Stochastik für Informatiker

In LaTeX von: Manuel Albert

Stand: 11.07.2011

Inhaltsverzeichnis

1 Wahrscheinlichkeitstheorie

1.1 Wahrscheinlichkeit

Der Begriff der Wahrscheinlichkeit (kurz: W-keit oder W.) geht auf die naive Vorstellung zurück, dass man unter verschiedenene Fällen, die in einer Situation eintreten können, manche als günstig (für ein bestimmtes "Ereignis") ansieht, und den Quotienten

$$W := \frac{Anzahl\ der\ guenstigen\ Faelle}{Anzahl\ aller\ moeglichen\ Faelle}\ (*)$$

bildet.
Die Zahl W interpretiert man dann als Wahrscheinlichkeit für das Eintreten dieses Ereignisses. Dies ist wirklich eine sehr naive Sicht, denn sie beruht auf der Annahme, dass diese verschiedenen Fälle in irgendeinem Sinn „gleich wahrscheinlich" sind. Diese Sicht ist aber zumindest für sogenannte Laplace'sche Wahrscheinlichkeitsräume gerechtfertigt.

Bsp.: Einmaliges Werfen eines unverfälschten Würfels

In diesem Fall ist die Wahrscheinlichkeit für jede der Zahlen von 1 bis 6 gleich $W = \frac{1}{6}$.
Nun sei ein realer, d.h. physischer Würfel, gegeben. Dann hat man folgende Möglichkeiten:

1. Man glaubt, der Würfel ist ideal, d.h. man wählt den idealen Würfel als Modell. Dann hat man das Problem, dass diese Annahme eventuell falsch ist.

2. Man würfelt, z.B. 10.000 mal, und bestimmt empirisch, die Wahrscheinlichkeit nach (*). Dann hat man das Problem, dass selbst bei 10.000 Versuchen die „wahren" Verhältnisse nur sehr ungenau bestimmt wurden, wenn sich diese Verhältnisse überhaupt für eine große Zahl von Versuchen stabilisieren.
   Es könnte ja sein, dass sich für gewisse Versuchsreihen ganz spezielle Zahlen ergeben, z.B. „ständig 6". Naiv würde man „ständig 6" als „ganz unwahrscheinlich" ansehen.

Man sieht sehr schnell ein, dass man mit dem Versuch, Wahrscheinlichkeit als „Grenzwert" von empirischer Häufigkeit zu definieren, in die Gefahr eines Kreisschlusses gerät.
In der Tat haben sich alle Versuche, Wahrscheinlichkeit auf elementaren Weg über relative Häufigkeiten zu erklären, als nicht praktikabel erwiesen. Letztlich hat sich ein nicht-empirischer Zugang als einfacher erwiesen. Dieser axiomatische Zugang geht auf Kolmogoroff zurück.
Zunächst wird eine einfache Situation betrachtet.

1.2 Laplace'sche Wahrscheinlichkeitsräume

Es sei ein Grundraum $\Omega = \{\omega_1, \ldots, \omega_n\}$ von Elementarereignissen ω_i gegeben.
Dabei ist $n \in \mathbb{N}$ eine natürliche Zahl.
Die Teilmengen $A \subseteq \Omega$ dieses Raumes heissen Ereignisse.
Die Menge aller Ereignisse ist also die Potenzmenge $\mathcal{P}(\Omega)$ von Ω.

Dabei sind $\emptyset$ (leere Menge) und Ω selbst auch Ereignisse.
Die Potenzmenge einer Menge von n Elementen hat 2^n Elemente.
Also gibt es 2^n Ereignisse in Ω.
Nun definiere eine Funktion $P\colon \mathcal{P}(\Omega) \to \mathbb{R}$ (P steht für "probability") wie folgt:
Für ein Elementarereignis $\{\omega_i\}$ setzt man $P(\{\omega_i\}) := \frac{1}{n}$ und schreibt dafür auch kurz $P(\omega_i)$. Alle Elementarereignisse haben also dieselbe Wahrscheinlichkeit $\frac{1}{n}$.
Für ein beliebiges Ereignis $A \in \mathcal{P}(\Omega) (\Leftrightarrow A \subseteq \Omega)$ definiere nun

$$P(A) := \frac{\textit{Anzahl der Elementarereignisse in } A}{n}$$

Insbesondere ist $P(\emptyset) = 0$ (mögliche Ereignis) und $P(\Omega) = 1$ (sichere Ereignis).
Die Menge Ω zusammen mit der Funktion P, also das Tupel (Ω, P) nennt man den Laplace'schen Wahrscheinlichkeitsraum mit n-Elementen.
P nennt man die Gleichverteilung auf Ω.
Da Ereignisse Teilmengen von Ω sind, kann man die Rechenregeln der Mengenlehre benutzen.
Man definiert Vereinigung $A \cup B$ und Schnitt $A \cap B$ von zwei Ereignissen $A, B \subseteq \Omega$ via:

$$A \cup B := \{\omega \mid \omega \in A \text{ oder } \omega \in B\}$$
$$A \cap B := \{\omega \mid \omega \in A \text{ und } \omega \in B\}$$

Dabei ist „oder" das einschließende „oder" (es tritt A oder auch B ein) und „und" heißt, dass wirklich A und zugleich B eintritt.
Analog definiert man für abzählbar unendlich viele $A_1, A_2, \ldots$
Vereinigung $\bigcup_{n=1,2,\ldots} A_n$ und Schnitt $\bigcap_{n=1,2,\ldots} A_n$.
Es gilt dann (nachrechnen oder siehe Bild) das Gesetz:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Speziell gilt:

$$P(A \cap B) = 0 \Rightarrow P(A \cup B) = P(A) + P(B)$$

Laplace'scher Wahrscheinlichkeitsraum:

$$P(A) = \frac{H \textit{ (Menge) der für } A \textit{ günstigen Fälle}}{n}$$

Beispiele für Laplace'sche Wahrscheinlichkeitsräume sind:

- idealer Würfel ($n = 6$)
  Hier kann man ide Potenzmenge, also die Menge aller Ereignisse, noch als Liste aufschreiben. Summe: $64 = 2^6$
- ideales Skat-Kartenspiel ($n = 32$)
  Im Fall des Kartenspieles, gibt es 2^{32} Ereignisse, es ist sicher nicht sinnvoll, diese alle

aufzuschreiben.

- ideales Roulette ($n = 37$)
- zweimaliges Werfen eines idealen Würfels ($n = 6^2 = 36$)
  Wie groß ist die Wahrscheinlichkeit, bei zwei Würfeln keine 6 zu bekommen?
  Dieses Problem kann man auf folgende Weise lösen. Man kann die 36 Elementarereignisse aufschreiben und abzählen, bei wie vielen „keine 6 " kommt. Das sind 25. Also ist die Wahrscheinlichkeit $\frac{25}{36}$. Man kann aber auch so argumentieren: Die Wahrscheinlichkeit beim ersten Wurf keine 6 zu bekommen ist $\frac{5}{6}$. Die Wahrscheinlichkeit beim zweiten Wurf keine 6 zu bekommen ist ebenfalls $\frac{5}{6}$. Also ist die gesuchte Wahrscheinlichkeit $\frac{5}{6} * \frac{5}{6} = \frac{25}{36}$.
  Wie groß ist die Wahrscheinlichkeit, bei zwei Würfen eine 6 und eine weitere gerade Zahl zu bekommen?
  Kann man abzählen: $(6,2), (6,4), (6,6), (4,6), (2,6)$.
  Also ist die Wahrscheinlichkeit gleich $\frac{5}{36}$.
  Es sei

$$A := \text{„mindestens eine 6“}, \qquad P(A) = \frac{11}{36}$$

$$B := \text{„zwei gerade Zahlen“}, \qquad P(B) = \left(\frac{3}{6}\right)^2 = \frac{1}{4}$$

  Nach obigem gilt

$$P(A \cap B) = \frac{5}{36} \Rightarrow P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{11}{36} + \frac{9}{36} - \frac{5}{36} = \frac{5}{12}$$

  Es ist also $P(\text{„mind. eine 6 oder zwei gerade Zahlen“}) = \frac{5}{36}$

1.3 Axiomatischer Zugang

Gegeben: Menge $\Omega \neq \emptyset$.
Diese nennt man Grundraum oder Ereignisraum.
Die Elemente von Ω heißen Elementarereignisse.
Sei $\mathcal{F}$ eine Familie von Teilmengen von Ω mit folgenden Eigenschaften:

1. $\Omega \in \mathcal{F}$
2. Mit $A \in \mathcal{F}$ ist auch das Komplement $\overline{A} := \{\Omega \setminus A\} \in \mathcal{F}$.
3. Gilt $A_1, A_2, \cdots \in \mathcal{F}$, so auch $\bigcup_{n=1,2,\ldots} A_n \in \mathcal{F}$ und $\bigcap_{n=1,2,\ldots} A_n \in \mathcal{F}$

Sei nun $P : \mathcal{F} \to \mathbb{R}$ eine Funktion mit folgenden Eigenschaften:

(i) $0 \leq P(A) \leq 1, \ \forall A \in \mathcal{F}$

(ii) $P(\emptyset) = 0$ und $P(\Omega) = 1$

(iii) Sind $A_1, A_2, \cdots \in \mathcal{F}$ mit $A_i \cap A_j = \emptyset \; \forall i \neq j$, so gilt:

$$P\left(\bigcup_{n=1,2,\ldots} A_n\right) = P(A_1) + P(A_2) + \cdots + A_n$$

Ein solches Tripel $(\Omega, \mathcal{F}, P)$ nennt man einen Wahrscheinlichkeitsraum.
Es gilt dann z.B. (der Beweis wird hier nicht gegeben) für beliebige

$$A, B \in \mathcal{F} : P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

1.4 Bedingte Wahrscheinlichkeit

Betrachte den Laplace-Raum mit n Elementarereignissen.
Für zwei Ereignisse A, B sei bekannt

$$P(A) = \frac{k}{n},\; P(B) = \frac{l}{n},\; P(A \cap B) = \frac{m}{n}$$

Dann definiert man die bedingte Wahrscheinlichkeit von B unter A durch:

$$P(B \mid A) := \frac{P(A \cap B)}{P(A)} = \frac{\frac{m}{n}}{\frac{k}{n}} = \frac{m}{k}$$

Nach naivem Verständnis ist dies die Wahrscheinlichkeit von B unter der Voraussetzung, dass A schon eingetreten ist.
Analog:

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

Nun gilt der Multiplikationssatz:

$$P(A \cap B) = P(A \mid B) * P(B) = P(B \mid A) * P(A)$$

Man kann diesen Satz benutzen, wenn man die bedingte Wahrscheinlichkeit $P(B \mid A)$ kennt, um $P(A \mid B)$ zu bestimmen.
Bsp.: Einmaliges Würfeln, $n = 6$.
Sei $A =$„≥ 2“, $B =$ „ungerade“
Dann gilt: $P(A) = \frac{5}{6}$, $P(B) = \frac{1}{2}$ und $P(B \mid A) = \frac{2}{5}$ Somit:

$$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)} = \frac{\frac{2}{5} * \frac{5}{6}}{\frac{1}{2}} = \frac{2}{3}$$

In vielen Fällen ist $P(B)$ im Nenner nicht direkt bekannt. Dann lässt sich diese oft mit dem Satz von der totalen Wahrscheinlichkeit

$$P(B) = P(B \mid A) * P(A) + P(B \mid \overline{A}) * P(\overline{A})$$

berechnen.
Dieser Satz gilt auch allgemeiner in der Form

$$P(B) = \sum_i P(B \mid A_i) * P(A_i)$$

sofern $(A_i)_i$ endlich viele oder sogar abzählbar unendlich viele Ereignisse sind, die eine Zerlegung von Ω bilden, d.h. $\Omega = \bigcup_i A_i$ und $A_i \cap A_j = \emptyset$, $\forall i \neq j$.

1.5 Unabhängigkeit

Definition

Zwei Ereignisse $A, B \in \mathcal{F}$ heißen unabhängig, wenn die Gleichheit

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

gilt.
Sind A und B unabhängig und gilt zusätzlich $P(B) > 0$, so folgt

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) * P(B)}{P(B)} = P(A),$$

d.h. die Wahrscheinlichkeit $P(A)$ hängt nciht davon ab, ob B eingetroffen ist.

Bsp.: Zweimaliges Ziehen einer Spielkarte

$A :=$ „Ass beim 1. Zug“
$B :=$ „Ass beim 2. Zug“

1. Ziehen und Zurücklegen
   Laplace-Raum mit $n = 32^2$.
   $P(A) = P(B) = \frac{1}{8}$. $P(A \cap B) = \frac{1}{64}$.
   A und B sind unabhängig.

2. Ziehen ohne Zurücklegen
   Dies ergibt einen Wahrscheinlichkeitsraum mit $32 * 31$ Elementarereignissen.
   Es ist $P(A) = P(B) = \frac{1}{8}$, aber $P(A \cap B) = \frac{1}{8} * \frac{3}{31}$, d.h. A und B sind nicht unabhängig.

1.6 Zufallsvariable

Das Konzept eines Wahrscheinlichkeitsraumes ist nützlich, um verschiedene Dinge zu berechnen, z.B. die Wahrscheinlichkeit für ein bestimmtes Ereignis.
Im Allgemeinen ist man jedoch an Aussagen interessiert, die mit beobachteten Größen verbunden sind. Man möchte Aussagen darüber treffen, wie sich eine zufällige Größe „im Mittel" verhält und man möchte die Variabilität dieser Größe beschreiben.
Dazu muss man die Elementarereignisse ω einem Zahlenwert $X(\omega)$ zuordnen.
In der Stochastik heißt eine solche Zuordnung

$$X\colon \Omega \to \mathbb{R},\ \omega \mapsto X(\omega),$$

eine Zufallsvariable (kurz: ZV), und nicht eine Zufallsfunktion, wie man denken könnte.
In vielen praktischen Fällen ist diese Zufallsvariable X eine Messung oder Beobachtung, z.B. das Atemvolumen eines Menschen in Ruhe, das Gewicht einer Frucht, uvm.
Beim Würfeln kann man jeder Augenzahl einen Verlust oder Gewinn zuordnen,
z.B. $X(1) = X(2) = \cdots = X(5) := -1$ (Verlust) und $X(6) := 1$ (Gewinn).
Auch die Augenzahl selber ist bereits eine Zufallsvariable $(X(\omega) := \omega)$.
Wenn die Zufallsvariable jedem Elementarereignis einen anderen Wert zuordnet, so geschieht noch nicht viel Interessantes. In der Regel wird jedoch die Zufallsvariable gelegentlich *verschiedenen* Elementarereignissen *gleiche* Werte zuordnen. Dann kann man alle Elementarereignisse, denen die Zufallsvariable einen bestimmten Wert zuordnet, zu einem Ereignis zusammenfassen und dessen Wahrscheinlichkeiten bestimmen.
Zu gegebenem $a \in \mathbb{R}$ definiert man das Ereignis

$$\{\omega \mid X(\omega) = a\},$$

das kurz mit

$$\{X = a\}$$

bezeichnet wird.
Wenn Ω überabzählbar viele Elemente hat, so wird i.A. dieses Ereignis die Wahrscheinlichkeit 0 haben. Dann bildet man zu Zahlen $a < b$ das Ereignis

$$A := \{\omega \mid a \le X(\omega) \le b\}$$

oder kurz

$$A = \{a \le X \le b\}$$

Zu jedem dieser Ereignisse A gibt es eine Wahrscheinlichkeit $P(A)$.
Die Gesamtheit all dieser Wahrscheinlichkeiten nennt man die Verteilung der Zufallsvariable X.
Die Verteilung einer Zufallsvariablen wird durch ihre Verteilungsfunktion $F\colon \mathbb{R} \to \mathbb{R}$ beschrie-

ben, die definiert ist durch

$$F(x) := P(X \leq x), \ \forall x \in \mathbb{R}$$

F ist eine monoton wachsende, rechtsseitig stetige Funktion mit Werten in $[0,1]$. Man kann zeigen, dass sich Verteilungen (von Zufallsvariablen) und Verteilungsfunktionen in eineindeutiger Weise entsprechen.
Es ist klar, dass die Definition des Wahrscheinlichkeitsraumes und die der Zufallsvariablen i.A. nicht zueinander passen. Z.B. wird i.A. eine Menge $\{a \leq X \leq b\}$ gar nicht in der Mengenfamilie $\mathcal{F}$ liegen. Spricht man von Zufallsvariablen, so setzt man dies jedoch voraus. Diese Eigenschaft nennt man Messbarkeit.
Eine genauere Betrachtung kann nur im Rahmen der Maßtheorie erfolgen, auf die hier nicht eingegangen wird.
Unabhängigkeit lässt sich nicht für Ereignisse, sondern auch für Zufallsvariablen definieren:

Definition

Zwei Zufallsvariablen X, Y heißen unabhängig, falls

$$P(\{X \leq a\} \cap \{Y \leq b\}) = P(X \leq a).P(Y \leq b), \ \forall a, b \in \mathbb{R}$$

1.7 Diskrete Zufallsvariablen

Definition

Eine Zufallsvariable X heißt diskret (verteilt), wenn sie endlich viele oder abzählbar unendlich viele Werte annimmt.
Im endlichen Fall sind dies also Werte $x_1, \ldots, x_n$, die jeweils mit Wahrscheinlichkeit $p_i := P(X = x_i)$, $i \in \{1, \ldots, n\}$, angenommen werden,
im anderen Fall bilden die $x_1, x_2, \ldots$ eine unendliche Folge. In jedem Fall gilt $p_i \geq 0$ und $\sum_i p_i = 1$.
Im abzählbaren Fall ist $\sum_i p_i$ also insbesondere eine konvergente Reihe.
Es kann natürlich sein, dass diese Konvergenz sehr langsam ist. Die Zahlenpaare $(x_i, p_i)_i$ bilden die Verteilung von X.
Es gilt

$$p_i = P(X = x_i)$$

In allen praktischen Fällen nimmt man an, dass die x_i aufsteigend geordnet sind. Die zugehörige Verteilungsfunktion F ist in diesem Fall eine monoton wachsende, rechtsseitig stetige Treppenfunktion. An der Stelle x_i hat sie einen Sprung der Höhe p_i.
Man will die Verteilung von X durch einige wichtige Zahlen (Momente) möglichst gut beschreiben. Eine der wichtigsten dieser Zahlen ist der Erwartungswert (1. Moment)

$$E(X) := \sum_i x_i * p_i$$

Er beschreibt in gewissem Sinne die „Mitte“ der Verteilung. Man kann leicht Beispiele angeben, in denen diese Reihe nicht absolut konvergiert (Übung).
Da solche Fälle aber nur geringe praktische Bedeutung haben, wird in Anwendungen oft davon ausgegangen, dass die Reihe konvergiert.
Eine zweite wichtige Kennzahl der Verteilung von X ist die Varianz

$$\mathrm{Var}(X) := \sum_i (x_i - E(X))^2 * p_i$$

Sie misst die Breite bzw. die Variabilität der Verteilung von X.
Man beachte, dass $E(X)$ und $\mathrm{Var}(X)$ verschiedene Dimension haben. Bezeichnet etwa X eine Länge in $\mathtt{cm}$, so hat $E(X)$ die Dimension $\mathtt{cm}$, aber $\mathrm{Var}(X)$ die Dimension $\mathtt{cm}^2$.
Daher führt man die Standardabweichung (oder Streuung)

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

ein.
Die Varianz lässt sich auf verschiedene Weise umformen, bzw. berechnen.
Es gilt nämlich

$$\begin{aligned}\mathrm{Var}(X) &= \sum_i \left((x_i)^2 - 2 * x_i * \mathrm{E}(X) + \mathrm{E}(X)^2\right) * p_i \\ &= \sum_i \left((x_i)^2 * p_i - 2 * \mathrm{E}(X)\right) * \sum_i \left(x_i * p_i + \mathrm{E}(X)^2\right) * \sum_i p_i \\ &= \mathrm{E}(X^2) - \mathrm{E}(X)^2\end{aligned}$$

Bsp.: Idealer Würfel

X zählt die Augen.
X nimmt die Werte $1, 2, 3, 4, 5, 6$ je mit Wahrscheinlichkeit $\frac{1}{6}$ an.
Also gilt

$$\mathrm{E}(X) = (1 + 2 + 3 + 4 + 5 + 6) * \frac{1}{6} = \frac{7}{2}$$

und

$$\mathrm{Var}(X) = (1 + 4 + 9 + 16 + 25 + 36) * \frac{1}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$$

Wenn X eine positive Varianz besitzt, kann man eine neue Zufallsvariable X^* bilden

$$X^* := \frac{X - \mathrm{E}(X)}{\sigma(X)}$$

Die Zufallsvariable X^* heißt die Standardisierte von X.
Es gilt

$$\mathrm{E}(X^*) = 0$$

und

$$\mathrm{Var}(X^*) = 1$$

Standardisieren ist nützlich, wenn man Verteilungen verschiedener Zufallsvariablen vergleichen will.

1.8 Binomialverteilung

Betrachte ein Experiment mit nur zwei Ausgängen: 1(Erfolg) und 0(Misserfolg).
Ein Erfolg trete mit Wahrscheinlichkeit p ein, wobei $0 \leq p \leq 1$.
Man bildet ein zusammengesetztes Experiment durch n-fache unabhängige Wiederholung. Das zusammengesetze Experiment hat offensichtlich 2^n Ausgänge, nämlich alle Folgen der Länge n mit Elementen 0 und 1.
Zugehöriger Grundraum:

$$\Omega = \{(a_1, \ldots, a_n) \mid a_i \in \{0, 1\},\ \forall\ 1 \leq i \leq n\}$$

Sei k die Anzahl der Einsen in der Folge ω.
Dann gilt:

$$P(\omega) = p^k * q^{n-k},\ q \coloneqq 1 - p$$

Betrachte nun $X \coloneqq$ „Anzahl der Einsen“.
Dadurch ist eine Zufallsvariable $X : \Omega \to \{0, \ldots, n\}$ definiert, die die Anzahl der Erfolge zählt.
Wie groß ist die Wahrscheinlichkeit genau k Einsen zu erhalten?
Da es genau $\binom{n}{k}$ mögliche Folgen $\omega = (a_1, \ldots, a_n) \in \Omega$ gibt, die genau k Einsen enthalten, folgt

$$P(X = k) = \binom{n}{k} * p^k * q^{n-k},\ k \in \{0, \ldots, n\} (*)$$

Definition

Eine Zufallsvariable X, die die Werte $k = 0, 1, \ldots, n$ mit den Wahrscheinlichkeiten $(*)$ annimmt heißt binomial-verteilt, kurz $B(n, p)$-verteilt.
Die Zahlen $n \in \mathbb{N} \coloneqq \{1, 2, \ldots\}$ und $p \in [0, 1]$ heißen die Parameter dieser Verteilung.
Im einfachsten Fall (ideale, unverfälschte Münze) ist $p = q = \frac{1}{2}$, also

$$P(X = k) = \binom{n}{k} * \left(\frac{1}{2}\right)^n$$

1.9 Erzeugende Funktionen

Diskrete Verteilungen führen schnell zu umständlichen Rechnungen. Z.B. ist es bereits etwas mühsam, für die Binomialverteilung den Erwartungswert

$$\mathrm{E}(X) = \sum_{k=0}^{n} k \cdot p_k$$

zu berechnen.
Man kann dieser Mühe etwas aus dem Weg gehen, indem man sich eine beliebige Zahl $s \in [0,1]$ wählt und dann mit den Wahrscheinlichkeiten p_k die Reihe(Summe)

$$f(s) := \sum_{k} p_k \cdot s^k$$

bildet.
Dies ergibt eine Funktion $f\colon [0,1] \to \mathbb{R}$, welche erzeugende Funktion der Verteilung $(p_k)_k$ genannt wird. Für die Binomialverteilung ergibt sich

$$f(s) = \sum_{k=0}^{n} \binom{n}{k} p^k \cdot q^{n-k} \cdot s^k = (p \cdot s + q)^n$$

Erwartungswert und Varianz einer diskreten Zufallsvariablen (mit Werten in $\mathbb{N}_0$) lassen sich mit Hilfe von f berechnen:
Es gilt

$$\begin{aligned} f(s) &= \sum_{k} p_k \cdot s^k, \\ f'(s) &= \sum_{k} k \cdot p_k \cdot s^{k-1}, \\ f''(s) &= \sum_{k} k \cdot (k-1) \cdot p_k \cdot s^{k-2} \end{aligned}$$

Setzt man speziell $s = 1$ ein, so folgt:

$$\begin{aligned} f(1) &= \sum_{k} p_k = 1 \\ f'(1) &= \sum_{k} k \cdot p_k = \mathrm{E}(X) \\ f''(1) &= \sum_{k} k \cdot (k-1) \cdot p_k = \left(\sum_{k} k^2 \cdot p_k \right) - \left(\sum_{k} k \cdot p_k \right) = \mathrm{E}(X^2) - \mathrm{E}(X) \end{aligned}$$

also

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 = f''(1) + f'(1) - f'(1)^2$$

Im Fall der Binomialverteilung ergibt sich

$$f'(s) = n \cdot (p \cdot s + q)^{n-1} \cdot p$$
$$\text{und}$$
$$f''(s) = n \cdot (n-1) \cdot (p \cdot s + q)^{n-2} \cdot p^2$$

also

$$\mathrm{E}(X) = f'(1) = np$$
$$\text{und}$$
$$f''(1) = n \cdot (n-1) \cdot p^2$$

und somit

$$\begin{aligned}\mathrm{Var}(X) &= f''(1) + f'(1) - f'(1)^2 \\ &= n \cdot (n-1) \cdot p^2 + np - (np)^2 \\ &= np \cdot ((n-1) \cdot p + 1 - np) = np(1-p) = npq\end{aligned}$$

Die Binomialverteilung hat also den Erwartungswert $n \cdot p$ und die Varianz $n \cdot p \cdot q$.
Man kann diese beiden Formeln auch unter Verwendung des folgenden, allgemeinen Satzes herleiten.

Satz

Seien X und Y zwei Zufallsvariablen mit existierenden Erwartungswerten und Varianzen. Dann gilt

$$\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y)$$

Sind X und Y unabhängig, so gilt

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

Wende diesen Satz an auf die Zufallsvariablen $X_1, \ldots, X_n$, definiert durch

$$X_i := \begin{cases} 1 & \textit{falls der i-te Versuch ein Erfolg ist}, \ 1 \le i \le 1 \\ 0 & \textit{sonst} \end{cases}$$

Es gilt dann $\mathrm{E}(X_i) \overset{\text{def}}{=} 1 \cdot p + 0 \cdot p = p$ und

$$\begin{aligned}\mathrm{Var}(X_i) &= \mathrm{E}(X_i^2) - (\mathrm{E}(X_i))^2 \\ &= \mathrm{E}(X_i) - (\mathrm{E}(X_i))^2 \\ &= p - p^2 = pq\end{aligned}$$

Beachtet man nun $X = X_1 + \cdots + X_n$, so folgt nach obigem Satz

$$\mathrm{E}(X) = \mathrm{E}(X_1 + \cdots + X_n) = \mathrm{E}(X_1) + \cdots + \mathrm{E}(X_n) = \underbrace{p + \cdots + p}_{n-mal} = np$$

Da die Zufallsvariablen $X_1, \ldots, X_n$ unabhängig sind, folgt

$$\mathrm{Var}(x) = \mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) = \underbrace{pq + \cdots + pq}_{n-mal} = npq$$

Damit haben wir zwei verschiedene Methoden zur Berechung des Erwartungswertes und der Varianz der Binomialverteilung gefunden.

1.10 Poisson-Verteilung

Die Poisson-Verteilung ist der Grenzfall der Binomialverteilung für den Fall seltener Ereignisse. In der Binomialverteilung lässt man $p \to 0$ und $n \to \infty$ streben derart, dass $np = \mu$ konstant ist.

$$\begin{aligned}
\binom{n}{k} \cdot p^k \cdot q^{n-k} &= \frac{n(n-1) \cdot \ldots \cdot (n-k+1)}{k!} \cdot \frac{p^k \cdot (1-p)^n}{(1-p)^k} \\
&= \frac{\mu^k \cdot n(n-1) \cdot \ldots \cdot (n-k+1)}{k! \cdot n^k} \cdot (1 - \frac{\mu}{n})^n \cdot \frac{1}{(1-p)^k} \\
&\to \frac{\mu^k}{k!} \cdot 1 \cdot \ldots \cdot 1 \cdot e^{-\mu} \cdot 1 = \frac{\mu^k}{k!} \cdot e^{-\mu}
\end{aligned}$$

Definition

Eine Zufallsvariable X, die gemäß

$$P(X = k) = \frac{\mu^k}{k!} \cdot e^{-\mu},\ k \in \{0, 1, 2, \ldots\} =: \mathbb{N}_0$$

verteilt ist, heißt Poisson-verteilt mit Parameter μ.
Die zugehörige Verteilung wird ofr mit $Pn(\mu)$ bezeichnet. Die erzeugende Funktion der Poisson-Verteilung ist (Übung):

$$f(s) = e^{\mu \cdot (s-1)}$$

Wegen $np = \mu$ und $np(1-p) \to \mu$ folgt $\mathrm{E}(x) = \mathrm{Var}(X) = \mu$, d.h. die Varianz ist gleich dem Erwartungswert.
Für große Werte von μ kann man die Poisson-Verteilung durch eine Normalverteilung (siehe 1.18) approximieren.

1.11 Geometrische Verteilung

Wie bei der Binomialverteilung wird wieder ein Einzelexperiment mit zwei möglichen Ausgängen, Erfolg und Misserfolg, betrachtet.
Die Wahrscheinlichkeit eines Erfolgs sei $p \in (0,1)$.
Die Wahrscheinlichkeit eines Misserfolges ist dann $q := 1 - p$.
Dieses Experiment wird so oft wiederholt, bis erstmals ein Erfolg eintritt. Die Zufallsvariable X zählt die Misserfolge. Die Wahrscheinlichkeit dafür, k Misserfolge zu haben, bevor der erste Erfolg eintritt, ist:

$$P(X = k) = p \cdot q^k,\ k \in \mathbb{N}_0\ (*)$$

Die Summe dieser Wahrscheinlichkeit ist:

$$\sum_{k \in \mathbb{N}_0}^{\infty} p \cdot q \cdot k = \frac{p}{1-q} = 1$$

Das Ereignis, dass niemals Erfolg eintritt, hat also Wahrscheinlichkeit 0.

Definition

Eine Zufallsvariable X, die gemäß (*) verteilt ist, heißt geometrisch verteilt mit Parameter p.
Die zugehörige Verteilung wird kurz mit $G(p)$ bezeichnet.
Die erzeugende Funktion der geometrischen Verteilung ist

$$f(s) = \sum_{k \in \mathbb{N}_0} p \cdot q^k \cdot s^k = \frac{p}{1 - q \cdot s}$$

Damit folgt

$$f'(s) = \frac{p \cdot q}{(1 - q \cdot s)^2}$$

und

$$f''(s) = \frac{2 \cdot p \cdot q^2}{(1 - q \cdot s)^3}$$

Insbesondere gilt

$$\mathrm{E}(X) = f'(1) = \frac{q}{P}$$

und

$$\begin{aligned}\mathrm{Var}(X) &= f''(1) + f'(1) - (f'(1))^2 \\ &= \frac{2q^2}{P^2} + \frac{q}{P} - \left(\frac{q}{P}\right)^2 \\ &= \frac{q}{P^2} \cdot (2q + P - 1) \\ &= \frac{q}{P^2}\end{aligned}$$

1.12 Zufallsvariablen mit Dichten

Definition

Eine Funktion $f\colon \mathbb{R} \to \mathbb{R}$ heißt Dichte (oder auch Wahrscheinlichkeitsdichte oder Verteilungsdichte), falls $f(x) \geq 0,\ \forall x \in \mathbb{R}$ und falls f integrierbar ist mit

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$$

Sei nun eine solche Dichte f gegeben. Falls X eine Zufallsvariable ist, deren Verteilungsfunktion

$$F(x) \coloneqq P(X \leq x),\ x \in \mathbb{R},$$

sich gemäß

$$F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t$$

berechnen lässt, so sagt man, dass f eine Dichte der Zufallsvariablen X (bzw. der Verteilung von X) ist.
In diesem Fall ist F also insbesondere eine Stammfunktion von f und es gilt (Hauptsatz der Differential- und Integralrechnung)

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a) = \int_{a}^{b} f(t)\mathrm{d}t$$

In Analogie zu diskreten Zufallsvariablen werden nun Erwartungswert und Varianz definiert:

Definition

Sei X eine Zufallsvariable mit Dichte f. Dann heißt die Zahl

$$\mathrm{E}(X) \coloneqq \int_{-\infty}^{\infty} x \cdot f(x)\mathrm{d}x$$

der Erwartungswert von X und die Zahl

$$\begin{aligned}\mathrm{Var}(X) &:= \int_{-\infty}^{\infty} (x - \mathrm{E}(X))^2 \cdot f(x)\mathrm{d}x \\ &= \int_{-\infty}^{\infty} x^2 \cdot f(x)\mathrm{d}x - (\mathrm{E}(X))^2\end{aligned}$$

Varianz von X.
Es gilt (wie bei diskreten Zufallsvariablen) die Formel

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2$$

Diese Definition ist nur sinnvoll, wenn die Integrale existieren. Wenn die Dichte f für $|\, x \,| \to \infty$ nicht rasch genug abklingt, kann es sein, dass die Varianz oder sogar der Erwartungswert nicht existieren.

1.13 Gleichverteilung

Eine Zufallsvariable X heißt gleichverteilt auf dem Intervall $[a, b]$, $a, b \in \mathbb{R}$, $a < b$, wenn

$$f(x) := \begin{cases} \frac{1}{b-a} & \textit{für } a \le x \le b \\ 0 & \textit{sonst} \end{cases}$$

eine Dichte von X ist.
Die Gleichverteilung auf $[a, b]$ wird kurz mit $U(a, b)$ bezeichnet. Es gilt dann

$$\begin{aligned}\mathrm{E}(X) &= \int_{-\infty}^{\infty} x \cdot f(x)\mathrm{d}x \\ &= \frac{1}{b-a} \int_{a}^{b} x \,\mathrm{d}x \\ &= \frac{1}{b-a} \cdot \frac{b^2 - a^2}{2} \\ &= \frac{a+b}{2}\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(X) &= \frac{1}{b-a}\int_a^b x^2 \mathrm{d}x - (\mathrm{E}(X))^2 \\ &= \frac{1}{b-a}\cdot\frac{b^3-a^3}{3} - \frac{(a+b)^2}{4} \\ &= \frac{a^2+ab+b^2}{3} - \frac{a^2+2ab+b^2}{4} \\ &= \frac{(b-a)^2}{12}\end{aligned}$$

1.14 Exponentialverteilung

Definition

Eine Zufallsvariable X heißt exponentialverteilt mit Parameter $\alpha > 0$, wenn sie die Verteilungsfunktion

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

besitzt.
Zusammenhang Dichte (f) $\leftrightarrow$ Verteilungsfunktion (F)

$$F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t$$

Offenbar ist dann

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

eine zugehörige Dichte.
Die Exponentialverteilung wird kurz mit $\mathrm{Exp}(\alpha)$ bezeichnet.
In Anwendungen wird z.B. die Lebensdauer X eines Individuums meist mittels einer Exponentialverteilung modelliert, da sie die Eigenschaft (leicht nachzurechnen)

$$P(X > x + t \mid X > t) = P(X > x),\ \forall x, t \in [0, \infty) (*)$$

besitzt, d.h. die Verteilung der verbleibenden Lebensdauer eines Individuums hängt nicht von seinem Alter ab. Man kann umgekehrt auch zeigen, dass jede Zufallsvariable X mit stetiger Verteilungsfunktion F, die (*), d.h.

$$\frac{1 - F(x+t)}{1 - F(t)} = 1 - F(x),\ \forall x, t \in [0, \infty)$$

erfüllt, exponentialverteilt ist (und zwar mit Parameter $\alpha := -\log P(X > 1)$).
In den Übungen wird gezeigt, dass eine $\mathrm{Exp}(\alpha)$-verteilte Zufallsvariable X Erwartungswert $\mathrm{E}(X) = \frac{1}{\alpha}$ und die Varianz $\mathrm{Var}(X) = \frac{1}{\alpha^2}$ besitzt.

1.15 Normalverteilung

Die Normalverteilung ist eine der wichtigsten Verteilungen überhaupt.
Ihre Dichte ist, als Graph (Gauß'sche Glockenkurve) und als Formel auf dem alten 10 DM Schein dargestellt, neben dem Portrait von Gauß. Man kann sich wundern, wie denn eine so spezielle Funktion in der Natur eine so wichtige Rolle spielen sollte (neben der Exponentialfunktion und den trigonometrischen Funktionen).
Genau wie einfache Annahmen über periodische Vorgänge zwangsläufig auf $\sin$, $\cos$ und Annahmen über proportionalen Zuwachs $(f' = af)$ auf die Exponentialfunktion führen, so führen Annahmen über den Einfluss von unabhängigen Zufallsvariablen $X_1, X_2, \ldots$ auf deren Summen $S_n := X_1 + \ldots + X_n$ bei großem n auf die Normalverteilung.
Solche Konvergenzaussagen werden als das Prinzip des „zentralen Grenzwertsatzes" bezeichnet, auf das später genauer eingegangen wird.

Definition

Eine Zufallsvariable X heißt normalverteilt mit den Parametern $\mu \in \mathbb{R}$ und $\sigma^2 > 0$, wenn

$$\varphi_{\mu,\sigma^2}(x) := \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ x \in \mathbb{R}$$

eine Dichte von X ist.
Man sagt dann kurz, X besitzt die Verteilung $\mathrm{N}(\mu, \sigma^2)$ oder auch X ist $\mathrm{N}(\mu, \sigma^2)$-verteilt.
Die Funktion φ_{μ,σ^2} ist wirklich eine Dichte, denn man kann nachrechnen, dass das Integral

$$\int_{-\infty}^{\infty} \varphi_{\mu,\sigma^2}(x)\mathrm{d}x$$

gleich 1 ist.
Die Dichte hängt von zwei Paramtern μ und σ^2 ab. Diese sind so gewählt, dass sie gleich dem Erwartungswert und der Varianz sind.

Satz

Hat die Zufallsvariable X eine $\mathrm{N}(\mu, \sigma^2)$-Verteilung, so gilt

$$\mathrm{E}(X) = \mu$$
$$\text{und}$$
$$\mathrm{Var}(X) = \sigma^2$$

Hat die Zufallsvariable X eine $\mathrm{N}(\mu, \sigma^2)$-Verteilung, so hat die linear transformierte Zufallsvariable $Y := aX + b$ eine $\mathrm{N}(a\mu + b, a^2\sigma^2)$-Verteilung.
Daraus folgt, dass die Standardisierte $X^* = \frac{X-\mu}{\sigma}$ eine $\mathrm{N}(0, 1)$-Verteilung besitzt.

Die $N(0,1)$-Verteilung wird daher Standard-Normalverteilung genannt. Die Verteilungsfunktion der $N(\mu,\sigma^2)$-Verteilung, also die Funktion

$$\begin{aligned}\Phi_{\mu,\sigma^2}(x) &:= \int_{-\infty}^{x} \varphi_{\mu,\sigma^2}(y)\mathrm{d}y \\ &= \frac{1}{\sigma\sqrt{2\pi}} \cdot \int_{-\infty}^{x} e^{-\frac{(y-\mu)^2}{2\sigma^2}}\mathrm{d}y \\ &\overset{\frac{y-\mu}{\sigma}=t}{=} \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-\frac{t^2}{2}}\mathrm{d}t \\ &= \int_{-\infty}^{\frac{x-\mu}{\sigma}} \varphi_{0,1}(t)\mathrm{d}t \\ &= \Phi_{0,1}\left(\frac{x-\mu}{\sigma}\right)\end{aligned}$$

lässt sich berechnen, sofern man die Funktion $\Phi := \Phi_{0,1}$, also die Verteilungsfunktion der Standard-Normalverteilung kennt.
Die Werte von Φ sind tabelliert. Einige Werte sind:

x	$\Phi(x)$
0	0,5
0,5	0,6915
1	0,8413
1,5	0,9332
1,6449	0,95
1,96	0,9750
2	0,9772
2,3263	0,99
2,5	0,9938
2,5758	0,9950
3	0,9987
3,0902	0,999

Wir wenden auf die Normalverteilung später bei der Behandlung des „Zentralen Grenzwertsatzes“ zurückkommen.

1.16 Tschebyscheff'sche Ungleichung

Sei Y eine Zufallsvariable mit Dichte g. Dann gilt für jedes $\varepsilon > 0$

$$\begin{aligned}\varepsilon^2 \cdot P(|Y| \geq \varepsilon) &= \int\limits_{\{|y| \geq \varepsilon\}} \varepsilon^2 \cdot g(y)\, \mathrm{d}y \\ &\leq \int\limits_{\{|y| \geq \varepsilon\}} y^2 \cdot g(y)\, \mathrm{d}y \\ &\leq \int y^2 \cdot g(y)\, \mathrm{d}y = \mathrm{E}(Y^2), \text{ also} \\ P(|Y| \geq \varepsilon) &\leq \frac{\mathrm{E}(Y^2)}{\varepsilon^2}, \ \forall \varepsilon > 0\end{aligned}$$

Man kann zeigen, dass diese Ungleichung für beliebige Zufallsvariablen, also auch für diskrete Zufallsvariablen gilt.
Sei nun X eine beliebige Zufallsvariable. Wir nehmen an, dass Erwartungswert und Varianz von X existieren. Wendet man obige Ungleichung auf $Y := X - \mathrm{E}(X)$ an, so folgt

$$P(|X - \mathrm{E}(X)| \geq \varepsilon) \leq \frac{\mathrm{Var}(X)}{\varepsilon^2}, \ \forall \varepsilon > 0$$

Dies ist die sogenannte Tschebyscheff'sche Ungleichung. Sie gibt eine obere Schranke für die Wahrscheinlichkeit an, dass die Zufallsvariable X um mindestens ε von ihrem Erwartungswert $\mathrm{E}(X)$ abweicht.

Beispiel

Wählt man z.B.

$$\varepsilon = 2 \cdot \sigma(x) = 2 \cdot \sqrt{\mathrm{Var}(X)},$$

so folgt

$$P(|X - \mathrm{E}(X)| \geq 2 \cdot \sigma(X)) \leq \frac{1}{4}$$

d.h. die Wahrscheinlichkeit, dass eine beliebige Zufallsvariable X um mindestens das Zweifache ihrer Standardabweichung von ihrem Erwartungswert abweicht, ist höchstens gleich $\frac{1}{4}$.
Man beachte, dass die Tschebyscheff Ungleichung für jede Zufallsvariable gilt (sofern $\mathrm{E}(X)$ und $\mathrm{Var}(X)$ existieren, d.h. sofern $\mathrm{E}(X^2) < \infty$). Für spezielle Zufallsvariablen, z.B. für normalverteilte Zufallsvariablen, kann man schärfere Ungleichungen herleiten.

1.17 Gesetz der großen Zahlen

Seien $X_1, X_2, \ldots$ unabhängige und identisch verteilte (i.i.d. = independent and identically distributed) Zufallsvariablen, d.h. jedes X_i besitze dieselbe Verteilung. Wir nehmen an, dass der Erwartungswert $\mu = \mathrm{E}(X_1)$ und die Varianz $\sigma^2 := \mathrm{Var}(X_1)$ existieren. Für eine natürliche Zahl $n \in \mathbb{N}$ betrachte die Summe $S_n = X_1 + \ldots + X_n$.

Es gilt

$$\mathrm{E}(S_n) = n \cdot \mu$$

und wegen der Unabhängigkeit der X_i gilt

$$\mathrm{Var}(S_n) = n \cdot \sigma^2$$

Wendet man die Tschebyscheff Ungleichung auf das arithmetische Mittel $\overline{X} := \frac{S_n}{n}$ der Zufallsvariablen $X_1, \ldots, X_n$ an, so folgt

$$P(|\overline{X} - \mu| \geq \varepsilon) \leq \frac{\mathrm{Var}(S_n)}{n^2 \cdot \varepsilon^2} = \frac{\sigma^2}{n \cdot \varepsilon^2}\ ,\ \forall \varepsilon > 0$$

Grenzübergang $n \to \infty$ liefert

$$\lim_{n\to\infty} P(|\overline{X} - \mu| \geq \varepsilon) = 0,\ \forall \varepsilon > 0$$

Dies ist das sogenannte (schwache) Gesetz der großen Zahlen.
Es sagt aus, dass die Wahrscheinlichkeit, dass das arithmetische Mittel $\overline{X} = \frac{S_n}{n} = \frac{X_1+\ldots+X_n}{n}$ der Zufallsvariablen $X_1, \ldots, X_n$ um mindestens ε von seinem Erwartungswert abweicht, mit wachsendem n beliebig klein wird.
Mit etwas mehr Aufwand kann man sogar eine stärkere Eigenschaft zeigen, nämlich dass die Zufallsvariable $\overline{X}\colon \Omega \to \mathbb{R}$ mit $n \to \infty$ fast sicher gegen μ konvergiert, d.h. es gibt eine Menge $A \subseteq \Omega$ mit $P(A) = 1$ und

$$\lim_{n\to\infty} \overline{X}(\omega) = \lim_{n\to\infty} \frac{X_1(\omega) + \ldots + X_n(\omega)}{n} = \mu,\ \forall \omega \in A$$

Dies nennt man das starke Gesetz der großen Zahlen.

1.18 Zentraler Grenzwertsatz

Seien $X_1, X_2, \ldots$ wie im vorigen Abschnitt. Wir haben gesehen, dass die Zufallsvariable $\overline{X} := \frac{X_1+\ldots+X_n}{n}$ für großes n in gewissem Sinne nur wenig von $\mu := \mathrm{E}(X_1)$ abweicht.
Man möchte nun wissen, wie $\overline{X}$ um μ herum verteilt ist. Diese Frage beantwortet einer der wichtigsten Sätze der Stochastik, der deshalb der Zentrale Grenzwertsatz genannt wird.

Satz (Zentraler Grenzwertsatz)

Seien $X_1, X_2, \ldots$ unabhängig, identisch verteilte (i.i.d.) Zufallsvariablen mit Erwartungswert $\mu\colon \mathrm{E}(X_1) \in \mathbb{R}$ und $0 < \sigma^2 := \mathrm{Var}(X_1) < \infty$
Für $n \in \mathbb{N}$ betrachte die Summe $S_n := X_1 + \ldots + X_n$ bzw. deren Standardisierte

$$S_n^* = \frac{X_1 + \ldots + X_n - n \cdot \mu}{\sigma \cdot \sqrt{n}}$$

Dann gilt für jedes $x \in \mathbb{R}$

$$\lim_{n\to\infty} P(S_n^* \leq x) = \Phi(x)$$

d.h. die Verteilungsfunktion der Standardisierten der Summe S_n konvergiert mit $n \to \infty$ punktweise gegen die Verteilungsfunktion der Standardnormalverteilung.
Man kann sogar zeigen, dass diese Konvergenz gleichmäßig in x ist, d.h.

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |P(S_n^* \leq x) - \Phi(x)| = 0$$

Daher lassen sich Verteilungen von Summen $S_n = X_1 + \cdots + X_n$ unabhängiger, identisch verteilter Zufallsvariablen durch Normalverteilungen approximieren, falls n hinreichend groß ist.

Beispiel

Wir wissen bereits, dass sich jede $B(n,p)$-verteilte (binomial-verteilte) Zufallsvariable X als Summe $X = X_1 + \cdots + X_n$ schreiben lässt, wobei $X_n, n \in \mathbb{N}$, unabhängig und identisch verteilt sind mit $\mathrm{E}(X_) = p$ und $\mathrm{Var}(X_n) = pq$, wobei $q := 1 - p$. Also lässt sich nach dem zentralen Grenzwertsatz die Binomialverteilung durch eine Normalverteilung approximieren. Da die Binomialverteilung den Erwartungswert np und die Varianz npq hat, gilt

$$\lim_{n\to\infty} P\left(\frac{X - np}{\sqrt{npq}} \leq x\right) = \Phi(x), \ \forall x \in \mathbb{R}$$

Wegen der gleichmäßigen Konvergenz, gilt somit für eine $B(n,p)$-verteilte Zufallsvariable X und für ganzzahlige Werte a, b mit $a \leq b$ die Approximation

$$P(a \leq X \leq b) = \sum_{k=a}^{b} \binom{n}{k} \cdot p^k \cdot q^{n-k} \approx \Phi\left(\frac{b - np}{\sqrt{npq}}\right) - \Phi\left(\frac{a - 1 - np}{\sqrt{npq}}\right)$$

sofern n hinreichend groß ist (Faustregel: $npq \geq 9$).
Genauere Untersuchungen zeigen, dass sich diese Approximation noch durch folgende Stetigkeitskorrektur (auch Diskretheitskorrektur genannt) verbessern lässt:

$$P(a \leq X \leq b) = \sum_{k=a}^{b} \binom{n}{k} \cdot p^k \cdot q^{n-k} \approx \Phi\left(\frac{b + \frac{1}{2} - np}{\sqrt{npq}}\right) - \Phi\left(\frac{a - \frac{1}{2} - np}{\sqrt{npq}}\right)$$

Man schreibt auch kurz $B(n,p) \approx N(np, npq)$, falls n hinreichend groß ist.
Analog lässt sich auch die Poisson-Verteilung approximieren: $P_n(\mu) \approx N(\mu, \mu)$, falls μ hinreichend groß ist.

Allgemein

Ist X approximativ normalverteilt und nimmt X nur ganzzahlige Werte an, so gilt für ganzzahlige a, b mit $a \leq b$ die Approximation

$$P(a \leq X \leq b) \approx \Phi\left(\frac{b + \frac{1}{2} - \mathrm{E}(X)}{\sqrt{\sigma(X)}}\right) - \Phi\left(\frac{a - \frac{1}{2} - \mathrm{E}(X)}{\sqrt{\sigma(X)}}\right)$$

1.19 Mehr zur Unabhängigkeit

In 1.5 wurde die Unabhängigkeit von Ereignissen definiert und am Ende von 1.6 die Unabhängigkeit von Zufallsvariablen auf der Unabhängigkeit von Ereignissen aufgebaut. Zur Wiederholung:
Zwei Zufallsvariablen X und Y (auf dem selben Grundraum) heißen unabhängig, falls

$$P(\{X \leq x\} \cap \{Y \leq y\}) = P(X \leq x) \cdot P(Y \leq y),\ \forall\, x, y \in \mathbb{R}$$

Die Unabhänigkeit von mehr als zwei Zufallsvariablen wird analog definiert. Wie wir bereits gesehen haben, gelten viele Aussagen der Stochastik (nur) unter der Voraussetzung der Unabhängigkeit der beteiligten Zufallsvariablen.

Definition

Seien X, Y zwei Zufallsvariablen. Dann heißt die Funktion $F\colon \mathbb{R}^2 \to \mathbb{R}$ definiert durch

$$F(x, y)\colon P(X \leq x, Y \leq y) \coloneqq P(\{X \leq x\} \cap \{Y \leq y\})$$

die gemeinsame Verteilungsfunktion von X und Y oder auch die Verteilungsfunktion des Paares (X, Y). Kennt man die gemeinsame Verteilungsfunktion F von (X, Y), so kennt man auch die Verteilungsfunktion F_1 von X und die Verteilungsfunktion F_2 von Y. Es gilt nämlich

$$F_1(x) = \lim_{y \to \infty} F(x, y) \text{ und } F_2(y) = \lim_{x \to \infty} F(x, y)$$

Das bedeutet: Kennt man die gemeinsame Verteilung von X und Y, so kennt man auch die Verteilung von X und die Verteilung von Y. Die Verteilung von X und die Verteilung von Y werden auch die beiden Randverteilungen des Paares (X, Y) genannt.
Vorsicht: Die Umkehrung gilt nicht!
Die Kenntnis der Verteilung(sfunktion F_1) von X und der Verteilung(sfunktion F_2) von Y genügt nicht, um die gemeinsame Verteilung(sfunktion F) von (X, Y) zu bekommen.
Allerdings gibt es spezielle Fälle, wo die Umkehrung gilt.
Die Zufallsvariablen X und Y sind z.B. genau dann unabhängig, wenn

$$F(x, y) = F_1(x) \cdot F_2(y),\ \forall\, x, y \in \mathbb{R} (*)$$

Bei Unabhängigkeit lässt sich also die gemeinsame Verteilungsfunktion F ganz einfach durch das Produkt von F_1 und F_2 berechnen. Zugleich kann man, sofern man die gemeinsame Verteilungsfunktion F kennt, mit Hilfe von (*) nachprüfen, ob zwei Zufallsvariablen X und Y unabhängig sind.

Situation 1

X und Y diskret. Die Zufallsvariable X nimmt dann endlich viele Werte $x_1, \ldots, x_n$ oder abzählber unendlich viele Werte $x_1, x_2, \ldots$ an.
Das sind Werte x_i ($i \in I$ mit $I := \{1, \ldots, n\}$ oder $I := \mathbb{N}$).
Analog seien y_j ($j \in J$ mit $J := \{1, \ldots, m\}$ oder $J := \mathbb{N}$) die Werte, die die Zufallsvariable Y annimmt.
Dann wird die gemeinsame Verteilung der beiden Zufallsvariablen durch die Wahrscheinlichkeiten

$$p_{ij} := P(X = x_i, Y = y_j) := P(\{X = x_i\} \cap \{Y = y_j\}),\ i \in I, j \in J$$

beschrieben.
Es gilt $p_{ij} \geq 0,\ \forall\, i \in I, j \in J,$ sowie $\sum\limits_{i \in I, j \in J} p_{ij} = 1$
Die (Rand-)Verteilung von X wird durch die Zahlen

$$p_{ij} := P(X = x_i) = \sum_{j \in J} P(X = x_i, Y = y_j) = \sum_{j \in J} p_{ij},\ i \in I$$

beschrieben. Analog ist die Verteilung von Y gegeben durch

$$q_j := P(Y = y_j) = \sum_{i \in I} p_{ij},\ j \in J$$

Man sieht leicht ein, dass X und Y genau dann unabhängig sind, wenn

$$p_{ij} = p_i \cdot q_j,\ \forall\, i \in I, j \in J$$

Übersetzt in die Sprache der linearen Algebra bedeutet dies, dass X und Y genau dann unabhängig sind, wenn sich die Matrix

$$P = (p_{ij})\ i \in I, j \in J$$

als „dyadisches Produkt" $P = p^T q$ darstellen lässt, wobei die Zeilenvektoren p und q definiert sind durch

$$p := (p_i)\ i \in I \text{ und } q := (q_j)\ j \in J$$

Situation 2

Man sagt, dass zwei Zufallsvariablen X, Y eine gemeinsame Dichte $f\colon \mathbb{R}^2 \to \mathbb{R}$ besitzen, falls sich die gemeinsame Verteilungsfunktion F von (X, Y) schreiben lässt als

$$F(x,y) = \int_{-\infty}^{x} \left(\int_{-\infty}^{y} f(s,t)\,\mathrm{d}t \right) \mathrm{d}s,\ \forall\, x, y \in \mathbb{R}$$

Man prüft leicht nach, dass dann

$$f_1(x) \coloneqq \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y$$

eine Dichte von X und

$$f_2(y) \coloneqq \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x$$

eine Dichte von Y ist.
Man nennt f_1 und f_2 die beiden Rand-Dichten von f. Man kann zeigen, dass X und Y genau dann unabhängig sind, wenn die Gleichung

$$f(x,y) = f_1(x) \cdot f_2(y)$$

für „fast alle (x, y)" gilt. Dabei versteht man unter „fast alle (x, y)" alle $(x, y) \in \mathbb{R}^2$ ausgenommen die, die in einer Menge $N \subset \mathbb{R}^2$ mit Oberfläche $O(N) = 0$ liegen. Solche Mengen N nennt man Nullmengen. Jeder Graph einer Funktion ist z.B. eine solche Nullmenge.

1.20 Kovarianz und Korrelation

Der Erwartungswert ist ein lineares Funktional, d.h. es gilt für beliebige Zufallsvariablen (auf dem selben Wahrscheinlichkeitsraum)

$$\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y)$$

sofern die beteiligten Erwartungswerte existieren.
Für die Varianz gilt eine solche Aussage nicht. Man sieht

$$\begin{aligned}
\mathrm{Var}(X+Y) &= \mathrm{E}((X+Y)^2) - (\mathrm{E}(X+Y))^2 \\
&= \mathrm{E}(X^2 + 2XY + Y^2) - (\mathrm{E}(X) + \mathrm{E}(Y))^2 \\
&= \mathrm{E}(X^2) + 2 \cdot \mathrm{E}(XY) + \mathrm{E}(Y^2) - (\mathrm{E}(X))^2 - (2 \cdot \mathrm{E}(X) \cdot \mathrm{E}(Y)) - (\mathrm{E}(Y))^2 \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2 \cdot (\mathrm{E}(XY) - \mathrm{E}(X) \cdot \mathrm{E}(Y)) \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2 \cdot \mathrm{Cov}(X,Y)
\end{aligned}$$

wobei die sogenannte Kovarianz $\mathrm{Cov}(X,Y)$ von X und Y definiert ist durch

$$\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\cdot\mathrm{E}(Y)$$

Gilt $\mathrm{Cov}(X,Y) = 0 (> 0, < 0)$ so sagt man, dass X und Y unkorreliert (positiv korreliert, negativ korreliert) sind. Man kann zeigen, dass unabhängige Zufallsvariablen X, Y stets unkorreliert sind.
Die Umkehrung ist i.A. falsch.
Sind X und Y unkorreliert oder sogar unabhängig, so gilt

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

Um die Kovarianzen verschiedener Paare von Zufallsvariablen vergleichbar zu machen, betrachtet man oft nicht die Kovarianz von X und Y selbst, sondern diejenige der Standardisierten

$$X^* := \frac{X - \mathrm{E}(X)}{\sigma(X)} \text{ und } Y^* := \frac{Y - \mathrm{E}(Y)}{\sigma(Y)}$$

also die Zahl

$$\rho(X,Y) := \mathrm{Cov}(X^*, Y^*) = \frac{\mathrm{Cov}(X,Y)}{\sigma(X)\cdot\sigma(Y)}$$

sofern $\sigma(X)$ und $\sigma(Y)$ existieren und beide von Null verschieden sind. Die Zahl $\rho(X,Y)$ heißt der Korrelationskoeffizient von X und Y. Mann kann zeigen, dass stets $-1 \le \rho(X,Y) \le 1$ gilt.
Genau dann gilt $|\rho(X,Y)| = 1$, wenn Konstanten $a, b \in \mathbb{R}$ mit $a \neq 0$ existieren mit $P(Y = a\cdot X + b) = 1$.
Dabei stimmt das Vorzeichen von a mit dem von $\rho(X,Y)$ überein.

Beispiel

Ein Kasten enthalte s schwarze und w weiße Kugeln.
Es werden zufällig $n (\le s+w)$ Kugeln ohne Zurücklegen entnommen. Für $i \in \{1,\dots,n\}$ sei

$$X_i := \begin{cases} 1 & falls die i-te gezogene Kugel schwarz ist \\ 0 & sonst \end{cases}$$

Dann kann man zeigen, dass die Verteilung von X_i nicht von i abhängt, d.h. es gilt insbesondere

$$\mathrm{E}(X_i) = \mathrm{E}(X_1) = 1\cdot P(X_1 = 1) = \frac{s}{s+w}$$

und

$$\mathrm{Var}(X_i) = \mathrm{Var}(X_1) = P(X_1 = 1)\cdot P(X_1 = 0) = \frac{s}{s+w}\cdot\frac{w}{s+w} = \frac{s\cdot w}{(s+w)^2}$$

Ähnlich gilt für $i \neq j$

$$\begin{aligned}\mathrm{E}(X_i \cdot X_j) = \mathrm{E}(X_1 \cdot X_2) &= 1 \cdot P(X_1 \cdot X_2 = 1) = P(X_1 = 1 = X_2) \\ &= P(X_1 = 1) \cdot P(X_2 = 1 \mid X_1 = 1) \\ &= \frac{s}{s+w} \cdot \frac{s-1}{s+w-1}\end{aligned}$$

Für die Kovarianz von X_i und X_j ($i \neq j$) gilt somit

$$\begin{aligned}\mathrm{Cov}(X_i, X_j) &= \mathrm{E}(X_i \cdot X_j) - \mathrm{E}(X_i) \cdot \mathrm{E}[X_j] \\ &= \frac{s(s-1)}{(s+w)(s+w-1)} - \frac{s^2}{(s+w)^2} \\ &= \frac{s}{s+w}\left(\frac{s-1}{s+w-1} - \frac{s}{s+w}\right) \\ &= \frac{s}{s+w}\frac{(s-1)(s+w) - s(s+w-1)}{(s+w)(s+w-1)} \\ &= -\frac{s \cdot w}{(s+w)^2(s+w-1)}\end{aligned}$$

Division durch

$$\sigma(X_i) \cdot \sigma(X_j) = \mathrm{Var}(X_1) = \frac{s \cdot w}{(s+w)^2}$$

liefert den Korrelationseffekt

$$\rho(X_i, X_j) = -\frac{1}{s+w-1} < 0$$

d.h. für $i \neq j$ sind die Zufallsvariablen X_i und X_j negativ korreliert.

2 Statistik

Die Statistik ist ein Konzept, das auf die Wahrscheinlichkeitstheorie aufbaut. Man geht davon aus, dass es eine Grundgesamtheit gibt, die sehr groß und daher nicht komplett zugänglich ist. Deren Struktur soll durch Daten (Stichprobe) aufgeklärt werden. Da die Stichprobe i.A. relativ klein ist, kann davon ausgegangen werden, dass die Stichprobe die Grundgesamtheit nicht beeinflusst.
Während die Wahrscheinlichkeitstheorie ohne jegliche Daten auskommt, ist es ein wesentliches Merkmal der Statistik, dass diese sehr stark mit Daten arbeitet.

2.1 Statistisches Modell und Stichprobe

Die Grundgesamtheit wird durch ein Modell in Form einer reellen Zufallsvariablen beschrieben. In der parametrischen Statistik wird angenommen, das die Verteilung P_θ dieser Zufallsvariablen durch einen Parameter θ beschrieben wird, wobei für θ alle Werte aus einer geeigneten Parametermenge Θ zugelassen wird.

Das statistische Modell $\mathcal{P}$ ist also von der Form

$$\mathcal{P} := \{P_\theta \mid \theta \in \Theta\}$$

Bezüglich dieses statistischen Modells ist der Parameter θ (oder Funktionen von θ) zu schätzen, Hypothesen zu testen usw.
Dazu werden Daten $x_1, \ldots, x_n$ erhoben. Man sagt auch, es wird eine Stichprobe vom Umfang n gezogen. Man geht davon aus, dass die Daten $x_1, \ldots, x_n$ die Realisierung von unabhängigen Zufallsvariablen $X_1, \ldots, X_n$ sind, die alle die Verteilung P_θ besitzen für ein geeignetes $\theta \in \Theta$. Aus der Stichprobe $\underline{x} := (x_1, \ldots, x_n)$ berechnet man Kennzahlen, von denen man annehmen kann, dass sie nicht nur für die Stichprobe, sondern sogar für die Grundgesamtheit charakteristisch sind, weitgehend unabhängig davon, welche Struktur die Grundgesamtheit hat. Die wichtigsten dieser Kennzahlen sind der Stichprobenmittelwert

$$\overline{x} := \frac{1}{n} \sum_{i=1}^{n} x_i \in \mathbb{R}$$

und die (unverzerrte) Stichprobenvarianz

$$s^2 := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

Diese beiden Zahlen sind Realisierungen (Auswertungen in einem Punkt $\omega \in \Omega$) der Zufallsvariablen

$$\overline{X} := \frac{1}{n} \sum_{i=1}^{n} X_i$$

und

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

Man sollte sich stets im Klaren sein, ob man von den - auf den konkreten Daten basierenden - reellen Zahlen $\overline{x}$ und s^2 spricht, oder von den auf dem statistischen Modell beruhenden Zufallsvariablen $\overline{X}$ und S^2. Offenbar gilt

$$\mathrm{E}(\overline{X}) = \mathrm{E}(X_1)$$

und

$$\mathrm{Var}(\overline{X}) = \frac{\mathrm{Var}(X_1)}{n}$$

Die letzte Gleichung besagt, dass durch Mitteln die Variablilität geringer wird. Es wird sich als sinnvoll erweisen, die Zufallsvariablen $X_1, \ldots, X_n$ mittels

$$\underline{X}(\omega) := (X_1(\omega), \ldots, X_n(\omega))$$

zu einer $\mathbb{R}^n$-wertigen Zufallsvariable $\underline{X}\colon \Omega \to \mathbb{R}^n$ zusammenfassen. Man schreibt auch kurz

$$\underline{X} := (X_1, \ldots, X_n)$$

und sowohl $\underline{x}$ als auch $\underline{X}$ wird Stichprobe vom Umfang n genannt.

2.2 Schätzen von Parametern

Gegeben sie ein statistisches Modell

$$\mathcal{P} := \{P_\theta \mid \theta \in \Theta\}$$

wobei angenommen wird, dass dieses Modell (approximativ) die Grundgesamtheit beschreibt. D.h., dass P_θ für ein geeignetes $\theta \in \Theta$ (approximativ) die Verteilung der Grundgesamtheit ist. Später wird behandelt wie eine solche Hypothese (z.B. approximative Normalverteilung) geprüft werden kann (χ^2-Test).
Mit Hilfe von Daten (Stichprobe) $x_1, \ldots, x_n$ soll der Parameter θ, oder allgemeiner eine Funktion $q := f(\theta)$ von θ, geschätzt werden. Natürlich ist jede Zahl eine solche Schätzung.

Definition

Gegeben sei eine Funktion $T\colon \mathbb{R}^n \to \mathbb{R}$
Dann heißt die Zufallsvariable

$$U_n := T(\underline{X}) := T \circ \underline{X}\colon \Omega \to \mathbb{R}$$

ein Schätzer für q.
Die Zahl $T(\underline{x})$ nennt man den Schätzwert für q, der auf der Stichprobe $\underline{x} = (x_1, \ldots, x_n)$ beruht.
Diese Definition besagt lediglich, dass jede Zahl eine Schätzung für q ist. Die Schätzung soll nun aber in gewissem Sinn gut sein. Was heißt nun „gut“? Dafür gibt es viele Ansätze:

Definition

a) Ein Schätzer U_n heißt erwartungstreu (oder auch unverzerrt) für q, falls

$$\mathrm{E}(U_n) = q = f(\theta),\ \forall \theta$$

b) Eine Folge $(U_n)_{n\in\mathbb{N}}$ heißt konsistent für q, falls

$$\lim_{n\to\infty} P(|U_n - q| \geq \varepsilon) = 0, \ \forall \varepsilon > 0$$

Man sagt dann auch, dass U_n mit $n \to \infty$ stochastisch gegen q konvergiert.

c) Konvergiert U_n mit $n \to \infty$ sogar fast sicher gegen q, d.h. gilt

$$\lim_{n\to\infty} U_n(\omega) = q, \ \forall \omega \in A$$

wobei A ein Ereignis mit $P(A) = 1$ ist, so heißt die Folge $(U_n)_{n\in\mathbb{N}}$ stark konsistent für q.

Beispiel

Verfälschte Münze
Die Zahl $p := P(\text{„Kopf“}) \in [0,1]$ ist ein unbekannter Parameter.
Wie schätzt man diesen sinnvoll?
Sieht man von dem unwahrscheinlichen Fall ab, dass die Münze auch auf der Kante stehen bleiben kann, so kann das Werfen der Münze durch eine $B(1,p)$-verteilte Zufallsvariable beschreiben werden, wobei der Ausgänge 1 für „Kopf“ und 0 für „Zahl“ stehen. Das statistische Modell lautet also

$$\mathcal{P} := \{B(1,p) \mid p \in [0,1]\}$$

Man geht nun ganz naiv vor. Man wirft die Münze n-mal und notiert sich für jeden Wurf, ob Kopf oder Zahl gefallen ist.

$$x_i := \begin{cases} 1 & \text{falls die Münze im } i\text{-ten Wurf Kopf zeigt} \\ 0 & \textit{sonst} \end{cases}$$

Dies ergibt die Stichprobe $\underline{x} := (x_1, \ldots, x_n)$, wobei diese als Realisierung von Zufallsvariablen $X_1, \ldots, X_n$ angesehen wird, die alle $B(1,p)$-verteilt sind.
Intuitiv wird man den Stichprobenmittelwert

$$T(\underline{x}) = \overline{x} := \frac{x_1, \ldots, x_n}{n} = \frac{\text{Anzahl der Würfe, die Kopf zeigen}}{\text{Anzahl aller Würfe}}$$

als Schätzwert für p verwenden, aber auch jede andere Zahl wäre ein Schätzwert für p
Um einzusehen, dass der Stichprobenmittelwert in diesem Beispiel ein „guter“ Schätzwert für p ist, betrachte nun den zugehörigen Schätzer

$$U_n = T \circ \underline{X} = \frac{x_1, \ldots, x_n}{n} = \overline{X}$$

Wegen

$$\mathrm{E}(U_n) = \mathrm{E}(\overline{X}) = \mathrm{E}(X_1) = p$$

ist U_n ein erwartungstreuer Schätzer für p.
Da die Zufallsvariablen $X_1, X_2, \ldots$ i.i.d. sind, lässt sich auf diese direkt das Gesetz der großen Zahlen (1.17) anwenden. Dies liefert

$$\lim_{n\to\infty} p\left(\mid U_n - p \mid \geq \varepsilon\right) = 0,\ \forall \varepsilon > 0$$

d.h. die Folge $(U_n)_{n\in\mathbb{N}}$ ist konsistent für p.
Das starke Gesetz der großen Zahlen (1.17) liefert sogar die starke Konsistenz von $(U_n)_{n\in\mathbb{N}}$.
Mit Hilfe des zentralen Grenzwertsatzes (1.18) lässt sich die Wahrscheinlichkeit
$P(\mid U_n - p \mid \geq \varepsilon)$ sogar approximativ berechnen. Die Standardisierte

$$S_n^* := \frac{(X_1 + \ldots + X_n) - np}{\sqrt{npq}} = \frac{n\left(U_n - p\right)}{\sqrt{npq}} = \frac{\sqrt{n}}{\sqrt{pq}} \cdot \left(U_n - p\right)$$

ist nämlich für großes n approximativ standardnormalverteilt. Somit gilt

$$P\left(\mid U_n - p \mid \geq \varepsilon\right) = P\left(|S_n^*| \geq \frac{\varepsilon\sqrt{n}}{\sqrt{pq}}\right) \approx 2 - 2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{pq}}\right)$$

2.3 Maximum-Likelihood-Methode und Momentenmethode

Nachdem einige wünschenswerte Eigenschaften von Schätzen dargestellt wurden, erhebt sich die Frage, wie man für einen gegebenen Parameter θ gute Schätzer erhalten kann. Ein möglicher Weg ist ein Optimalitäts-Prinzip, das Maximum-Likelihood-Prinzip bezeichnet wird:
Man nimmt an, dass die beobachtete Zufallsvariable gemäß P_θ verteilt ist, wobei $\theta \in \Theta$ unbekannt, also zu schätzen ist. Dazu sammelt man Daten (Stichprobe) $\underline{x} = (x_1, \ldots, x_n)$, wobei jedes x_i als Realisierung einer Zufallsvariablen mit Verteilung P_θ aufgefasst wird.
Man schätzt nun θ so, d.h. man wählt einen Schätzwert $\hat{\theta} = T\left(\underline{x}\right)$ für θ derart, dass die Wahrscheinlichkeit für die tatsächlich beobachteten Daten $\underline{x}$ maximal wird. Dies heißt also, dass man die Maxima einer Funktion $L : \Theta \to \mathbb{R}$ zu suchen hat, die vom Parameter θ (und auch von dem gegebenen Daten $\underline{x}$) abhängt.
Die Frage ist nun, wie man L sinnvoll wählt. Falls P_θ eine diskrete Verteilung ist, dann wird sich für L die Funktion

$$L(\theta) \ := \ P_\theta(x_1) \ldots p_\theta(x_n)$$

eignen. Falls hingegen P_θ eine Dichte $f_\theta \colon \mathbb{R} \to \mathbb{R}$ besitzt, so wählt man

$$L(\theta) \ := \ f_\theta(x_1) \ldots f_\theta(x_n)$$

Die so definierte Funktion L heißt Likelihood-Funktion.
Es wird nochmals darauf hingewiesen, dass bei diesem Verfahren die Daten $x_1, \ldots, x_n$ gegeben - also fest - sind. Maximiert wird bzgl. der Variablen θ, d.h. $T\left(\underline{x}\right) = \hat{\theta}$, wobei $\hat{\theta}$ derart, dass

$$L\left(\hat{\theta}\right) = \max_{\theta\in\Theta} L(\theta)$$

Beispiel

Beobachtet werden n „Sterbedaten“ $x_1, \ldots, x_n$, wobei davon ausgegangen wird, dass die Lebenszeit durch eine exponential-verteilte Zufallsvariable mit Parameter $\alpha > 0$ modelliert wird.
Das statistische Modell ist also

$$\mathcal{P} = \{\exp(\alpha) \mid \alpha > 0\}$$

Da $f_\alpha(x) := \alpha e^{-\alpha x} (x \geq 0)$ eine Dichte der Exponentialverteilung ist,
ist nach dem Maximum-Likelihood-Prinzip von der Funktion L, definiert durch

$$\begin{aligned} L(\alpha) &= f_\alpha(x_1) \ldots f_\alpha(x_n) \\ &= \alpha e^{-\alpha x_1} \ldots \alpha e^{-\alpha x_n} \\ &= \alpha^n e^{-\alpha(x_1 + \ldots + x_n)} \\ &= \alpha^n e^{-\alpha n \bar{x}} \end{aligned}$$

(bei festen x_i und variablem α) das Maximim zu suchen. Der Wert $L(\alpha)$ ist für $\alpha > 0$ positiv, er ist klein für α nahe bei Null und für sehr großes α.
Die Ableitung L' von L

$$\begin{aligned} L'(\alpha) &= n \cdot \alpha^{n-1} \cdot e^{-\alpha n \bar{x}} + \alpha^n \cdot (-n\bar{x}) \cdot e^{-\alpha n \bar{x}} \\ &= n \cdot \alpha^{n-1} \cdot e^{-\alpha n \bar{x}} \cdot (1 - \alpha \bar{x}) \end{aligned}$$

hat genau eine Nullstelle bei $\hat{\alpha} = \frac{1}{\bar{x}}$
Also ist

$$T(\underline{x}) := \frac{1}{\bar{x}} = \frac{n}{x_1 + \cdots + x_n}$$

der Maximum-Likelihood-Schätzwert für α, der auf den Daten $x_1, \ldots, x_n$ beruht.
Der zugehörige Maximum-Likelihood-Schätzer (ML-Schätzer) ist

$$U_n = T(\underline{X}) = \frac{1}{\bar{X}}$$

Oft ist es technisch einfacher, die sogenannte Log-Likelihood-Funktion

$$\mathcal{L}(\theta) := \log L(\theta)$$

zu maximieren. Da die log-Funktion streng monoton wachsend ist, liefert dies dieselben Schätzwerte. Im obigen Beispiel ist etwa

$$\mathcal{L}(\alpha) = n \log \alpha - (\alpha n \bar{x})$$

Die Ableitung

$$\mathcal{L}'(\alpha) = \frac{n}{\alpha} - n\bar{x}$$

hat offenbar eine Nullstelle bei $\hat{\alpha} = \frac{1}{\overline{x}}$ und wegen

$$\mathcal{L}''(\alpha) = -\left(\frac{n}{\alpha^2}\right) < 0$$

liegt an der Stelle $\hat{\alpha} = \frac{1}{\overline{x}}$ ein Maximum vor.
Ein weiteres Verfahren ist die sogenannte Momenten-Methode.
Diese basiert darauf, dass $\overline{X}$ ein erwartungstreuer Schätzer für den Erwartungswert (erstes Moment) $\mathrm{E}(X_1)$ von X_1 ist.
Da $\mathrm{E}(X_1) = \mathrm{E}_\theta(X_1)$ vom Parameter θ abhängt, macht man den Ansatz

$$\overline{x} \;=\; \mathrm{E}_{\hat{\theta}}(X_1)$$

In vielen Fällen lässt sich diese Gleichung nach $\hat{\theta}$ auflösen. Eine solche Lösung $\hat{\theta} = \hat{\theta}(\underline{x})$ nennt man einen Schätzwert für θ nach der Momenten-Methode.

Beispiel

In vorigem Beispiel ist $\mathrm{E}_\alpha(X_1) = \frac{1}{\alpha}$ der Erwartungswert der Exponentialverteilung mit Parameter α.
Der Ansatz

$$\overline{x} = \mathrm{E}_{\hat{\alpha}}(X_1) = \frac{1}{2}$$

ergibt (aufgelöst nach $\hat{\alpha}$) den Schätzwert $\hat{\alpha} = \frac{1}{\overline{x}}$,
der in diesem Beispiel mit dem nach der Maximum-Likelihood-Methode gefundenen Schätzwert übereinstimmt.

Bemerkung

Im Allgemeinen stimmt ein nach der Momenten-Methode gefundener Schätzwert nicht mit dem Maximum-Likelihood-Schätzwert überein.
Betrachte dazu z.B. eine Rayleigh-verteilte Zufallsvariable X_1, d.h. eine Zufallsvariable X_1 mit Dichte

$$f(x) = \theta x e^{-\left(\frac{\theta}{2}\right)x^2}, x > 0$$

abhängig von einem Parameter $\theta > 0$.

2.4 Konfidenzintervalle und Statistik normalverteilter Zufallsvariablen

Sei $\underline{X} = (X_1, \ldots, X_n)$ eine Stichprobe und $\mathcal{P} := \{P_\theta \mid \theta \in \Theta\}$ das zugehörige statistische Modell.
Oft möchte man θ gar nicht mit Hilfe eines einzigen Schätzers $U_n = T(\underline{X})$ schätzen,
sondern vielmehr zwei Schätzer angeben, die mit einer gewissen Sicherheit
(vorgegebene Konfidenzzahl γ) unterhalb bzw. oberhalb des wahren Parameters θ liegen.
Man möchte also zwei Schätzer U_n und O_n derart finden, dass die Wahrscheinlichkeit,
dass U_n kleiner oder gleich und O_n größer oder gleich dem wahren Parameter θ ist,

möglichst groß, d.h. z.B. mindestens gleich γ ist.

Man definiert daher:

Definition (Konfidenzintervall)

Sei eine Konfidenzzahl (Vertrauenszahl) $0 < \gamma < 1$ vorgegeben.
Ein Intervall $[U_n, O_n]$ heißt γ-Konfidenzintervall für θ,
falls U_n und O_n zwei Schätzer sind mit der Eigenschaft

$$P(U_n \leq \theta \leq O_n) \geq \gamma, \ \forall\, \theta \in \Theta$$

Beispiel

Betrachte das statistische Modell $\mathcal{P} := \{N(\mu, \sigma^2) \mid \mu \in \mathbb{R}\}$ mit bekannter Varianz $\sigma^2 > 0$ und unbekanntem Erwartungswert μ.
Diese Situation tritt z.B. dann auf, wenn bei einem Produktionsprozess die Toleranz aus längerer Erfahrung bekannt ist,
aber die aktuelle Einstellung immer wieder justiert werden muss.
Bekannt: $\overline{X}$ ist ein erwartungstreuer Schätzer für $\mathrm{E}(X_1) = \mu$.
Zu vorgegebenem γ (z.B. $\gamma = 0.95$) bestimme das sogenannte

$$\frac{1+\gamma}{2}\text{ - Quantil } z$$

d.h. die Zahl z mit $\Phi(z) = \frac{1+\gamma}{2}$.
Z.B. ist $z \approx 1.96$ für $\gamma = 0.95$. Dann ist durch

$$U_n := \overline{X} - z \cdot \frac{\sigma}{\sqrt{n}}$$

und

$$O_n := \overline{X} + z \cdot \frac{\sigma}{\sqrt{n}}$$

ein γ-Konfidenzintervall $[U_n, O_n]$ für μ gegeben, da die Standardisierte

$$S_n^* = \frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} = \frac{\sqrt{n}}{\sigma} \cdot \left(\overline{X} - \mu\right)$$

standardnormalverteilt ist, d.h.

$$\begin{aligned} P(U_n \leq \mu \leq O_n) &= P\left(\mid \overline{X} - \mu \mid \leq\ z\frac{\sigma}{\sqrt{n}}\right) \\ &= P(\mid S_n^* \mid \leq z) && = \Phi(z) - \Phi(-z) \\ &= 2 \cdot \Phi(z) - 1 && = \gamma \end{aligned}$$

Ist die Varianz σ^2 unbekannt, so sehen die Konfidenzintervalle für μ komplizierter aus.

Vorbereitend sind zwei weitere Verteilungen einzufürhen.

Definition (Chi-Quadrat-Verteilung)

Seien $Y_1, \ldots, Y_n$ unabhängige, $N(0,1)$-verteilte Zufallsvariablen.
Dann heißt die Verteilung der Zufallsvariablen $Y := Y_1^2 + \cdots + Y_n^2$
die Chi-Quadrat-Verteilung mit n Freiheitsgraden, kurz: χ_n^2
Durch Induktion nach n zeigt man, dass die Verteilung von Y Dichte

$$g_n(x) := \begin{cases} \frac{1}{2^{(\frac{n}{2})} \cdot \Gamma(\frac{n}{2})} \cdot x^{\frac{n}{2}-1} \cdot e^{-\frac{x}{2}} & für\ x > 0 \\ 0 & für\ x \leq 0 \end{cases}$$

hat.
Die Chi-Quadrat-Verteilung tritt bei vielen statistischen Problemen auf.

Definition (t-Verteilung)

Seien Y und Z unabhängige Zufallsvariablen mit $Z \stackrel{d}{=} N(0,1)$ und $Y \stackrel{d}{=} \chi_n^2$.
Dann heißt die Verteilung der Zufallsvariablen

$$T := \frac{Z}{\sqrt{\frac{y}{n}}} = \sqrt{n} \cdot \frac{Z}{\sqrt{y}}$$

die Students-t-Verteilung mit n Freiheitsgraden, kurz: t_n.
Auch T besitzt eine Dichte f_n, gegeben durch

$$f_n(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi} \cdot \Gamma(\frac{n}{2})} \cdot \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

Für großes n sieht f_n etwa so aus wie ϕ (Gauß'sche Glockenkurve).
Für kleines n (z.B. $n = 5$) unterscheidet sich f_n von ϕ dadurch,
dass $f_n(x)$ für $x \to \pm\infty$ weniger rasch gegen Null geht.
Da der Graph von f_n symmetrisch zur y-Achse ist, kann man mit der t-Verteilung fast so rechnen wie mit der Standardnormalverteilung.
Bezeichnet etwa

$$F_n(x) := \int_{-\infty}^{\infty} f_n(t) \mathrm{d}t$$

die Verteilungsfunktion der t-Verteilung,
so gilt z.B. $F_n(-x) = 1 - F_n(x)$.
Mit der t-Verteilung lassen sich Konfidenzintervalle für den Mittelwert μ normalverteilter

Zufallsvariablen bei unbekannter Varianz σ^2 angeben.
Das unbekannte σ^2 darf nun nicht mehr in den Formeln der Konfidenzintervalle auftreten, wie es bei bekannter Varianz der Fall war.
Die auf Gosset (der das Pseudonym „Student“ wählte) zurückgehende Idee ist,
σ^2 durch die Zufallsvariable S^2 erwartungstreu zu schätzen,
d.h. ein Konfidenzintervall $[U_n, O_n]$ mit

$$U_n := \overline{X} - t \cdot \frac{S}{\sqrt{n}}$$

$$O_n := \overline{X} + t \cdot \frac{S}{\sqrt{n}}$$

zu wählen.
Die Aufgabe reduziert sich also darauf, die Konstante t so zu bestimmen, dass

$$P(U_n \leq \mu \leq O_n) = \gamma$$

gilt.

Bestimmung von t:
$\overline{X}$ hat Erwartungswert

$$\mathrm{E}(\overline{X}) = \mathrm{E}(X_1) = \mu$$

und

$$\mathrm{Var}(\overline{X}) = \frac{1}{n} \cdot \mathrm{Var}(X_1) = \frac{\sigma^2}{n}$$

Die Standardisierte

$$Z := \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}$$

von $\overline{X}$ ist $N(0,1)$-verteilt.
Man kann zeigen, dass

$$Y := \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma} \right)^2 = \frac{n-1}{\sigma^2} \cdot S^2$$

unabhängig von Z ist und dass $Y \stackrel{d}{=} \chi^2_{n-1}$.
Also ist

$$T := \sqrt{n-1} \cdot \frac{Z}{\sqrt{Y}} \stackrel{\downarrow}{=} \frac{\sqrt{n}}{S} \left(\overline{X} - \mu \right)$$

t-verteilt mit $n-1$ Freiheitsgraden.

$$\Rightarrow P(U_n \leq \mu \leq O_n) = P(\mid \overline{X} - \mu \mid \leq t \cdot \frac{S}{\sqrt{n}})$$
$$= P(\mid T \mid \leq t)$$

Also ist t so zu wählen, dass

$$\begin{aligned}\gamma &= P(\mid T \mid \leq t) \\ &= P(T \leq t) - P(T \leq -t) \\ &= 2 \cdot P(T \leq t) - 1\end{aligned}$$

bzw. $P(T \leq t) = \frac{1+\gamma}{2}$
Für t hat man also das $\frac{1+\gamma}{2}$ - Quantil der t-Verteilung mit $n-1$ Freiheitsgraden zu wählen.

Wir fassen zusammen:

Satz

(Konfidenzintervall für den Mittelwert normalverteilter Zufallsvariablen
bei unbekannter Varianz)
Gegeben sei eine Stichprobe von Umfang n bzgl. einer Normalverteilung $N(\mu, \sigma^2)$ mit unbekanntem Mittelwert μ und unbekannter Varianz σ^2.
Zu einer vorgegebenen Konfidenzzahl $0 < \gamma < 1$ sei t das
$\frac{1+\gamma}{2}$ - Quantil der t-Verteilung mit $n-1$ Freiheitsgraden.
Dann ist durch

$$\left[\overline{X} - t \cdot \frac{S}{\sqrt{n}}\ ,\ \overline{X} + t \cdot \frac{S}{\sqrt{n}}\right]$$

ein γ-Konfidenzintervall für μ gegeben.
Die Werte der Verteilungsfunktion und der α - Quantile der t-Verteilung sind tabelliert, oder sie können mit Rechnern ermittelt werden.

α	1	2	3	4	5	6	7	8	9	10	∞
0,95	6,314	2,920	2,353	2,132	2,015	1,943	1,895	1,860	1,833	1,812	1,6449
0,975	12,706	4,303	3,182	2,776	2,571	2,447	2,365	2,306	2,262	2,228	1,96

Es gilt also z.B. $F_8(x) = P(T_8 \leq x) = 0{,}975$ für $x \approx 2{,}306$

2.5 Testen von Hypothesen

Die wahre Verteilung P eines bestimmten Zufallsmechanismus sei nicht bekannt, aber man geht davon aus, dass P zu einer Menge $\{P_\theta \mid \theta \in \Theta\}$ von möglichen Verteilungen gehören.
Man teilt nun Θ auf in zwei Bereiche $\Theta_H \subset \Theta$ und $\theta_K := \Theta \setminus \Theta_H$

Dabei beschreibt Θ_H die zu testende Hypothese, d.h.

$$\text{Hypothese } H\colon P \in \{P_\theta \mid \theta \in \Theta_H\}$$

$$\text{Alternative } K\colon P \in \{P_\theta \mid \theta \in \Theta_K\}$$

Die Ausgänge des Zufallsmechanismus, also die Daten $\underline{x} \coloneqq (x_1, \ldots, x_n)$,
lassen in machnen Fällen schon intuitiv Rückschlüsse auf die Gültigkeit der Hypothese zu.
Ziel der Testtheorie ist es, solche Rückschlüsse auf objektivem Weg zu erhalten.

Definition (Test)

Ein Test ist eine Abbildung $\phi\colon \mathbb{R}^n \to \{0,1\}$ mit folgender Interpretation.
Gilt $\phi(\underline{x}) = 1$, so lehnt man H ab, anderenfalls nimmt man H an.

$$\text{Verwerfungsbereich von } \phi\colon V \coloneqq \left\{\underline{x} \in \mathbb{R}^n \mid \phi(\underline{x}) = 1\right\}$$

$$\text{Annahmebereich von } \phi\colon A \coloneqq \left\{\underline{x} \in \mathbb{R}^n \mid \phi(\underline{x}) = 0\right\}$$

Fehlermöglichkeiten:

- Fehler 1. Art: H liegt vor, wird aber abgehlehnt
- Fehler 2. Art: H liegt nicht vor, wird aber angenommen

Ziel ist es, den Test ϕ bzw. (äquivalent hierzu) den Verwerfungsbereich V so zu wählen,
dass beide Fehler möglichst geringe Wahrscheinlichkeit haben.
Problem: Vergrößern von V erhöht die Wahrscheinlichkeit für den Fehler 1. Art und verringert die Wahrscheinlichkeit für den Fehler 2. Art.
Daher ist eine Kompromiss nötig. Häufig ist die Problemstellung so, dass auf jeden Fall die Wahrscheinlichkeit für den Fehler 1. Art hinreichend klein sein soll.
Suche dann unter dieser Prämisse V so, dass dann die Warhscheinlichkeit für dne Fehler 2. Art möglichst klein ist.

Definition (Niveau)

Die Zahl $\sup\limits_{\theta \in \Theta_H} P(\phi(\underline{X}) = 1)$ heißt das Niveau des Tests ϕ.

Bemerkung

Die Wahrscheinlichkeit $P(\phi(\underline{X}) = 1)$ hängt von ϕ ab, da jedes X_i die Verteilung P_θ besitzt.
Kleines Niveau bedeutet also kleine Wahrscheinlichkeit für den Fehler 1. Art.
Wie findet man nun einen Test ϕ, dessen Niveau klein ist, d.h. z.B. ein vorgegebenes Signifikanzniveau α (z.B. $\alpha = 0{,}05$) nicht überschreitet?

Hierzu ist es zweckmäßig, die sogenannte Gütefunktion eines Testes einzuführen:

Definition (Gütefunktion)

Die Funktion $\pi_\phi \colon \Theta \to [0,1]$, definiert durch

$$\pi_\phi := P(\phi(\underline{x}) = 1),\ \forall \theta \in \Theta$$

heißt die Gütefunktion des Tests ϕ.
Ein Test hat also das Niveau α, falls $\sup\limits_{\theta \in \Theta_H} \pi_\phi(\theta) = \alpha$.

Beispiel: Zweiseitiger Gauß-Test

Es werden n Messungen $x_1, \ldots, x_n$ durchgeführt, um eine physikalische Größe μ zu bestimmen.
Man nimmt an, dass die Messungen Realisierungen von n unabhängigen, $N(\mu, \sigma^2)$-verteilten Zufallsvariablen $X_1, \ldots, X_n$ sind mit bekannter Varianz $\sigma^2 > 0$.
Es ist zu testen, ob μ gleich einem vorgegebenen μ_0 ist, oder nicht. Also wird man für die Hypothese H den Bereich $\Theta_H := \{\mu\}$ und für die Alternative K den Bereich $\Theta_K := \mathbb{R} \setminus \{\mu_0\}$ wählen.
Man sagt auch kurz, es liegt die Testsituation

$$H \colon \mu = \mu_0 \text{ gegen } K \colon \mu \neq \mu_0$$

vor.
Es ist plausibel, H zu verwerfen, wenn $\underline{x}$ weit von μ_0 entfernt ist.
Als Verwerfungsbereich eignet sich also

$$V := \left\{ \underline{x} \in \mathbb{R}^n : \ \mid \overline{x} - \mu_0 \mid \geq c \right\}$$

und der zugehörige Test lautet:

$$\phi(\underline{x}) := \begin{cases} 1 & \text{falls } \mid \overline{x} - \mu_o \mid \geq c \\ 0 & \text{sonst} \end{cases}$$

Nun ist c so zu bestimmen, dass das Niveau des Tests ϕ ein vorgegebenes Signifikanzniveau α nicht überschreitet.
Die Gütefunktion lautet

$$\begin{aligned} \pi_\phi(\mu) &= P(\phi(\underline{X}) = 1) \\ &= P(\mid \overline{X} - \mu_0 \mid \geq c) \\ &= P(\overline{X} \leq \mu_0 - c) + P(\overline{X} \geq \mu_0 + c) \end{aligned}$$

Wegen $\overline{X} \stackrel{d}{=} N(\mu, \frac{\sigma^2}{n})$ folgt

$$\begin{aligned}\pi_\phi(c) &= \Phi_{\mu,\frac{\sigma^2}{n}}(\mu_0 - c) + 1 - \Phi_{\mu,\frac{\sigma^2}{n}}(\mu_0 + c) \\ &= \Phi\left(\frac{\mu_0 - c - \mu}{\frac{\sigma}{\sqrt{n}}}\right) + 1 - \Phi\left(\frac{\mu_0 + c - \mu}{\frac{\sigma}{\sqrt{n}}}\right) \\ &= 2 - \Phi\left(\frac{c - (\mu - \mu_0)}{\frac{\sigma}{\sqrt{n}}}\right) - \Phi\left(\frac{c - (\mu - \mu_0)}{\frac{\sigma}{\sqrt{n}}}\right)\end{aligned}$$

Insbesondere ist π_ϕ symmetrisch um $\mu = \mu_0$ und ϕ hat das Niveau

$$\pi_\phi(\mu_0) = 2 - 2\Phi\left(\frac{c}{\frac{\sigma}{\sqrt{n}}}\right)$$

Dieses Niveau soll nun höchstens gleich α sein, d.h. c ist so zu bestimmen,
dass $2 - 2\Phi\left(\frac{c}{\frac{\sigma}{\sqrt{n}}}\right) = \alpha$
Auflösen nach c ergibt

$$c = z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

wobei $z_{1-\frac{\alpha}{2}}$ das $(1 - \frac{\alpha}{2})$-Quantil von $N(0,1)$ ist, d.h. $\Phi(z_{1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$
Damit haben wir den sogenannten zweiseitigen Gauß-Test zum Niveau α gefunden.
Man berechnet aus den Daten $\underline{x} = (x_1, \ldots, x_n)$ den Wert $\phi(\underline{x})$
Falls dieser Wert gleich 1 ist, d.h. falls

$$| \overline{x} - \mu_0 | \geq z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

so lehnt man H (auf dem Signifikanzniveau α) ab.
Ist hingegen $\phi(\underline{x}) = 0$, so kann H (auf dem Signifikanzniveau α) nicht abgelehnt werden.