Gewichtsfunktionen der Randverteilungen als Zeilen- bzw. Spaltensummen der gemeinsamen Gewichtsfunktionen, wie das folgende Bespiel illustriert:

x \ y	0	1	2	3	$p_X(x)$
0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	0	$\frac{1}{2}$
1	0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{1}{2}$
$p_Y(y)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	

Aus den Randverteilungen kann man jedoch nicht ohne Weiteres die gemeinsame Verteilung herleiten, dazu fehlt Information über die *Abhängigkeitsstruktur* der Zufallsvariable.

Def. 2.7 (**Unabhängigkeit**). Zufallsvariablen $X_1, \dots, X_n$ heissen *unabhängig*, falls gilt

$$F(x_1, \dots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n)$$

Folgendes Lemma gibt den Zusammenhang zu unabhängigen Ereignissen:

Lemma 2.2. Die diskreten Zufallsvariablen $X_1, \dots, X_n$ sind unabhängig
$\iff$ für beliebige Teilmengen $B_i \subseteq \mathcal{W}(X_i), i = 1 \dots n$ sind die Ereignisse $A_i := \{X_i \in B_i\}$ für $i = 1 \dots n$ unabhängig
$\iff$ für beliebige Teilmengen $B_i \subseteq \mathcal{W}(X_i), i = 1 \dots n$ gilt:

$$P[X_1 \in B_1, \dots, X_n \in B_n] = \prod_{i=1}^{n} P[X_i \in B_i]$$

Satz 2.3 (**Funktionen auf Zufallsvariablen**). Seien $X_1, \dots, X_n$ diskrete unabhängige Zufallsvariablen und $f_i : \mathbb{R} \to \mathbb{R}$ irgendwelche Funktionen. Sei weiter $Y_i := f_i(X_i)$ für $1 \leq 1 \leq n$. Dann sind die Zufallsvariablen $Y_1, \dots, Y_n$ ebenfalls unabhängig.

2.4 Funktionen von mehreren Zufallsvariablen

Sind $X_1, \dots, X_n$ diskrete Zufallsvariablen, dann ist $Y = g(X_1, \dots, X_n)$ wieder eine Zufallsvariable für eine Funktion $g : \mathbb{R}^n \to \mathbb{R}$.

Satz 2.4. Seien $X_1, \dots, X_n$ diskrete Zufallsvariablen mit endlichen Erwartungswerten. Sei $Y = a + \sum_{i=0}^{n} b_i X_i$ für Konstanten a, b_i. Dann gilt:

$$\mathbb{E}[Y] = a + \sum_{i=0}^{n} b_i \mathbb{E}[X_i]$$

Def. 2.8 (**Kovarianz**). Seien X, Y Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, P)$ mit endlichen Erwartungswerten. Dann ist die *Kovarianz* definiert als

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

Def. 2.9 (**Korrelation**). Die Korrelation von X und Y ist definiert durch

$$\rho(X,Y) := \begin{cases} \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)} & \text{falls } \sigma(X)\sigma(Y) > 0 \\ 0 & \text{sonst} \end{cases}$$

Satz 2.5 (**Wertebereich der Korrelation**). Seien X, Y wie in der Definition der Kovarianz, dann folgt aus der Cauchy-Schwarz Ungleichung, dass $|Cov(X,Y)| \leq \sigma(X)\sigma(Y)$, und damit folgt für die Korrelation

$$-1 \leq \rho(X,Y) \leq 1$$

Wir haben bereits gesehen, dass der Erwartungswert linear ist. Für die Varianz ist dies nicht ganz so einfach. Es gilt:

Korollar 2.1 (**Summenformel für Varianzen**).

$$\text{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \text{Var}[X_i] + 2 \cdot \sum_{i<j} Cov(X_i, X_j)$$

Ist $Cov(X,Y) = 0$, so nennt man X und Y **unkorreliert**. $\implies$ Linearität der Varianz gilt nur für unkorrelierte Zufallsvariablen. Für Produkte von Zufallsvariablen gilt:

Satz 2.6 (**Produkte von Zufallsvariablen**). Seien $X_1, \ldots, X_n$ diskrete Zufallsvariablen mit endlichen Erwartungswerten. Falls $X_1, \ldots, X_n$ unabhängig sind, dann gilt

$$\mathbb{E}\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} \mathbb{E}[X_i]$$

Insbesondere sind dann $X_1, \ldots, X_n$ paarweise unkorreliert und daher gilt

$$\text{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \text{Var}[X_i]$$

sofern die Varianzen existieren und endlich sind.

Bemerkung: Es gilt die Implikationskette: unabhängig $\implies$ paarweise unabhängig $\implies$ unkorreliert

Bemerkung: Es gibt keine allgemeine Produktregel für Varianzen!

Faltung

Seien X, Y diskrete Zufallsvariablen mit gemeinsamer Gewichtsfunktion $p(x,y)$. Dann ist auch ihre Summe $Z := X + Y$ diskret. Damit können wir die Gewichtsfunktion von Z beschreiben durch

$$p_Z(z) = P[Z = z] = \sum_{x_k \in \mathcal{W}(X)} P[X = x_k, Y = z - x_k] = \sum_{x_k \in \mathcal{W}(X)} p(x_k, z - x_k)$$

oder analog via Symmetrie $= \sum_{y_j \in \mathcal{W}(Y)} p(z - y_j, y_j)$. Dies ist ein völlig allgemeines Resultat. Sind nun X und Y unabhängig, dann gilt bekanntlich $p(x,y) = p_X(x) \cdot p_Y(y)$. Damit folgt die bekannte *Faltung* der Gewichtsfunktionen p_X und p_Y:

$$p_Z(z) = \sum_{x_k \in \mathcal{W}(X)} p_X(x_k) \cdot p_Y(z - x_k) = \sum_{y_j \in \mathcal{W}(Y)} p_X(z - y_j) \cdot p_Y(y_j)$$

und schreiben dies kurz als $p_Z = p_X * p_Y = p_Y * p_X$.

2.5 Bedingte Verteilungen

Hier haben wir die gemeinsame Verteilung zweier Zufallsvariablen und wollen Informationen, die wir über eine der beiden Zufallsvariablen haben, ausnutzen um eine genauere Aussage über die andere Zufallsvariable zu machen.

Def. 2.10 (**bedingte Gewichtsfunktion**). X, Y diskrete ZV mit gemeinsamer Gewichtsfunktion $p(x,y)$. Die *bedingte Gewichtsfunktion* von X, gegeben dass $Y = y$, ist definiert als

$$p_{X \mid Y}(x \mid y) := P[X = x \mid Y = y] = \frac{P[X = x, Y = y]}{P[Y = y]} = \frac{p(x,y)}{p_Y(y)}$$

für $p_Y(y) > 0$ und 0 sonst.

Lemma 2.3 (**Kriterium für Unabhängigkeit**)**.** Aus der Charakterisierung der Unabhängigkeit folgt sofort:
X und Y sind unabhängig $\Longleftrightarrow$ für alle y mit $p_Y(y) > 0$ gilt: $p_{X \mid Y}(x \mid y) = p_X(x) \quad \forall x \in \mathcal{W}(X)$.

Eine symmetrische Aussage gilt natürlich, wenn X und Y vertauscht werden.

Bemerkung: Man kann auch auf ein Ereignis bedingen, welches man dann mithilfe einer Indikatorvariable in eine Zufallsvariable verwandelt (siehe Beispiel Seite 64)

3 Wichtige Diskrete Verteilungen

3.1 Diskrete Gleichverteilung

Die *diskrete Gleichverteilung* existiert nur auf einer endlichen Menge. Sie gehört zu einer ZV X mit Wertebereich $\mathcal{W}$ und Gewichtsfunktion

$$p_X(x_k) = P[X = x_k] = \frac{1}{N} \text{ für } k = 1, \ldots, N$$

3.2 Unabhängige 0-1 Experimente

Wir betrachten eine Folge gleichartiger Experimente, die alle nur mit Erfolg oder Misserfolg enden können und betrachten die Ereignisse $A_i = \{\text{Erfolg beim } i\text{-ten Experiment}\}$. Wir nehmen an, dass alle A_i unabhängig sind und dass $P[A_i] = p$ für alle i. Wir können nun eine Indikatorfunktion $Y_i = I_{A_i}$ für jedes i definieren, und danach die Folge von Ereignissen als Folge von 0 und 1 codieren. Dies werden wir für die nächsten Verteilungen brauchen.

4 Allgemeine Zufallsvariablen

4.1 Grundbegriffe

Def. 4.1 (Zufallsvariable). Sein $(\Omega, \mathcal{F}, P)$ ein Wahrscheinlichkeitsraum. Eine *Zufallsvariable* (ZV) auf Ω ist eine messbare Funktion $X : \Omega \to \mathbb{R}$, das bedeutet, dass die Menge $\{X \leq t\} = \{\omega \mid X(\omega) \leq t\}$ für jedes t ein beobachtbares Ereignis, also $\in \mathcal{F}$ sein muss.
Die *Verteilungsfunktion* (VF) von X ist die Abbildung $F_X : \mathbb{R} \to [0, 1]$ mit

$$t \mapsto F_X(t) := P[X \leq t] := P[\{\omega \mid X(\omega) \leq t\}]$$

Wir betrachten nur messbare Zufallsvariablen in dieser Vorlesung.

Satz 4.1 (Eigenschaften der Verteilungsfunktion). F_X hat folgende Eigenschaften:

(i) F_X ist *wachsend* und *rechtsstetig*: $F_X(s) \leq F_X(t)$ für $s \leq t$ und $F_X(u) \to F_X(t)$ für $u \to t$ mit $u > t$.

(ii) $lim_{t\to-\infty} F_X(t) = 0$ und $\lim_{t\to\infty} F_X(t) = 1$

Das stochastische Verhalten einer ZV X wird durch die *Verteilung* beschrieben, d.h. das Wahrscheinlichkeitsmass μ_X, welches durch $\mu_X(B) = P[X \in B]$ definiert ist. Sobald die Verteilungsfunktion F_X bekannt ist, ist das Mass μ_X festgelegt, nämlich durch den Zusammenhang

$$F_X(t) = \mu_X((-\infty, t])$$

Anstelle der Gewichtsfunktion aus dem diskreten Fall verwenden wir die *Dichtefunktion*, sofern diese existiert.

Def. 4.2 (Dichtefunktion). Eine ZV X mit Verteilungsfunktion $F_X(t) = P[X \leq t]$ heisst *(absolut) stetig* mit Dichtefunktion $f_X : \mathbb{R} \to [0, \infty)$, falls gilt

$$F_X(t) = \int_{-\infty}^{t} f_X(s)\, ds \qquad \text{für alle } t \in \mathbb{R}.$$

Bemerkung: X heisst stetig, falls F_X nur stetig ist. Eine ZV X mit einer Dichte hat aber eine VF F_X, die fast überall differenzierbar ist. Dafür verwenden wir den Begriff *stetig mit Dichte*.

Satz 4.2 (Eigenschaften der Dichte). Die Dichtefunktion f_X hat folgende Eigenschaften:

(i) $f_X \geq 0$ und $f_X = 0$ ausserhalb des Wertebereichs $\mathcal{W}(X)$

(ii) $\int_{-\infty}^{\infty} f_X(s)\, ds = 1$ (dies folgt aus Eigenschaft (ii), 2. GW von der Verteilungsfunktion.)

In beinahe allen praktischen Beispielen ist f_X zusätzlich stetig oder zumindest stückweise stetig.

Die Dichtefunktion ist beinahe analog zur Gewichtsfunktion für diskrete Zufallsvariablen, jedoch unterscheidet sie sich in Punktwahrscheinlichkeiten. Es gilt

$$P[a < X \leq b] = P[X \leq b] - P[X \leq a] = F_X(b) - F_X(a) = \int_a^b f_X(s)\, ds$$

$$\implies P[X \in B] = \int_B f_X(s)\, ds$$

und betrachtet man nun einen Grenzwert, so erhält man

$$\lim_{\varepsilon\to 0^+} P[t - \varepsilon < X \leq t + \varepsilon] = \lim_{\varepsilon\to 0^+} \int_{t-\varepsilon}^{t+\varepsilon} f_X(s)\, ds = 0 = P[X = t]$$

Damit ist die Punktwahrscheinlichkeit an jedem Punkt $= 0$. Jedoch gilt für kleine ε (wir verwenden hier $\varepsilon = dt$) das Folgende:

$$P[X \in (t, t + dt]] = f_X(t)dt$$

In allen vernünftigen Situationen gilt also der folgende Zusammenhang zwischen Dichtefunktion und Verteilung:

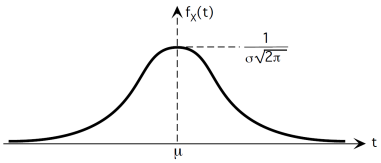
$$\text{Dichtefunktion} = \text{Ableitung der Verteilungsfunktion}$$

Vom diskreten zum stetigen Fall kommt man , indem Summen durch Integrale und die Gewichtsfunktion durch die Dichte ersetzt.

4.2 Normalverteilung

Normalverteilung oder *Gauss-Verteilung* nimmt zwei Parameter $\mu \in \mathbb{R}$, $\sigma^2 > 0$. Ihre Dichte ist symmetrisch um μ und hat eine glockenförmige Gestalt.

- **Wertebereich:** $\mathcal{W}(X) = \mathbb{R}$
- **Dichtefunktion:** $f_X(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$ für $t \in \mathbb{R}$
- **Erwartungswert:** $\mathbb{E}[X] = \mu$ und **Varianz:** $\mathrm{Var}[X] = \sigma^2$
- **Verteilungsfunktion:** entspricht dem Integral von der Dichtefunktion über dem Intervall $[-\infty, t)$, es existiert jedoch kein geschlossener Term.
- **Notation:** $X \sim \mathcal{N}(\mu, \sigma^2)$



Mit einer Normalverteilung können z.b: die Streuung von Messwerten um ihren Mittelwert, Gewichte bzw. Grössen in Bevölkerungen, Leistungen in IQ-Tests und viele mehr modelliert werden. Der Grund für die Wichtigkeit der Normalverteilung liegt im *Zentralen Grenzwertsatz*, der in Kapitel 5 besprochen wird.

4.2.1 Standard-Normalverteilung

Die *Standard-Normalverteilung* gibt die beiden Parameter vor: $\mu = 0$ und $\sigma^2 = 1$.

- **Dichtefunktion:** $\varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$
- **Verteilungsfunktion:** Wieder existiert kein geschlossener Ausdruck, jedoch ist das Integral *tabelliert*:

$$\Phi(t) = \int_{-\infty}^{t} \varphi(s)\, ds = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{s^2}{2}}\, ds$$

Wichtig: $X \sim \mathcal{N}(\mu, \sigma^2) \implies \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$. Daraus folgt unmittelbar, dass es ausreicht, nur die Werte von $\Phi(t)$ zu tabellieren, denn es gilt:

$$F_X(t) = P[X \le t] = P\left[\frac{X-\mu}{\sigma} \le \frac{t-\mu}{\sigma}\right] = \Phi\left(\frac{t-\mu}{\sigma}\right)$$

4.3 Erwartungswerte

Eine beliebige reellwertige ZV X kann immer durch eine Folge diskreter ZV approximiert werden. Ist bspw. $X \ge 0$, dann kann man

$$X_N := \sum_{k=1}^{n2^n} \frac{k-1}{2^n} I_{\{\frac{k-1}{2^n} \le X \le \frac{k}{2^n}\}} + n I_{\{X \ge n\}}$$

für $X_n \nearrow X$ wählen und erhält den Erwartungswert als

$$\mathbb{E}[X] := \lim_{n\to\infty} \mathbb{E}[X_n]$$

Für allgemeine Zufallsvariablen zerlegt man $X = X^+ - X^- := \max(X, 0) - \max(-X, 0)$ mit $X^+, X^- \ge 0$ und setzt dann $\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-]$. Sind diese beiden Erwartungswerte nicht endlich, so existiert der Erwartungswert von X nicht (in $\mathbb{R}$).

Erwartungswert berechnen: Ist X stetigt mit einer Dichte $f_X(x)$, so gilt (sofern konvergent):

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f_X(x)\, dx$$

Cauchy-Verteilung: $\mathcal{W}(X) = \mathbb{R}$ mit Dichte $f_X(x) = \frac{1}{\pi}\frac{1}{1+x^2}$ und Verteilung $F_X(x) = \frac{1}{2} + \frac{1}{\pi}\arctan(x)$. Es gilt, dass für zwei unabhängige, $\mathcal{N}(0,1)$-verteilte ZV X, Y ihr Quotient $Z := X/Y$ gerade *Cauchy-verteilt* ist. Die Charakteristik liegt darin, dass die Dichte für $|x| \to \infty$ sehr langsam gegen 0 geht, d.h. auch sehr grosse Werte nocht mit substantieller Wahrscheinlichkeit angenommen werden. Ein Erwartungswert existiert nicht.

Satz 4.3. Seien X und $Y = g(X)$ zwei ZV. Ist X stetig mit Dichte $f_X(x)$ dann gilt (sofern das Integral konvergiert)

$$\mathbb{E}[Y] = \mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f_X(x)\, dx$$

Weitere Eigenschaften für Erwartungswerte gelten analog zum diskreten Fall, einzig die konkreten Berechnungen unterscheiden sich.

4.4 Momente & Absolute Momente

Def. 4.3 (**Moment**). Sei X eine Zufallsvariable und $p \in R_+$. Wir definieren:

- *das p-te absolute Moment von X* durch $M_o := \mathbb{E}[|X|^p]$ (kann ∞ sein)
- falls $M_n < \infty$ für ein n, dann ist das *n-te (rohe) Moment von X* durch $m_n := \mathbb{E}[X^n]$ definiert.
- Das *n-te zentralisierte Moment von X* durch $m_n := \mathbb{E}[(X - \mathbb{E}[X])^n]$ definiert.

Damit folgt sofort:

Korollar 4.1. $M_n < \infty$ für $n \in \mathbb{N} \implies |m_n| \leq M_n$

Hat X eine Dichte f_X, dann gilt zudem für das absolute Moment

$$M_p = \int_{-\infty}^{\infty} |x|^p f_X(x)\ dx$$

Gilt dann $M_n < \infty$ für ein $n \in \mathbb{N}$, dann können wir auch das n-te Moment per Integral bestimmen:

$$m_n = \int_{-\infty}^{\infty} x^n f_X(x)\ dx$$

Satz 4.4. Sei X ZV und $p, q \in R_+$. Dann:

$$p \leq q\ \wedge\ M_q < \infty\ \implies\ M_p < \infty$$

4.5 Gemeinsame Verteilungen, Unabhängige Zufallsvariablen

Def. 4.4 (**Gemeinsame Verteilung**). Die *gemeinsame Verteilungsfunktion* von n ZV $X_1, \ldots, X_n$ ist die Abbildung $F : \mathbb{R}^n \to [0, 1]$ mit:

$$(x_1, \ldots, x_n) \mapsto F(x_1, \ldots, x_n) := P[X_1 \leq x_1, \ldots, X_n \leq x_n]$$

Lässt sich F für eine Funktion $f : \mathbb{R}^n \to [0, \infty)$ schreiben als

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(t_1, \ldots, t_n) dt_n \ldots dt_1$$

dann heisst $f(x_1, \ldots, x_n)$ die *gemeinsame Dichte* von $X_1, \ldots, X_n$.

Korollar 4.2 (**Eigenschaften der Dichte**). Für die gemeinsame Dichte von $X_1, \ldots, X_n$ gilt:

(i) $f(x_1, \ldots, x_n) \geq 0$ und $= 0$ ausserhalb $\mathcal{W}(X_1, \ldots, X_n)$

(ii) $\iiint_{\mathbb{R}^n} f(x_1, \ldots, x_n) dx_n \ldots dx_1 = 1$

(iii) $P[(X_1, \ldots, X_n) \in A] = \iiint_{(x_1, \ldots, x_n) \in A} f(x_1, \ldots, x_n) dx_n \ldots dx_1$ für $A \subseteq \mathbb{R}^n$.

Def. 4.5 (**Randverteilung**). Haben X, Y die gemeinsame Verteilungsfunktion F, dann sind $F_X : \mathbb{R} \to [0, 1]$ und $F_Y : \mathbb{R} \to [0, 1]$ die Verteilungsfunktionen der *Randverteilung* von X bzw. Y und sind definiert als:

$$x \mapsto F_X(x) := P[X \leq x] = P[X \leq x, Y < \infty] = \lim_{y \to \infty} F(x, y)$$

$$y \mapsto F_Y(y) := P[Y \leq y] = P[X < \infty, Y \leq y] = \lim_{x \to \infty} F(x, y)$$

Haben X, Y eine gemeinsame Dichte f, dann haben auch die Randverteilungen Dichten $f_X : \mathbb{R} \to [0, \infty)$ und $f_Y : \mathbb{R} \to [0, \infty)$ mit

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy \qquad f_Y(y) = \int_{-\infty}^{\infty} f(x, y)dx$$

Def. 4.6 (**Unabhängigkeit**). Die ZV $X_1, \ldots, X_n$ heissen *unabhängig* $\iff F(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n)$.
Hat man stetige Zufallsvariablen mit Dichten, dann ist die gemeinsame Dichtefunktion das Produkt der Randdichten, also

$$f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$$

4.6 Bedingte Verteilungen usw

Def. 4.7 (**Bedingte Dichte, Verteilungsfunktion und Erwartungswert**).

$$f_{X_1|X_2}(x_1 \mid x_2) = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)}$$

$$P(Y > t \mid Y < a) = \frac{P[t < Y < a]}{P[Y < a]}$$

$$E[X_1 \mid X_2](x_2) = \int x_1 f_{X_1|X_2}(x_1 \mid x_2)\ dx_1$$

Mit Trick:

$$\begin{aligned} E[X_1] = E[E[X_1 \mid X_2]] &= \int E[X_1 \mid X_2](x_2) f_{X_2}(x_2) dx_2 \\ &= \int\int x_1 f_{X_1,X_2}(x_1, x_2) dx_1 dx_2 \end{aligned}$$

Anm. $E[X_1 \mid X_2](x_2) = E[X_1 \mid X_2 = x_2]$

4.7 Funktionen und Transformationen von Zufallsvariablen

Summen

Für $Z = X + Y$ suchen wir die Verteilungsfunktion $F_Z(z) = P[Z \leq z] = P[X + Y \leq z]$. Dies kann man als Punktemenge im $\mathbb{R}^2$ auffassen, nämlich $A_z := \{(x, y) \in \mathbb{R}^2 \mid x + y \leq z\}$. Damit ist $F_Z(z) = P[(X, Y) \in A_z]$. Damit erhält man

$$F_Z(z) = \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f(x, y)\ dy\ dx$$

Substituiere nun $v = x + y \Rightarrow y = v - x, dy = dv$ so erhält man

$$F_Z(z) = \int_{-\infty}^{\infty} \int_{-\infty}^{z} f(x, v - x)\ dv\ dx = \int_{-\infty}^{z} \int_{-\infty}^{\infty} f(x, v - x)\ dx\ dv$$

$$\implies f_Z(z) = \frac{d}{dz} F_Z(z) = \int_{-\infty}^{\infty} f(x, z - x)\ dx = \int_{-\infty}^{\infty} f(z - y, y)\ dy$$

womit wir also auch die Dichte erhalten haben. Das letzte Gleichheitszeichen gilt wegen Symmetrie zwischen X, Y. Sind X, Y unabhängig, so gilt $f(x, y) = f_X(x) \cdot f_Y(y)$ und dann ist f_Z die *Faltung* von f_X und f_Y.

Transformationen

Sei X ZV mit Verteilung und Dichte. Sei $g : \mathbb{R} \to \mathbb{R}$ messbare Funktion. Betrachte $Y = g(X)$, wir suchen Verteilung und Dichte (falls existent) von Y. Allgemein löst man dieses Problem wie folgt:

$$F_Y(t) = P[Y \leq t] = P[g(X) \leq t] = \int_{A_g} f_X(s)\ ds$$

mit $A_g := \{s \in \mathbb{R} \mid g(s) \leq t\}$. Die Dichtefunktion (falls existent) erhält man dann durch Ableiten der Verteilung.

Awendung der Transformation

Satz 4.5. Sei F stetige, streng-monoton wachsende Verteilungsfunktion mit Umkehrfunktion F^{-1}. Dann:

$$X \sim \mathcal{U}(0,1) \quad \wedge \quad Y = F^{-1}(X) \implies Y \text{ hat Verteilungsfunktion } F.$$

Dieser Satz erlaubt die Konstruktion einer Zufallsvariablen Y mit einer gewünschten Verteilungsfunktion F, wenn man eine Zufallsvariable $X \sim \mathcal{U}(0,1)$ zur Hand hat. Damit kann man beispielsweise eine Verteilung mit einem Computer simulieren. Ein Zufallszahlengenerator produziert in einem gewissen Sinn eine Folge von $\mathcal{U}(0,1)$-verteilten Zufallsvariablen. $\implies$ F^{-1}(Zufallszahlengenerator) simuliert also die Verteilung F.

5 Ungleichungen und Grenzwertsätze

5.1 Wahrscheinlichkeit & Konvergenz

Def. 5.1 (**Konvergenz in Wahrscheinlichkeit**). Sei $X_1, X_2, \dots$ und Y ZV auf gemeinsamen Wahrscheinlichkeitsraum.

(i) $X_1, X_2, \dots$ konvergiert gegen Y *in Wahrscheinlichkeit* falls

$$\forall \varepsilon > 0. \quad \lim_{n\to\infty} P[|X_n - Y| > \varepsilon] = 0$$

(ii) Für $p > 0$ konvergiert die Folge $X_1, X_2, \dots$ gegen Y *in* L^p falls

$$\lim_{n\to\infty} \mathbb{E}[|X_n - Y|^p] = 0$$

(iii) $X_1, X_2, \dots$ konvergiert gegen Y *P-fast sicher* falls

$$P\left[\lim_{n\to\infty} X_n = Y\right] = P\left[\left\{\omega \in \Omega \mid \lim_{n\to\infty} X_n(\omega) = Y(\omega)\right\}\right] = 1$$

Def. 5.2 (**Konvergenz in Verteilung**). Seien $X_1, X_2, \dots,$ und Y ZV auf möglicherweise verschiedenen Wahrscheinlichkeitsräumen mit Verteilungsfunktionen $F_1, F_2, \dots$ und F_Y. Dann konvergiert $X_1, X_2, \dots$ gegen Y *in Verteilung* falls

$$\lim_{n\to\infty} F_n(x) = F_Y(x) \qquad \text{für alle } x \in R, \text{wo } F_Y \text{ stetig ist}$$

Satz 5.1. Es gilt folgende Äquivalenz:

$$X_1, X_2, \dots \text{ konvergiert in Verteilung gegen } Y \iff$$

$$\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(Y)] \text{ für jedes beschränkte stetige } f : \mathbb{R} \to \mathbb{R}$$

5.2 Ungleichungen

Satz 5.2 (**Markov-Ungleichung**). Sei X eine Zufallsvariable und $g : \mathcal{W}(X) \to [0, \infty)$ eine wachsende Funktion. Für jedes $c \in \mathbb{R}$ mit $g(c) > 0$ gilt dann:

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

Bemerkung: Insbesondere gilt der satz für die Identitätsfunktion $g = id$. Daraus folgt unmittelbar:

Satz 5.3 (**Chebyshev-Ungleichung**). Sei Y Zufallsvariable mit endlicher Varianz. Für jedes $b > 0$ gilt dann:

$$P\left[|Y - \mathbb{E}[Y]| \geq b\right] \leq \frac{\mathrm{Var}[Y]}{b^2}$$

Beweis. Wähle $X := |Y - \mathbb{E}[Y]|$ und $g(x) = x^2$ für $x \geq 0 \implies \mathbb{E}[g(Y)] = \mathrm{Var}[Y]$. □

5.3 Gesetz der grossen Zahlen

Wir betrachten nun Folgen von Zufallsvariablen mit dem gleichen Erwartungswert und der gleichen Varianz. Uns interessiert das Verhalten des arithmetischen Mittel dieser Folge von Zufallsvariablen.

Satz 5.4 (**Schwaches Gesetz der grossen Zahlen**). Sei $X_1, X_2, \dots$ eine Folge von unabhängigen ZV mit $\mathbb{E}[X_i] = \mu$ und Varianz $\mathrm{Var}[X_i] = \sigma^2$. Sei $\overline{X_n} = \frac{1}{n}\sum_{i=1}^n X_i$. Dann konvergiert $\overline{X_n}$ für $n \to \infty$ in Wahrscheinlichkeit/stochastisch gegen μ.

Beweis. Betrachte Linearität des EW: $\mathbb{E}[\overline{X_n}] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] = \mu$. Da die ZV paarweise unkorreliert sind, gilt auch die Linearität der Varianz und somit $\mathrm{Var}[\overline{X_n}] = \frac{1}{n}\sum_{i=1}^n \mathrm{Var}[X_i] = \frac{\sigma^2}{n}$. Die Chebyshev-Ungleichung liefert damit:

$$P\left[|\overline{X_n} - \mu| > \varepsilon\right] \leq \frac{\mathrm{Var}[\overline{X_n}}{\varepsilon^2} = \frac{\sigma^2}{n \cdot \varepsilon^2}$$

Dieser Term geht für jedes beliebige $\varepsilon > 0$ gegen 0, was *Def. 5.1 (i)* entspricht. □

Bemerkung 1: Es genügt bereits, wenn X_i nur paarweise unkorreliert sind.
Bemerkung 2: Die Existenz des Erwartungswerts ist essentiell, damit das Gesetz gilt: So existiert bspw kein Erwartungswert für die bereits eingeführte *Cauchy-Verteilung*. Damit konvergiert $n \mapsto \overline{X_n}(\omega)$ nicht, denn Summen von Cauchy-verteilten Zufallsvariablen sind wiederum Cauchy-verteilt.

Monte-Carlo-Integration

Wir wollen für $h : [0,1]^d \to \mathbb{R}$ ein Integral $I := \int_{[0,1]^d} h(\vec{x})\ d\vec{x}$ berechnen, welches auch numerisch schwer lösbar ist. Dafür können wir I als einen Erwartungswert auffassen. Sei $d = 1$. Ist $U \sim \mathcal{U}(0,1)$, dann gilt

$$\mathbb{E}[h(U)]) = \int_{\mathbb{R}} h(x) f_U(x)\ dx = \int_0^1 h(x)\ dx = I$$

Die letzte Gleichheit gilt, weil die Dichte von U auf $[0,1]$ konstant 1 ist, und sonst 0. Deshalb können wir mit einem Zufallszahlengenerator eine Folge $U_1, U_2, \dots$ generieren mit $U_i \sim \mathcal{U}(0,1)$ und den Wert von I mit dem schwachen GGZ approximieren:

$$\overline{h(U_n)} = \frac{1}{n} \sum_{i=1}^{n} h(U_i)$$

Damit ist aber auch gleich klar, wieso man eine stärkere Aussage möchte, denn der berechnete Wert liegt nur mit grosser Wahrscheinlichkeit sehr nahe bei I, aber man weiss nicht, ob eine feste Realisierung ω in dieser guten Approximationsmenge liegt.

Satz 5.5 (**Starkes Gesetz der grossen Zahlen**). Sei $X_1, X_2, \dots$ eine Folge von unabhängigen Zufallsvariablen mit gleicher Verteilung und EW μ endlich. Für das arithmetische Mittel $\overline{X_n} := \frac{1}{n} \sum_{i=1}^{n} X_i$ gilt dann, dass $\overline{X_n}$ *P-fast sicher* (P.f.s.) gegen μ konvergiert, also

$$P\left[\left\{\omega \in \Omega \mid \overline{X_n}(\omega) \underset{n\to\infty}{\longrightarrow} \mu\right\}\right]$$

Für die Monte-Carlo Integration bedeutet dies, dass unserer berechneter Wert mit Wahrscheinlichkeit 1 nahe bei I liegt. Schlechte Approximationen sind zwar möglich, aber mit Wahrscheinlichkeit 0.

5.4 Zentraler Grenzwertsatz

Wir bezeichnen unabhängige gleichverteilte Zufallsvariablen als *i.i.d.* für *independent identically distributed*.

Satz 5.6 (**Zentraler Grenzwert**). Sei $X_1, X_2, \dots$ eine Folge von i.i.d. ZV mit EW μ und Varianz σ^2. Für die Summe $S_n = \sum_{i=1}^{n} X_i$ gilt dann:

$$\lim_{n\to\infty} P\left[\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right] = \Phi(x) \qquad \forall x \in \mathbb{R}$$

Für praktische Anwendungen existieren zwei alternative Notationen:

- $P[S_n^* \leq x] \approx \Phi(x)$ für n gross
- $S_n^* \overset{\text{approx.}}{\sim} \mathcal{N}(0,1)$ für n gross

wobei S_n^* die *Standardisierung von* S_n gennant wird:

$$S_n^* = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathrm{Var}[S_n]}}$$

Daraus folgt $S_n \sim \mathcal{N}(n\mu, n\sigma^2)$ und $\overline{X_n} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$, wobei beide Verteilungen nur approximativ gelten.

Häufige Anwendung: *Approximation der Binomialverteilung durch Normalverteilung* weil die Binomialverteilung mühsam zu berechnen ist. Ist $S_n \sim Bin(n,p)$ dann können wir approxmativ sagen, dass $S_n \sim \mathcal{N}(np, np(1-p))$. Fügen wir noch eine additiven Konstante $+\frac{1}{2}$ dazu, die sogenannte *Kontinuitätskorrektur*, so wird das Resultat noch genauer. Dies lässt sich intuitiv dadurch rechtfertigen, dass sich die Binomialverteilung besser approximieren lässt, wenn man die Normalverteilungsdichte unter den "Stäbenßentriert, statt am linken/rechten Rand zu betrachten. Damit gilt:

Korollar 5.1. Dieses Korollar braucht man eigentlich überhaupt nicht.

$$P[a < S_n \leq b] = P\left[\frac{a-np}{\sqrt{np(1-p)}} < S_n^* \leq \frac{b-np}{\sqrt{np(1-p)}}\right]$$
$$\approx \Phi\left(\frac{b+\frac{1}{2}-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a+\frac{1}{2}-np}{\sqrt{np(1-p)}}\right)$$

5.5 Grosse Abweichungen & Chernoff-Schranken

Def. 5.3 (**momenterzeugende Funktion**). Für eine Zufallsvariable X ist die *momenterzeugende Funktion* definiert als

$$M_X(t) := \mathbb{E}[e^{tX}] \quad \text{für } t \in \mathbb{R}$$

Diese ist wohldefiniert auf $[0, \infty]$, kann aber den Wert unendlich annehmen.

Satz 5.7. Seien $X_1, \ldots, X_n$ i.i.d. für welche die momenterzeugende Funktion $M_X(t)$ für alle $t \in \mathbb{R}$ endlich ist. Dann gilt für jedes $b \in \mathbb{R}$:

$$P[S_n \geq b] \leq \exp\left(\inf_{t \in \mathbb{R}}(n \log M_X(t) - tb)\right)$$

Diese Aussage ist zwar stark und liefert ziemlich genaue Abschätzungen, ist allerdings nicht praktisch wegen der momenterzeugenden Funktion. Diese schätzen wir im folgenden Satz nach oben ab:

Satz 5.8 (**Chernoff Schranken**). Seien $X_1, \ldots, X_n$ unabhängig mit $X_i \sim Be(p_i)$ und $S_n = \sum_{i=1}^n X_i$. Sei $\mu_n := \mathbb{E}[S_n] = \sum_{i=1}^n p_i$ und $\delta > 0$. Dann gilt:

$$P[S_n \geq (1+\delta)\mu_n] \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^{\mu_n}$$

Teil II
Statistik

6 Statistische Grundideen

Man unterscheidet im Grunde zwei Formen der Statistik:

- Die *deskriptive Statistik* beschäftigt sich hauptsächlich mit graphischer Aufbereitung der Daten etc.
- Die *induktive Statistik* sucht für eine gesammelte Menge an Daten ein passendes (Verteilungs-)Modell

Wir unterscheiden *Daten* $x_1, \dots, x_n$ (generell Zahlen) und den generierenden Mechanismus $X_1, \dots, X_n$ (Zufallsvariablen, also Funktionen auf Ω). Die Gesamtheit der Beobachtungen $x_1, \dots, x_n$ oder Zufallsvariablen $X_1, \dots, X_n$ nennt man oft *Stichprobe* mit *Stichprobenumfang* n.

Ausgangspunkt ist oft ein Datensatz $x_1, \dots, x_n$ aus einer Stichprobe $X_1, \dots, X_n$, für die wir ein Modell suchen. $\Longrightarrow$ durch Parameter $\vartheta \in \Theta$ (möglicherweise hoch-dimensional). Dazu betrachtet man einge ganze Familie von Wahrscheinlichkeitsräumen. Der Grundraum $(\Omega, \mathcal{F})$ ist fest und für jeden Parameter ϑ aus dem Parameterraum Θ hat man ein Wahrscheinlichkeitsmass P_ϑ auf dem Grundraum. Dies gibt uns also einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, P_\vartheta)$ für jedes $\vartheta \in \Theta$. Wir betrachten dann die Daten $x_1, \dots, x_n$ als Ergebnisse von Zufallsvariablen $X_1, \dots, X_n$ und versuchen daraus Rückschlüsse über ϑ zu ziehen.

Das Vorgehen erfolgt in 5 Schritten:

1. Deskriptive Statistik um sich einen Überblick zu verschaffen
2. Wahl eines (parametrischen) Modells $\rightarrow$ spezifiziere eine Parametermenge Θ und die Familie $(P_\vartheta)_{\vartheta \in \Theta}$
3. Schätzung der Parameter aufgrund der Daten mithilfe eines *Schätzers*
4. Kritische Modellüberprüfung und Anpassung $\rightarrow$ überprüft ob Daten gut zu gewähltem Paramter ϑ passen mittels geeignetem statistischen Test
5. Aussagen über die Zuverlässigkeit $\rightarrow$ wie gut passt das Modell? kann auch *Konfidenzbereich* anstelle eines einzelnen Parameters angeben.

Dieses Vorgehen nennt man *parametrische statistische Analyse*.

7 Schätzer

Wir suchen ein Modell für eine Stichprobe $X_1, \dots, X_n$ und haben einen Parameterraum Θ (oft $\subseteq \mathbb{R}^m$) und für jedes $\vartheta \in \Theta$ einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, P_\vartheta)$. Wir wollen daher die Paramter $\vartheta_1, \dots, \vartheta_m$ bestimmen.

Def. 7.1 (**Schätzer**). Ein *Schätzer* T_j für einen Parameter ϑ_j ist eine Zufallsvariable der Form $T_j := t_j(X_1, \dots, X_n)$ für eine *Schätzfunktion* $t_j : \mathbb{R}^n \to \mathbb{R}$.

Def. 7.2 (**Schätzwert**). Ein Schätzwert ist das Ergenis einer konkreten Berechnung, eine Zahl. Sie entsteht durch Einsetzen konkreter Daten in einen Schätzer: $T_j(\omega) = t_j(x_1, \dots, x_n)$ und liefert damit einen Wert für genau einen Parameter ϑ_j.

Damit ist ein Schätzer also eine Funktion, die eine Berechnungs*methode* angibt und ein Schätzwert ist ein Ergebnis einer solchen *konkreten Berechnung*. Der Einfachheit halber schreiben wir oft $T = (T_1, \dots, T_m)$ und $\vartheta = (\vartheta_1, \dots, \vartheta_m)$. Wir betrachten nun einige wünschenswerte Eigenschaften für Schätzer:

Def. 7.3 (**Eigenschaften von Schätzern**). Sei T ein Schätzer.

- T ist **erwartungstreu**, falls $\mathbb{E}_\vartheta[T] = \vartheta$ gilt. T schätzt im Mittel also richtig
- der **Bias** ist definiert als $\mathbb{E}_\vartheta[T] - \vartheta \implies$ ein erwartungstreuer Schätzer hat keinen Bias.
- der **mean-squared-error (MSE)** ist definiert als
  $\mathrm{MSE}_\vartheta[T] := \mathbb{E}_\vartheta[(T-\vartheta)^2] = \mathrm{Var}_\vartheta[T] + (\mathbb{E}_\vartheta[T] - \vartheta)^2$
  $\implies$ für erwartungstreue Schätzer ist MSE = Varianz
- eine Folge $T^{(n)}$ von Schätzern heisst **konsistent** für ϑ, falls $T^{(n)}$ für $n \to \infty$ in P_ϑ-Wahrscheinlichkeit gegen ϑ konvergiert, d.h. für jedes $\vartheta \in \Theta$ gilt:

$$lim_{n\to\infty} P_\vartheta \left[|T^{(n)} - \vartheta| > \varepsilon \right] = 0 \qquad \forall \varepsilon > 0$$

7.1 Maximum-Likelihood Methode

Man unterscheidet den diskreten und stetigen Fall. Wir betrachten hier nur den stetigen Fall, der diskrete Fall verläuft analog (man verwendet Gewichtsfunktion statt Dichtefunktion).

In einem Modell P_ϑ sind dann die Zufallsvariablen $X_1, \dots, X_n$ stetig mit einer gemeinsamen Dichtefunktion $f(x_1, \dots, x_n; \vartheta)$. Oft sind die X_i sogar i.i.d. mit individueller Dichtefunktion $f_X(x; \vartheta)$ und man erhält die gemeinsame Dichtefunktion als Produkt (dies wird später nützlich):

$$f(x_1, \dots, x_n; \vartheta) = P_\vartheta[X_1 = x_1, \dots, X_n = x_n] = \prod_{i=1}^{n} f_X(x_i; \vartheta)$$

Beachte, dass die erste Gleichheit auch im allgemeinen Fall gilt, während die zweite Gleichheit nur für i.i.d. ZV gilt.

Def. 7.4 (**Likelihood-Funktion**). Die *Likelihood*-Funktion L ist definiert durch

$$L(x_1, \dots, x_n; \vartheta) := \begin{cases} p(x_1, \dots, x_n; \vartheta) & \text{diskreter Fall} \\ f(x_1, \dots, x_n; \vartheta) & \text{stetiger Fall} \end{cases}$$

Die Funktion $\log L(x_1, \dots, x_n; \vartheta)$ ist dann die *log-Likelihood-Funktion* (natürlicher Logarithmus)

Für eine Stichprobe $X_1, \dots, X_n$ gibt die Likelihood-Funktion die Wahrscheinlichkeit, dass im Modell P_ϑ unsere Stichprobe gerade die Werte $x_1, \dots, x_n$, die wir beobachtet haben, liefert. Die Idee der *Maximum-Likelihood* Funktion besteht nun darin, dass wir die beobachteten Werte $x_1, \dots, x_n$ als sehr wahrscheinlich betrachten. Konkret "definieren" wir diese Ergebnis als das wahrscheinlichste Ergebnis, das auftauchen kann. Aus diesem Grund maximieren wir die Likelihood-Funktion nach dem Parameter ϑ:

Def. 7.5 (**Maximum-Likelihood-Schätzer**). Der *ML-Schätzer* T für ϑ ist dadurch definiert, dass er die Funktion $\vartheta \mapsto L(X_1, \dots, X_n; \vartheta)$ als Funktion von ϑ maximiert.

Bemerkung: Normalerweise arbeiten wir mit i.i.d. Zufallsvariablen $X_i \implies$ die Likelihood-Funktion L ist ein Produkt. Verwenden wir aber $\log L$, so können wir die log-Likelihood-Funktion als Summe schreiben, was das Differenzieren erleichtert. Dies funktioniert, da $\log : (0, \infty) \to \mathbb{R}$ streng monoton wachsend ist. Das bedeutet konkret, dass jedes Maximum/Minimum von L auch eines von $\log L$ ist.

Im Allgemeinen versucht man, dises Maximum analytisch zu finden, z.B. durch Differenzieren. Es kann aber auch vorkommen, dass die Likelihood-Funktion nicht differenzierbar ist. In diesem Fall muss man iterativ vorgehen, z.B. mit der Newton-Methode als Iterationsverfahren.

7.2 Momentenmethode

Der *Momentenmethode* liegt die Idee zugrunde, dass die Momente einer Zufallsvariable bzw. einer Wahrscheinlichkeitsverteilung durch Stichprobenmomente geschätzt werden können.

Sei dazu $X_1, \dots, X_n$ eine Stichprobe und $\Theta \subseteq \mathbb{R}^m$ der Parameterraum. Für jeden Parameter $\vartheta = (\vartheta_1, \dots, \vartheta_m) \in \Theta$ sei $X_1, \dots, X_n$ i.i.d. unter dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, P_\vartheta)$.

Def. 7.6 (**Empirisches Moment**). Für $k \in \{1, \dots, m\}$ sei das k-te *empirische Moment* oder *Stichprobenmoment* $\widehat{m}_k$ der Realisierungen $(x_1, \dots, x_n)$ definiert durch

$$\widehat{m}_k(x_1, \dots, x_n) := \frac{1}{n} \sum_{i=1}^{n} x_i^k$$

Annahmen

(i) $\mathbb{E}_\vartheta[|X_1|^m] < \infty$ für jedes $\vartheta \in \Theta$

(ii) Für jedes $k \in \{1, \dots, m\}$ ist das k-te Moment $m_k^\vartheta := \mathbb{E}_\vartheta[X_1^k]$ der Stichprobenvariablen eine bekannte Funktion des Parametervektors ϑ. Konkret:

$$\forall k \in \{1, \dots, m\}.\ \exists\ g_k : \Theta \to \mathbb{R} \text{ (borel-messbar)}.\ \forall \vartheta \in \Theta.\quad m_k^\vartheta = g_k(\vartheta_1, \dots, \vartheta_m)$$

Beachte, dass wir aufgrund der Tatsache, dass die X_i i.i.d. sind, diese Eigenschaften nur für X_1 überprüfen müssen. Sind diese Annahmen erfüllt, so kann man die Momentenmethode nach dem folgenden Schema anwenden.

Methode

1. Für gegebene Realisierungen $x_1, \dots, x_n$ bestimmen für jedes $k \in \{1, \dots, m\}$ das k-te empirische Moment.
2. Stelle ein Gleichungssystem für die Unbekannten Paramter $\vartheta_1, \dots, \vartheta_m$ auf, in dem das k-te empirische Moment dem k-ten Moment gleichgesetzt wird, also:

$$\widehat{m}_k(x_1, \dots, x_n) = g_k(\vartheta_1, \dots, \vartheta_m) \qquad k = 1, \dots, m$$

3. Überprüfe, ob dieses LGS eine eindeutige Lösung besitzt. Dann entspricht die Lösung $\widehat{\vartheta} = \widehat{\vartheta}(x_1, \dots, x_n) \in \Theta$ unserer Schätzung für die Paramter ϑ.

Def. 7.7 (**Momenten-Schätzer**). Der Vektor $\widehat{\vartheta}(X_1, \dots, X_n)$ heisst *Momenten-Schätzer* des Parameters ϑ.

Beispiel: Normalverteilte Stichprobenvariablen

Sei $X_1, \dots, X_n$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$-verteilt mit unbekanntem Parameter $\vartheta = (\mu, \sigma^2)$ und in diesem Fall gilt $g_1(\mu, \sigma^2) = \mu$ und $g_2(\mu, \sigma^2) = \mu^2 + \sigma^2$. Damit berechnen wir den ML-Schätzer für $\vartheta = (\mu, \sigma^2)$:

$$\begin{aligned} T_1 &= \frac{1}{n} \sum_{i=1}^{n} X_i =: \overline{X_n} \\ T_2 &= \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X_n})^2 \end{aligned}$$

Dieser Schätzer $T = (T_1, T_2)$ ist im Allgemeinen der Momementenschätzer für $(E_\vartheta[X], \mathrm{Var}_\vartheta[X])$. Dieser ist aber nicht erwartungstreu, denn es gilt $\mathbb{E}_\vartheta[T_2] = \frac{n-1}{n} \mathrm{Var}_\vartheta[X]$. Man kann aber durch eine kleine Modifikation einen erwartungstreuen Schätzer $T' = (T_1', T_2')$ mit $T_1' = T_1$ und $T_2' = S^2$, der *empirischen Stichprobenvarianz*.

$$S^2 := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X_n})^2$$

7.3 Verteilungsaussagen

Es gibt sehr wenige allgemeingültige Aussagen über Verteilungen von Schätzern. Da diese aber von grosser Wichtigkeit in der Statistik sind, verschafft man sich einen approxmativen Zugang über die Normalverteilung. Schätzer sind nämlich häufig Funktion einer Summe von i.i.d. Zufallsvariablen im Modell P_ϑ. Diese Summe ist nach dem ZGS approximativ normalverteilt unter P_ϑ. Für normalverteilte Stichproben existieren nämlich exakte Aussagen. Zuerst führen wir aber zwei neue Verteilungen ein:

χ^2-Verteilung

Die χ^2-Verteilung mit n Freiheitsgraden (bezeichnet mit χ^2_n) ist eine stetige Verteilung einer Zufallsvariablen X. Es gibt folgenden Zusammenhang mit der Normalverteilung:

Lemma 7.1. $(\forall i \in \{1,\dots,n\}. \quad Z_i \sim \mathcal{N}(0,1) \ \wedge \ Z_i \text{ i.i.d.}) \implies \left(\sum_{i=1}^n Z_i^2\right) \sim \chi^2_n$

Zudem ist die χ^2-Verteilung ein Spezialfall der Gamma-Verteilung, es gilt nämlich:

Lemma 7.2. $X \sim \chi^2_n \iff X \sim Ga(\frac{n}{2}, \frac{1}{2})$

Damit ist eine χ^2_2-Verteilung gerade die Exponentialverteilung mit $\lambda = \frac{1}{2}$. Sei $X \sim \chi^2_n$, dann gilt:

- **Wertebereich:** $\mathcal{W}(X) = \mathbb{R}_0^+$
- **Erwartungswert:** $\mathbb{E}[X] = n$
- **Varianz:** $\mathrm{Var}[X] = 2n$
- **Dichtefunktion:**

$$f_X(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} y^{\frac{n}{2}-1} e^{-\frac{1}{2}y} & \text{für } x \geq 0 \\ 0 & \text{für } x < 0 \end{cases}$$

Die χ^2-Verteilung ermöglicht ein Urteil über die Kompabilität eines funktionalen Zusammenhangs mit empirischen Messpunkten. So kann bspw. bestimmt werden, ob eine Gerade, Logarithmhus oder eine Parabel die gesammelten Daten am besten erklärt.

t-Verteilung

Die t-Verteilung mit n Freiheitsgraden gehört zu einer stetigen Zufallsvariablen Z. Sie entsteht durch die standarisierte Schätzfunktion des Stichprobenmittelwerts normalverteilter Daten, wenn bei der Standarisierung des Mittelwerts die Varianz (weil sie nicht bekannt ist) durch die *Stichprobenvarianz* abgeschätzt werden muss. Die standardisierte Schätzfunktion ist dann nicht mehr normalverteilt, sondern folgt der t-Verteilung.

Sei $Z \sim t_n$. Dann hat Z folgende Eigenschaften:

- **Dichtefunktion:**

$$f_Z(z) = \frac{\Gamma\left(\frac{n+1}{2}\right))}{\sqrt{n\pi} \cdot \Gamma(\frac{n}{2})} \left(1 + \frac{z^2}{n}\right)^{-\frac{n+1}{2}} \qquad z \in \mathbb{R}$$

  $\implies$ für $n = 1$ ist dies eine *Cauchy-Verteilung* $\implies$ Erwartungswert existiert für $n = 1$ nicht.
- für $n \to \infty$ erhält man eine $\mathcal{N}(0,1)$-Verteilung
- **Erwartungswert:** für $n > 1$ gilt: $\mathbb{E}[Z] = 0$
- **Varianz:** für $n > 2$ gilt: $\mathrm{Var}[Z] = \frac{n}{n-2}$
- *Faustregel:* ab $n = 30$ Freiheitsgraden kann man die t-Verteilung durch die Normalverteilung approximieren

Die t-Verteilung kann auch anders hergeleitet werden, Seien $X \sim \mathcal{N}(0,1)$ und $Y \sim \chi^2_n$ unabhängig. Dann ist $Z := \frac{X}{\sqrt{\frac{1}{n}Y}}$ t-verteilt mit n Freiheitsgraden.

Satz 7.1 (Normalverteilte Stichproben). Seien $X_1,\dots,X_n$ i.i.d. $\sim \mathcal{N}(\mu,\sigma^2)$. Dann gilt:

(i) $\overline{X_n} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ und normalisiert $\frac{\overline{X_n}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$

(ii) $\frac{n-1}{\sigma^2}S^2 = \left(\frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \overline{X_n})^2\right) \sim \chi^2_{n-1}$

(iii) $\overline{X_n}$ und S^2 sind unabhängig.

(iv) $\frac{\overline{X_n}-\mu}{S/\sqrt{n}} = \frac{\frac{\overline{X_n}-\mu}{\sigma/\sqrt{n}}}{S/\sigma} = \frac{\frac{\overline{X_n}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{1}{n-1}\frac{n-1}{\sigma^2}S^2}} \sim t_{n-1}$

Die Hauptaussage dieses Satzes ist (iii). (i) ist schon bekannt und (iv) folgt unmittelbar aus der Herleitung der t-Verteilung.

8 Tests

Ausganspunkt: Stichprobe $X_1, \dots, X_n$ und Familie von Wahrscheinlichkeiten P_ϑ mit $\vartheta \in \Theta$ die unsere möglichen Modelle beschreiben. $\implies$ Grundproblem besteht darin, Entscheidung zwischen zwei konkurrierenden Modelkassen zu treffen: der *Hypothese* oder *Nullhypothese* $\Theta_0 \subset \Theta$ oder der *Alternative* $\Theta_A \subseteq \Theta$. Dabei muss zwingend $\Theta_0 \cap \Theta_A = \emptyset$ gelten. Man Schreibt $H_0 \ : \ \vartheta \in \Theta_0$ und $H_A \ : \ \vartheta \in \Theta_A$. Falls keine Alternative explizit definiert ist, so wählen wir $\Theta_A = \Theta \setminus \Theta_0$. Wir unterscheiden:

- *einfache Hyptohesen* bestehen aus einem einzelnen Wert, also z.B. $\Theta_0 = \{\vartheta_0\}$
- *zusammengesetzte Hypothesen* bestehen aus mehreren Werten

Ein *Test* ist im Allgemeinen eine Entscheidungsregel, die zu gegebenen Daten $x_1, \dots, x_n$ einen Wert $\{0, 1\}$ liefert und dieser ist $1 \iff$ die Nullhypothese soll abgelehnt werden. Formal:

Def. 8.1 (**Test, Teststatistik**)**.** Ein *Test* besteht aus

- einer Abbildung $t : \mathbb{R}^n \to \mathbb{R}$, $(x_1, \dots, x_n) \mapsto t(x_1, \dots, x_n)$
- und einem *kritischen Bereich* oder *Verwerfungsbereich* $K \subseteq \mathbb{R}$.

Die Zufallsvariable $T = t(X_1, \dots, X_n)$ heisst *Teststatistik*. Die Entscheidungsregel ist definiert durch die Zufallsvariable

$$I_{\{t(x_1,\dots,x_n)\in K\}}$$

d.h. man verwirft die Hypothese genau dann, wenn der realisierte Wert $t(x_1, \dots, x_n)$ im Verwerfungsbereich K liegt.

Für eine Realisierung ω gilt $t(x_1, \dots, x_n) = t(X_1(\omega), \dots, X_n(\omega)) = T(\omega)$. Weil T eine Zufallsvariable ist, ist der Raum $\{T \in K\} \subseteq \Omega$ messbar. Damit kann für jedes Modell P_ϑ die Wahrscheinlichkeit $P_\vartheta[T \in K]$ betrachtet werden.

Arten von Fehlern

- *Fehler 1. Art:* Hypothese zu Unrecht abgelehnt $\implies$ $\vartheta \in \Theta_0$ und $T \in K$
- *Fehler 2. Art:* Hypothese zu Unrecht nicht verworfen, d.h. die Hypothese wird akzeptiert obwohl sie falsch ist. $\implies$ $\vartheta \in \Theta_A$ und $T \notin K$.

$\implies$ man würde gerne beide Fehler-Wahrscheinlichkeiten minimieren. Dazu sollte $\vartheta \mapsto P_\vartheta[T \in K]$ auf Θ_0 möglichst klein sein, aber gleichzeitig möglichst gross in Θ_A. $\implies$ oft nicht möglich, deshalb folgendes Verfahren:

1. Man wählt ein *Signifikanzniveau* $\alpha \in (0, 1)$ und kontrolliert die Wahrscheinlichkeit eines Fehlers erster Art durch α:
$$\sup_{\vartheta\in\Theta_0} P_\vartheta[T \in K] \leq \alpha$$
2. Man versucht die Wahrscheinlichkeit für einen Fehler zweiter Art $P_\vartheta[T \notin K]$ für $\vartheta \in \Theta_A$ zu minimieren. Dazu maximiert man die *Macht des Tests*
$$\beta : \Theta_A \to [0, 1] \qquad \vartheta \mapsto \beta(\vartheta) := P_\vartheta[T \in K]$$
Damit ergibt sich der Zusammenhang $1 - \beta(\vartheta) = P_\vartheta[T \in K]$.

$\implies$ asymmetrisches Vorgehen führt dazu, dass es schwieriger ist, eine Hypothese zu verwerfen, als diese zu behalten. Das führt zu folgendem Verhalten in der Statistik:

In einem Test verwendet man als Hypothese immer die Negation der eigentlich gewünschten Aussage.

Aufgrund der Asymmetrie kann es durchaus vorkommen, dass bei Vertauschen von *Hypothese* und *Alternative* unterschiedlich entschieden wird.

8.1 Konstruktion von Tests

Def. 8.2 (**Likelihood-Quotient**)**.** Sei $L(x_1, \dots, x_n; \vartheta)$ die Likelihood Funktion und $\vartheta_0 \in \Theta_0$ und $\vartheta_A \in \Theta_A$. Dann definieren wir den Likelihood-Quotienten als

$$R(x_1, \dots, x_n; \vartheta_0, \vartheta_a) := \frac{L(x_1, \dots, x_n; \vartheta_0)}{L(x_1, \dots, x_n; \vartheta_A)}$$

Je kleiner dieser Quotient wird, desto wahrscheinlicher sind die Beobachtungen im Modell P_{ϑ_a} im Gegensatz zum Modell P_{ϑ_0}. $\implies$ wähle als Teststatistik $T = R(X_1, \dots, X_n; \vartheta_0, \vartheta_A)$ und als kritischen Bereich $K := [0, c)$. Sind Hypothese und Alternative jeweils einfach, so ist diesr Test optimal:

Satz 8.1 (Neyman-Pearson-Lemma). $\Theta_0 = \{\vartheta_0\}, \Theta_A = \{\vartheta_A\}$. Sei die Teststatistik $T := (X_1, \ldots, X_n; \vartheta_0, \vartheta_A)$ mit $K := [0, c)$ und sei $\alpha^* := P_{\vartheta_0}[T \in K] = P_{\vartheta_0}[T < c]$. Dann ist der *Likelihood-Quotienten-Test* mit T und K im folgenden Sinne optimal:

jeder andere Test mit Signifikanzniveau $\alpha \leq \alpha^*$ hat kleinere *Macht des Tests*,

was bedeutet, dass die Wahrscheinlichkeit für einen Fehler 2. Art grösser ist. Etwas formaler bedeutet dies für jeden anderen Test (T', K'):

$$P_{\vartheta_0}[T' \in K] \leq \alpha^* \implies P_{\vartheta_A}[T' \in K] \leq P_{\vartheta_A}[T \in K]$$

In den allermeisten Fällen sind weder Hypothese noch Alternative einfach. Um dennoch ein systematisches Vorgehen zu liefern, verallgemeinern wir zuerst den Likelihood-Quotienten:

$$\begin{aligned} R(x_1, \ldots, x_n) &:= \frac{\sup_{\vartheta \in \Theta_0} L(x_1, \ldots, x_n; \vartheta)}{\sup_{\vartheta \in \Theta_A} L(x_1, \ldots, x_n; \vartheta)} \\ \widetilde{R}(x_1, \ldots, x_n) &:= \frac{\sup_{\vartheta \in \Theta_0} L(x_1, \ldots, x_n; \vartheta)}{\sup_{\vartheta \in (\Theta_A \cup \Theta_0)} L(x_1, \ldots, x_n; \vartheta)} \end{aligned}$$

Nun wählt man eine dieser beiden Quotienten als Teststatistik T_0 mit einem kritischen Bereich $K_0 := [0, c_0)$. C_0 muss dabei so gewählt werden, dass der Test ein gewähltes Signifikanzniveau einhält.

Oft kann man auch durch Umformen eine einfachere Teststatistik finden, in dem man versucht, eine Beziehung der Art "Quotient klein *genau dann, wenn ...* "herzuleiten. Diese Bedingung kann man dann als Teststatistik verwenden. Schlussendlich braucht man noch die Verteilung von T unter der Hypothese H_0, um den kritischen Bereich K passend zum gewünschten Signifikanzniveau zu finden.

8.2 *p*-Wert

Def. 8.3 (*p*-Wert). ei $\Theta_0 = \{\vartheta_0\}$. Dann ist der *p-Wert* die Wahrscheinlichkeit, einen mindestens so extremen Wert der Teststatistik zu erhalten, falls die Nullhypothese wahr ist. Die Alternativhypothese bestimmt dabei, was als "extremer"gilt.

Haben wir also Daten $(x_1, \ldots, x_n)$ gesammelt und betrachten wir den Wert der Teststatistik $t(x_1, \ldots, x_n)$, so interessiert es uns, wie extrem dieser Wert unter Annahme der Nullhypothese ist.

Bemerkung: Der *p-Wert* gibt **nicht** an, wie wahrscheinlich die Nullhypothese bei Erhalt dieses Wertes ist!

Lemma 8.1. Am p-Wert kann direkt der Testentscheid abgelesen werden, liegt er unter dem Signifikanzniveau α, wird die Nullhypothese verworfen, ansonsten nicht.

Dies lässt sich wie folgt begründen: Ist der p-Wert kleiner als α, dann liegt der beobachtete Wert der Teststatistik sicher im Verwerfungsbereich.

8.3 *z*-Test

Test für den Erwartungswert einer Normalverteilung mit bekannter Varianz der Grundgesamtheit. Seien also $X_1, \ldots, X_n \sim \mathcal{N}(\vartheta, \sigma^2)$-verteilt (i.i.d.) für bekanntes $\sigma > 0$.

- **Hypothese:** $H_0 : \vartheta = \vartheta_0$
- **Teststatistik:**
$$T = \frac{\overline{X}_n - \vartheta_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1) \qquad \text{unter } P_{\vartheta_0}$$
- **Kritische Bereiche** (zum Signifikanzniveau $\alpha \in (0, 1)$) kann au Tabelle abgelesen werden: Dabei bezeichnet z_α das

Alternative H_A	**Kritischer Bereich**
$\vartheta < \vartheta_0$	$(-\infty, z_\alpha)$
$\vartheta > \vartheta_0$	$(z_{1-\alpha}, \infty)$
$\vartheta \neq \vartheta_0$	$(-\infty, z_{\alpha/2}) \cup (z_{1-\alpha/2}, \infty)$

α-Quantil der Standardnormalverteilung. Man findet es, indem man in der Tabelle der Standardnormalverteilung nach $\Phi^{-1}(\alpha)$ sucht. Aus Symmetriegründen gilt $z_\alpha = -z_{1-\alpha}$.

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

Rezept Fehler 2. Art berechnen:
Nehme an: einseitiger z-Test, $T = \frac{\overline{X_n}-\mu_0}{\sigma/\sqrt{n}}$, $\mu_0 = 70$

$H_0 : \mu = \mu_0$; $H_A : \mu < \mu_0$.

Kritischer Bereich mit 5%-Niveau: $K = (-\infty, -1.645)$

Objective: Fehler 2. Art finden für $\mu_A = 69.5$. Wir nehmen an, dass $T = \frac{\overline{X_n}-\mu_A}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ unter P_{μ_A}

$$\begin{aligned}
\text{Fehler 2. Art} &= P_{\mu_A}[T \notin K] \\
&= P_{\mu_A}[T > -1.645] \\
&= P_{\mu_A}\left[\frac{\overline{X_n} - \mu_0}{\sigma/\sqrt{n}} > -1.645\right] \\
&= P_{\mu_A}\left[\frac{\overline{X_n} - \mu_A}{\sigma/\sqrt{n}} > \frac{\mu_0 - \mu_A}{\sigma/\sqrt{n}} - 1.645\right] \quad \text{mit addition} \\
&= 1 - P_{\mu_A}\left[\frac{\overline{X_n} - \mu_A}{\sigma/\sqrt{n}} \leq \frac{\mu_0 - \mu_A}{\sigma/\sqrt{n}} - 1.645\right] \\
&= 1 - \Phi\left(\frac{\mu_0 - \mu_A}{\sigma - \sqrt{n}} - 1.645\right) \quad \text{weil } \sim \mathcal{N}(0,1)
\end{aligned}$$

8.4 *t*-Test

Test für den Erwartungswert einer Normalverteilung mit unbekannter Varianz. Seien also $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$-verteilt (i.i.d.) für unbekanntes $\sigma > 0$.

- **Hypothese:** $H_0 : \mu = \mu_0$. Formal präziser wäre $\Theta_0 = \{\vartheta = (\mu_0, \sigma) \mid \sigma > 0\}$
- **Teststatistik:**

$$T = \frac{\overline{X}_n - \mu_0}{S/\sqrt{n}} \sim t_{n-1} \qquad \text{unter } P_{\mu_0}, \text{ wobei } S^2 := \text{empirische Stichprobenvarianz}$$

- **Kritische Bereiche** (zum Signifikanzniveau $\alpha \in (0,1)$) kann aus Tabelle abgelesen werden:

Alternative H_A	**Kritischer Bereich**
$\mu < \mu_0$	$(-\infty, t_{n-1,\alpha})$
$\mu > \mu_0$	$(t_{n-1,1-\alpha}, \infty)$
$\mu \neq \mu_0$	$(-\infty, t_{n-1,\alpha/2}) \cup (t_{n-1,1-\alpha/2}, \infty)$

Dabei bezeichnet $t_{m,\alpha}$ das α-Quantil der t_m-Verteilung. Aus Symmetriegründen gilt $t_{m,\alpha} = -t_{m,1-\alpha}$:

$$\int_{-\infty}^{t_{m,\alpha}} f_m(x)\, dx = \alpha$$

wobei f_m die Dichte der t_m Verteilung ist. Diesen Wert erhält man aus einer Tabelle zur t-Verteilung.

8.5 Gepaarte Zweistichproben-Tests für Normalverteilungen

Seien $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ Zufallsvariablen, so dass (X_i, Y_i) natürliche Paare bilden. Bezeichnen wir nun $Z_i := X_i - Y_i$.

- *bekannte Varianz:* Falls $Z_1, \ldots, Z_n \sim \mathcal{N}(\vartheta, \sigma^2)$ (i.i.d.) für bekanntes $\sigma > 0$, dann kann z-Test analog zu Kapitel 8.3 angewendet werden.
- *unbekannte Varianz:* Falls $Z_1, \ldots, Z_n \sim \mathcal{N}(\mu, \sigma^2)$ (i.i.d.) für unbekanntes $\sigma > 0$, dann kann t-Test analog zu Kapitel 8.4 angewendet werden.

8.6 Ungepaarte Zweistichproben-Tests für Normalverteilungen

Seien $X_1, \ldots, X_n \sim \mathcal{N}(\mu_X, \sigma_X^2)$ (i.i.d.) und $Y_1, \ldots, Y_m \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ (i.i.d.), so dass alle X_i, Y_j unabhängig.

8.6.1 Normalverteilungen mit bekannten Varianzen

Seien also σ_X, σ_Y bekannt.

- **Hypothese:** $H_0 : \mu_X - \mu_Y = \mu_0$ (bspw. $\mu_0 = 0$)
- **Teststatistik:**

$$T = \frac{\overline{X}_n - \overline{Y}_m - \mu_0}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim \mathcal{N}(0,1) \qquad \text{für } P_{\mu_0}$$

  Die kritischen Bereiche zum Signifikanzniveau sind analog zur Tabelle aus Kapitel 8.3.

8.6.2 Normalverteilungen mit unbekannten aber gleichen Varianzen

Sei also $\sigma_X = \sigma_Y = \sigma$ für $\sigma > 0$ unbekannt.

- **Hypothese:** $\mu_X - \mu_Y = \mu_0$ (bspw. $\mu_0 = 0$)
- **Teststatistik:**

$$T = \frac{\overline{X}_n - \overline{Y}_m - \mu_0}{S\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2} \qquad \text{unter } P_{\mu_0}$$

- **Kritische Bereiche:** analog zu Tabellae aus Kapitel 8.4, jedoch ist nun die Anzahl der Freiheitsgrade $n + m - 2$ und nicht mehr $n - 1$.

Dabei benutzen wir für die Varianz ein gewichtetes Mittel aus den Stichprobenvarianzen S_X, S_Y, definiert als

$$S^2 := \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$

9 Konfidenzbereiche

Wir suchen aus einer Familie $(P_\vartheta)_{\vartheta\in\Theta}$ von Modellen eines, welches zu unserern Daten passt. Da es aber extrem schwierig ist, einen Parameter ϑ genau zu schätzen, suchen wir nun eine (zufällige) Teilmenge des Parameterbereichs, der hoffentlich den wahren Parameter enthält.

Def. 9.1 (Konfidenzbereich). Ein *Konfidenzbereich* für ϑ zu Daten $x_1,\dots,x_n$ ist eine Menge $C(x_1,\dots,x_n)\subseteq\Theta$. Damit ist $C(X_1,\dots,X_n)$ eine zufällige Teilmenge Θ. Dieses C heisst Konfidenzbereich *zum Niveau* $1-\alpha$, falls für alle $\vartheta\in\Theta$ gilt:

$$P_\vartheta[\vartheta\in C(X_1,\dots,X_n)]\geq 1-\alpha$$

Rezept (angewendeter t-Test): Konfidenzintervall mit Niveau $1-\alpha$, n Stichproben, Stichprobenmittel $\overline{X_n}$, Stichprobenvarianz s_X^2; beidseitig

$$I_{(1-\alpha)}=[\overline{X_n}-\frac{s_X}{\sqrt{n}}\cdot t_{n-1,1-\frac{\alpha}{2}},\ \overline{X_n}+\frac{s_X}{\sqrt{n}}\cdot t_{n-1,1-\frac{\alpha}{2}}]$$

Die bedeutet intuitiv, dass man in jedem Modell den wahren Parameter mit grosser Wahrscheinlichkeit erwischt. Kennt man die Verteilung genau genug, so kann man exakte Konfidenzintervalle zu einem Signfikanzniveau angeben. Oft ist dies jedoch nicht der Fall und man kann nur approximative Angaben machen, z.B. mit dem *Zentralen Grenzwertsatz*

9.1 Zusammenhang von Kondifenzbereichen und Tests

Wir zeigen im Folgenden, dass beide Konzept grundlegend zusammenhängen und ineinander überführt werden können.

Sei $C(X_1,\dots,X_n)$ ein Konfidenzbereich für ϑ zum Niveau $1-\alpha$. Wir wollen die Hypothese $H_0:\vartheta=\vartheta_0$ testen. Dazu definieren wir einen Test

$$I_{\{\vartheta_0\notin C(X_1,\dots,X_n)\}}$$

der H_0 ablehnt $\iff$ v_0 liegt nicht in $C(X_1,\dots,X_n)$. Damit folgt aus der Einfacheit von $\Theta_0=\{\vartheta_0\}$ für jedes $\vartheta\in\Theta_0$:

$$P_\vartheta[\vartheta_0\notin C(X_1,\dots,X_n)]=1-P_\vartheta[\vartheta_0\in C(X_1,\dots,X_n)]\leq\alpha$$

Dieser Test hat also gerade Signifikanzniveau α. Aus dem Konfidenzbereich für ϑ erhalten wir also eine Familie von Tests, nämlich für jede einfache Hypothese $\Theta_0=\{\vartheta_0\}$ mit $\vartheta_0\in\Theta$ genau einen Test.

10 Einfache Lineare Regression

Wir betrachten lediglich zwei Zufallsvariablen X, Y, zwischen denen wir einen linaren Zusammenhang vermuten. Dies schreiben wir durch ein Modell der Form

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

für konstante β_0, β_1 und eine von X unabhängige Zufallsvariable ε mit $\mathbb{E}[\varepsilon] = 0$ und endlicher Varianz.

Seien (x_i, y_i) für $i \in \{1, \dots, n\}$ unabhängige Realisierungen von (X, Y), also von uns beobachtete Daten. Wir suchen die Paramter β_0, β_1 mittels der Methode der kleinsten Quadrate, d.h. wir minimieren

$$f(\beta_0, \beta_1) := \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i - y_i)^2$$

in dem wir nach β_0 und β_1 ableiten, und den Gradient $\nabla f = 0$ setzen. Zuerst definieren wir aber

$$\overline{x} := \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \overline{y} := \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad \overline{x^2} := \frac{1}{n}\sum_{i=1}^{n} x_i^2, \qquad \overline{xy} := \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

Nun finden wir den Gradienten von f und setzen ihn $= 0$:

$$\nabla f(\beta_0, \beta_1) = \begin{pmatrix} 2\sum_{i=1}^{n} (\beta_0 + \beta_1 x_i - y_i) \\ 2\sum_{i=1}^{n} x_i(\beta_0 + \beta_1 x_i - y_i) \end{pmatrix} \overset{!}{=} 0 \implies \beta_0 + \beta_1 \overline{x} = \overline{y} \quad \wedge \quad \beta_0 \overline{x} + \beta_1 \overline{x^2} = \overline{xy}$$

Dieses Gleichungssystem kann man nun auflösen und erhält:

$$\beta_1 = \frac{\operatorname{cov}(x, y)}{\operatorname{var}(x)} \qquad\qquad \beta_0 = \overline{y} - \frac{\operatorname{cov}(x, y)}{\operatorname{var}(x)} \overline{x}$$

wobei $\operatorname{var}(x)$ und $\operatorname{cov}(x, y)$ die *Stichprobenvarianz* bzw. *Strichprobenkovarianz* bezeichnet. Beachte, dass diese Schätzungen nicht erwartungstreu sind, und man deshalb oft die korrigierten Varianten benutzt:

$$\operatorname{var}(x) := \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2 \quad \text{vs.} \quad \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})^2$$

$$\operatorname{cov}(x, y) := \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) \quad \text{vs.} \quad \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

Für grosse n wird dieser Unterschied jedoch zunehmened geringer und es spielt dann keine Rolle mehr, welche Formel verwendet wird.