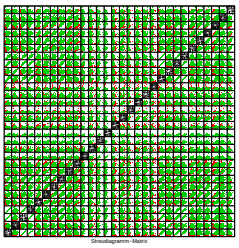
der Bundeswehr
Universität München

Dr. Robert Schmied

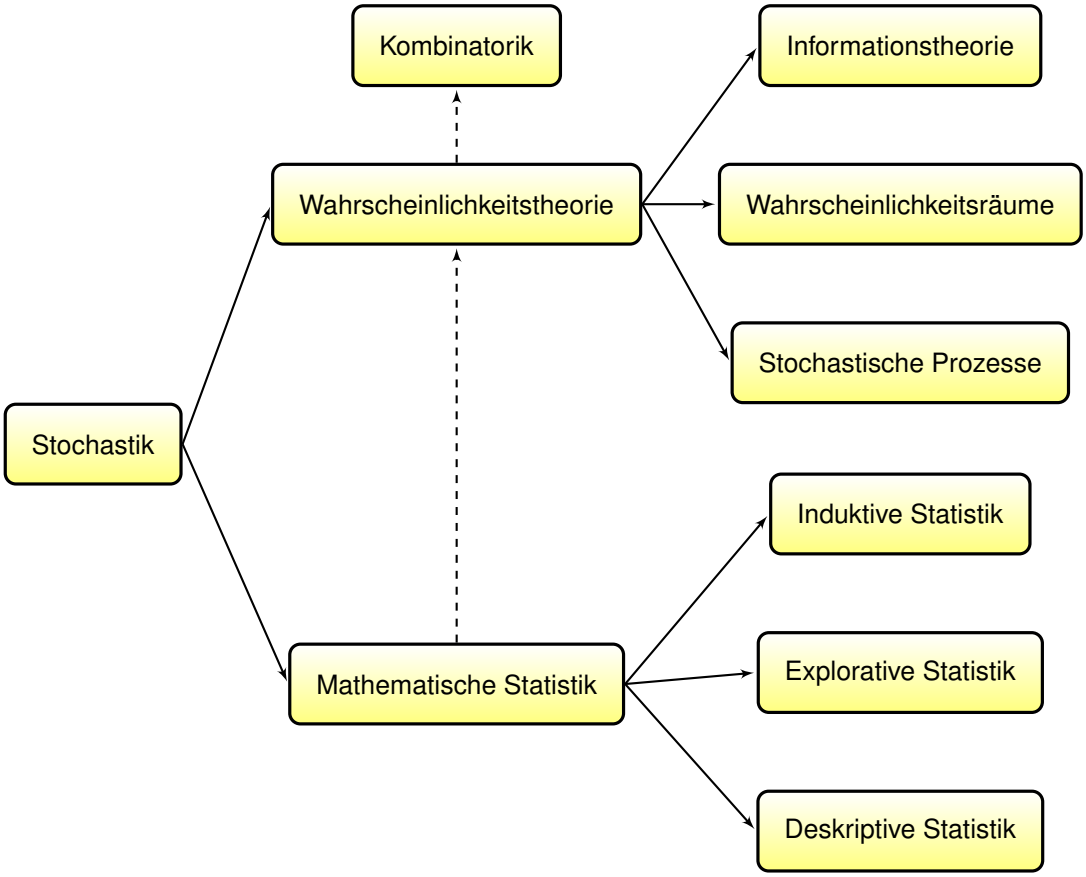


Professur für Mathematik
Fakultät für Elektrotechnik
und Informationstechnik
85577 Neubiberg
Tel.: 089/6004-3931
robert.schmied@unibw.de

Statistik für Ingenieure

5. Januar 2024

Inhaltsverzeichnis


Kombinatorik
Informationstheorie
Wahrscheinlichkeitstheorie
Wahrscheinlichkeitsräume
Stochastische Prozesse
Stochastik
Induktive Statistik
Mathematische Statistik
Explorative Statistik
Deskriptive Statistik

Teil I.

Wahrscheinlichkeitstheorie

1. Kombinatorik

1.1. Experimente mit abzählbaren Ausgängen

Die Durchführung von Experimenten kann häufig nicht als rein deterministischer Vorgang gesehen werden. Deshalb werden sie auf Basis stochastischer Ansätze modelliert. In der Stochastik sind Ergebnismengen von zentraler Bedeutung. Mit der Festlegung von Ergebnismengen wird die Grundlage zur theoretischen Behandlung von Experimenten und praktischen Verarbeitung von Daten aus Experimenten geschaffen.


Cantor
1845-1918

Erläuterung

Wir unterscheiden zwischen endlichen, abzählbaren und überabzählbaren Mengen. Endliche Mengen lassen sich in dem Sinne hinschreiben, dass eine vollständige Aufzählung z.B. in Form einer Nummerierung der einzelnen Elemente möglich ist. Dagegen sind abzählbare und überabzählbare Mengen unendliche (nicht endliche) Mengen und so beschaffen, dass sich nicht alle Elemente hinschreiben lassen. Bei abzählbaren Mengen ist eine Nummerierung der einzelnen Elemente möglich. Doch es lässt sich keine vollständige Auflistung der Elemente durchführen. Die „kleinste“ abzählbare Menge ist die Menge $\mathbb{N}$ der natürlichen Zahlen. Sie besitzt unendlich viele Elemente, was durch $|\mathbb{N}| = \aleph_0$ bezeichnet wird. Der Index 0 bei $\aleph_0$ deutet bereits an, dass es Mengen gibt, die eine andere Art von unendlich vielen Elementen enthalten. So gilt etwa $|\mathbb{R}| = \aleph_1$. Wären $\mathbb{N}$ und $\mathbb{R}$ gleichmächtig, müsste es eine bijektive Abbildung zwischen beiden Mengen geben, was aber durch den Beweis Georg Cantors[a] widerlegt ist. Cantor, einer der Studenten von Weierstraß, war es, der die Mengenlehre begründete. Dabei war seine Mengenlehre nicht unumstritten. Insbesondere seine Theorien zu den Mächtigkeiten von Mengen musste sich erst durchsetzen.

[a] Cantors zweites Diagonalargument, [1]

Definition 1.1

Eine Menge A heißt endlich, wenn es eine natürliche Zahl $n \in \mathbb{N}$ gibt, so dass es eine bijektive Abbildung

$$f : A \to \{0, 1, \ldots, n-1\}$$

zwischen A und $\{0, 1, \ldots, n-1\}$ gibt. Die Menge A besteht aus n Elementen, man sagt, A habe die Mächtigkeit $|A| = n$. Allgemein heißt eine Menge A gleichmächtig zu einer Menge B, wenn es eine bijektive Abbildung zwischen beiden Mengen A und B gibt.

Eine überabzählbare Menge ist die Menge $\mathbb{R}$ der reellen Zahlen, die Menge $\mathbb{N}$ der natürlichen Zahlen ist abzählbar und $\{0, 1, 2\}$ ist endlich.

Beispiel 1.2

Die Menge $\mathbb{Z}$ der ganzen Zahlen ist abzählbar. Mit der Bijektion

$$\mathbb{N}_0 \to \mathbb{Z},\ n \mapsto (-1)^n \cdot \left\lceil \frac{n}{2} \right\rceil$$

ist $|\mathbb{Z}| = \aleph_0$.

Bijektive Abbildungen endlicher Mengen auf sich selbst spielen in der Kombinatorik eine wichtige Rolle.

Definition 1.3

Ist A endlich, so heißt eine bijektive Abbildung $\pi : A \to A$ Permutation.

Eine Permutation π einer Menge $A = \{a_1, \ldots, a_n\}$ lässt sich in der Form

$$\begin{pmatrix} a_1 & a_2 & \ldots & a_n \\ \pi(a_1) & \pi(a_2) & \ldots & \pi(a_n) \end{pmatrix}$$

schreiben.

Beispiel 1.4

Die Menge

$$F = \{\text{rot,grün,blau,gelb,schwarz,weiß}\}$$

lässt sich durch

$$F \to \{0,1,2,3,4,5\},$$
$$\text{rot} \mapsto 0, \text{grün} \mapsto 1, \text{blau} \mapsto 2, \text{gelb} \mapsto 3, \text{schwarz} \mapsto 4 \text{ und weiß} \mapsto 5$$

bijektiv auf die Menge $\{0,1,2,3,4,5\}$ abbilden. Es ist $|F| = 6$. Eine Permutation der Menge F ist etwa

$$\begin{pmatrix} \text{rot} & \text{grün} & \text{blau} & \text{gelb} & \text{schwarz} & \text{weiß} \\ \text{grün} & \text{weiß} & \text{rot} & \text{gelb} & \text{schwarz} & \text{blau} \end{pmatrix}.$$

Da die Menge A endlich ist mit $|A| = n$, lässt sie sich durch die Menge $\{1, \ldots, n\}$ identifizieren. Per vollständiger Induktion nach n lässt sich zeigen, dass es genau $n!$ verschiedene Permutationen gibt, wobei $n!$ durch die Abbildung

$$\mathbb{N}_0 \to \mathbb{N},\ n \mapsto n! := n \cdot (n-1) \cdot \ldots \cdot 1 = n \cdot (n-1)!,\ 0! := 1 \quad \text{(Fakultät)}$$

gegeben ist.

Fasst man ein Element einer Menge als Ergebnis eines Experiments auf, so werden oft die Ergebnisse eines mehrmals wiederholten Experiments oder die Ergebnisse verschiedener Experimente in Form eines k-Tupels zusammengeführt.

Erläuterung

Ein k-Tupel ist eine Anordnung von k Elementen und unterscheidet sich damit grundlegend von einer Menge, bei der die Reihenfolge ihrer Elemente keine Rolle spielt. Sind A, B zwei Mengen, so wird (a, b) mit $a \in A$ und $b \in B$ ein Paar genannt. Das a steht vor dem b, hier ist die Reihenfolge wichtig. Im Unterschied dazu ist bei der Menge $\{a, b\}$ keine Reihenfolge für die Elemente a und b festgelegt. Wichtig: im Allgemeinen ist $(a, b) \neq (b, a)$ und stets ist $\{a, b\} = \{b, a\}$. Allgemein ist beim k-Tupel $(a_1, \ldots, a_k)$, $k > 1$, die Reihenfolge wichtig. Sie wird nach Kazimierz Kuratowski durch $(a_1, \ldots, a_{k+1}) = ((a_1, \ldots, a_k), a_{k+1})$ induktiv festgelegt. Das kartesische Produkt liefert die allgemeine Definition einer angeordneten Ansammlung von Elementen.

Definition 1.5

Kuratowski
1896-1980

Das kartesische Produkt der Mengen $A_1, \ldots, A_k$ ist die Gesamtheit aller k-Tupel $(a_1, \ldots, a_k)$ mit $a_i \in A_i$:

$$\begin{aligned} A_1 \times A_2 &:= \{(a_1, a_2);\ a_1 \in A_1 \text{ und } a_2 \in A_2\}, \\ A_1 \times \ldots \times A_k &:= (A_1 \times \ldots \times A_{k-1}) \times A_k \\ &= \{(a_1, \ldots, a_k);\ a_i \in A_i \text{ für } i = 1, \ldots, k\}. \end{aligned}$$

Sind $A_1, \ldots, A_k$ endliche Mengen, so stellt sich die Frage nach der Mächtigkeit der Menge $A_1 \times \ldots \times A_k$ aller k-Tupel $(a_1, \ldots, a_k)$ mit $a_i \in A_i$.

Satz 1.6: Abzählprinzip

Man betrachte eine Serie von k Experimenten mit endlichen Ergebnismengen A_1 bis A_k und $a_i \in A_i$ sei das Ergebnis des i-ten Experiments. Gibt es dafür n_i mögliche Ausgänge, d.h. $|A_i| = n_i$ unabhängig vom Ausgang anderer Experimente, so gilt

$$|A| := |A_1 \times \ldots \times A_k| = n_1 \cdot \ldots \cdot n_k.$$

Beweis.

Um das zu zeigen, wird eine Induktion nach der Anzahl k der Mengen durchgeführt. Es sei $n_i = |A_i|$ für jeden Index i.
$k = 2$: $|A| = |A_1 \times A_2| = |\{(a_1, a_2);\ a_1 \in A_1, a_2 \in A_2\}| = |A_1| \cdot\ |A_2| = n_1 \cdot n_2$ $\surd$
$k \to k + 1$:

$$\begin{aligned} |A| &= |\{(a_1, \ldots, a_k, a_{k+1});\ a_i \in A_i\}| = |A_1 \times \ldots \times A_k \times A_{k+1}| \\ &= |(A_1 \times \ldots \times A_k) \times A_{k+1}| \overset{\text{I.A.}}{=} |A_1 \times \ldots \times A_k| \cdot |A_{k+1}| \\ &\overset{\text{I.V.}}{=} (n_1 \cdot \ldots \cdot n_k) \cdot n_{k+1} = n_1 \cdot \ldots \cdot n_k \cdot n_{k+1}. \end{aligned}$$

□

Mit $n_i = |A_i|$ gibt es $|A_1 \times \ldots \times A_k| = n_1 \cdot \ldots \cdot n_k$ verschiedene k-Tupel. Ein Einzelexperiment kann auf verschiedene Arten durchgeführt werden. Das Ergebnis aller k Einzelexperimente wird in einem k-Tupel notiert.

(1) Jedes Einzelexperiment wird stets unter den gleichen Voraussetzungen durchgeführt. Die Ergebnisse werden ablaufgetreu notiert.

(2) Jedes Einzelexperiment wird so durchgeführt, dass das jeweilige Ergebnis aller vorherigen Einzelexperimente ausgeschlossen wird. Die Ergebnisse werden ablaufgetreu notiert.

(3) Jedes Einzelexperiment wird so durchgeführt, dass das jeweilige Ergebnis aller vorherigen Einzelexperimente ausgeschlossen wird. Die Einzelergebnisse werden am Ende aufsteigend sortiert.

(4) Jedes Einzelexperiment wird stets unter den gleichen Voraussetzungen durchgeführt. Die Einzelergebnisse werden am Ende aufsteigend sortiert.

Jede einzelne Situation wird im Folgenden anhand von Beispielen erklärt und hinsichtlich der möglichen Anzahl an verschiedenen Ergebnissen untersucht.

1.2. Urnenmodelle

Erläuterung: Urnenmodell

Viele Überlegungen der Kombinatorik lassen sich in Form eines Urnenmodells beschreiben. Dabei werden in einem nicht einsehbaren Behälter (Urne) Kugeln platziert und diese dann nach einer festgelegten Vorschrift gezogen. Meist wird von n Kugeln in der Urne und vom Ziehen von k Kugeln aus der Urne ausgegangen. Eine gezogene Kugel lässt sich zurücklegen oder nicht, die Reihenfolge der Züge kann wesentlich oder unwichtig sein.

Werden aus einer Urne mit $n \in \mathbb{N}$ Kugeln nacheinander $k \in \mathbb{N}$ Kugeln herausgezogen, ist das Ergebnis eine Auswahl (Stichprobe) von k Kugeln aus der Gesamtheit (Grundgesamtheit) aller n Kugeln. Ist dabei die Reihenfolge der gezogenen Kugeln von Bedeutung, heißt die Stichprobe geordnet (Permutation), ansonsten ungeordnet (Kombination). Zudem muss im Ablauf unterschieden werden, ob eine gezogene Kugel vor der nächsten Ziehung wieder in die Urne zurückgelegt wird oder nicht. Im ersten Fall wird von einer Stichprobe mit, im zweiten Fall von einer Stichprobe ohne Zurücklegen gesprochen. Die vier Varianten sind in folgender Tabelle noch einmal zusammengefasst.

Urnenmodelle: Ziehen von k Kugeln aus n Kugeln

	mit Zurücklegen	ohne Zurücklegen
geordnet	(n,k)-Permutation mit Wiederholung (1)	(n,k)-Permutation ohne Wiederholung (2)
ungeordnet	(n,k)-Kombination mit Wiederholung (4)	(n,k)-Kombination ohne Wiederholung (3)

Beispiel 1.7

(1) Ein Fahrradschloss aus vier unabhängig drehbaren Rädern mit je zehn Ziffern öffnet sich nur bei genau einer Ziffernfolge. Es handelt sich dabei um die Situation einer geordneten Stichprobe mit Zurücklegen. Das Ergebnis ist ein Element der Menge

$$\Omega_{(1)} = \{0, 1, 2, \ldots, 9\}^4.$$

So könnte die Kombination $(5, 6, 7, 8) \in \Omega_{(1)}$ das Fahrradschloss öffnen.

(2) Bei einem 100-m-Lauf ist die Reihenfolge der acht Läufer im Zieleinlauf von entscheidender Bedeutung, es handelt sich um eine geordnete Stichprobe ohne Zurücklegen. Für die ersten drei Plätze sind sämtliche 3-Tupel möglich, wobei die Komponenten paarweise verschieden sind:

$$\Omega_{(2)} = \{(1, 2, 3), (1, 2, 4), \ldots, (1, 2, 8), (1, 3, 2), \ldots, (5, 7, 8), (6, 7, 8)\}.$$

Sind die drei Erstplatzierten unter den acht Läufern in dieser Reihenfolge die Läufer 4,3 und 5, so schreibt man $(4, 3, 5) \in \Omega_{(2)}$.

(3) Bei der Ziehung der Lottozahlen 6 aus 49 spielt die Reihenfolge der gezogenen Kugeln keine Rolle, es handelt sich um eine ungeordnete Stichprobe ohne Zurücklegen. Möglich sind sämtliche 6-elementige Teilmengen der Menge aller Kugeln,

$$\Omega_{(3)} = \{\{k_1, \ldots, k_6\};\ k_i \in \{1, 2, \ldots, 49\},\ k_i \neq k_j \text{ für } i \neq j\}.$$

Werden die Zahlen $15, 2, 11, 34, 20, 24$ gezogen, wird das Ergebnis in der Form $\{2, 11, 15, 20, 24, 34\} \in \Omega_{(3)}$ dargestellt. Wie $\Omega_{(3)}$ als Menge von k-Tupeln geschrieben wird, muss noch geklärt werden.

(4) Im Supermarkt sollen fünf Pizzen auf drei Kühlregale verteilt werden. Für jede Pizza wird ein Kühlregal ausgewählt, wobei in einem Kühlregal auch mehrere Pizzen liegen dürfen und die Reihenfolge der Pizzen nicht berücksichtigt wird. Es liegt der Fall einer ungeordneten Stichprobe mit Zurücklegen vor. Jede Pizza wird in eines der drei Kühlregale gelegt. Die Notation als (ungeordnete) Menge klappt hier nicht, da bei mehrmaligem Verwenden eines Kühlregals (z.B. dreimal das Kühlregal 2) als Ergebnis $\{1, 2, 2, 2, 3\}$ geschrieben werden müsste. In einer Menge unterscheiden sich sämtliche Elemente, demnach hier auch die drei Zweien. Die Ergebnismenge $\Omega_{(4)}$ kann auch hier als Menge von k-Tupeln und in sehr ähnlicher Weise wie $\Omega_{(3)}$ geschrieben werden.

Zu jedem der Beispiele stellt sich die Frage nach der Anzahl der möglichen Ergebnisse, d.h. nach der Anzahl der Elemente in $\Omega_{(i)}$. Zur vereinfachten Notation benötigen wir die Abkürzung

$$\binom{n}{k} := \frac{n!}{k!(n-k)!} \quad \text{(Binomialkoeffizient)}.$$

1.2.1. (n, k)-Permutation mit Wiederholung (1)

Beim Fahrradschloss gibt es für jedes der Räder unabhängig voneinander zehn verschiedene Möglichkeiten. Nach dem Abzählprinzip insgesamt eine Anzahl von $10^4 = 10000$ Möglichkeiten.

Zu Beginn befinden sich alle Kugeln in der Urne. Wird eine gezogene Kugel im Urnenmodell nach der ersten Ziehung wieder zurückgelegt (geordnete Stichprobe mit Zurücklegen), gibt es für den nächsten Zug wiederum n Möglichkeiten, weiter ist dies auch nach der zweiten, dritten bzw. k-ten Ziehung so. Insgesamt gibt es demnach

$$\underbrace{n \cdot n \cdot \ldots \cdot n}_{n\text{-mal}} = n^k, \quad k \in \mathbb{N}$$

verschiedene Möglichkeiten. Ist A die Ergebnismenge des Einzelexperiments, so ist

$$\Omega_{(1)} = \{(a_1, \ldots, a_k);\ a_i \in A \text{ für } i = 1, \ldots, n\} = A^k \text{ und } |\Omega_{(1)}| = |A|^k.$$

Beispiel 1.8

Es wird fünfmal nacheinander ein Würfel geworfen. Ein möglicher Ausgang des Experiments könnte

sein. In dieser Situation ist das Ergebnis von

zu unterscheiden, da nacheinander das Ergebnis der Einzelexperimente notiert wird. Für jeden Wurf sind sechs verschiedene Ergebnisse möglich und somit ergeben sich insgesamt $6^5 = 7776$ verschiedene Möglichkeiten. Wir notieren dies in der Form

$$(3, 6, 2, 6, 3) \in \{(a_1, \ldots, a_5);\ a_i \in \{1, 2, 3, 4, 5, 6\}\} = \{1, 2, 3, 4, 5, 6\}^5.$$

Beispiel 1.9

Es sollen fünf Murmeln auf sechs Schachteln verteilt werden, wobei die Reihenfolge der Wahl der Schachteln festgehalten werde. Wie viele verschiedene Möglichkeiten gibt es? Angenommen, die erste Murmel landet in Schachtel drei, die zweite in der sechsten, die dritte in der zweiten, die vierte in der sechsten und die fünfte in der dritten. Indem die Schachteln durchnummeriert werden, entsteht die Situation wie bei den Würfeln, so lautet das Ergebnis $(3, 6, 2, 6, 3)$ und es gibt 7776 verschiedene Möglichkeiten.

Beispiel 1.10

Aus einer Urne, die sechs verschiedenfarbige Kugeln enthält, werden nacheinander fünf Kugeln herausgezogen, jeweils die Farbe notiert und die Kugel wieder hineingelegt. Die Kugeln haben die Farben gelb, blau, rot, schwarz, weiß und grün. Ein mögliches Ergebnis könnte

(rot,grün,blau,grün,rot) sein. Die Menge $F = \{$gelb,blau,rot,schwarz,weiß,grün$\}$ lässt sich durch $F \to \{1,2,3,4,5,6\}$, gelb $\mapsto 1$, blau $\mapsto 2$, rot $\mapsto 3$, schwarz $\mapsto 4$, weiß $\mapsto 5$ und grün $\mapsto 6$ bijektiv auf die Menge $\{1,2,3,4,5,6\}$ abbilden. Somit erhält man wiederum die Würfelsituation mit 7776 verschiedenen Ergebnissen.

Was ändert sich, wenn in der Urne von den sechs Kugeln jeweils zwei rot, grün und blau sind? Oder, wenn der Würfel je zweimal eins, zwei und drei als Augenzahl hat? Dann gibt es für jedes Einzelexperiment genau drei mögliche Ausgänge (rot, grün, blau bzw. 1, 2, 3), insgesamt somit $3^5 = 243$ mögliche Ergebnisse.

1.2.2. (n,k)-Permutation ohne Wiederholung (2)

Beim 100-m-Lauf gibt es für den Sieger acht, für den Zweitplatzierten noch sieben und für den Drittplatzierten noch sechs Möglichkeiten, zusammen somit $8 \cdot 7 \cdot 6 = 336$ Möglichkeiten.

Ist allgemein die Ergebnismenge $A = A_1$ mit $|A_1| = n$ für das erste Einzelexperiment gegeben, so gibt es für das zweite Einzelexperiment mit der Ergebnismenge A_2 noch $|A_2| = n - 1 = n - 2 + 1$ verschiedene Ausgänge. Für das k-te Einzelexperiment gibt es für die Ergebnismenge A_k noch $|A_k| = n - k + 1$ mögliche Ausgänge. Das Gesamtergebnis stammt aus der Menge

$$\Omega_{(2)} = \{(a_1, \ldots, a_k);\ a_i \in A,\ i = 1, \ldots, k;\ a_j \neq a_l, j \neq l\},$$

und es gilt nach dem Abzählprinzip

$$\begin{aligned} |\Omega_{(2)}| &= |A_1| \cdot \ldots \cdot |A_k| = n \cdot (n-1) \cdot \ldots \cdot (n-k+1) \\ &\overset{k \leq n}{=} \frac{n \cdot \ldots \cdot (n-k+1) \cdot (n-k)!}{(n-k)!} \\ &= \frac{n!}{(n-k)!} = \frac{n!}{(n-k)! \cdot k!} \cdot k! \\ &= \binom{n}{k} \cdot k! \end{aligned}$$

Für $k = n$ gibt es somit $n \cdot (n-1) \cdot \ldots \cdot 2 \cdot 1 = n!$ verschiedene Möglichkeiten. Dies ist die Anzahl der Permutationen einer n-elementigen Menge entsprechend der Beschreibung nach Beispel 1.4 und beschreibt die Anzahl der Möglichkeiten, n unterscheidbare Kugeln anzuordnen. Für $k > n$ (d.h. $k \geq n+1$) kommt in dem Produkt der Faktor $n-(n+1)+1 = 0$ vor und damit ergibt sich in diesem Fall der Wert Null. Es können schließlich nicht mehr als n Kugeln entnommen werden.

Beispiel 1.11

Werden fünf Murmeln auf sechs Schachteln verteilt, wobei jede Schachtel nur einmal benutzt werden darf, so stehen zunächst sechs Schachteln zur Auswahl, dann fünf usw. Insgesamt gibt es $6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 = 720$ Möglichkeiten, die Murmeln aufzuteilen. Da nur noch eine Schachtel übrig bleibt, entspricht das der Anzahl Permutationen der sechs Schachteln.

Beispiel 1.12

Aus einer Urne, die sechs verschiedenfarbige Kugeln enthält, werden nacheinander fünf Kugeln herausgezogen, die Farbe notiert und die Kugel zur Seite gelegt. Die Kugeln haben die Farben gelb, blau, rot, schwarz, weiß und grün. Ein mögliches Ergebnis könnte (rot,grün,blau,weiß,gelb) sein. Auch hier gibt es $6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 = 720$ verschiedene Möglichkeiten.

Die MISSISSIPPI-Fragestellung
Was ändert sich, wenn in der Urne von den sechs Kugeln jeweils zwei rot, grün und blau sind? Dazu ist zunächst eine Vorüberlegung anzustellen. Wir überlegen uns, wie es für den Fall aussieht, wenn alle Kugeln gezogen werden und zunächst alle Kugeln unterschieden werden können, indem beispielsweise je eine rote, grüne und blaue Kugel einen weißen Rand erhält. Dann gibt es zunächst $6! = 720$ Permutationen, eine davon ist etwa

Wird von der künstlichen Unterscheidung abgesehen, spielt es keine Rolle, welche der roten Kugeln an welcher Position kommt. Für die roten Kugeln gibt es $2! = 2$ Permutationen. Die Gesamtzahl an Möglichkeiten wird durch diesen Faktor dividiert. Ebendasselbe gilt für die grünen und blauen Kugeln. Das ergibt insgesamt $\frac{6!}{2! \cdot 2! \cdot 2!} = 90$ verschiedene Möglichkeiten.

Bei einer Ergebnismenge A mit $|A| = n$, von deren Elementen jeweils $n_1, \ldots, n_l$ mit $n = n_1 + n_2 + \ldots + n_l$ gleich sind, gibt es $\frac{n!}{n_1! \cdot \ldots \cdot n_l!}$ Permutationen. Die Buchstaben des Wortes MISSISSIPPI können auf $\frac{11!}{1!4!4!2!} = 34650$ verschiedene Arten angeordnet werden.

Um den Fall der Ziehung von $k < n$ Kugeln aus der Urne behandeln zu können, muss dafür zunächst die Situation (3) als Spezialfall der Permutationen betrachtet werden.

1.2.3. (n, k)-Kombination ohne Wiederholung (3)

Sind beim Lotto sechs Kugeln aus den 49 Kugeln entnommen und werden diese markiert, lassen sich die markierten und die nicht markierten Kugeln auf $\frac{49!}{6! \cdot 43!} = 13982816$ verschiedene Arten anordnen.

Bei der ungeordneten Stichprobe ohne Zurücklegen kommt es nicht auf die Reihenfolge der k gezogenen Kugeln an. Für die Reihenfolge der k Kugeln gibt es $k!$ Permutationen, deren Unterscheidung nicht von Belang ist. Damit muss es

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

Möglichkeiten geben. Ist A mit $|A| = n$ die Ergebnismenge für jedes Einzelexperiment, so stammt das Gesamtergebnis aus der Menge

$$\Omega_{(3)} = \{\{a_1, \ldots, a_k\};\ a_i \in A,\ a_i \neq a_j \text{ für } i \neq j,\ 1 \leq i, j \leq k\}\,.$$

Erläuterung

Um ein mögliches Ergebnis $\{a_1, \ldots, a_k\}$ als k-Tupel $(a_1, \ldots, a_k)$ zu schreiben, überlegen wir uns, dass zu jeder der $k!$ Permutationen eine gehört, bei der die Elemente aufsteigend sortiert sind. Dies lässt sich formell durch eine Äquivalenzrelation auf $\Omega_{(2)}$ beschreiben:

Zwei k-Tupel $(a_1, \ldots, a_k), (\tilde{a}_1, \ldots, \tilde{a}_k) \in \Omega_{(2)}$ seien äquivalent,

$$(a_1, \ldots, a_k) \sim (\tilde{a}_1, \ldots, \tilde{a}_k),$$

genau dann wenn es eine Permutation $\pi : \{1, \ldots, k\} \to \{1, \ldots, k\}$ gibt mit $\tilde{a}_i = a_{\pi(i)}$.

Betrachtet man die Äquivalenzklasse $[(a_1, \ldots, a_k)]_\sim$, so gibt es einen Repräsentanten

$$(\tilde{a}_1, \ldots, \tilde{a}_k) \in [(a_1, \ldots, a_k)]_\sim$$

mit $\tilde{a}_j < \tilde{a}_l$ für $j < l$.

Durch Sortierung der Elemente der Menge $\{a_1, \ldots, a_k\}$ zu einem k-Tupel lässt sich $\Omega_{(3)}$ schreiben als

$$\Omega_{(3)} = \{(\tilde{a}_1, \ldots, \tilde{a}_k);\ \tilde{a}_i \in A,\ \tilde{a}_j < \tilde{a}_l \text{ für } j < l\}$$

und es gilt $|\Omega_{(3)}| = \binom{n}{k}$.

Beispiel 1.13

Aus einer Urne, die sechs verschiedenfarbige Kugeln enthält, werden gleichzeitig fünf Kugeln herausgezogen und die Farbe notiert. Die Kugeln haben die Farben rot, grün, blau, gelb, schwarz und weiß. Ein mögliches Ergebnis könnte die Menge {blau,weiß,grün,rot,gelb} sein, die eine 5-elementige Teilmenge der Ergebnismenge ist. Schreiben wir sämtliche Elemente der Ergebnismenge auf und markieren diejenigen Elemente, die gezogen wurden, durch ein Plus, alle anderen durch ein Minus,

rot	grün	blau	gelb	schwarz	weiß
+	+	+	+	-	+

so stellt jede mögliche Permutation der k Pluszeichen und $n - k$ Minuszeichen nach der allgemeinen Permutationsregel $\frac{n!}{k! \cdot (n-k)!} = \binom{n}{k}$ mit $n_1 = k$ und $n_2 = n - k$ die Anzahl der k-elementigen Teilmengen einer n-elementigen Menge dar.

Erläuterung: Additionsregel

Oftmals gibt es kombinatorische Fragestellungen, die nicht durch eine einzige Formel gelöst werden können. Die Fragestellung kann zunächst so in Teilprobleme zerlegt werden, dass die für die einzelnen Probleme resultierenden Ergebnismengen untereinander disjunkt und deren Mächtigkeit mit Hilfe einfacher Formeln bestimmbar sind. Es handelt sich dabei um eine Partition des Gesamtergebnisses. Die Anzahl der Möglichkeiten der kombinatorischen Fragestellung ergibt sich dann als Summe der Anzahlen für die einzelnen Teilprobleme[a].

[a] z.B. für zwei Teilprobleme mit Ergebnismengen A und B: $A \cap B = \emptyset \Rightarrow |A \cup B| = |A| + |B|$

Sollen aus einer Urne mit zwei roten, zwei grünen und zwei blauen Kugeln fünf Kugeln ohne Zurücklegen und ohne Reihenfolge entnommen werden, können zunächst die möglichen Farbkombinationen, die entstehen können, bestimmt werden: zweimal rot, zweimal grün und einmal blau; zweimal rot, einmal grün und zweimal blau; einmal rot, zweimal grün und zweimal blau. Die Teilergebnisse sind disjunkt zueinander und deswegen kann das Gesamtergebnis additiv zusammengesetzt werden. Es geht dreimal um die Frage, wie viele Permutationen es in einer Menge mit fünf Elementen gibt, von denen jeweils zwei ununterscheidbar sind:

$$\frac{5!}{2!2!1!} + \frac{5!}{2!1!2!} + \frac{5!}{1!2!2!} = 3 \cdot 5 \cdot 3 \cdot 2 = 90.$$

1.2.4. (n, k)-Kombination mit Wiederholung (4)

Sollen fünf Pizzen auf drei Kühlregale verteilt werden, so gibt es nachfolgende Anzahl an Möglichkeiten. Dabei steht $x|y|z$ für x Pizzen in Regal 1, y in Regal 2 und z Pizzen in Regal 3, $x + y + z = 5$.

$$\begin{aligned}
&5|0|0, 4|0|1, 4|1|0, 3|0|2, 3|2|0, 3|1|1, 2|0|3,\\
&2|3|0, 2|1|2, 2|2|1, 1|0|4, 1|4|0, 1|1|3, 1|3|1,\\
&1|2|2, 0|0|5, 0|5|0, 0|1|4, 0|4|1, 0|2|3, 0|3|2.
\end{aligned}$$

Insgesamt also gibt es 21 Möglichkeiten, die Pizzen zu verteilen.

Bei der ungeordneten Stichprobe ohne Zurücklegen kommt es nicht auf die Reihenfolge der k gezogenen Kugeln an. Aber es ist zu berücksichtigen, dass eine Kugel mehrfach gezogen werden darf.

Erläuterung

Um den Ergebnisraum $\Omega_{(4)}$ erhalten zu können, betrachtet man die Äquivalenzrelation in (3) auf der Menge $\Omega_{(1)}$, da es Wiederholungen geben darf. Es sei für die Untersuchung der Mächtigkeit von $\Omega_{(4)}$ eine Menge $A = \{1, 2, \ldots, n\}$ gegeben. Wird aus jeder Äquivalenzklasse ein Repräsentant $(a_1, \ldots, a_k)$ mit $a_1 \leq a_2 \leq \ldots \leq a_k$ bestimmt, so lässt sich $\Omega_{(4)}$ schreiben als

$$\Omega_{(4)} = \{(a_1, \ldots, a_k);\ a_i \in A,\ a_j \leq a_l \text{ für } j \leq l\}.$$

Die Abbildung $f : \Omega_{(4)} \to \tilde{\Omega}_{(3)}$ ordne jedem k-Tupel $(a_1, \ldots, a_k)$ ein k-Tupel $(\tilde{a}_1, \ldots, \tilde{a}_k)$ mittels $\tilde{a}_i = a_i + i - 1$ zu, d.h. $\tilde{a}_{k,\max} = n + k - 1$. Dabei sei $\tilde{\Omega}_{(3)} := \{(\tilde{a}_1, \ldots, \tilde{a}_k);\ \tilde{a}_i \in \{1, 2, \ldots, n-1+k\},\ \tilde{a}_j < \tilde{a}_l \text{ für } j < l\}$. Dann ist f bijektiv und es gilt $|\Omega_{(4)}| = \binom{n-1+k}{k}$.

f ist injektiv: Seien $a, b \in \Omega_{(4)}$ mit $a \neq b$ beliebig. Es gibt wenigstens einen Index i mit $a_i \neq b_i$. Sei i der kleinste dieser Indizes. Dann gilt $f(a_i) = a_i + i - 1 \neq b_i + i - 1 = f(b_i)$. Also ist f injektiv.

f ist surjektiv: Sei $(\tilde{a}_1, \ldots, \tilde{a}_k) \in \tilde{\Omega}_{(3)}$. Wegen $\tilde{a}_i = a_i + i - 1$ schreibe man $a_i = \tilde{a}_i - i + 1$ und bilde das k-Tupel

$$(\tilde{a}_1 - 1 + 1, \tilde{a}_2 - 2 + 1, \ldots, \tilde{a}_k - k + 1) = (\tilde{a}_1, \tilde{a}_2 - 1, \ldots, \tilde{a}_k - k + 1).$$

Wegen $\tilde{a}_1 = a_1 \geq 1$, $\tilde{a}_k \leq n - 1 + k$ d.h. $a_k \leq n$ und $a_i \leq a_{i+1} \Leftrightarrow \tilde{a}_i - i + 1 \leq \tilde{a}_{i+1} - (i+1) + 1 \Leftrightarrow \tilde{a}_i \leq \tilde{a}_{i+1} + 1 \Leftrightarrow \tilde{a}_i < \tilde{a}_{i+1}$ und $f(a_1, \ldots, a_k) = (\tilde{a}_1, \ldots, \tilde{a}_k)$ ist f wohldefiniert und surjektiv.

Zusammen ist f bijektiv. Damit sind beide Mengen gleichmächtig. $\tilde{\Omega}_{(3)}$ ist von der gleichen Struktur wie $\Omega_{(3)}$ und besitzt deshalb $\binom{n-1+k}{k}$ Elemente.

Die Abbildung f erzeugt eine aufeinanderfolgende Anordnung gleicher Ergebnisse und führt zwischen zwei verschiedenen Ergebnissen a_i und $a_{i+1} = a_i + 1$ durch den Index i einen Zwischenwert $a_i + i$ ein. Denn für $a_i < a_{i+1}$ gilt

$\tilde{a}_i = a_i + i - 1 < a_i + i < a_{i+1} + i = a_{i+1} + (i+1) - 1 = \tilde{a}_{i+1}$ mit $\tilde{A} = \{1, \ldots, n-1+k\}$.

Das Gesamtergebnis stammt aus der Menge

$$\tilde{\Omega}_{(4)} = \{(\tilde{a}_1, \ldots, \tilde{a}_k);\ \tilde{a}_i \in \tilde{A},\ \tilde{a}_j \leq \tilde{a}_l \text{ für } j \leq l\}$$

und es gilt $|\tilde{\Omega}_{(4)}| = \binom{n-1+k}{k}$ gemäß Abschnitt 1.2.3. Somit können die $k = 5$ Pizzen in $n = 3$ Kühlregale auf $\binom{3-1+5}{5} = \binom{7}{5} = 21$ verschiedene Arten einsortiert werden.

Beispiel 1.14

Aus einer Urne mit vier Kugeln $A = \{\text{rot, grün, blau, gelb}\}$ werden drei Kugeln mit Zurücklegen gezogen. Da die Reihenfolge im Gesamtergebnis keine Rolle spielen soll, werden die gezogenen Farben in alphabetischer Reihenfolge sortiert. Zwischen jeweils zwei Farben wird ein Trennstrich eingefügt und anschließend jede gezogene Kugel eingetragen, z. B.

Es gibt $n - 1$ Trennstriche und k Kugeln. Auf wie viele Arten lassen sich Kugeln und Trennstriche anordnen? Jede mögliche Ziehung ist jede mögliche Permutation der Kugeln und Trennstriche, insgesamt eine Anzahl $\binom{4-1+3}{3} = 20$ verschiedener Anordnungen.

Beispiel 1.15

Auf wie viele Arten lassen sich 20 Murmeln auf 3 Schachteln verteilen? Unter der Annahme, dass in jeder Schachtel genügend Platz vorhanden ist, ergeben sich $\binom{3-1+20}{20} = \frac{22 \cdot 21}{2} = 231$ verschiedene Möglichkeiten.

Beispiel 1.16: Bose-Einstein-Verteilung: Anzahl der Systemzustände

Man betrachte ein System bestehend aus k gleichartigen und unterscheidbaren Teilchen, die sich in einem Zustandsraum mit n möglichen Zuständen befinden. Dann lässt sich ein Zustand des Systems durch ein n-Tupel $(a_1, \ldots, a_n)$ mit $\sum_{i=1}^{n} a_i = k$ beschreiben. Der Zustandsraum ist die Menge aller solcher Tupel:

$$\Omega_{(4)} = \left\{ (a_1, \ldots, a_n); \sum_{i=1}^{n} a_i = k \right\}.$$

Es gibt $|\Omega_{(4)}| = \binom{n-1+k}{k}$ verschiedene Zustände.

2. Ideen der Wahrscheinlichkeitstheorie

2.1. Wahrscheinlichkeitsräume

Experimente,

- die nach einer bestimmten Vorschrift durchgeführt werden,
- die beliebig oft wiederholbar sind,
- deren Ausgang nicht vorhergesagt werden kann,

werden Zufallsexperimente genannt. Zu einem Zufallsexperiment gehören

- ein Ergebnisraum in Form einer Menge Ω,
- ein Elementarereignis $\omega \in \Omega$.
- Ereignisse.

Jedes Elementarereignis kann das Ergebnis des Zufallsexperiments sein. Ein Ereignis ist eine Menge $A \subseteq \Omega$ und tritt ein, wenn das Ergebnis ω des Zufallsexperiments in A enthalten ist, $\omega \in A$. Alle benötigten Ereignisse eines Zufallsexperiments werden zur Ereignismenge $\mathcal{F}$ zusammengefasst. Tabelle 2.1 zeigt einige oft benötigte Arten von Ereignissen.

Tabelle 2.1.: Beispiele für Ereignisse

Ereignis	Bezeichnung	Tore beim Fußball
Ω	Sicheres Ereignis	$\mathbb{N}_0$
$\emptyset$	Unmögliches Ereignis	$\emptyset$
$\bar{A} := \Omega \setminus A$	Das zu A komplementäre Ereignis	$A = \{0, 1, 2\}$, $\bar{A} = \{3, 4, 5, \ldots\}$
B mit $B \cap A = \emptyset$	Ein zu A disjunktes Ereignis	$A = \{1, 2\}$, $B = \{3, 4\}$
B mit $B \subseteq A$	Ein A implizierendes Ereignis	$A = \{0, 1, 2, 3, 4\}$, $B = \{0, 1, 2\}$

Die Sicherheit des Eintretens von Ereignissen wollen wir einschätzen. Dies soll über eine Funktion $\mathbb{P} : \mathcal{F} \to [0, 1]$ geschehen, an die wir gewisse Forderungen stellen wollen:

- $\mathbb{P}(\emptyset) = 0,\ \mathbb{P}(\Omega) = 1,\ \mathbb{P}(A) \geq 0$,
- $1 = \mathbb{P}(\Omega) = \mathbb{P}(A \cup \bar{A}) = \mathbb{P}(A) + \mathbb{P}(\bar{A})$,
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ für $A \cap B = \emptyset$,
- $\mathbb{P}\left(\bigcup_{i \in \mathbb{N}} A_i\right) = \sum_{i \in \mathbb{N}} \mathbb{P}(A_i)$ für paarweise disjunkte A_i.

- Falls $B \subseteq A$, soll $\mathbb{P}(A) = \mathbb{P}(B \cup (A \setminus B)) = \mathbb{P}(B) + \mathbb{P}(A \setminus B)$ und damit $\mathbb{P}(B) \leq \mathbb{P}(A)$ sein.

Das unmögliche Ereignis tritt sicher nie, das sichere Ereignis immer ein. Alle anderen Ereignisse sollen dazwischenliegende Bewertungen erhalten. Die Bewertungen zweier komplementärer Ereignisse sollen addiert Eins ergeben, da eines der beiden Ereignisse sicher eintritt. Dies soll noch verallgemeinert werden. Für zwei disjunkte Ereignisse sollen sich die addierten Bewertungen so verhalten, wie wenn die beiden Ereignisse zusammen betrachtet werden. Das soll nicht nur für zwei sondern immer dann gelten, wenn die Ereignisse abgezählt werden können. Die Bewertung eines Ereignisses, das Teilmenge eines anderen Ereignisses ist, soll entsprechend kleiner sein. Wenn eine Funktion diese Kriterien erfüllt, ordnet sie Ereignissen eine Wahrscheinlichkeit zu.

2.1.1. Das Maßproblem und Wahrscheinlichkeitsmaße

Leider gibt es für den wichtigen Standardfall $\Omega = \mathbb{R}$ und die Potenzmenge $\mathcal{F} = \mathcal{P}(\mathbb{R})$ keine solche Funktion $\mathbb{P}$, siehe [7]. Dennoch ist der Ansatz der Abbildung von Ereignissen in das Intervall $[0, 1]$ nachvollziehbar. Die geforderten Eigenschaften scheinen plausibel. Deshalb scheint es eine sinnvolle Idee zu sein, die Ereignismenge $\mathcal{F}$ so einzuschränken, dass die Eigenschaften erhalten bleiben. Eine geeignete Wahl für Ereignismengen sind so genannte σ-Algebren.

Definition 2.1: σ-Algebra

Ein Mengensystem $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ bestehend aus Teilmengen des Ergebnisraums Ω heißt σ-Algebra über Ω, wenn es die drei folgenden Eigenschaften erfüllt:

- $\Omega \in \mathcal{F}$,
- $\bar{A} \in \mathcal{F}$ für alle $A \in \mathcal{F}$,
- $\bigcup_{i \in \mathbb{N}} A_i \in \mathcal{F}$ für alle $A_i \in \mathcal{F}$.

Auf Basis einer σ-Algebra können wir eine den Ereignissen eines Zufallsexperiments Wahrscheinlichkeiten zuordnende Funktion definieren.

Definition 2.2: Wahrscheinlichkeitsmaß

Sei $\mathcal{F}$ eine σ-Algebra über einem Ergebnisraum Ω. Eine Funktion $\mathbb{P} : \mathcal{F} \to \mathbb{R}$, die jedem Ereignis A aus $\mathcal{F}$ eine reelle Zahl zuordnet, heißt Wahrscheinlichkeitsmaß auf $\mathcal{F}$, wenn die drei folgenden Axiome erfüllt sind:

- $\mathbb{P}(A) \geq 0$ für alle $A \in \mathcal{F}$,
- $\mathbb{P}(\Omega) = 1$,
- $\mathbb{P}\left(\bigcup_{i \in \mathbb{N}} A_i\right) = \sum_{i \in \mathbb{N}} \mathbb{P}(A_i)$ für paarweise disjunkte A_i.

Die Idee der Axiome der Wahrscheinlichkeitstheorie stammt vom sowjetischen Mathematiker Andrei Nikolajewitsch Kolmogorow. Aus den Axiomen lassen sich verschiedene Rechenregeln für Ereignisse folgern. Als Beispiel dient der folgende Satz.


Andrei N. Kolmogorov
1903-1987

Satz 2.3: Additionssatz für Ereignisse

Für zwei Ereignisse A und B gilt

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

Beweis.

Man betrachte folgende disjunkte Zerlegungen zweier Ereignisse A und B und deren Vereinigungsmenge:

$$\begin{array}{rcll} A & = & (A \cap B) \cup (A \cap \bar{B}) & \Rightarrow \quad \mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap \bar{B}) \\ B & = & (B \cap A) \cup (B \cap \bar{A}) & \Rightarrow \quad \mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}(B \cap \bar{A}) \\ A \cup B & = & (A \cap B) \cup (A \cap \bar{B}) \cup (B \cap \bar{A}) & \end{array}$$

Damit gilt

$$\begin{aligned} \mathbb{P}(A \cup B) &= \mathbb{P}(A \cap B) + \mathbb{P}(A \cap \bar{B}) + \mathbb{P}(B \cap \bar{A}) \\ &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B). \end{aligned}$$

□

Das Zufallsexperiment wird nun durch das Festlegen dreier Bestandteile modelliert. Dabei ist die Ergebnismenge der erste Baustein, auf den die anderen aufsetzen.

Definition 2.4: Wahrscheinlichkeitsraum, Träger

$(\Omega, \mathcal{F}, \mathbb{P})$ ist ein Wahrscheinlichkeitsraum, falls

- Ω ein Ergebnisraum,
- $\mathcal{F}$ eine σ-Algebra über Ω,
- $\mathbb{P}$ ein Wahrscheinlichkeitsmaß auf $\mathcal{F}$ ist.

Gibt es eine abzählbare Menge $T \in \mathcal{F}$ mit $\mathbb{P}(T) = 1$, so heißt T der Träger von $\mathbb{P}$ und der Wahrscheinlichkeitsraum diskret, ansonsten stetig. Das Paar $(\Omega, \mathcal{F})$ heißt Messraum.

Bemerkung.
(1) Das Wahrscheinlichkeitsmaß eines diskreten Wahrscheinlichkeitsraums heißt diskretes Wahrscheinlichkeitsmaß.
(2) Durch $f : T \to [0,1]$ mit $\mathbb{P}(\{\omega\}) = f(\omega)$, $\sum_{\omega \in T} f(\omega) = 1$ wird das diskrete Wahrscheinlichkeitsmaß vollständig charakterisiert. Dies wird als Zähldichte oder diskrete Dichte bezeichnet.

(3) Ein Ereignis $A \in \mathcal{F}$ heißt $\mathbb{P}$-fast sicher, falls $\mathbb{P}(A) = 1$. Das bedeutet jedoch nicht, dass das Ereignis auch eintritt. Ist $\mathbb{P}(A) = 0$, heißt A eine $\mathbb{P}$-Nullmenge.

2.1.2. Diskrete Wahrscheinlichkeitsräume

In Anwendungen tritt immer wieder die Situation auf, dass zwar nicht das exakte Ergebnis eines Experiments vorhergesagt werden kann, alle möglichen Ergebnisse endlicher Zahl aber näherungsweise gleichwahrscheinlich sind. Die Modellierung gleichwahrscheinlicher Ergebnisse ist dann oftmals eine zufriedenstellende Beschreibung der Wirklichkeit.


Pierre-Simon Laplace
1749-1827

Definition 2.5

Ein Zufallsexperiment mit endlich vielen, gleichwahrscheinlichen Elementarereignissen heißt Laplace-Experiment.

Laplace-Experimente werden durch diskrete Wahrscheinlichkeitsräume modelliert. Dabei wird als σ-Algebra über dem Ergebnisraum Ω oft deren Potenzmenge $\mathcal{P}(\Omega)$ benutzt. Ein sehr bekanntes Beispiel für ein Laplace-Experiment ist das Werfen eines fairen Würfels. Das Ergebnis eines Wurfs ist die Augenzahl.

Beispiel 2.6: fairer Würfel

Das Werfen eines Würfels ist ein sehr komplexer Vorgang, der sich nicht exakt beschreiben lässt. Wir nehmen nun an, dass beim Werfen eines Würfels jede mögliche Augenzahl gleichwahrscheinlich ist. Bei sechs möglichen Augenzahlen ergibt sich jeweils die Wahrscheinlichkeit $1/6$. Die möglichen Ergebnisse fassen wir in der Menge $\Omega = \{1, 2, 3, 4, 5, 6\}$ zusammen und schreiben $p_\omega = 1/6$ für jedes $\omega \in \Omega$. Der dazugehörige Wahrscheinlichkeitsraum lautet $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ mit

$$\mathbb{P} : \mathcal{P}(\Omega) \to [0, 1], A \mapsto \mathbb{P}(A) := \sum_{\omega \in A} p_\omega.$$

Wegen $\mathbb{P}(\{\omega\}) = p_\omega$ gilt $\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{|A|}{6}$. Die Funktion $f : \Omega \to [0, 1]$, $\omega \mapsto p_w$ ist hierbei eine Zähldichte.

Sind wir im letzten Beispiel 2.6 nur daran interessiert, ob die Augenzahl gerade oder ungerade ist, können wir statt der Potenzmenge die Menge $\mathcal{F} = \{\emptyset, \Omega, \{1, 3, 5\}, \{2, 4, 6\}\}$ als Ereignismenge wählen. Das Wahrscheinlichkeitsmaß wird dabei lediglich auf $\mathcal{F}$ eingeschränkt.

Beispiel 2.7: einfaches Urnenmodell

In einer Urne befinden sich $n \in \mathbb{N}$ durchnummerierte Kugeln. Das Ziehen einer Kugel aus der Urne lässt sich als Laplace-Experiment modellieren und liefert für jede Kugel $\omega \in \Omega = \{1, 2, \ldots, n\}$ die Wahrscheinlichkeit $p_\omega = 1/n$. Mit dem entsprechend zum fairen Würfel modellierten Wahrscheinlichkeitsraum gilt $\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{|A|}{n}$.

Die beiden Mengen Ω aus den Beispielen 2.6 und 2.7 besitzen $|\Omega| = 6$ bzw. $|\Omega| = n$ Elemente. Sehr häufig wird die Mächtigkeit einer Menge oder einer Teilmenge der Menge für Fragestellungen in der Wahrscheinlichkeitstheorie oder der Statistik benötigt.

Definition 2.8

Ein Laplace-Experiment mit $|\Omega| = n$ und $p_\omega = 1/n$ sei durch den Wahrscheinlichkeitsraum $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ mit

$$\mathbb{P} : \mathcal{P}(\Omega) \to [0,1], A \mapsto \mathbb{P}(A) := \sum_{\omega \in A} p_\omega.$$

modelliert. Die Wahrscheinlichkeit

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

wird als Laplace-Wahrscheinlichkeit bezeichnet. $|A|$ ist die Anzahl günstiger Fälle für das Ereignis A und $|\Omega|$ ist die Gesamtzahl möglicher Fälle.

Beispiel 2.9

Beim Kniffel werden fünf gleiche Würfel gleichzeitig geworfen. Eine große Straße erhält man, wenn entweder $\{1,2,3,4,5\}$ oder $\{2,3,4,5,6\}$ geworfen wird. Wie groß ist die Wahrscheinlichkeit dafür? Dem Elementarereignis $A = \{(1,1,1,1,1)\}$ in der Menge der $(6,5)$-Kombinationen mit Wiederholung entspricht genau das Elementarereignis $\{(1,1,1,1,1)\}$ in der Menge der $(6,5)$-Permutationen mit Wiederholung. Im Gegensatz dazu entsprechen dem Elementarereignis $B = \{(1,2,3,4,5)\}$ in der Menge der $(6,5)$-Kombinationen mit Wiederholung aber $5! = 120$ Elementarereignisse in der Menge der $(6,5)$-Permutationen mit Wiederholung, beispielsweise $\{(1,2,3,5,4)\}$ oder $\{(5,3,4,2,1)\}$, vereinigt zum Ereignis $\tilde{B}$. Im dazugehörigen Laplace-Experiment für $(6,5)$-Permutationen mit Wiederholung und $\Omega_{(1)} = \{1,2,3,4,5,6\}^5$ gilt

$$\mathbb{P}_{\Omega_{(1)}}(B) = \frac{1}{6^5} \neq \frac{5!}{6^5} = \mathbb{P}_{\Omega_{(1)}}(\tilde{B}).$$

Die Modellierung als Laplace-Experiment auf Grundlage von $\Omega_{(4)}$ und der $(6,5)$-Kombinationen mit Wiederholung würde beiden Elementarereignissen die gleiche Wahrscheinlichkeit zuweisen, was dem ersten Modell widerspricht. $\Omega_{(4)}$ eignet sich daher nicht zur Modellierung eines Laplace-Raumes.

Man modelliert durch $\Omega_{(1)} = \{1,2,3,4,5,6\}^5$ ein Laplace-Experiment. Die Wahrscheinlichkeit für ein Elementarereignis ist dann $p_\omega = \frac{1}{6^5}$ für jedes $\omega \in \Omega_{(1)}$. Die Anzahl für die Fragestellung „günstiger Fälle“ bekommt man dadurch, dass sowohl die Variante $\{1,2,3,4,5\}$ als auch die Variante $\{2,3,4,5,6\}$ auf $5! = 120$ verschiede-

ne Arten erzielt werden kann. Mit den disjunkten Ereignissen

$$\begin{aligned} A_1 &= \{(a_1,a_2,a_3,a_4,a_5);\ a_i \in \{1,2,3,4,5\},\ a_j \neq a_l \text{ für } j \neq l\} \\ A_2 &= \{(a_1,a_2,a_3,a_4,a_5);\ a_i \in \{2,3,4,5,6\},\ a_j \neq a_l \text{ für } j \neq l\} \end{aligned}$$

und der Additionsregel aus Abschnitt 1.2.3 gilt

$$\mathbb{P}(A_1 \cup A_2) = \frac{5!}{6^5} + \frac{5!}{6^5} = \frac{2 \cdot 5!}{6^5} = \frac{240}{6^5} = \frac{5}{162} = 0.031.$$

Beispiel 2.10

Wie wahrscheinlich ist es, beim Zahlenlotto vier richtige Zahlen zu haben? Beim Zahlenlotto geht es um das Ziehen ohne Zurücklegen. Bei Berücksichtigung der Reihenfolge wird $\Omega_{(2)}$ zugrundegelegt. Ist die Reihenfolge ohne Bedeutung nimmt man $\Omega_{(3)}$. Einem Elementarereignis $A = \{(a_1,a_2,a_3,a_4,a_5,a_6)\} \in \Omega_{(3)}$ bei den $(49,6)$-Kombinationen ohne Wiederholung entsprechen bei den $(49,6)$- Permutationen ohne Wiederholung $6! = 720$ Elementarereignisse, vereinigt zum Ereignis $\tilde{B}$. Für ein Laplace-Experiment auf Basis von $\Omega_{(2)}$ gilt

$$\mathbb{P}_{\Omega_{(2)}}(\tilde{B}) = \frac{6!}{49 \cdot 48 \cdot \ldots \cdot 44} = \frac{1}{\binom{49}{6}}.$$

Mit dem Ergebnisraum $\Omega_{(3)}$ erhält man das identische Resultat:

$$\mathbb{P}_{\Omega_{(3)}}(A) = \frac{1}{\binom{49}{6}}.$$

Hier sind beide Ansätze möglich. Der Ansatz mit $\Omega_{(2)}$ ist dann sinnvoll, wenn es um Positionen einzelner Kugeln bei der Ziehung geht. Deswegen benutzt man für die gegebene Fragestellung $\Omega_{(3)}$. Was ist die Anzahl günstiger Ereignisse? Aus der Menge $\Omega_{(3)}$ ist ein Element C (der Tipp) ausgwählt. Das heißt, dass von 49 Zahlen sechs gewählt und 43 nicht gewählt sind. Von den sechs Zahlen sollen vier tatsächlich ohne Wiederholung gezogen werden, egal in welcher Reihenfolge. Aus den verbleibenden 43 Kugeln sollen zwei tatsächlich ohne Wiederholung und ohne Beachtung der Reihenfolge gezogen werden. Somit ergibt sich die hypergeometrische Verteilung

$$\mathbb{P}_{\Omega_{(3)}}(C) = \frac{\binom{6}{4} \cdot \binom{43}{2}}{\binom{49}{6}} = \frac{3 \cdot 5! \cdot 5 \cdot 3 \cdot 43 \cdot 42}{49 \cdot 48 \cdot 47 \cdot 46 \cdot 45 \cdot 44} = \frac{645}{665896} = 0.00097$$

Beispiel 2.11: Bose-Einstein-Verteilung

Werden die Systemzustände $(a_1,\ldots,a_n)$ eines Zustandsraums mit k Teilchen und n möglichen Zuständen (siehe Beispiel 1.16) und $\sum_{i=1}^{n} a_i = k$ als gleichwahrscheinlich angenommen, gilt $\mathbb{P}(\{(a_1,\ldots,a_n)\}) = 1/\binom{n-1+k}{k}$.

Beispiel 2.12: Poissonverteilung

Die Anzahl erzielter Tore bei einem Fußballspiel kann über den Wahrscheinlichkeitsraum $(\mathbb{N}_0, \mathcal{P}(\mathbb{N}_0), \mathbb{P})$ mit

$$\mathbb{P} : \mathcal{P}(\mathbb{N}_0) \to [0,1],\ A \mapsto \mathbb{P}(A) := \sum_{\omega \in A} e^{-\lambda} \cdot \frac{\lambda^\omega}{\omega!}$$

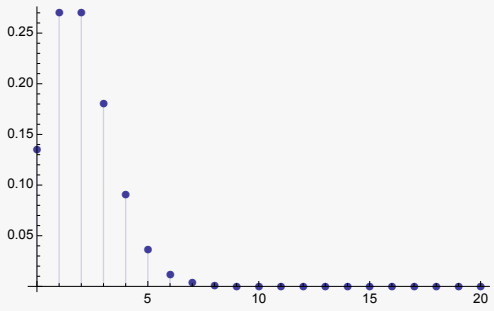
mit einem Parameter $\lambda > 0$, der die erwartete Trefferzahl beschreibt, modelliert werden.

Im Beispiel wird die Zähldichte $\mathbb{P}(\{\omega\}) = f(\omega) := e^{-\lambda} \cdot \frac{\lambda^\omega}{\omega!}$ benutzt. Das diskrete Wahrscheinlichkeitsmaß im Beispiel 2.12 heißt Poisson-Verteilung

Siméon D. Poisson
1781-1840

Beispiel 2.13

Annahme: Die erwartete Anzahl erzielter Treffer bei einem Fußballspiel ist 2. Wir setzen $\lambda = 2$ und betrachten die dazugehörige Zähldichte der Poisson-Verteilung.



$$\mathbb{P}(\{0,1,2,3\}) = \sum_{\omega=0}^{3} e^{-2} \cdot \frac{2^\omega}{\omega!} = \frac{19}{3e^2} = 0.86.$$

Sei T Träger von $\mathbb{P}$ eines diskreten Wahrscheinlichkeitsraums, dessen Elemente geordnet sind. Auf Grundlage einer diskreten Dichte $f : T \to [0,1]$ lässt sich eine Funktion

$$F_\mathbb{P} : T \to [0,1],\ t \mapsto F_\mathbb{P}(t) := \sum_{\substack{\omega \le t \\ \omega \in T}} f(\omega)$$

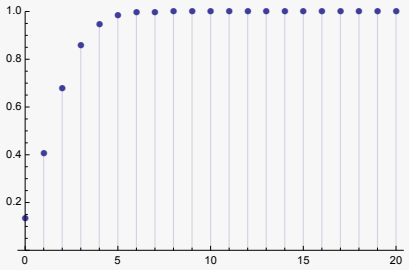
definieren, bei der Wahrscheinlichkeiten aufsummiert werden. In diesem Fall sprechen wir von einer diskreten Verteilungsfunktion. Für die im Beispiel berechnete Wahrscheinlichkeit gilt

$$\mathbb{P}(\{0,1,2,3\}) = F_\mathbb{P}(3) = 0.86.$$

Wegen $\mathbb{P}(T) = 1$ gilt $0 \le F_\mathbb{P}(t) \le 1$ für jedes $t \in T$. Zudem muss $F_\mathbb{P}$ monoton wachsend sein, da $\mathbb{P}$ ein Wahrscheinlichkeitsmaß ist und somit nur nicht-negative Werte annimmt.

Beispiel 2.14

Die diskrete Verteilungsfunktion zur Poisson-Verteilung im letzten Beispiel sieht folgendermaßen aus:



Die Differenz $F_{\mathbb{P}}(t) - F_{\mathbb{P}}(t-1)$ entspricht der Wahrscheinlichkeit $\mathbb{P}(\{t\})$ und führt zu den sichtbaren Wertesprüngen im Graphen der diskreten Verteilungsfunktion.

Eine diskrete Verteilungsfunktion lässt sich für $T \subset \mathbb{R}$ ohne Einschränkung auf ganz $\mathbb{R}$ definieren, $F_{\mathbb{P}} : \mathbb{R} \to [0,1]$. Gibt es nur endlich viele Sprünge in einer solchen Verteilungsfunktion, liegt eine Treppenfunktion vor. Die einfachste Treppenfunktion erhalten wir, wenn es nur einen Sprung gibt. Ein diskreter Wahrscheinlichkeitsraum $(T, \mathcal{F}, \mathbb{P})$ genügt der Einpunktverteilung in $a \in T$, wenn für die Verteilungsfunktion

$$F_{\mathbb{P}}(x) = \begin{cases} 0, & x < a, \\ 1, & x \geq a, \end{cases}$$

für alle $x \in T$ gilt. Die nachfolgend aufgeführten beiden diskreten Verteilungen auf Basis einer speziellen Ergebnismenge Ω spielen im Weiteren eine wichtige Rolle.


Jakob I Bernoulli
1655-1705

Definition 2.15: Bernoulli-Verteilung

Sei $\Omega = \{0,1\}$ und $0 < p < 1$. Ein diskreter Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P})$ genügt der Bernoulli-Verteilung, wenn für die Verteilungsfunktion

$$F_{\mathbb{P}} : \mathbb{R} \to [0,1],\ x \mapsto F_{\mathbb{P}}(x) = \begin{cases} 0, & x < 0, \\ 1-p, & 0 \leq x < 1, \\ 1, & x \geq 1, \end{cases}$$

gilt.

Definition 2.16: Diskrete Gleichverteilung

Sei $\Omega = \{\omega_1, \ldots, \omega_k\} \subset \mathbb{R}$ mit $\omega_i < \omega_j$ für $i < j$. Ein diskreter Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P})$ genügt der diskreten Gleichverteilung, wenn für die Verteilungsfunktion gilt:

$$F_{\mathbb{P}} : \mathbb{R} \to [0,1],\ x \mapsto F_{\mathbb{P}}(x) = \frac{|\{i;\ \omega_i \leq x\}|}{k}.$$

Das Urnenmodell basiert auf der Annahme, dass jede Kugel mit gleicher Wahrscheinlichkeit gezogen werden kann. Je nach Vorgang entsteht dann aus der diskreten Gleichverteilung eine andere diskrete Verteilung.

2.1.3. Stetige Wahrscheinlichkeitsräume


Emile Borel
1871-1956

Kehren wir zu der Frage aus dem Maßproblem zurück, welche Ereignismenge für den Ergebnisraum $\mathbb{R}$ die Bildung eines Wahrscheinlichkeitsmaßes ermöglicht. Dazu bedarf es einer Menge, die alle „interessanten“ Mengen wie z.B. Einpunktmengen, beliebige Intervalle oder abzählbare Vereinigungen oder endliche Durchschnitte von Intervallen enthält. Solche Mengen werden oft in Anwendungen benötigt. Die Borelsche σ-Algebra $\mathcal{B}$ ist die Ereignismenge der reellen Zahlen und erfüllt diese Anforderungen. Ein Wahrscheinlichkeitsmaß $\mathbb{P} : \mathcal{B} \to [0,1]$ bestimmt z.B.

- $\mathbb{P}(]-\infty, a])$,

und mit Hilfe der disjunkten Zerlegung $]-\infty, b] =]-\infty, a] \cup]a, b]$ (für $a < b$) und des Additionssatzes

- $\mathbb{P}(]a, b]) = \mathbb{P}(]-\infty, b] \setminus]-\infty, a]) = \mathbb{P}(]-\infty, b]) - \mathbb{P}(]-\infty, a])$,

d.h. $]-\infty, a],]a, b],]-\infty, b] \in \mathcal{B}$ für $a, b \in \mathbb{R}$. Ist der Ergebnisraum Ω lediglich eine überabzählbare Teilmenge der reellen Zahlen, so lässt sich die Borelsche σ-Algebra durch $\mathcal{B}(\mathbb{R} \cap \Omega)$ beschreiben. Analog zu diskreten Wahrscheinlichkeitsmaßen heißt das Wahrscheinlichkeitsmaß eines stetigen Wahrscheinlichkeitsraums stetiges Wahrscheinlichkeitsmaß. Ebenso gibt es wie im diskreten Fall mit der diskreten Dichte eine stetige Dichte.

Definition 2.17: Verteilungsfunktion, Dichte

Ist $\mathbb{P} : \mathcal{B} \to [0,1]$ ein Wahrscheinlichkeitsmaß auf $\mathbb{R}$ mit der Borelschen σ-Algebra $\mathcal{B}$, so heißt

$$F_{\mathbb{P}} : \mathbb{R} \to [0,1],\ x \mapsto F_{\mathbb{P}}(x) := \mathbb{P}(]-\infty, x])$$

die Verteilungsfunktion von $\mathbb{P}$. Eine Funktion $f : \mathbb{R} \to \mathbb{R}_0^+$, für die

$$\int_{-\infty}^{\infty} f(x)dx = 1 \text{ und } F_{\mathbb{P}}(x) = \int_{-\infty}^{x} f(x)dx$$

gilt, heißt stetige Dichte.

Bemerkung.
(1) $F_{\mathbb{P}}$ ist monoton wachsend, rechtsseitig stetig und es gelten die Grenzwerte

$$\lim_{x \to \infty} F_{\mathbb{P}}(x) = 1 \text{ und } \lim_{x \to -\infty} F_{\mathbb{P}}(x) = 0.$$

(2) Oft interessieren uns lediglich die Verteilungsfunktion oder die stetige Dichte und nicht das eigentliche Wahrscheinlichkeitsmaß. Dann schreiben wir F anstelle von $F_{\mathbb{P}}$.
(3) Ist der Träger einer diskreten Dichte eine Teilmenge der reellen Zahlen, so lässt sich die dazugehörige diskrete Verteilungsfunktion durch eine Verteilungsfunktion repräsentieren. Ein Beispiel dafür ist die in Definition 2.15 eingeführte Verteilungsfunktion für die Bernoulli-Verteilung.

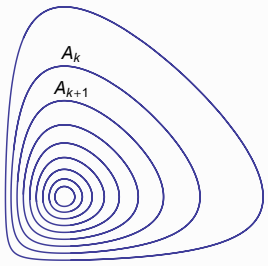
Satz 2.18

Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Ist $A_1 \supset A_2 \supset \ldots$ eine monoton fallende Folge $(A_k)_{k\in\mathbb{N}}$ von Ereignissen in $\mathcal{F}$, und $A = \lim\limits_{k\to\infty} A_k = \bigcap\limits_{k\in\mathbb{N}} A_k$ deren Durchschnitt, so gilt

$$\mathbb{P}(A) = \lim_{k\to\infty} \mathbb{P}(A_k)$$

Beweis.

Für jedes k ist A_k die disjunkte Vereinigung von A und den Differenzmengen $D_i = A_i \setminus A_{i+1} \in \mathcal{F}$ mit $i = k, k+1, \ldots$.



Damit gilt $\mathbb{P}(A_k) = \mathbb{P}(A) + \sum\limits_{i=k}^{\infty} \mathbb{P}(D_i)$. Da die D_i disjunkt sind, konvergiert die Reihe $\sum\limits_{i=1}^{\infty} \mathbb{P}(D_i)$ wegen

$$\sum_{i=1}^{\infty} \mathbb{P}(D_i) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} D_i\right) \text{ und } \bigcup_{i=1}^{\infty} D_i \in \mathcal{F}.$$

Aus der Konvergenz der Reihe folgt, dass die Folge $(\mathbb{P}(D_i))_{i\in\mathbb{N}}$ eine Nullfolge ist[a] und so $\sum\limits_{i=k}^{\infty} \mathbb{P}(D_i) \overset{k\to\infty}{\longrightarrow} 0$. Damit gilt

$$\lim_{k\to\infty} \mathbb{P}(A_k) = \mathbb{P}(A) + \lim_{k\to\infty} \sum_{i=k}^{\infty} \mathbb{P}(D_i) = \mathbb{P}(A).$$

□

[a] vgl. [4], S. 60

Eine stetige Dichte besitzt nicht dieselben Eigenschaften wie eine diskrete Dichte. Ist $(a_k)_{k\in\mathbb{N}}$ eine monoton wachsende Folge mit $a_k < b$ und $\lim\limits_{k\to\infty} a_k = b$, so ist $\{b\} = \bigcap\limits_{k=1}^{\infty}]a_k, b]$. Somit folgt mit Satz 2.18

$$\begin{aligned} \mathbb{P}(]a_k, b]) &= \int_{-\infty}^{b} f(x)dx - \int_{-\infty}^{a_k} f(x)dx = \int_{a_k}^{b} f(x)dx \\ \mathbb{P}(\{b\}) &= \lim_{k\to\infty} \mathbb{P}(]a_k, b]) = 0 \end{aligned}$$

Jedes Elementarereignis besitzt die Wahrscheinlichkeit Null und jedes Ereignis $\Omega \setminus \{b\}$ ist $\mathbb{P}$-fast sicher. Es lassen sich damit auch Wahrscheinlichkeiten für z.B. abgeschlossene Intervalle bestimmen:

$$\mathbb{P}([a,b]) = \mathbb{P}(\{a\} \cup]a,b]) = \mathbb{P}(\{a\}) + \mathbb{P}(]a,b]) = \mathbb{P}(]a,b]).$$

Mit einer dazugehörigen Verteilungsfunktion gilt dann

$$\mathbb{P}([a,b]) = \mathbb{P}(]a,b]) = \mathbb{P}(]-\infty,b]) - \mathbb{P}(]-\infty,a]) = F_{\mathbb{P}}(b) - F_{\mathbb{P}}(a).$$

Definition 2.19: Stetige Gleichverteilung

Ein stetiger Wahrscheinlichkeitsraum mit $\Omega = \mathbb{R}$ genügt der stetigen Gleichverteilung auf $[a,b] \subset \mathbb{R}$, wenn für die Verteilungsfunktion

$$F_{\mathbb{P}}(x) = \begin{cases} 0, & x < a, \\ \frac{x-a}{b-a}, & x \in [a,b], \\ 1, & x > b, \end{cases}$$

gilt.

Sei $[c,d] \subseteq [a,b]$. Für die stetige Gleichverteilung gilt dann $\mathbb{P}([c,d]) = F_{\mathbb{P}}(d) - F_{\mathbb{P}}(c) = \frac{d-c}{b-a}$.

Beispiel 2.20: Normalverteilung

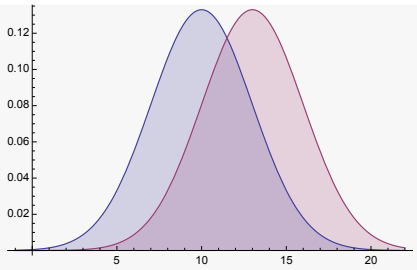
Eine der wichtigsten stetigen Dichten ist durch die Normalverteilung über

$$f : \mathbb{R} \to \mathbb{R}^+,\ x \mapsto f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\cdot\frac{(x-\mu)^2}{\sigma^2}}$$

mit Parametern $\mu \in \mathbb{R}$ und $\sigma^2 \in \mathbb{R}^+$ gegeben. Für $\mu = 0$ und $\sigma^2 = 1$ sprechen wir von der Standardnormalverteilung. Für die Verteilungsfunktion $F_{(\mu,\sigma^2)} := F_{\mathbb{P}}$ gibt es lediglich Näherungswerte. Es sei $\Phi = F_{(0,1)}$. Eine beliebige Normalverteilung lässt sich aus der Standardnormalverteilung mittels der Substitution

$$F_{(\mu,\sigma^2)}(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\cdot\frac{(x-\mu)^2}{\sigma^2}}\, dx \overset{u=\frac{x-\mu}{\sigma}}{=} \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\cdot u^2}\, du = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

gewinnen. Das Bild zeigt den Graphen der Normalverteilung für die Parameter $(\mu,\sigma^2) = (10,9)$ bzw. $(\mu,\sigma^2) = (13,9)$.



Eine Verteilungsfunktion $F_{\mathbb{P}} : M \to [0,1]$ hilft uns dabei, verschiedene Fragestellungen zu beantworten. Neben dem Wert von $F_{\mathbb{P}}$ an verschiedenen Stellen $x \in M$, z.B. für die Standardnormalverteilung

$$\Phi(1) = 0.8413, \Phi(-1) = 0.1587 \Rightarrow \Phi(1) - \Phi(-1) = 0.6827, \tag{2.1}$$

lässt sich umgekehrt die Frage stellen, für welches $x \in M$ denn $F_{\mathbb{P}}(x) = \alpha$ für $0 < \alpha < 1$ gilt. Da es ein solches x nicht geben muss oder es nicht eindeutig bestimmt sein muss, nennen wir

$$F_{\mathbb{P}}^{-1} : (0,1) \to M,\ \alpha \mapsto F_{\mathbb{P}}^{-1}(\alpha) := \inf\{x \in M;\ F_{\mathbb{P}}(x) \geq \alpha\} \tag{2.2}$$

die Quantilfunktion von $\mathbb{P}$ und $F_{\mathbb{P}}^{-1}(\alpha)$ das α-Quantil von $\mathbb{P}$. Bei der Standardnormalverteilung wird ein beliebiges α-Quantil oft mit $z_\alpha := \Phi^{-1}(\alpha)$ abgekürzt.

Beispiel 2.21: Median

Ein besonders wichtiges α-Quantil ist der Median $F_{\mathbb{P}}^{-1}(0.5)$. Für die Standardnormalverteilung gilt $z_{0.5} = 0$. Für die allgemeine Normalverteilung ist

$$F_{(\mu,\sigma^2)}(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) = 0.5 \Leftrightarrow \frac{x-\mu}{\sigma} = 0 \Leftrightarrow x = \mu.$$

Für $x = \mu$ erhalten wir den maximalen Wert der Dichtefunktion der Normalverteilung. Denn, da die Dichtefunktion differenzierbar ist, folgt

$$\begin{aligned}
\frac{d}{dx}\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\cdot\frac{(x-\mu)^2}{\sigma^2}}\right) &= -\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\cdot\frac{(x-\mu)^2}{\sigma^2}} \cdot \frac{(x-\mu)}{\sigma^2} = 0 \Leftrightarrow x = \mu, \\
\frac{d^2}{dx^2}\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\cdot\frac{(x-\mu)^2}{\sigma^2}}\right) &= \frac{1}{\sigma^3\sqrt{2\pi}} e^{-\frac{1}{2}\cdot\frac{(x-\mu)^2}{\sigma^2}} \cdot \left(\frac{(x-\mu)^2}{\sigma^2} - 1\right) \overset{x=\mu}{<} 0
\end{aligned}$$

Der maximale endliche Wert - falls es ihn gibt - einer diskreten oder stetigen Dichte heißt Modalwert, der oder die dazugehörigen Werte der Definitionsmenge der Dichte heißen Modus. Der Modalwert der Normalverteilung ist $f(\mu) = \frac{1}{\sigma\sqrt{2\pi}}$, das 0.95-Quantil der Standardnormalverteilung lautet $z_{0.95} = \Phi^{-1}(0.95) = 1.6449$.

2.2. Bedingte Wahrscheinlichkeiten und Unabhängigkeit

Bei Zufallsexperimenten spielt es oftmals eine Rolle, mit welcher Wahrscheinlichkeit ein Ereignis unter der Bedingung eintritt, dass ein anderes Ereignis eingetreten ist.

Beispiel 2.22

Von Steinsalzkörnern[a] werde der Korndurchmesser D (in Millimetern [mm]) und die Druckfestigkeit F (in Newton pro Quadratmeter $[N/mm^2]$) erfasst. Mögliche Ereignisse in dieser Situation könnten $A = \{(d, f);\ d \in D,\ f \in F,\ f < 40\ [N/mm^2]\}$ oder $B = \{(d, f);\ d \in D,\ f \in F,\ d < 3\ [\text{mm}]\}$ sein. Wie wahrscheinlich ist es, dass unter der Bedingung des Eintretens des Ereignisses B das Ereignis A eintritt?

[a] Beispiel aus [9]

Beispiel 2.23

Das Werfen zweier Würfel kann durch den Ergebnisraum $\Omega_{(1)} = \{(i, j);\ i, j = 1, \ldots, 6\}$ beschrieben werden. Ist $C \subseteq \Omega_{(1)}$, so wird $\mathbb{P}(C)$ durch die Laplace-Wahrscheinlichkeit $\mathbb{P}(C) = \frac{|C|}{36}$ bestimmt. Dem Ereignis „Augensumme größer als neun“ entspricht die Menge

$$A = \{(4, 6), (5, 6), (6, 6), (5, 5), (6, 5), (6, 4)\}$$

und dem Ereignis „beide Würfel zeigen dieselbe Augenzahl“ die Menge

$$B = \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\}.$$

Unter der Bedingung, dass das Ergebnis in B liegt, wie wahrscheinlich ist es, dass das Ergebnis auch in A liegt?[a]

[a] Beispiel aus [12]

Sei $\Omega = \{a_1, \ldots, a_n\}$, $|\Omega| = n$, ein Ergebnisraum mit den Elementarereignissen $\{a_i\}$ und zudem $|A| = n_A$ bzw. $|B| = n_B$ für zwei Ereignisse A und B. Es sei weiter $|A \cap B| = n_{AB}$ die Anzahl Elementarereignisse, die sowohl zu A als auch zu B gehören. Auf Basis von Laplace-Wahrscheinlichkeiten ist zunächst

$$\mathbb{P}(A) = \frac{n_A}{n},\ \mathbb{P}(B) = \frac{n_B}{n} \text{ und } \mathbb{P}(A \cap B) = \frac{n_{AB}}{n}.$$

Nach Durchführung des Experiments sei bekannt, dass B eingetreten ist. Der zu betrachtende Ergebnisraum verringert sich auf alle Elemente von B. Nur noch diejenigen Elementarereignisse n_{AB}, die sowohl zu A als auch zu B gehören, führen zum Eintreten des Ereignisses A:

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

Dies führt zu folgender Definition.

Definition 2.24: Bedingte Wahrscheinlichkeit

Sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein beliebiger Wahrscheinlichkeitsraum. Sind $A, B \in \mathcal{F}$ beliebige Ereignisse mit $\mathbb{P}(B) > 0$, so heißt

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{2.3}$$

die bedingte Wahrscheinlichkeit von A unter B.

Beispiel 2.25

Beim Beispiel 2.23 mit den Würfeln hat man zunächst $\mathbb{P}(A) = \frac{1}{6}$ und $\mathbb{P}(B) = \frac{1}{6}$. Das Ereignis $A \cap B = \{(5,5),(6,6)\}$ hat die Wahrscheinlichkeit $\mathbb{P}(A \cap B) = \frac{2}{36} = \frac{1}{18}$. Unter der Bedingung, dass das Ereignis B eingetreten ist, ergibt sich die Wahrscheinlichkeit, dass das Ereignis A eintritt zu

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{1/18}{1/6} = \frac{6}{18} = \frac{1}{3}.$$

Angenommen, beide Ereignisse A und B sind eingetreten, wie wahrscheinlich ist es, dass zweimal die Sechs gewürfelt wurde?

$$\mathbb{P}(\{(6,6)\}|A \cap B) = \frac{\mathbb{P}(\{(6,6)\} \cap (A \cap B))}{\mathbb{P}(A \cap B)} = \frac{\mathbb{P}(\{(6,6)\})}{\mathbb{P}(\{(5,5),(6,6)\})} = \frac{1}{2}.$$

Eine Erweiterung bedingter Wahrscheinlichkeiten auf n Ereignisse ist folgendermaßen möglich: Zunächst formt man die Gleichung (2.3) etwas um zu

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

Dann betrachtet man drei Ereignisse $A_1, A_2, A_3 \in \mathcal{F}$ mit $\mathbb{P}(A_1 \cap A_2) > 0$ und erhält die bedingte Wahrscheinlichkeit

$$\begin{aligned} \mathbb{P}(A_3|A_1 \cap A_2) \cdot \mathbb{P}(A_1 \cap A_2) &= \mathbb{P}(A_3|A_1 \cap A_2) \cdot \mathbb{P}(A_2|A_1) \cdot \mathbb{P}(A_1) \\ &= \mathbb{P}(A_3 \cap (A_1 \cap A_2)). \end{aligned}$$

Iterativ setzt man fort mit $A_1, A_2, \ldots, A_n \in \mathcal{F}$ und

$$\mathbb{P}(A_1) > \mathbb{P}(A_1 \cap A_2) > \ldots > \mathbb{P}(A_1 \cap A_2 \cap \ldots \cap A_{n-2}) > \mathbb{P}(A_1 \cap A_2 \cap \ldots \cap A_{n-1}) > 0.$$

Satz 2.26: Multiplikationssatz

In einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P})$ gilt für Ereignisse $A_1, \ldots, A_n \in \mathcal{F}$

$$\mathbb{P}(A_1 \cap \ldots \cap A_n) = \mathbb{P}(A_n|A_1 \cap \ldots \cap A_{n-1}) \cdot \mathbb{P}(A_{n-1}|A_1 \cap \ldots \cap A_{n-2}) \cdot \ldots \cdot \mathbb{P}(A_2|A_1) \cdot \mathbb{P}(A_1).$$

Beweis.

$$\begin{aligned}
\mathbb{P}(A_1 \cap \ldots \cap A_n) &= \mathbb{P}(A_n \cap (A_1 \cap \ldots \cap A_{n-1})) \\
&= \mathbb{P}(A_n | A_1 \cap \ldots \cap A_{n-1}) \cdot \mathbb{P}(A_1 \cap \ldots \cap A_{n-1}) \\
&= \mathbb{P}(A_n | A_1 \cap \ldots \cap A_{n-1}) \cdot \mathbb{P}(A_{n-1} | A_1 \cap \ldots \cap A_{n-2}) \\
&\quad \cdot \mathbb{P}(A_1 \cap \ldots \cap A_{n-2}) \\
&= \mathbb{P}(A_n | A_1 \cap \ldots \cap A_{n-1}) \cdot \mathbb{P}(A_{n-1} | A_1 \cap \ldots \cap A_{n-2}) \\
&\quad \cdot \ldots \cdot \mathbb{P}(A_2 | A_1) \cdot \mathbb{P}(A_1)
\end{aligned}$$

□

In Anwendungen wird der Ergebnisraum häufig partitioniert in Ereignisse. So können etwa in der Wahlforschung die Wähler nach ihrer Wahlentscheidung aufgeteilt werden.

Definition 2.27

Sei $(A_i)_{i \in \mathbb{N}}$ mit $A_i \in \mathcal{F}$ für alle $i \in \mathbb{N}$ eine Folge paarweise disjunkter Ereignisse in $\mathcal{F}$ mit $\bigcup_{i=1}^{\infty} A_i = \Omega$ und $\mathbb{P}(A_i) > 0$, so heißt $(A_i)_{i \in \mathbb{N}}$ eine messbare Zerlegung von Ω.

Mit Hilfe einer messbaren Zerlegung des Ergebnisraums, der endlich, abzählbar oder überabzählbar sein kann, lässt sich die Wahrscheinlichkeit für ein beliebiges einzelnes Ereignis zusammensetzen.

Satz 2.28: Satz der totalen Wahrscheinlichkeit

Für ein Ereignis $A \in \mathcal{F}$ und eine messbare Zerlegung $(A_i)_{i \in \mathbb{N}}$ von Ω gilt

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) \cdot \mathbb{P}(A | A_i)$$

Beweis.

$$\begin{aligned}
\mathbb{P}(A) = \mathbb{P}(A \cap \Omega) &= \mathbb{P}\left(A \cap \left(\bigcup_{i=1}^{\infty} A_i\right)\right) \\
&= \mathbb{P}\left(\bigcup_{i=1}^{\infty} (A \cap A_i)\right) = \sum_{i=1}^{\infty} \mathbb{P}(A \cap A_i) \\
&= \sum_{i=1}^{\infty} \mathbb{P}(A_i) \cdot \frac{\mathbb{P}(A \cap A_i)}{\mathbb{P}(A_i)} = \sum_{i=1}^{\infty} \mathbb{P}(A_i) \cdot \mathbb{P}(A | A_i)
\end{aligned}$$

□

Sind $\mathbb{P}(A), \mathbb{P}(B) > 0$, so lassen sich die Gleichungen für die bedingten Wahrscheinlichkeiten $\mathbb{P}(A|B)$ und $\mathbb{P}(B|A)$ wie oben umformen zu

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

Löst man die Gleichung nach $\mathbb{P}(B|A)$ auf, ergibt sich

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

Das ist der so genannte Satz von Bayes in seiner einfachsten Form.


Thomas Bayes
1701-1761

Satz und Definition 2.29: Bayessche Formel

Seien $(\Omega, \mathcal{F}, \mathbb{P})$ ein beliebiger Wahrscheinlichkeitsraum und $(A_i)_{i\in\mathbb{N}}$ mit $A_i \in \mathcal{F}$ für alle $i \in \mathbb{N}$ eine messbare Zerlegung von Ω. Ist $\mathbb{P}(A) > 0$, dann folgt der so genannte Satz von Bayes

$$\mathbb{P}(A_i|A) = \frac{\mathbb{P}(A_i) \cdot \mathbb{P}(A|A_i)}{\sum_{i=1}^{\infty} \mathbb{P}(A_i) \cdot \mathbb{P}(A|A_i)}.$$

Diese Wahrscheinlichkeit wird auch als a-posteriori Wahrscheinlichkeit von A_i unter der Bedingung A bezeichnet, während $\mathbb{P}(A_i)$ a-priori Wahrscheinlichkeit und $\mathbb{P}(B|A_i)$ Modellwahrscheinlichkeit (Likelihood) genannt wird.

Beweis.

$$\mathbb{P}(A_i|A) = \frac{\mathbb{P}(A_i \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A_i) \cdot \mathbb{P}(A|A_i)}{\sum_{i=1}^{\infty} \mathbb{P}(A_i) \cdot \mathbb{P}(A|A_i)}.$$

□

Beispiel 2.30

Für ein Fotogeschäft arbeiten zwei Labors (Beispiel aus [10]). Eine Fotoarbeit wird zufällig ausgewählt und auf ihre Qualität hin untersucht. Man betrachte folgende Ereignisse: A_i seien die zufälligen Ereignisse „Fotoarbeit stammt aus dem Labor i“, $i = 1, 2$. B sei das Ereignis „Fotoarbeit ist einwandfrei“. Dann ist $\Omega = A_1 \cup A_2$ mit $A_1 \cap A_2 = \emptyset$ (das ist dann eine Partitionierung von Ω). Es seien $\mathbb{P}(A_1) = 0.7$, $\mathbb{P}(A_2) = 0.3$, $\mathbb{P}(B|A_1) = 0.8$ und $\mathbb{P}(B|A_2) = 0.9$ gegeben. Wie wahrscheinlich ist es,

- dass die Fotoarbeit einwandfrei ist?

$$\mathbb{P}(B) = \mathbb{P}(B|A_1) \cdot \mathbb{P}(A_1) + \mathbb{P}(B|A_2) \cdot \mathbb{P}(A_2) = 0.8 \cdot 0.7 + 0.9 \cdot 0.3 = 0.83,$$

- dass die Fotoarbeit von Labor 1 stammt und einwandfrei ist?

$$\mathbb{P}(A_1 \cap B) = \mathbb{P}(A_1) \cdot \mathbb{P}(B|A_1) = 0.7 \cdot 0.8 = 0.56,$$

- dass die Fotoarbeit von Labor 2 stammt und einwandfrei ist?

$$\mathbb{P}(A_2 \cap B) = \mathbb{P}(A_2) \cdot \mathbb{P}(B|A_2) = 0.3 \cdot 0.9 = 0.27,$$

- dass eine einwandfreie Fotoarbeit aus Labor 1 stammt?

$$\mathbb{P}(A_1|B) = \frac{\mathbb{P}(A_1) \cdot \mathbb{P}(B|A_1)}{\mathbb{P}(B)} = \frac{0.56}{0.83} = 0.6747,$$

- dass eine einwandfreie Fotoarbeit aus Labor 2 stammt?

$$\mathbb{P}(A_2|B) = \frac{\mathbb{P}(A_2) \cdot \mathbb{P}(B|A_2)}{\mathbb{P}(B)} = \frac{0.27}{0.83} = 0.3253,$$

- dass eine fehlerhafte Fotoarbeit aus Labor 1 stammt?

$$\begin{aligned} \mathbb{P}(A_1|\bar{B}) &= \frac{\mathbb{P}(A_1) \cdot \mathbb{P}(\bar{B}|A_1)}{\mathbb{P}(\bar{B})} \\ &= \frac{\mathbb{P}(A_1) \cdot \mathbb{P}(\bar{B}|A_1)}{\mathbb{P}(A_1) \cdot \mathbb{P}(\bar{B}|A_1) + \mathbb{P}(A_2) \cdot \mathbb{P}(\bar{B}|A_2)} \\ &= \frac{0.2 \cdot 0.7}{0.2 \cdot 0.7 + 0.1 \cdot 0.3} = 0.8235, \end{aligned}$$

- dass eine fehlerhafte Fotoarbeit aus Labor 2 stammt?

$$\begin{aligned} \mathbb{P}(A_2|\bar{B}) &= \frac{\mathbb{P}(A_2) \cdot \mathbb{P}(\bar{B}|A_2)}{\mathbb{P}(\bar{B})} \\ &= \frac{\mathbb{P}(A_2) \cdot \mathbb{P}(\bar{B}|A_2)}{\mathbb{P}(A_1) \cdot \mathbb{P}(\bar{B}|A_1) + \mathbb{P}(A_2) \cdot \mathbb{P}(\bar{B}|A_2)} \\ &= \frac{0.1 \cdot 0.3}{0.2 \cdot 0.7 + 0.1 \cdot 0.3} = 0.1765. \end{aligned}$$

Im Beispiel 2.30 lassen sich die Wahrscheinlichkeiten in einer sogenannten Kontingenztabelle darstellen. Dazu werde die Notation $A = A_1$ und $\bar{A} = A_2$ verwendet.

Tabelle 2.2.: Kontingenztabelle für das Beispiel 2.30

	A	$\bar{A}$	
B	0.56	0.27	0.83
$\bar{B}$	0.14	0.03	0.17
	0.7	0.3	1

Die orangefarbenen Zellen der Tabelle geben die Wahrscheinlichkeiten für die Durchschnitte der Ereignismengen an. Die grauen Zellen am Rand geben Randwahrscheinlichkeiten an und resultieren jeweils als Zeilen- oder Spaltensumme der Wahrscheinlichkeiten in den inneren Zellen. Eine Randwahrscheinlichkeit ist die Wahrscheinlichkeit für das Ereignis der jeweiligen Zeile bzw. Spalte. Die Summe aller Spalten- oder Zeilen-Randwahrscheinlichkeiten ergibt Eins.

	A	$\bar{A}$	
B	$\mathbb{P}(B \cap A)$	$\mathbb{P}(B \cap \bar{A})$	$\mathbb{P}(B)$
$\bar{B}$	$\mathbb{P}(\bar{B} \cap A)$	$\mathbb{P}(\bar{B} \cap \bar{A})$	$\mathbb{P}(\bar{B})$
	$\mathbb{P}(A)$	$\mathbb{P}(\bar{A})$	1

Sind $A, B, C \in \mathcal{F}$ drei Ereignisse und deren Gegenereignisse $\bar{A}, \bar{B}$ und $\bar{C}$. Mit Hilfe bedingter Wahrscheinlichkeiten gilt dann:

$$\begin{aligned}
\mathbb{P}(A|B) &= \mathbb{P}(A \cap \Omega|B) = \mathbb{P}(A \cap (C \cup \bar{C})|B) \\
&= \mathbb{P}((A \cap C) \cup (A \cap \bar{C})|B) \\
&= \mathbb{P}(A \cap C|B) + \mathbb{P}(A \cap \bar{C}|B) \\
&= \frac{\mathbb{P}(A \cap B \cap C)}{\mathbb{P}(B)} + \frac{\mathbb{P}(A \cap B \cap \bar{C})}{\mathbb{P}(B)} \\
&= \mathbb{P}(A|B \cap C) \cdot \frac{\mathbb{P}(B \cap C)}{\mathbb{P}(B)} + \mathbb{P}(A|B \cap \bar{C}) \cdot \frac{\mathbb{P}(B \cap \bar{C})}{\mathbb{P}(B)} \\
&= \mathbb{P}(A|B \cap C) \cdot \mathbb{P}(C|B) + \mathbb{P}(A|B \cap \bar{C}) \cdot \mathbb{P}(\bar{C}|B)
\end{aligned} \tag{2.4}$$

Analog folgt $\mathbb{P}(A|\bar{B}) = \mathbb{P}(A|\bar{B} \cap C) \cdot \mathbb{P}(C|\bar{B}) + \mathbb{P}(A|\bar{B} \cap \bar{C}) \cdot \mathbb{P}(\bar{C}|\bar{B})$. Die bedingten Wahrscheinlichkeiten für das Ereignis A unter B $(\bar{B})$ sind gewichtete Summen von Wahrscheinlichkeiten für A unter B $(\bar{B})$ und C bzw. B $(\bar{B})$ und $\bar{C}$. Das Einbeziehen eines dritten Ereignisses spielt beim Simpsonschen Paradoxon (nach Edward Hugh Simpson, 1922-2019) eine wichtige Rolle.

In manchen Situationen hat das Eintreten eines Ereignisses B keinen Einfluss auf die Wahrscheinlichkeiten für ein anderes Ereignis A. Das bedeutet, dass

$$\mathbb{P}(A) = \mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \text{ und } \mathbb{P}(A) = \mathbb{P}(A|\bar{B}) = \frac{\mathbb{P}(A \cap \bar{B})}{\mathbb{P}(\bar{B})}.$$

ist. Dies lässt sich umformen zu

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B) \text{ und } \mathbb{P}(A \cap \bar{B}) = \mathbb{P}(A) \cdot \mathbb{P}(\bar{B}).$$

Beispiel 2.31

Werfen eines fairen Würfels: $A = \{1, 2\}$ und $B = \{1, 3, 5\}$. Damit:

$$\mathbb{P}(A) = \frac{1}{3},\ \mathbb{P}(B) = \frac{1}{2},\ \mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{1/6}{1/2} = \frac{1}{3} \Rightarrow \mathbb{P}(A) = \mathbb{P}(A|B).$$

Definition 2.32: Stochastische Unabhängigkeit von Ereignissen

Sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein beliebiger Wahrscheinlichkeitsraum. Die Ereignisse $(A_i)_{i\in I},\ A_i \in \mathcal{F}$ heißen stochastisch unabhängig, wenn für jede nicht-leere endliche Teilmenge $J \subseteq I$ gilt:

$$\mathbb{P}\left(\bigcap_{j\in J} A_j\right) = \prod_{j\in J} \mathbb{P}(A_j).$$

Daraus folgt unmittelbar für zwei stochastisch unabhängige Ereignisse A und B mit $\mathbb{P}(B) > 0$ die obige Überlegung, dass

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

Beispiel 2.33

In einer Urne befinden sich k schwarze und $n-k$ weiße Kugeln $(1 \leq k \leq n-1)$. Zwei Kugeln werden herausgezogen. Sei A das Ereignis „Die erste Kugel ist schwarz“ und B das Ereignis „Die zweite Kugel ist schwarz“. Sind die beiden Ereignisse stochastisch unabhängig? Der Ergebnisraum wird durch $\Omega = \{s, w\}^2 = \{(w,w),(s,w),(w,s),(s,s)\}$ modelliert.

- 1. Fall: Die gezogene Kugel wird nicht zurückgelegt. Dann gilt:

$$\begin{aligned}
\mathbb{P}(A) &= \frac{k}{n} \\
\mathbb{P}(B) &= \frac{n-k}{n} \cdot \frac{k}{n-1} + \frac{k}{n} \cdot \frac{k-1}{n-1} = \frac{k}{n} \\
\mathbb{P}(A \cap B) &= \frac{k}{n} \cdot \frac{k-1}{n-1} = \frac{k(k-1)}{n(n-1)} \\
\mathbb{P}(A \cap B) &\neq \mathbb{P}(A) \cdot \mathbb{P}(B),
\end{aligned}$$

  Die Ereignisse sind nicht stochastisch unabhängig.

- 2. Fall: Die gezogene Kugel wird zurückgelegt. Dann gilt:

$$\begin{aligned}
\mathbb{P}(A) &= \frac{k}{n} \\
\mathbb{P}(B) &= \frac{n-k}{n} \cdot \frac{k}{n} + \frac{k}{n} \cdot \frac{k}{n} = \frac{k}{n} \\
\mathbb{P}(A \cap B) &= \frac{k}{n} \cdot \frac{k}{n} = \frac{k^2}{n^2} \\
\mathbb{P}(A \cap B) &= \mathbb{P}(A) \cdot \mathbb{P}(B),
\end{aligned}$$

  Die Ereignisse sind stochastisch unabhängig.

Am Beispiel 2.33 zeigt sich, dass die Unabhängigkeit nicht nur eine Eigenschaft von Ereignissen ist, sondern vom zugrundeliegenden Wahrscheinlichkeitsmaß abhängt.

Beispiel 2.34

Mittels einer Kontigenztabelle lässt sich die stochastische Unabhängigkeit von Ereignissen überprüfen, indem das Produkt einer Zeilen- und Spaltenwahrscheinlichkeit mit dem dazu korrespondierenden Zellenwert der Tabelle verglichen wird. So ist im Fotobeispiel $0.83 \cdot 0.7 = 0.581 \neq 0.56$. Damit sind A und B nicht stochastisch unabhängig.

Dass die Definition 2.32 eine schärfere Bedingung als die paarweise Unabhängigkeit von Ereignissen darstellt, zeigt

Beispiel 2.35

Das zweimalige Werfen einer fairen Münze wird durch den Ergebnisraum $\Omega = \{\mathsf{K},\mathsf{Z}\}^2$ mit den Laplace-Wahrscheinlichkeiten $\mathbb{P}(\{\mathsf{K}\}) = \mathbb{P}(\{\mathsf{Z}\}) = \frac{1}{2}$ und $\mathbb{P}(\{\mathsf{K},\mathsf{K}\}) = \mathbb{P}(\{\mathsf{K},\mathsf{Z}\}) = \mathbb{P}(\{\mathsf{Z},\mathsf{K}\}) = \mathbb{P}(\{\mathsf{Z},\mathsf{Z}\}) = \frac{1}{4}$ beschrieben. Werden die Ereignisse

$$\begin{aligned} A &= \{\mathsf{Z}\} \times \{\mathsf{K},\mathsf{Z}\}, \\ B &= \{\mathsf{K},\mathsf{Z}\} \times \{\mathsf{Z}\}, \\ C &= \{(\mathsf{K},\mathsf{K}),(\mathsf{Z},\mathsf{Z})\}, \end{aligned}$$

betrachtet, so ergeben sich folgende Wahrscheinlichkeiten:

$$\begin{aligned} \mathbb{P}(A) &= \frac{1}{2}, \\ \mathbb{P}(B) &= \frac{1}{2}\cdot\frac{1}{2} + \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{2}, \\ \mathbb{P}(C) &= \frac{1}{2}\cdot\frac{1}{2} + \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{2}, \\ \mathbb{P}(A\cap B) &= \frac{1}{4} = \mathbb{P}(A)\cdot\mathbb{P}(B), \\ \mathbb{P}(A\cap C) &= \frac{1}{4} = \mathbb{P}(A)\cdot\mathbb{P}(C), \\ \mathbb{P}(B\cap C) &= \frac{1}{4} = \mathbb{P}(B)\cdot\mathbb{P}(C), \\ \mathbb{P}(A\cap B\cap C) &= \frac{1}{4} \neq \frac{1}{8} = \mathbb{P}(A)\cdot\mathbb{P}(B)\cdot\mathbb{P}(C). \end{aligned}$$

Die drei Ereignisse sind jeweils paarweise stochastisch unabhängig, aber nicht stochastisch unabhängig.

2.3. Zufallsvariablen

Oft sind wir nicht nur an einem Zufallsexperiment interessiert, sondern an einer Verknüpfung mehrerer möglicherweise identischer Zufallsexperimente oder an einer Weiterverarbeitung des Ergebnisses eines Zufallsexperiments. Dazu brauchen wir so genannte Zufallsvariablen.

Definition 2.36: Zufallsvariable

Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $(\Psi, \mathcal{G})$ ein Tupel bestehend aus einer Menge Ψ und einer dazugehörigen σ-Algebra $\mathcal{G}$ auf Ψ. Eine Abbildung $X : \Omega \to \Psi$ heißt Zufallsvariable, wenn für jedes $G \in \mathcal{G}$ gilt $X^{-1}(G) \in \mathcal{F}$.

Bemerkung.
(1) Die Eigenschaft $X^{-1}(G) \in \mathcal{F}$ für jedes $G \in \mathcal{G}$ wird auch als $\mathcal{F}$-$\mathcal{G}$-Messbarkeit von X bezeichnet.
(2) Auf $\mathcal{G}$ ist ein Wahrscheinlichkeitsmaß $\mathbb{P}_X : \mathcal{G} \to [0,1]$ durch $\mathbb{P}_X(G) = \mathbb{P}(X^{-1}(G))$ definiert. $(\Psi, \mathcal{G}, \mathbb{P}_X)$ ist damit ein Wahrscheinlichkeitsraum.
(3) Eigenschaften eines Wahrscheinlichkeitsmaßes $\mathbb{P}_X : \mathcal{G} \to [0,1]$ werden als Eigenschaften der Zufallsvariablen X deklariert. Man schreibt z.B.

- $\mathbb{P}_X(A) = \mathbb{P}(\{\omega \in \Omega;\ X(\omega) \in A\}) = \mathbb{P}(X \in A)$ für ein Ereignis $A \in \mathcal{G}$,
- $\mathbb{P}_X(]-\infty, x]) = \mathbb{P}(\{\omega \in \Omega;\ X(\omega) \in]-\infty, x]\}) = \mathbb{P}(X \leq x)$ für ein Ereignis $]-\infty, x] \in \mathcal{B}(\mathbb{R}) = \mathcal{G}$,
- $\mathbb{P}_X(\{a\}) = \mathbb{P}(X = a)$ für ein Ereignis $\{a\} \in \mathcal{G}$,
- $\mathbb{P}_X([a,b]) = \mathbb{P}(X \in [a,b]) = \mathbb{P}(a \leq X \leq b)$ für das Ereignis $[a,b] \in \mathcal{B}(\mathbb{R}) = \mathcal{G}$,
- $\mathbb{P}_X(]-\infty, -\epsilon[\cup]\epsilon, \infty[) = 1 - \mathbb{P}(-\epsilon \leq X \leq \epsilon) = 1 - \mathbb{P}(|X| \leq \epsilon) = \mathbb{P}(|X| > \epsilon)$ für das Ereignis $]-\infty, -\epsilon[\cup]\epsilon, \infty[\in \mathcal{B}(\mathbb{R}) = \mathcal{G}$,
- $X \sim \mathcal{N}(\mu, \sigma^2)$, wenn das Wahrscheinlichkeitsmaß $\mathbb{P}_X$ auf $\mathcal{G}$ einer Normalverteilung mit Parametern μ und σ^2 entspricht,
- $X \sim \mathcal{P}o(\lambda)$, wenn das Wahrscheinlichkeitsmaß $\mathbb{P}_X$ auf $\mathcal{G}$ einer Poisson-Verteilung mit Parameter λ entspricht.

Beispiel 2.37

Bei einem Münzwurf sei $\Omega = \{\mathsf{K},\mathsf{Z}\}$ gegeben. Es sei $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ der dazugehörige Wahrscheinlichkeitsraum mit $\mathbb{P}(\{\mathsf{Z}\}) = \mathbb{P}(\{\mathsf{K}\}) = \frac{1}{2}$. Sei $\Psi = \{0,1\}$. Die Abbildung

$$X : \Omega \to \{0,1\},\ X(\mathsf{Z}) = 0,\ X(\mathsf{K}) = 1,$$

ist eine Zufallsvariable, da jedes Urbild eines Elements der Potenzmenge von Ψ unter X in der Potenzmenge von Ω liegt. X ordnet den Ergebnissen des Zufallsexperiments Zahlen zu, um damit weiterrechnen zu können. Es ist $\mathbb{P}(X = 0) = \mathbb{P}(X = 1) = \frac{1}{2}$.

Zu einem Wahrscheinlichkeitsmaß $\mathbb{P}$ auf $\mathbb{R}$ kann die Verteilungsfunktion $F_{\mathbb{P}}$ bestimmt werden. Die Verteilungsfunktion einer Zufallsvariablen ist analog dazu die Funktion

$$F_{\mathbb{P}} : \mathbb{R} \to [0,1],\ x \mapsto F_{\mathbb{P}}(x) = \mathbb{P}_X(]-\infty, x]) = \mathbb{P}(X \leq x).$$

Will man die Anzahl Tore zweier Fußballspiele untersuchen, so lässt sich jedes Spiel für sich alleine mittels einer Poisson-Verteilung modellieren. Gibt es eine Möglichkeit beide

Spiele zusammen zu betrachten und zu bestimmen, mit welcher Wahrscheinlichkeit eine bestimmte Anzahl an Toren erzielt wird? Zunächst betrachtet man einen Wahrscheinlichkeitsraum

$$(\mathbb{N}_0^2, \mathcal{P}(\mathbb{N}_0^2), \mathbb{P}).$$

Ein Ergebnis könnte etwa $\omega = (2,3) \in \mathbb{N}_0^2$ sein, im ersten Spiel fallen zwei und im zweiten drei Tore. Nun interessiert die Gesamtzahl der Tore. Man modelliert dies mit $(\mathbb{N}_0, \mathcal{P}(\mathbb{N}_0))$ über die Zufallsvariable

$$X : \mathbb{N}_0^2 \to \mathbb{N}_0,\ (\omega_1, \omega_2) \mapsto \omega_1 + \omega_2.$$

Das Wahrscheinlichkeitsmaß $\mathbb{P}_X$ bekommt man, wenn man das Wahrscheinlichkeitsmaß $\mathbb{P}$ kennt. Doch wie sieht das aus? Um diese Frage zu beantworten benötigt man noch etwas Vorarbeit.

2.3.1. Mehrdimensionale Zufallsvariablen

Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Sind

$$X_1 : \Omega \to V_1, \ldots, X_n : \Omega \to V_n$$

Zufallsvariablen, so heißt eine Funktion $f_{X_1 \ldots X_n} = f_{X_1 \ldots X_n}(x_1, \ldots, x_n)$ mit

- diskreter Fall:
$$f_{X_1 \ldots X_n}(x_1, \ldots, x_n) = \mathbb{P}(X_i = x_i;\ i = 1, \ldots, n)$$
- stetiger Fall:
$$\int_{-\infty}^{b_1} \ldots \int_{-\infty}^{b_n} f_{X_1 \ldots X_n}(x_1, \ldots, x_n) dx_n \ldots dx_1 = \mathbb{P}(X_i \in (-\infty, b_i];\ i = 1, \ldots, n)$$

gemeinsame Wahrscheinlichkeitsdichtefunktion von $X_1, \ldots, X_n$. Sind $X_1, \ldots, X_n$ Zufallsvariablen mit einer gemeinsamen Wahrscheinlichkeitsdichtefunktion

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

so heißt eine Funktion f_{X_i} mit

- diskreter Fall:
$$f_{X_i}(x_i) = \mathbb{P}(X_i = x_i) = \sum_{x_1 \in V_1} \ldots \sum_{x_{i-1} \in V_{i-1}} \sum_{x_{i+1} \in V_{i+1}} \ldots \sum_{x_n \in V_n} f_{X_1 \ldots X_n}(x_1, \ldots, x_n)$$
- stetiger Fall:
$$f_{X_i}(x_i) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f_{X_1 \ldots X_n}(x_1, \ldots, x_n) dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_n$$

Randverteilung von $X_i,\ i = 1, \ldots, n$.

Definition 2.38: Stochastische Unabhängigkeit

Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Zufallsvariablen

$$X_1 : \Omega \to V_1, \ldots, X_n : \Omega \to V_n$$

mit einer gemeinsamen Wahrscheinlichkeitsdichtefunktion

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

heißen stochastisch unabhängig, wenn

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

gilt.

Bemerkung.
Oft wird angenommen, dass n Zufallsvariablen $X_1, \ldots, X_n$ stochastisch unabhängig und identisch verteilt sind (Schreibweise: u.i.v.).

Beispiel 2.39

Sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Mit den beiden Zufallsvariablen $X_1 : \Omega \to \mathbb{N}_0$ und $X_2 : \Omega \to \mathbb{N}_0$ soll die Anzahl Tore bei zwei Fußballspielen modelliert werden. Angenommen, die beiden Zufallsvariablen sind identisch Poisson-verteilt mit Parameter λ und stochastisch unabhängig, d.h. für die gemeinsame Wahrscheinlichkeitsdichtefunktion $f_{X_1 X_2}(x_1, x_2)$ von X_1 und X_2 gelte

$$f_{X_1 X_2}(x_1, x_2) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) = e^{-\lambda} \cdot \frac{\lambda^{x_1}}{x_1!} \cdot e^{-\lambda} \cdot \frac{\lambda^{x_2}}{x_2!} = e^{-2\lambda} \cdot \frac{\lambda^{x_1 + x_2}}{x_1! \cdot x_2!}.$$

Sei

$$(\mathbb{N}_0^2, \mathcal{P}(\mathbb{N}_0^2), \mathbb{P})$$

der dazugehörige Wahrscheinlichkeitsraum mit $\mathbb{P}(\{x_1, x_2\}) = f_{X_1 X_2}(x_1, x_2)$. Ein Ergebnis könnte etwa $\omega = (2, 3) \in \mathbb{N}_0^2$ sein, im ersten Spiel fallen zwei und im zweiten drei Tore. Nun sei nach der Summe der Tore gefragt. Man modelliert dies mit $(\mathbb{N}_0, \mathcal{P}(\mathbb{N}_0))$ über die Zufallsvariable

$$X : \mathbb{N}_0^2 \to \mathbb{N}_0, \ (\omega_1, \omega_2) \mapsto \omega_1 + \omega_2.$$

Das Wahrscheinlichkeitsmaß $\mathbb{P}_X$ bekommt man durch folgende Überlegung:

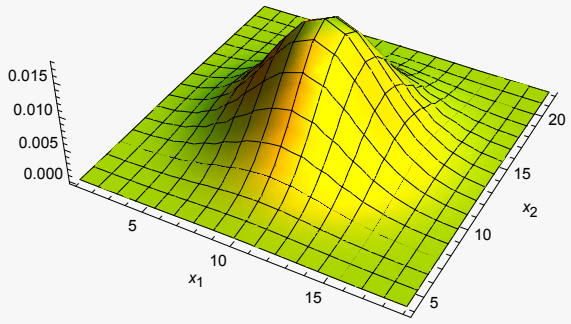
$$\begin{aligned}
\mathbb{P}(X=n) = \mathbb{P}_X(\{n\}) &= \mathbb{P}(X^{-1}(\{n\})) = \mathbb{P}(\{(x_1,x_2)\in\mathbb{N}_0^2;\ x_1+x_2=n\}) \\
&= \mathbb{P}(\{(x_1,n-x_1);\ x_1\in\{0,\dots,n\}\}) \\
&= \sum_{x_1=0}^{n} e^{-2\lambda}\cdot\frac{\lambda^{x_1+n-x_1}}{x_1!\cdot(n-x_1)!} \\
&= \frac{e^{-2\lambda}}{n!}\sum_{x_1=0}^{n}\frac{n!}{x_1!\cdot(n-x_1)!}\lambda^{x_1}\cdot\lambda^{n-x_1} \\
&= \frac{e^{-2\lambda}}{n!}\cdot(2\lambda)^n.
\end{aligned}$$

Es ist somit $X\sim\mathcal{Po}(2\lambda)$.

Beispiel 2.40: Zweidimensionale Normalverteilung

Seien $X_1\sim\mathcal{N}(10,9)$ und $X_2\sim\mathcal{N}(13,9)$ stochastisch unabhängig. Die gemeinsame Wahrscheinlichkeitsdichtefunktion lautet

$$f_{X_1X_2}(x_1,x_2)=\frac{1}{3\cdot3\cdot2\pi}e^{-\frac{1}{2}\left(\frac{(x_1-10)^2}{9}+\frac{(x_2-13)^2}{9}\right)}.$$



Sind (μ_1,σ_1^2) und (μ_2,σ_2^2) die Parameter von X_1 bzw. X_2, so kann mit einer etwas aufwändigeren Rechnung analog zu Beispiel 2.39 gezeigt werden, dass für die Zufallsvariable $X:\mathbb{R}^2\to\mathbb{R}$, $(x_1,x_2)\mapsto X(x_1,x_2)=x_1+x_2$ unter Annahme der stochastischen Unabhängigkeit die stetige Dichte

$$f_X(x)=\frac{1}{\sqrt{\sigma_1^2+\sigma_2^2}\sqrt{2\pi}}\cdot e^{-\frac{1}{2}\frac{(x-(\mu_1+\mu_2))^2}{\sigma_1^2+\sigma_2^2}}$$

resultiert. Damit folgt $X\sim\mathcal{N}(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$.

2.3.2. Erwartungswert und Varianz von Zufallsvariablen


Huygens
1629-1695

Die Parameter einer Verteilung haben meist eine ganz bestimmte Bedeutung. Sie charakterisieren eine Verteilung. Sei $X : \Omega \to \Psi \subseteq \mathbb{R}$ eine diskrete Zufallsvariable. Wir sagen, X besitze einen Erwartungswert, wenn

$$\sum_{x \in \Psi} |x| \mathbb{P}_X(\{x\}) < \infty.$$

Ebenso besitze eine reelle (stetige) Zufallsvariable $X : \Omega \to \mathbb{R}$ mit stetiger Dichte f_X einen Erwartungswert, wenn

$$\int_{-\infty}^{\infty} |x| f_X(x) dx < \infty.$$

X ist dann absolut summierbar bzw. absolut integrierbar. Im diskreten, endlichen Fall, d.h. $\Psi = \{x_1, \ldots, x_r\}$, ist nun folgende gewichtete Summe

$$\sum_{\omega \in \Omega} X(\omega) P(\{\omega\}) = \sum_{j=1}^{r} \sum_{\omega : X(\omega) = x_j} x_j P(\{\omega\}) = \sum_{j=1}^{r} x_j P_X(\{x_j\})$$

von Interesse. Die „Umsortierung“ nach dem ersten Gleichheitszeichen klappt im Allgemeinen, nicht-endlichen Fall nur, wenn die Summe absolut summierbar ist. Die Summe entspricht anschaulich dem Schwerpunkt einer diskreten Massenverteilung. Dies führt zu folgender Definition.

Definition 2.41: Erwartungswert und Varianz von Zufallsvariablen

Es sei f_X eine diskrete oder stetige Dichte einer absolut summierbaren bzw. integrierbaren reellen Zufallsvariablen $X : \Omega \to \Psi \subseteq \mathbb{R}$ mit Parametern $\vartheta \in \Theta$ in einem Parameterraum Θ. Dann heißt

- diskret: $\mathbb{E}_\vartheta[X] := \sum_{x \in \Psi} x \cdot f_X(x)$,
- stetig: $\mathbb{E}_\vartheta[X] := \int_{-\infty}^{\infty} x \cdot f_X(x) dx$

der Erwartungswert von X und

- diskret: $\mathbb{V}_\vartheta[X] := \sum_{x \in \Psi} \underbrace{(x - \mathbb{E}_\vartheta[X])^2}_{q(x)} \cdot f_X(x)$,
- stetig: $\mathbb{V}_\vartheta[X] := \int_{-\infty}^{\infty} \underbrace{(x - \mathbb{E}_\vartheta[X])^2}_{q(x)} \cdot f_X(x) dx$

die Varianz von X

Bemerkung.
(1) Bei der Poisson-Verteilung ist $\Theta = \mathbb{R}^+$, bei der Normalverteilung $\Theta = \mathbb{R} \times \mathbb{R}^+$.

(2) Allgemein lässt sich in Analogie zur Definition der Varianz für eine beliebige Zufallsvariable $Y = q(X)$ unter den angegebenen Voraussetzungen der Erwartungswert bestimmen: $\mathbb{E}_\vartheta[Y] = \sum\limits_{x\in\Psi} q(x)f(x)$ (hier nur für den diskreten Fall aufgeschrieben). Denn für $X \in \{x_1, \ldots, x_r\}$ und $Y(\omega) = q(X(\omega))$ ist

$$\begin{aligned}\mathbb{E}_\vartheta[Y] &= \sum_{\omega\in\Omega} Y(\omega)P(\{\omega\}) = \sum_{\omega\in\Omega} q(X(\omega))P(\{\omega\}) \\ &= \sum_{j=1}^{r} \sum_{\omega: X(\omega)=x_j} q(x_j)P(\{\omega\}) = \sum_{j=1}^{r} q(x_j)P_X(\{x_j\}).\end{aligned}$$

(3) Es gilt: $\mathbb{E}_\vartheta[aX + b] = a\mathbb{E}_\vartheta[X] + b,\ a, b \in \mathbb{R}$, $(Y = aX + b)$.
(4) Für die Varianz $\mathbb{V}_\vartheta[X]$ gilt mit $\mathbb{E}_\vartheta[X] = \mu$:

$$\begin{aligned}\mathbb{V}_\vartheta[X] &= \mathbb{E}_\vartheta[(X-\mu)^2] \\ &= \mathbb{E}_\vartheta[X^2 - 2\mu X + \mu^2] \\ &= \mathbb{E}_\vartheta[X^2] - 2\mu\mathbb{E}_\vartheta[X] + \mathbb{E}_\vartheta[X]^2 \\ &= \mathbb{E}_\vartheta[X^2] - 2\mathbb{E}_\vartheta[X]^2 + \mathbb{E}_\vartheta[X]^2 \\ &= \mathbb{E}_\vartheta[X^2] - \mathbb{E}_\vartheta[X]^2.\end{aligned}$$

Beispiel 2.42

Sei X eine Zufallsvariable „Augenzahl beim einmaligen fairen Würfelwurf“ mit Werten in $\Psi = \{1, 2, 3, 4, 5, 6\}$. Für den Erwartungswert ergibt sich

$$\mathbb{E}_\vartheta[X] = \sum_{i=1}^{6} i \cdot \frac{1}{6} = \frac{21}{6} = \frac{7}{2}.$$

Damit lässt sich die Varianz zu

$$\mathbb{V}_\vartheta[X] = \sum_{i=1}^{6} i^2 \cdot \frac{1}{6} - \left(\frac{7}{2}\right)^2 = \frac{182 - 147}{12} = \frac{35}{12}.$$

berechnen.

Satz 2.43

Seien $X : \Omega \to \Psi, Y : \Omega \to \Psi$ zwei Zufallsvariablen mit $Y \leq X$, d.h. $Y(\omega) \leq X(\omega)$ für alle $\omega \in \Omega$, und existierenden Erwartungswerten. Dann gilt

$$\mathbb{E}_\vartheta[Y] \leq \mathbb{E}_\vartheta[X].$$

Beweis.

Der Beweis erfolgt hier nur für den diskreten Fall.
Wegen $Y \leq X$ gilt $\mathbb{P}(Y = y|X = x) = 0$ für $x < y$ und somit ist

$$\begin{aligned}
\mathbb{E}_\vartheta[Y] &= \sum_{y\in\Phi} y\mathbb{P}(Y = y) \\
&\overset{\text{Satz } 2.28}{=} \sum_{y\in\Phi} y \cdot \sum_{x\in\Psi} \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y|X = x) \\
&\overset{\text{Def. } 2.24}{=} \sum_{y\in\Phi} y \cdot \sum_{x\in\Psi} \mathbb{P}(X = x) \cdot \frac{\mathbb{P}(Y = y)\mathbb{P}(X = x|Y = y)}{\mathbb{P}(X = x)} \\
&= \sum_{x\in\Psi}\sum_{y\in\Phi} y \cdot \mathbb{P}(X = x) \cdot \frac{\mathbb{P}(Y = y)\mathbb{P}(X = x|Y = y)}{\mathbb{P}(X = x)} \\
&\leq \sum_{x\in\Psi}\sum_{y\in\Phi} x \cdot \mathbb{P}(X = x) \cdot \frac{\mathbb{P}(Y = y)\mathbb{P}(X = x|Y = y)}{\mathbb{P}(X = x)} \\
&= \sum_{x\in\Psi} x \cdot \sum_{y\in\Phi} \mathbb{P}(Y = y)\mathbb{P}(X = x|Y = y) \\
&= \sum_{x\in\Psi} x \cdot \mathbb{P}(X = x) \\
&= \mathbb{E}_\vartheta[X].
\end{aligned}$$

□

Definition 2.44: Kovarianz zweier Zufallsvariablen

Sind f, g zwei entweder diskrete oder stetige Dichten der Zufallsvariablen X und Y mit Parametern $\vartheta \in \Theta$ und $\psi \in \Psi$, so wird die Kovarianz zwischen X und Y durch

$$\begin{aligned}
\text{Cov}_{(\vartheta,\psi)}[X, Y] &:= \mathbb{E}_{(\vartheta,\psi)}[(X - \mathbb{E}_\vartheta[X])(Y - \mathbb{E}_\psi[Y])] \\
&= \mathbb{E}_{(\vartheta,\psi)}[XY] - \mathbb{E}_\vartheta[X]\mathbb{E}_\psi[Y].
\end{aligned}$$

definiert.

Es ist $\mathbb{V}_\vartheta[X] = \text{Cov}_\vartheta[X, X]$. Sind X und Y zwei unabhängige Zufallsvariablen mit der gemeinsamen Wahrscheinlichkeitsdichtefunktion $f_{XY}(x, y)$, so gilt für $Z = X \cdot Y$:

$$\begin{aligned}
\mathbb{E}_{(\vartheta,\psi)}[Z] = \mathbb{E}_{(\vartheta,\psi)}[XY] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{XY}(x, y)dxdy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_X(x)f_Y(y)dxdy \\
&= \int_{-\infty}^{\infty} xf_X(x)dx \int_{-\infty}^{\infty} yf_Y(y)dy \\
&= \mathbb{E}_\vartheta[X]\mathbb{E}_\psi[Y].
\end{aligned}$$

Damit ist die Kovarianz dieser beiden unabhängigen Zufallsvariablen gleich Null.

Für Erwartungswerte und Kovarianzen gelten folgende Eigenschaften:

- $\mathrm{Cov}_{(\vartheta,\psi)}[aX+b, cY+d] = ac \cdot \mathrm{Cov}_{(\vartheta,\psi)}[X,Y],\ a,b,c,d \in \mathbb{R}$,
- $\mathbb{V}_\vartheta[aX+b] = \mathrm{Cov}_\vartheta[aX+b, aX+b] = a^2\mathbb{V}_\vartheta[X],\ a,b \in \mathbb{R}$.

Sind zwei Zufallsvariablen X und Y unabhängig, ist die Varianz der Summe der beiden gleich der Summe der Varianzen:


Bienaymé
1796-1878

$$\begin{aligned}
\mathbb{V}_{\vartheta,\psi}[X+Y] &= \mathbb{E}_{(\vartheta,\psi)}[(X+Y)^2] - \mathbb{E}_{(\vartheta,\psi)}[X+Y]^2 \\
&= \mathbb{E}_{(\vartheta,\psi)}[X^2+Y^2+2XY] - (\mathbb{E}_\vartheta[X]+\mathbb{E}_\psi[Y])^2 \\
&= \mathbb{E}_\vartheta[X^2]+\mathbb{E}_\psi[Y^2]+2\underbrace{\mathbb{E}_{(\vartheta,\psi)}[XY]}_{\mathbb{E}_\vartheta[X]\mathbb{E}_\psi[Y]} \\
&\quad -\mathbb{E}_\vartheta[X]^2-\mathbb{E}_\psi[Y]^2-2\mathbb{E}_\vartheta[X]\mathbb{E}_\psi[Y] \\
&= \mathbb{V}_\vartheta[X]+\mathbb{V}_\psi[Y]
\end{aligned}$$

Diese Gleichung geht (in etwas allgemeinerer Form) auf Irénée-Jules Bienaymé zurück.

Satz 2.45

Sei $X \sim \mathcal{N}(\mu,\sigma^2)$, d.h. X ist normalverteilt mit Parametern μ und σ^2. Für den Erwartungswert von X gilt $\mathbb{E}_{(\mu,\sigma^2)}[X] = \mu$.

Beweis.

Zunächst ist

$$\begin{aligned}
\int x \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx &= \int (x-\mu+\mu)\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \\
&= \int (x-\mu)\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \\
&\quad + \int \mu\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \\
&\overset{u=\left(\frac{x-\mu}{\sigma}\right)^2}{=} \int \frac{\sigma}{2\sqrt{2\pi}} e^{-\frac{1}{2}u} du \\
&\quad +\mu\int \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \\
&= -\frac{\sigma}{\sqrt{2\pi}} e^{-\frac{1}{2}u} + c \\
&\quad +\mu\int \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx.
\end{aligned}$$

Damit folgt $\mathbb{E}_{(\mu,\sigma^2)}[X] = \int_{-\infty}^{\infty} x \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = 0 + \mu \cdot 1 = \mu.$

□

Satz 2.46

Sei $X \sim \mathcal{N}(\mu, \sigma^2)$, d.h. X ist normalverteilt mit Parametern μ und σ^2. Für die Varianz von X gilt $\mathbb{V}_{(\mu,\sigma^2)}[X] = \sigma^2$.

Beweis.

$$\begin{aligned}
\mathbb{V}_{(\mu,\sigma^2)}[X] &= \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \\
&\overset{u=\frac{x-\mu}{\sigma}}{=} \int_{-\infty}^{\infty} \sigma^3 u^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du \\
&= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u \cdot u e^{-\frac{1}{2}u^2} du \\
&\overset{p.I.}{=} \frac{\sigma^2}{\sqrt{2\pi}} \left[-u e^{-\frac{1}{2}u^2} \Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-\frac{1}{2}u^2} du \right] \\
&= \frac{\sigma^2}{\sqrt{2\pi}} \left[0 + \sqrt{2\pi} \right] = \sigma^2
\end{aligned}$$

□

Bemerkung.
Für $X \sim \mathcal{Po}(\lambda)$ lässt sich zeigen[1], dass $\mathbb{E}_\lambda[X] = \lambda = \mathbb{V}_\lambda[X]$. Die Parameter einer Verteilung stehen oft für den Erwartungswert oder die Varianz der Zufallsvariablen.

2.3.3. Standardisierung von Zufallsvariablen

In Satz 2.46 haben wir im zweiten Schritt des Beweises im Integral mittels $u = \frac{x-\mu}{\sigma^2}$ eine Substitution durchgeführt, wodurch sich im Integranden die Dichte der Standardnormalverteilung ergab. Das Integral haben wir „standardisiert“. Dieser Ansatz soll nun auf Zufallsvariablen angewandt werden.

[1] siehe etwa [7] bzw. Beispiel 2.58

Definition 2.47

Eine Zufallsvariable X heißt standardisiert, wenn gilt:

$$\mathbb{E}_\vartheta[X] = 0 \text{ und } \mathbb{V}_\vartheta[X] = 1.$$

Sei X eine Zufallsvariable. Für

$$Y = \frac{X - \mathbb{E}_\vartheta[X]}{\sqrt{\mathbb{V}_\vartheta[X]}}.$$

und $\mathbb{E}_\vartheta[X] = \mu$ und $\mathbb{V}_\vartheta[X] = \sigma^2$ gilt

$$\begin{aligned}
\mathbb{E}_\vartheta[Y] &= \mathbb{E}_\vartheta\left[\frac{X - \mathbb{E}_\vartheta[X]}{\sqrt{\mathbb{V}_\vartheta[X]}}\right] \\
&= \mathbb{E}_\vartheta\left[\frac{X - \mu}{\sigma}\right] \\
&= \frac{1}{\sigma}\left(\mathbb{E}_\vartheta[X] - \mu\right) \\
&= 0 \quad \text{bzw.} \\
\mathbb{V}_\vartheta[Y] &= \mathbb{V}_\vartheta\left[\frac{X - \mathbb{E}_\vartheta[X]}{\sqrt{\mathbb{V}_\vartheta[X]}}\right] \\
&= \mathbb{V}_\vartheta\left[\frac{X - \mu}{\sigma}\right] \\
&= \frac{1}{\sigma^2}\mathbb{V}_\vartheta[X] \\
&= 1.
\end{aligned}$$

Y ist eine standardisierte Zufallsvariable.

Sind $X_1, \ldots, X_n$ u.i.v. Zufallsvariablen mit $\mathbb{E}_\vartheta[X_i] = \mu$ und $\mathbb{V}_\vartheta[X_i] = \sigma^2$, so definieren wir durch Bildung des arithmetischen Mittels

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

eine neue Zufallsvariable. Dann gilt

$$\mathbb{E}_\vartheta[\bar{X}] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_\vartheta[X_i] = \frac{1}{n}\cdot n\mu = \mu, \tag{2.5}$$

$$\mathbb{V}_\vartheta[\bar{X}] = \frac{1}{n^2}\mathbb{V}_\vartheta\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}_\vartheta[X_i] = \frac{\sigma^2}{n}. \tag{2.6}$$

Beispiel 2.48

Beim n-maligen Werfen eines fairen Würfels gilt $\mathbb{E}_\vartheta[\bar{X}] = \frac{7}{2}$ und $\mathbb{V}_\vartheta[\bar{X}] = \frac{35}{12n}$.

Mit Hilfe des Erwartungwertes und der Varianz einer Zufallsvariablen X lässt sich ohne Kenntnisse der Verteilung der Zufallsvariablen die Wahrscheinlichkeit abschätzen, mit der X außerhalb eines um den Erwartungswert symmetrischen Bereichs liegt. Da sehr wenig Information hineingesteckt wird, ist die Abschätzung jedoch recht grob.

Tschebyschev
1821-1894

Satz 2.49: Ungleichung von Tschebyschev

Sei X eine beliebige Zufallsvariable mit Erwartungswert $\mathbb{E}_\vartheta[X] = \mu$ und Varianz $\mathbb{V}_\vartheta[X] = \sigma^2$. Dann gilt mit $c \in \mathbb{R}^+$

$$\mathbb{P}(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}.$$

Beweis.

Es sei

$$Y = \begin{cases} 0, & |X - \mu| < c, \\ c^2, & |X - \mu| \geq c, \end{cases}$$

wobei

$$\begin{aligned} \mathbb{P}(|X - \mu| < c) &= p_1, \\ \mathbb{P}(|X - \mu| \geq c) &= p_2, \end{aligned}$$

gelte. Wegen $0 \leq |X - \mu|^2 < c^2$ im ersten und $c^2 \leq |X - \mu|^2$ im zweiten Fall ist stets $Y \leq |X - \mu|^2$. Damit folgt

$$\mathbb{E}_\vartheta[Y] \leq \mathbb{E}_\vartheta[|X - \mu|^2] = \mathbb{E}_\vartheta[(X - \mu)^2] = \mathbb{V}_\vartheta[X].$$

Weiter ist $\mathbb{E}_\vartheta[Y] = 0 \cdot p_1 + c^2 \cdot p_2 = c^2 \mathbb{P}(|X - \mu| \geq c)$ und so folgt zusammen

$$c^2 \mathbb{P}(|X - \mu| \geq c) \leq \mathbb{V}_\vartheta[X].$$

□

Mit $\mathbb{P}(|X - \mu| \geq c) = 1 - \mathbb{P}(|X - \mu| < c)$ lässt sich die Ungleichung schreiben als

$$\mathbb{P}(|X - \mu| < c) \geq 1 - \frac{\mathbb{V}_\vartheta[X]}{c^2}.$$

Beispiel 2.50

Eine Zufallsvariable besitze den Erwartungswert $\mathbb{E}_\vartheta[X] = \frac{7}{2}$ und die Varianz $\mathbb{V}_\vartheta[X] = \frac{35}{12}$. Für das arithmetische Mittel $\bar{X}$ bei $n = 50$ unabhängigen Wiederholungen gilt

$$\mathbb{P}\left(\left|\bar{X} - \frac{7}{2}\right| < 1\right) \geq 1 - \frac{35}{12n} \overset{n=50}{=} \frac{565}{600} = 0.942.$$

Für $c = k\sigma$ ergeben sich spezielle Abschätzungen mit Hilfe der Tschebychev-Ungleichung:

$$\begin{aligned}
\mathbb{P}\left(|X - \mu| < k\sigma\right) &\geq 1 - \frac{1}{k^2} \\
\mathbb{P}\left(\mu - \sigma < X < \mu + \sigma\right) &\overset{k=1}{\geq} 1 - 1 = 0 \\
\mathbb{P}\left(\mu - 2\sigma < X < \mu + 2\sigma\right) &\overset{k=2}{\geq} 1 - \frac{1}{4} = \frac{3}{4} \\
\mathbb{P}\left(\mu - 3\sigma < X < \mu + 3\sigma\right) &\overset{k=3}{\geq} 1 - \frac{1}{9} = \frac{8}{9}.
\end{aligned}$$

Bemerkung.
Unter Kenntnis der Verteilung von X lässt sich oft eine deutlich bessere Abschätzung finden.

2.3.4. Monotone Transformation von Zufallsvariablen

Ist $X : \Omega \to M$ eine Zufallsvariable und $g : M \to N$ eine invertierbare und stetig differenzierbare Funktion, so ist $Y = g(X)$ ebenfalls eine Zufallsvariable. Bei diskreten Zufallsvariablen genügt die Invertierbarkeit.

Beispiel 2.51

Es sei $X : \Omega \to \{4, 6\}$ eine Zufallsvariable mit $\mathbb{P}(X = 4) = 0.4$ und $\mathbb{P}(X = 6) = 0.6$. Sei $Y = 2X$. Wie sehen die diskrete Dichte bzw. die Verteilungsfunktion aus? Y kann zwei Werte annehmen, acht und zwölf. Dabei ist

$$\begin{aligned}
\mathbb{P}(Y = 8) &= \mathbb{P}(2X = 8) = \mathbb{P}(X = 4) = 0.4 \text{ und} \\
\mathbb{P}(Y = 12) &= \mathbb{P}(2X = 12) = \mathbb{P}(X = 6) = 0.6.
\end{aligned}$$

Mit der Verteilungsfunktion $F_X(x)$ für X,

$$F_X : \mathbb{R} \to [0, 1],\ x \mapsto F_X(x) = \begin{cases} 0, & x < 4, \\ 0.4, & 4 \leq x < 6, \\ 1, & 6 \leq x, \end{cases}$$

ergibt sich die Verteilungsfunktion $F_Y(y)$ von Y mit $F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(2X \leq$

$y) = \mathbb{P}(X \leq \frac{y}{2}) = F_X(\frac{y}{2})$ zu

$$F_Y : \mathbb{R} \to [0,1],\ y \mapsto F_Y(y) = \begin{cases} 0, & y < 8, \\ 0.4, & 8 \leq y < 12, \\ 1, & 12 \leq y. \end{cases}$$

Es sei $Y = aX + b$ für $a \neq 0$ eine affin-lineare Transformation der reellwertigen Zufallsvariablen X, so gilt für die Verteilungsfunktion F_Y von Y

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(aX + b \leq y) = \mathbb{P}(aX \leq y - b).$$

Nun müssen zwei Fälle unterschieden werden:
1. Fall: $a > 0$: $F_Y(y) = \mathbb{P}(X \leq \frac{y-b}{a}) = F_X(\frac{y-b}{a})$,
2. Fall:
$a < 0$: $F_Y(y) = \mathbb{P}(X \geq \frac{y-b}{a}) = 1 - \mathbb{P}(X < \frac{y-b}{a}) = \begin{cases} 1 - F_X(\frac{y-b}{a}), & X \text{stetig}, \\ 1 - F_X(\frac{y-b}{a}) + \mathbb{P}(X = \frac{y-b}{a}), & X \text{diskret}. \end{cases}$
Ist X eine stetige Zufallsvariable mit gegebener Dichte, so gilt zudem

$$f_Y(y) = F_Y'(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right).$$

Satz 2.52

Seien $X : \Omega \to M \subseteq \mathbb{R}$ eine stetige Zufallsvariable mit der stetigen Dichte f_X und $g : M \to N \subseteq \mathbb{R}$ eine stetig differenzierbare, streng monotone Funktion. Dann gilt für die Zufallsvariable $Y = g(X)$, dass für $y_1 < y_2$

$$\mathbb{P}(y_1 \leq Y \leq y_2) = \int_{y_1}^{y_2} f_X(g^{-1}(y)) \cdot \frac{1}{|g'(g^{-1}(y))|} dy.$$

Beweis.

Es seien $u = \min\{g^{-1}(y_1), g^{-1}(y_2)\}$ und $v = \max\{g^{-1}(y_1), g^{-1}(y_2)\}$. Nun ist g aufgrund der strengen Monotonie invertierbar und es ist $g'(x) \neq 0$ für alle $x \in M$. Mit der Substitution $x = g^{-1}(y)$ und $\frac{d}{dy} g^{-1}(y) = \frac{1}{g'(g^{-1}(y))}$ ergibt sich

$$\begin{aligned} \mathbb{P}(y_1 \leq Y \leq y_2) &= \mathbb{P}(y_1 \leq g(X) \leq y_2) = \mathbb{P}(u \leq X \leq v) \\ &= \int_{y_1}^{y_2} f_X(g^{-1}(y)) \cdot \frac{1}{|g'(g^{-1}(y))|} dy. \end{aligned}$$

Der Ausdruck $f_Y(y) = f_X(g^{-1}(y)) \cdot \frac{1}{|g'(g^{-1}(y))|}$ ist die Dichte von Y.

□

Beispiel 2.53

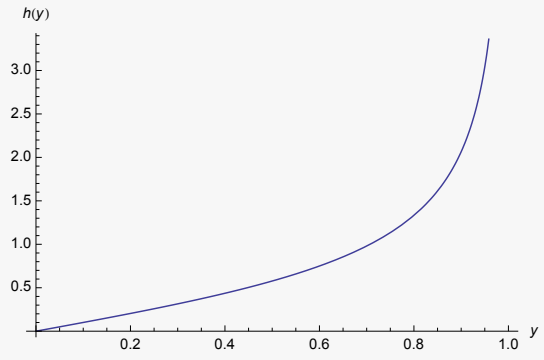
Sei $X \sim \mathcal{U}(0,1)$ eine auf dem Intervall $[0,1]$ stetig gleichverteilte Zufallsvariable und eine Funktion g gegeben durch $g : [0,1] \to [0,1]$, $g(x) = \sqrt{1-x^2}$. Dann ergibt sich die Dichte von $Y = g(X)$ mit $g^{-1}(y) = \sqrt{1-y^2}$ und $g'(x) = \frac{-x}{\sqrt{1-x^2}}$, $x \neq 1$, zu

$$\begin{aligned} f_Y(y) &\overset{y\neq 0}{=} f_X(g^{-1}(y)) \cdot \frac{1}{|g'(g^{-1}(y))|} \\ &= 1 \cdot \frac{1}{\left|\frac{-\sqrt{1-y^2}}{\sqrt{1-(\sqrt{1-y^2})^2}}\right|} \\ &= \frac{y}{\sqrt{1-y^2}} \end{aligned}$$

Wegen $\lim\limits_{y\to 0^+} f_Y(y) = 0$ legt man $f_Y(0) = 0$ fest. Es gilt

$$\begin{aligned} \int_0^1 \frac{y}{\sqrt{1-y^2}} dy &\overset{u=\sqrt{1-y^2}}{=} \int_1^0 (-1) du = 1 \\ \mathbb{E}[Y] = \int_0^1 \frac{y^2}{\sqrt{1-y^2}} dy &\overset{\text{p.I.}}{=} \int_0^1 \sqrt{1-y^2} dy \\ &\overset{y=\sin(\phi)}{=} \int_0^{\pi/2} \cos(\phi)^2 d\phi \\ &\overset{\text{p.I.}}{=} \frac{1}{2} \cdot \frac{\pi}{2} = \frac{\pi}{4}. \end{aligned}$$

Für die Varianz erhält man $\mathbb{V}[Y] = \frac{2}{3} - \frac{\pi^2}{16}$.



Satz 2.54

Sind $X : \Omega \to M_1$ und $Y : \Omega_2 \to M_2$ zwei unabhängige Zufallsvariablen und seien $V = g_1(X)$ bzw. $W = g_2(Y)$ mit $g_1 : M_1 \to N_1$ und $g_1 : M_2 \to N_2$ Zufallsvariablen. Dann sind V und W ebenfalls unabhängig.

Beweis.

Für $v \in N_1$ und $w \in N_2$ gilt:

$$\begin{aligned}\mathbb{P}(V \leq v, W \leq w) &= \mathbb{P}(g_1(X) \leq v, g_2(Y) \leq w) \\ &= \mathbb{P}(X \in g_1^{-1}(\{x \in N_1;\ x \leq v\})), Y \in g_2^{-1}(\{x \in N_2;\ x \leq w\})) \\ &= \mathbb{P}(X \in g_1^{-1}(\{x \in N_1;\ x \leq v\})) \cdot \mathbb{P}(Y \in g_2^{-1}(\{x \in N_2;\ x \leq w\})) \\ &= \mathbb{P}(V \leq v) \cdot \mathbb{P}(W \leq w)\end{aligned}$$

□

Beispiel

Sind X_1 und X_2 unabhängige Zufallsvariablen, so können für jedes $t \in \mathbb{R}$ durch $Y_1 = e^{itX_1}$ und $Y_2 = e^{itX_2}$ komplexwertige Zufallsvariablen erzeugt werden. Y_1 und Y_2 sind unabhängig.

2.3.5. Charakteristische Funktion

Ein wichtiges Hilfsmittel in der Stochastik sind charakteristische Funktionen. Insbesondere bei der Betrachtung von Summen unabhängiger Zufallsvariablen kommt diesen Funktionen eine überragende Bedeutung zu.

Definition 2.55

Als charakteristische Funktion der Zufallsvariablen X bezeichnen wir die Funktion $\phi_X : \mathbb{R} \to \mathbb{C}$,

$$\phi_X(t) = \mathbb{E}[e^{itX}].$$

Die charakteristische Funktion ist bis auf das Vorzeichen im Exponenten gleich der Fourier-Transformierten. Ist $X : \Omega \to M$ mit $M = \{x_1, x_2, \ldots\}$ eine diskrete Zufallsvariable mit $\mathbb{P}(X = x_j) = p_j$, erhalten wir

$$\phi_X(t) = \sum_j e^{itx_j} \cdot p_j.$$


Fourier
1768-1830

Mit $|e^{itx_j}| = 1$ für alle t und alle x_j und wegen $\sum_{j=1}^{k} p_j = 1$ ist die Reihe absolut (und gleichmäßig) konvergent und der Erwartungswert existiert. Die charakteristische Funktion ϕ als Summe einer gleichmäßig konvergenten Reihe von stetigen Funktionen ist somit für jedes $t \in \mathbb{R}$ stetig.

Beispiel 2.56

Eine Zufallsvariable X mit Werten in $\{0,1\}$ heißt Bernoulli-verteilt, falls $\mathbb{P}_X(\{0\}) = 1-p$ und $\mathbb{P}_X(\{1\}) = p$ für ein $p \in [0,1]$ ist. Notation: $X \sim \mathcal{B}er(p)$. Die charakteristische Funktion von X ergibt sich zu

$$\phi_X(t) = e^{it\cdot 1}\cdot p + e^{it\cdot 0}\cdot(1-p) = 1 - p + pe^{it}.$$

Ist X eine stetige Zufallsvariable mit der Dichte $f_X(x)$, dann bestimmt sich die charakteristische Funktion durch

$$\phi_X(t) = \int_{-\infty}^{\infty} f_X(x)e^{itx}dx.$$

Wegen $\int_{-\infty}^{\infty} f_X(x)|e^{itx}|dx = 1$ konvergiert das Integral absolut und gleichmäßig und somit ist ϕ stetig.

Beispiel 2.57

Es sei $X \sim \mathcal{U}(a,b)$. Dann gilt für $t \neq 0$

$$\phi_X(t) = \int_a^b \frac{1}{b-a}e^{itx}dx = \frac{1}{b-a}\left[\frac{e^{itx}}{it}\right]_a^b = \frac{1}{(b-a)it}(e^{itb} - e^{ita}).$$

Die charakteristische Funktion besitzt einige Eigenschaften, z.B.:

$$\begin{aligned}
\phi_X(0) &= \mathbb{E}[e^0] = \mathbb{E}[1] = 1\\
|\phi_X(t)| &= |\mathbb{E}[e^{itX}]| \leq \mathbb{E}[|e^{itX}|] = 1\\
\phi_X(-t) &= \mathbb{E}[e^{-itX}] = \mathbb{E}[\cos(tX) - i\sin(tX)]\\
&= \mathbb{E}[\cos(tX)] - i\mathbb{E}[\sin(tX)]\\
&= \overline{\mathbb{E}[\cos(tX)] + i\mathbb{E}[\sin(tX)]}\\
&= \overline{\mathbb{E}[\cos(tX) + i\sin(tX)]}\\
&= \overline{\mathbb{E}[e^{itX}]}\\
&= \overline{\phi_X(t)}.
\end{aligned}$$

Diese Eigenschaften sind notwendige Bedingungen einer charakteristischen Funktion, allerdings sind sie zusammen nicht hinreichend. Es gibt noch weitere Eigenschaften einer charakteristischen Funktion. Existiert etwa für eine Zufallsvariable X der Erwartungswert

$m_d := \mathbb{E}[X^d]$, $(d \in \mathbb{N}_0)$, so gilt

$$\begin{aligned}
\frac{d}{dt}\mathbb{E}[e^{itX}] &= \frac{d}{dt}\mathbb{E}\left[\sum_{j=0}^{\infty}\frac{(itX)^j}{j!}\right] \\
&= \frac{d}{dt}\sum_{j=0}^{\infty}\frac{(it)^j}{j!}\mathbb{E}[X^j] \\
\Rightarrow \phi_X^{(d)}(t) &= i^d\sum_{j=d}^{\infty}\frac{(it)^{j-d}}{(j-d)!}\mathbb{E}[X^j] \\
&\overset{l=j-d}{=} i^d\sum_{l=0}^{\infty}\frac{(it)^l}{l!}\mathbb{E}[X^{l+d}]
\end{aligned}$$

Für $t = 0$ ergibt sich $\phi_X^{(d)}(0) = i^d\mathbb{E}[X^d]$ und so

$$m_d = \mathbb{E}[X^d] = \frac{\phi_X^{(d)}(0)}{i^d}. \tag{2.7}$$

m_d heißt d-tes Moment von X. Mit Hilfe der ersten beiden d-ten Momente ergeben sich der Erwartungswert und die Varianz einer Zufallsvariablen.

Beispiel 2.58

Es sei $X \sim \mathbb{Po}(\lambda)$. Es ist

$$\begin{aligned}
\phi_X(t) &= \sum_{k=0}^{\infty}e^{itk}e^{-\lambda}\frac{\lambda^k}{k!} \\
&= e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda e^{it})^k}{k!} \\
&= e^{-\lambda}e^{\lambda e^{it}} = e^{\lambda(e^{it}-1)}.
\end{aligned}$$

Damit gilt $\phi_X'(t) = \lambda i e^{it}e^{\lambda(e^{it}-1)}$ und $\phi''(t) = \lambda i^2 e^{it}e^{\lambda(e^{it}-1)}(\lambda e^{it}+1)$ und somit

$$\begin{aligned}
\mathbb{E}_\lambda[X] &= \frac{\phi'(0)}{i} = \frac{\lambda i}{i} = \lambda \text{ und} \\
\mathbb{V}_\lambda[X] &= \mathbb{E}_\lambda[X^2] - \mathbb{E}_\lambda[X]^2 = \frac{\phi''(0)}{i^2} - \lambda^2 = \lambda(\lambda+1) - \lambda^2 = \lambda
\end{aligned}$$

Beispiel 2.59

Für $X \sim \mathcal{N}(0,1)$ gilt

$$\begin{aligned}\phi_X(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx} e^{-\frac{x^2}{2}}\, dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x-it)^2}{2}} e^{-\frac{t^2}{2}}\, dx \\ &= e^{-\frac{t^2}{2}}.\end{aligned}$$

Mit $\phi'_X(t) = -te^{-\frac{t^2}{2}}$ und $\phi''_X(t) = e^{-\frac{t^2}{2}}(t^2-1)$ ergibt sich

$$\begin{aligned}\mathbb{E}_{(\mu,\sigma^2)}[X] &= \frac{\phi'_X(0)}{i} = 0 \text{ und} \\ \mathbb{V}_{(\mu,\sigma^2)}[X] &= \mathbb{E}_{(\mu,\sigma^2)}[X^2] - \mathbb{E}_{(\mu,\sigma^2)}[X]^2 = \frac{\phi''_X(0)}{i^2} = 1\end{aligned}$$

Sei $Y = aX + b$ für beliebige reelle Konstanten a und b. Dann gilt

$$\begin{aligned}\phi_Y(t) &= \mathbb{E}[e^{itY}] \\ &= \mathbb{E}[e^{it(aX+b)}] \\ &= \mathbb{E}[e^{itaX} e^{itb}] \\ &= e^{itb}\phi_X(at).\end{aligned}$$

Beispiel 2.60

Es sei X eine Zufallsvariable mit Erwartungswert μ und Varianz σ^2. Für $Y = \frac{X-\mu}{\sigma}$ erhalten wir

$$\phi_Y(t) = e^{-\frac{it\mu}{\sigma}} \phi_X\left(\frac{t}{\sigma}\right).$$

Beispiel 2.61

Für eine normalverteilte Zufallsvariable $X \sim \mathcal{N}(\mu, \sigma^2)$ gilt $X = \sigma Y + \mu$ für $Y \sim \mathcal{N}(0,1)$ und somit

$$\phi_X(t) = e^{it\mu}\phi_Y(\sigma t) = e^{it\mu} e^{-\frac{(\sigma t)^2}{2}}$$

Der folgende Satz liefert uns die zentrale Aussage über die charakteristische Funktion der Summe unabhängiger Zufallsvariablen.

Satz 2.62

Sind $X_1, \ldots, X_n$ unabhängige Zufallsvariablen, so gilt für die charakteristische Funktion ϕ_Y von $Y = \sum_{i=1}^{n} X_i$

$$\phi_Y(t) = \phi_{X_1}(t) \cdot \ldots \cdot \phi_{X_n}(t).$$

Beweis.

Seien $X_1, \ldots, X_n$ für $n \in \mathbb{N}$ unabhängige Zufallsvariablen und $Y = \sum_{i=1}^{n} X_i$. Dann gilt

$$\begin{aligned}
\phi_Y(t) &= \mathbb{E}[e^{itY}] \\
&= \mathbb{E}\left[e^{it\sum_{i=1}^{n} X_i}\right] \\
&= \mathbb{E}\left[e^{itX_1} \cdot \ldots \cdot e^{itX_n}\right] \\
&\overset{\text{Satz (2.54)}}{=} \mathbb{E}[e^{itX_1}] \cdot \ldots \cdot \mathbb{E}[e^{itX_n}] \\
&= \phi_{X_1}(t) \cdot \ldots \cdot \phi_{X_n}(t).
\end{aligned}$$

□

Beispiel 2.63

Für die Summe von n unabhängig poissonverteilten Zufallsvariablen mit Parameter λ_j ergibt sich

$$\phi_Y(t) = \prod_{j=1}^{n} e^{\lambda_j(e^{it}-1)} = e^{\sum_{j=1}^{n} \lambda_j(e^{it}-1)} = e^{(e^{it}-1)\sum_{j=1}^{n} \lambda_j}, \quad Y \sim Po\left(\sum_{j=1}^{n} \lambda_j\right).$$

Beispiel 2.64

Sind $X_i \sim \mathbb{B}er(p)$ u.i.v. Zufallsvariablen, so ist die charakteristische Funktion der Summe $Y = \sum_{i=1}^{n} X_i$ gegeben durch

$$\phi_Y(t) = \prod_{i=1}^{n} (1 - p + pe^{it}) = (1 - p + pe^{it})^n.$$

Beispiel 2.65

Die Summe $Y = \sum_{i=1}^{n} X_i$ von n unabhängigen und normalverteilten Zufallsvariablen $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ ist wegen

$$\phi_Y(t) = \prod_{j=1}^{n} e^{it\mu_j} e^{-\frac{\sigma_j^2 t^2}{2}} = e^{it\sum_{j=1}^{n}\mu_j} e^{-\frac{t^2}{2}\sum_{j=1}^{n}\sigma_j^2}$$

normalverteilt mit $Y \sim \mathcal{N}\left(\sum_{i=1}^{n}\mu_i, \sum_{i=1}^{n}\sigma_i^2\right)$. Für das arithmetische Mittel $Z = \frac{Y}{n}$ der X_i ergibt sich $Z \sim \mathcal{N}\left(\frac{1}{n}\sum_{i=1}^{n}\mu_i, \frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2\right)$ wegen

$$\phi_Z(t) = \phi_Y\left(\frac{t}{n}\right).$$

Besitzen die X_i zudem dieselben Parameter μ und σ^2, so ist Z gemäß $Z \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ verteilt.

Doch es gibt auch nichtlineare Zusammenhänge, die zu interessanten charakteristischen Funktionen führen.

Beispiel 2.66

Es sei $X \sim \mathcal{U}(0,1)$. Dann ist für $Y = -\frac{1}{\lambda}\ln(X)$ die charakteristische Funktion gegeben durch

$$\begin{aligned}\phi_Y(t) &= \mathbb{E}[e^{itY}] = \mathbb{E}\left[e^{-it\frac{1}{\lambda}\ln(X)}\right] = \mathbb{E}\left[X^{-\frac{it}{\lambda}}\right] \\ &= \int_0^1 x^{-\frac{it}{\lambda}}\,dx = \left.\frac{x^{-\frac{it}{\lambda}+1}}{-\frac{it}{\lambda}+1}\right|_0^1 \\ &= \frac{\lambda}{\lambda - it}.\end{aligned}$$

Zwei Zufallsvariablen besitzen dieselbe Verteilung dann und genau dann, wenn ihre charakteristischen Funktionen gleich sind. Nach der Definition wird die charakteristische Funktion durch die Verteilung eindeutig bestimmt. Die Umkehrung folgt aus dem nächsten Satz.

Satz 2.67

Seien F die Verteilungsfunktion und ϕ die charakteristische Funktion einer Zufallsvariablen X. Sind $x_0 + h$ und $x_0 - h$ für $h > 0$ beliebige Stetigkeitsstellen der Verteilungsfunktion F, dann gilt

lungsfunktion F, so ist

$$F(x_0+h)-F(x_0-h)=\lim_{T\to\infty}\frac{1}{\pi}\int_{-T}^{T}\frac{\sin(ht)}{t}e^{-itx_0}\phi(t)dt.$$

Beweis.

Beweis siehe [3].

Somit ist für beliebige Stetigkeitsstellen $x_1=x_0-h$ und $x_2=x_0+h$ von F die Wahrscheinlichkeit $\mathbb{P}(x_1\leq X\leq x_2)=F(x_2)-F(x_1)$ festgelegt. Für X stetig gilt weiter

$$f_X(x)=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-itx}\phi_X(t)dt.$$

Ist X dagegen diskret, ist

$$p_j=\frac{1}{2\pi}\int_{-\pi}^{\pi}e^{-itx_j}\phi_X(t)dt.$$

Beispiel 2.68

- Das Beispiel 2.63 ergibt für die Summe von n unabhängig poissonverteilten Zufallsvariablen mit Parameter λ_j eine poissonverteilte Zufallsvariable mit Parameter $\sum_{j=1}^{n}\lambda_j$.
- In Beispiel 2.64 ergab sich für die Summe von n u.i.v. Bernoulli-verteilten Zufallsvariablen mit Parameter p die charakteristische Funktion $\phi(t)=(1-p+pe^{it})^n$. Dies ist aber genau die charakteristische Funktion einer binomialverteilten Zufallsvariablen mit Parametern n und p. Die Summe ist somit binomialverteilt. Die Verteilungsfunktion der Binomialverteilung ist

  $$F(x)=\sum_{k=0}^{x}\binom{n}{k}p^k(1-p)^{n-k}.$$

  Wir schreiben $X\sim\mathbb{B}in(n,p)$.
- Die Summe $Y=X_1+X_2$ zweier unabhängiger binomialverteilter Zufallsvariablen X_1,X_2 mit gleichem Parameter p und Anzahlen n_1 bzw. n_2 ist wegen

  $$\phi_Y(t)=(1-p+pe^{it})^{n_1}(1-p+pe^{it})^{n_2}=(1-p+pe^{it})^{n_1+n_2}$$

  wiederum binomialverteilt mit Parametern n_1+n_2 und p.

Für ein beliebiges $p > 0$ heißt

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

die Gammafunktion. Sie besitzt einige Eigenschaften, von denen hier die folgenden erwähnt seien:

(1) $\Gamma(1) = \int_0^\infty e^{-x} dx = 1$, $\Gamma(\frac{1}{2}) = \int_0^\infty x^{-\frac{1}{2}} e^{-x} dx = \sqrt{\pi}$,

(2) $\Gamma(p+1) = p\Gamma(p)$,

(3) $\Gamma(n+1) = n!$ für $n \in \mathbb{N}$,

(4) $\frac{\Gamma(p)}{a^p} = \int_0^\infty y^{p-1} e^{-ay} dy$ aufgrund der Substitution $y = \frac{x}{a}$, $a = b + ic \in \mathbb{C}$ und $b > 0$,

$$\Gamma(p) = \int_0^\infty x^{p-1} e^{-x} dx = \int_0^\infty (ay)^{p-1} e^{-ay} \cdot a dy = a^p \int_0^\infty y^{p-1} e^{-ay} dy.$$

Definition 2.69: Gammaverteilung

Seien $p > 0$ und $b > 0$. Eine Zufallsvariable $X : \Omega \to \mathbb{R}$ besitzt die Gammaverteilung, wenn für die Dichte

$$f_X(x) = \begin{cases} 0, & x < 0, \\ \frac{b^p}{\Gamma(p)} \cdot x^{p-1} e^{-bx}, & x \geq 0, \end{cases}$$

gilt, $X \sim \Gamma(p, b)$.

Dabei handelt es sich wegen

$$\int_{-\infty}^\infty f_X(x) dx = \int_0^\infty \frac{b^p}{\Gamma(p)} \cdot x^{p-1} e^{-bx} dx \overset{(4)}{=} 1$$

tatsächlich um eine Dichte. Zur Bestimmung des Erwartungswerts und der Varianz betrachten wir die charakteristische Funktion der Gammverteilung:

$$\begin{aligned} \phi_X(t) = \mathbb{E}[e^{itX}] &= \int_{-\infty}^\infty e^{itx} f_X(x) dx = \frac{b^p}{\Gamma(p)} \cdot \int_0^\infty x^{p-1} e^{-(b-it)x} dx \overset{(4)}{=} \frac{b^p}{\Gamma(p)} \cdot \frac{\Gamma(p)}{(b-it)^p} \\ &= \left(1 - \frac{it}{b}\right)^{-p}. \end{aligned}$$

Die charakteristische Funktion ist beliebig oft differenzierbar mit

$$\phi_X^{(k)}(t) = \frac{p(p+1)(p+2) \cdot \ldots \cdot (p+k-1)}{b^k} \cdot i^k \cdot \left(1 - \frac{it}{b}\right)^{-(p+k)}.$$

und somit erhalten wir mit $m_d = \frac{\phi_X^{(d)}(0)}{i^d} = \frac{p(p+1)(p+2)\cdot\ldots\cdot(p+d-1)}{b^d}$ und $d \in \mathbb{N}_0$

$$\mathbb{E}_{(p,b)}[X] = m_1 = \frac{p}{b} \text{ und } \mathbb{V}_{(p,b)}[X] = m_2 - m_1^2 = \frac{p(p+1)}{b^2} - \frac{p^2}{b^2} = \frac{p}{b^2}$$

Die Summe zweier unabhängiger Gamma-verteilter Zufallsvariablen X_1 und X_2 mit gleichem Parameter b folgt wegen

$$\phi_{X_1+X_2}(t) = \left(1 - \frac{it}{b}\right)^{-p_1} \cdot \left(1 - \frac{it}{b}\right)^{-p_2} = \left(1 - \frac{it}{b}\right)^{-(p_1+p_2)}$$

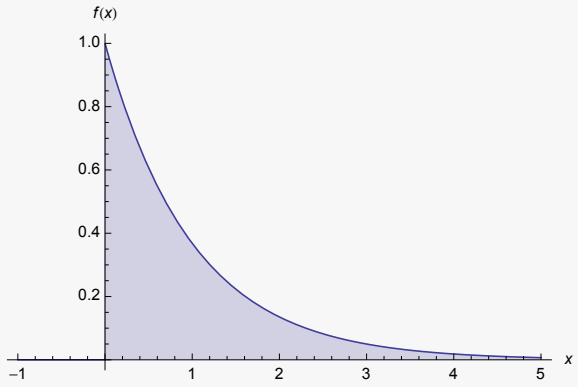
wiederum einer Gammaverteilung mit den Parametern $p_1 + p_2$ und b (Additionssatz).

Beispiel 2.70: Exponentialverteilung

Wir betrachten eine Zufallsvariable mit Gammverteilung und den Parametern $p = 1$ und $b = \lambda > 0$. Dann hat X die Dichte

$$f_X(x) = \begin{cases} 0, & x < 0, \\ \lambda e^{-\lambda x}, & x \geq 0, \end{cases}$$

mit Erwartungswert $\frac{1}{\lambda}$ und Varianz $\frac{1}{\lambda^2}$.



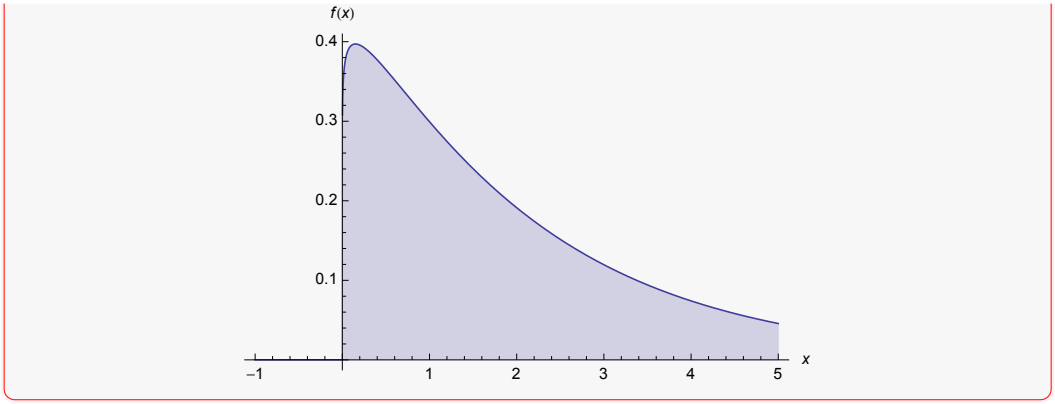
Wir sagen, X ist exponentialverteilt mit Parameter λ, $X \sim \mathcal{Exp}(\lambda)$.

Beispiel 2.71: χ^2-Verteilung

Wir betrachten jetzt eine Zufallsvariable mit Gammverteilung und den Parametern $p = \frac{1}{2}$ und $b = \frac{1}{2}$. Dann hat X die Dichte

$$f_X(x) = \begin{cases} 0, & x < 0, \\ \frac{1}{\sqrt{2\pi}} x^{-\frac{1}{2}} e^{-\frac{x}{2}}, & x \geq 0, \end{cases}$$

mit Erwartungswert 1 und Varianz 2.



χ^2-Verteilung
Es seien $X_i \sim \mathcal{N}(0,1)$ u.i.v. standardnormalverteilte Zufallsvariablen, $i = 1, \ldots, n$. Zu bestimmen ist die Verteilung von $Y = \sum_{i=1}^{n} X_i^2$. Die X_i besitzen die Dichte

$$f_{X_i}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\cdot x^2}.$$

Wir wollen zunächst die Verteilung von Y für $n = 1$ betrachten. Für die Verteilungsfunktion F_Y einer Zufallsvariablen $Y = X^2$ gilt allgemein

$$F_Y(y) \quad = \quad \begin{cases} 0, & y < 0, \\ \mathbb{P}(Y \leq y) = \mathbb{P}(X^2 \leq y), & y \geq 0. \end{cases}$$

Betrachten wir die Verteilungsfunktion für nicht-negative y weiter, erhalten wir

$$\mathbb{P}(X^2 \leq y) = \mathbb{P}(-\sqrt{y} < X < \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}),$$

und damit bei gegebener Dichte für nicht-negative y

$$f_Y(y) = \frac{F_X'(\sqrt{y}) + F_X'(-\sqrt{y})}{2\sqrt{y}} = \frac{f_X(\sqrt{y}) + f_X(-\sqrt{y})}{2\sqrt{y}}.$$

Ist X normalverteilt mit $\mu = 0$, dann besitzt $Y = X^2$ die Dichte

$$f_Y(y) \quad = \quad \begin{cases} 0, & y < 0, \\ \frac{1}{\sqrt{2\pi}} \cdot y^{-\frac{1}{2}} e^{-\frac{1}{2}\cdot y}, & y \geq 0. \end{cases}$$

Dies ist aber genau eine χ^2-Verteilung gemäß Beispiel 2.71. Die Summe n unabhängiger χ^2-verteilter Zufallsvariablen ist Gamma-verteilt mit $p = \frac{n}{2}$ und $b = \frac{1}{2}$. Deren Erwartungswert ist n und deren Varianz ergibt sich zu $2n$.

Definition 2.72: χ^2-Verteilung

Die Verteilung einer Zufallsvariablen X, die Gamma-verteilt mit Parametern $p = \frac{n}{2}$ und $b = \frac{1}{2}$ ist, d.h. X besitzt die Dichte

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

heißt χ^2-Verteilung mit n Freiheitsgraden, $X \sim \chi^2(n)$.

Eine χ^2-verteilte Zufallsvariable mit n Freiheitsgraden entspricht damit der Summe von n standardnormalverteilten Zufallsvariablen.

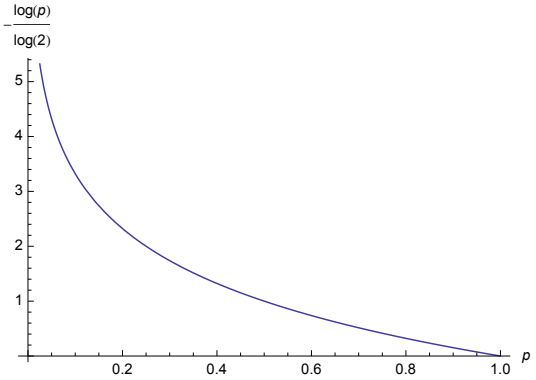
2.4. Entropie: Ein Maß der Unsicherheit

Ein Zufallsexperiment $(\Omega, \mathcal{F}, \mathbb{P})$ birgt eine gewisse Unsicherheit hinsichtlich seines Ausgangs. Es stellt sich die Frage, wie diese Unsicherheit beurteilt werden kann. Zunächst ist es unzweifelhaft, dass die Unsicherheit am kleinsten ist, wenn das Ergebnis eines Zufallsexperiments von Anfang an klar ist, d.h. es gibt ein Elementarereignis $\omega \in \Omega$, das in jedem Fall eintritt: $\mathbb{P}(A) = 1$ für jedes Ereignis A mit $\omega \in A$. Das Ergebnis überrascht nicht und liefert keinerlei Information. Ist andererseits jedes mögliche Ergebnis gleich wahrscheinlich, wäre die Unsicherheit und damit die Information, die im Ergebnis steckt, intuitiv am größten. Es spielt somit eine Rolle, mit welcher Wahrscheinlichkeit ein Ereignis auftritt. Zudem soll sich für zwei unabhängige Ereignisse die Information aufsummieren. Möchte man die Unsicherheit messen, kann deshalb zunächst eine Abbildung

$$u : [0, 1] \to [0, m)$$

angesetzt werden, die jeder Wahrscheinlichkeit eine nicht-negative Zahl zuordnet, wobei m noch zu bestimmen ist. Diese Abbildung beschreibe die Informationsmenge für eine gegebene Wahrscheinlichkeit. Aus den obigen Überlegungen fordern wir, dass $u(1) = 0$ und $u(p \cdot q) = u(p) + u(q)$ ist. Nehmen wir die plausible Forderung, dass u stetig ist (eine beliebig kleine Veränderung der Wahrscheinlichkeit soll auch eine beliebig kleine Änderung der Informationsmenge bewirken) und die „Normierungsforderung“ $u(\frac{1}{2}) = 1$ hinzu, ergibt sich, dass die Abbildung eindeutig durch den negativen dualen Logarithmus festgelegt ist[2]. Dabei gilt $\lim_{x \to 0^+} -\log_2(x) = \infty$. Somit ist $m = \infty$ zu setzen.

[2] siehe [6]



In einem diskreten Wahrscheinlichkeitsraum lässt sich jedem Elementarereignis $\omega \in \Omega$ dessen Wahrscheinlichkeit $\mathbb{P}(\{\omega\})$ zuordnen. Das gewichtete arithmetische Mittel der negativen dualen Logarithmen $-\log_2(\mathbb{P}(\{\omega\}))$ für die Elementarereignisse mit den Gewichten $\mathbb{P}(\{\omega\})$ führt zur mittleren Informationsmenge.


Shannon
1916-2001

Definition 2.73: Shannon-Entropie

Es sei $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ ein diskreter Wahrscheinlichkeitsraum. Dann wird die mittlere Informationsmenge $\mathbb{S}_\mathbb{P}$ gegeben durch das gewichtete arithmetische Mittel

$$\mathbb{S}_\mathbb{P} := -\sum_{\omega\in\Omega} \mathbb{P}(\{\omega\}) \cdot \log_2(\mathbb{P}(\{\omega\})).$$

als (Shannon-)Entropie bezeichnet. Dabei gelte $0 \cdot \log_2(0) := 0$.

Die Shannon-Entropie $\mathbb{S}_\mathbb{P}$ kann als Erwartungswert interpretiert werden, als mittlere Informationsmenge einer diskreten Verteilung $\mathbb{P}$. Da der duale Logarithmus zugrunde liegt, kann auch von der benötigten mittleren Anzahl Bits zur Beschreibung der Zufallsvariablen gesprochen werden. Wie wir bald sehen werden, gilt $0 \leq \mathbb{S}_\mathbb{P} \leq \log_2(n)$. Durch Skalieren mit dem Faktor $\frac{1}{\log_2(n)}$ wird $\mathbb{S}_\mathbb{P}$ auf das Intervall $[0, 1]$ normalisiert. Der errechnete Wert kann als Maß für die in einer diskreten Verteilung erwartete Unsicherheit hinsichtlich einer Realisierung gesehen werden. Intuitiv wird die mittlere Informationsmenge am größten, wenn eine Gleichverteilung vorliegt. um das einzusehen, betrachten wir zunächst ein einfaches Beispiel.

Beispiel 2.74

Beim Werfen einer Münze, bei der mit Wahrscheinlichkeit p Kopf (K) und mit Wahrscheinlichkeit $1 - p$ Zahl (Z) fällt, gilt für die Shannon-Entropie:

$$\begin{aligned} \mathbb{S}_\mathbb{P} &= -\mathbb{P}(\{\mathsf{K}\}) \cdot \log_2(\mathbb{P}(\{\mathsf{K}\})) - \mathbb{P}(\{\mathsf{Z}\}) \cdot \log_2(\mathbb{P}(\{\mathsf{Z}\})) \\ &= -p \cdot \log_2(p) - (1-p) \cdot \log_2(1-p). \end{aligned}$$

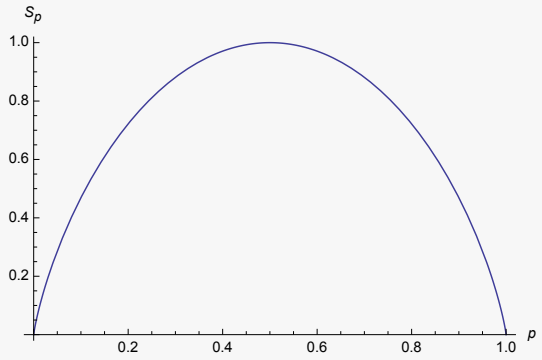
Wegen

$$\frac{d\mathbb{S}_{\mathbb{P}}}{dp} = -\frac{1}{\log(2)} \cdot (\log(p) - \log(1-p)) \stackrel{!}{=} 0 \Leftrightarrow p = \frac{1}{2}$$

und

$$\frac{d^2\mathbb{S}_{\mathbb{P}}}{dp^2} = -\frac{1}{\log(2)} \cdot \left(\frac{1}{p(1-p)}\right) < 0$$

für $p \in (0,1)$ ist $\mathbb{S}_{\mathbb{P}}$ konkav und damit liegt bei $p = \frac{1}{2}$ ein Maximum vor. Es ist in diesem Fall $\mathbb{S}_{\mathbb{P}} = 1$. Der Graph der Entropie abhängig von p sieht folgendermaßen aus:



Im Beispiel ergibt sich für $|\Omega| = 2$, dass die Gleichverteilung die maximale Entropie $\mathbb{S}_{\mathbb{P}} = \frac{\log(2)}{\log(2)} = 1$ liefert. Wie sieht es allgemein bei $|\Omega| = n \in \mathbb{N}$ aus?

$$\mathbb{S}_{\mathbb{P}} = \sum_{i=1}^{n}(-1) \cdot \frac{1}{n} \cdot \log_2\left(\frac{1}{n}\right) = -\frac{1}{\log(2)} \cdot \frac{n}{n} \cdot \log\left(\frac{1}{n}\right) = \frac{\log(n)}{\log(2)} = \log_2(n).$$

Satz und Definition 2.75: Ungleichung von Jensen

Es sei f eine konvexe Funktion, d.h. es gelte für $\lambda \in [0,1]$

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y).$$

Dann ist für nicht-negative λ_i mit $\sum_{i=1}^{n} \lambda_i = 1$

$$f\left(\sum_{i=1}^{n} \lambda_i x_i\right) \leq \sum_{i=1}^{n} \lambda_i f(x_i).$$

Beweis.

Induktion nach n:
$n = 1$: $\lambda_1 = 1$, $f(x_1) \leq f(x_1)$,

$n \to n+1$: Gegeben seien nicht-negative λ_i mit $\sum_{i=1}^{n+1} \lambda_i = 1$. Dann ist zunächst

$$\sum_{i=1}^{n} \lambda_i = 1 - \lambda_{n+1}.$$

$$\begin{aligned}
f\left(\sum_{i=1}^{n+1} \lambda_i x_i\right) &= f\left(\sum_{i=1}^{n} \lambda_i x_i + \lambda_{n+1} x_{n+1}\right) \\
&= f\left((1-\lambda_{n+1}) \sum_{i=1}^{n} \frac{\lambda_i}{1-\lambda_{n+1}} x_i + \lambda_{n+1} x_{n+1}\right) \\
&\overset{\text{Def. 2.75}}{\leq} (1-\lambda_{n+1}) f\left(\sum_{i=1}^{n} \frac{\lambda_i}{1-\lambda_{n+1}} x_i\right) + \lambda_{n+1} f(x_{n+1}) \\
&\overset{\text{I.V.}}{\leq} (1-\lambda_{n+1}) \sum_{i=1}^{n} \frac{\lambda_i}{1-\lambda_{n+1}} f(x_i) + \lambda_{n+1} f(x_{n+1}) \\
&= \sum_{i=1}^{n+1} \lambda_i f(x_i).
\end{aligned}$$

Satz 2.76

Es sei $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ ein diskreter Wahrscheinlichkeitsraum mit $|\Omega| = n \in \mathbb{N}$. Dann gilt $0 \leq \mathbb{S}_{\mathbb{P}} \leq \log_2(n)$. Die Gleichheit gilt genau dann, wenn $\mathbb{P}(\{\omega\}) = \frac{1}{n}$ für alle $\omega \in \Omega$.

Beweis.

Die Abschätzung $0 \leq \mathbb{S}_{\mathbb{P}}$ ist wegen der Definition von $\mathbb{S}_{\mathbb{P}}$ klar. Wir betrachten die Funktion

$$f : [0,1] \to \mathbb{R}, \quad x \mapsto \begin{cases} 0, & x = 0, \\ x \log_2(x), & x \neq 0. \end{cases}$$

Wegen $f''(x) = \frac{1}{x \log(2)} > 0$ ist f strikt konvex. Damit folgt mit der Ungleichung von Jensen ($\lambda_i = \frac{1}{n}$)

$$f\left(\frac{1}{n} \sum_{i=1}^{n} x_i\right) \leq \frac{1}{n} \sum_{i=1}^{n} f(x_i), \quad x_1, \ldots, x_n \in [0,1].$$

Für $x_1 = x_2 = \ldots = x_n$ gilt die Gleichheit. Mit $x_i = \mathbb{P}(\{\omega_i\})$, $\omega_i \in \Omega$ folgt nun

$$\begin{aligned} & f\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(\{\omega_i\})\right) \leq \frac{1}{n}\sum_{i=1}^{n} f(\mathbb{P}(\{\omega_i\})) \\ \Leftrightarrow \quad & \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(\{\omega_i\})\log_2\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(\{\omega_i\})\right) \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(\{\omega_i\})\log_2(\mathbb{P}(\{\omega_i\})) \\ \Leftrightarrow \quad & \frac{1}{n}\log_2\left(\frac{1}{n}\right) \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(\{\omega_i\})\log_2(\mathbb{P}(\{\omega_i\})) \\ \Leftrightarrow \quad & \log_2(n) \geq -\sum_{i=1}^{n}\mathbb{P}(\{\omega_i\})\log_2(\mathbb{P}(\{\omega_i\})). \end{aligned}$$

Da f strikt konvex ist, wird der Maximalwert der Funktion nur für die Gleichverteilung. Also gilt die Gleichheit genau dann, wenn $\mathbb{P}(\{\omega_i\}) = \frac{1}{n}$ für alle $i = 1, \ldots, n$ ist.

□

Eine Möglichkeit, die Unähnlichkeit zweier diskreter Verteilungen $\mathbb{P} : \mathcal{F} \to [0,1]$ und $\mathbb{Q} : \mathcal{F} \to [0,1]$ auf dem selben Messraum $(\Omega, \mathcal{F})$ zu untersuchen, besteht in der Berechnung der Kullback-Leibler Divergenz

$$KL(\mathbb{P}||\mathbb{Q}) = \sum_{i=1}^{n}\mathbb{P}(\{\omega_i\})\log_2\left(\frac{\mathbb{P}(\{\omega_i\})}{\mathbb{Q}(\{\omega_i\})}\right). \tag{2.8}$$

Dabei gehen wir davon aus, dass aus $\mathbb{Q}(F) = 0$ für ein $F \in \mathcal{F}$ stets $\mathbb{P}(F) = 0$ („$\mathbb{P}$ ist absolutstetig bzgl. $\mathbb{Q}$“) und setzen $0\log_2(\frac{0}{0}) = 0$. Weiter gilt

$$KL(\mathbb{P}||\mathbb{Q}) = \sum_{i=1}^{n}\mathbb{P}(\{\omega_i\})\log_2(\mathbb{P}(\{\omega_i\})) - \sum_{i=1}^{n}\mathbb{P}(\{\omega_i\})\log_2(\mathbb{Q}(\{\omega_i\})) = -\mathbb{S}_{\mathbb{P}} + \mathbb{S}_{\mathbb{P},\mathbb{Q}},$$

wobei

$$\mathbb{S}_{\mathbb{P},\mathbb{Q}} = -\sum_{i=1}^{n}\mathbb{P}(\{\omega_i\})\log_2(\mathbb{Q}(\{\omega_i\}))$$

die Kreuzentropie von $\mathbb{P}$ und $\mathbb{Q}$ genannt wird. Die Kullback-Leibler Divergenz ist ein Maß für die Ineffizienz für die Annahme der Verteilung $\mathbb{Q}$, falls $\mathbb{P}$ die wahre Verteilung ist. Dabei ist KL nicht symmetrisch, d.h. i.A. ist $KL(\mathbb{P}||\mathbb{Q}) \neq KL(\mathbb{Q}||\mathbb{P})$:

Beispiel 2.77

Sei $\Omega = \{0,1\}$ gegeben. Mit $\mathbb{P}(\{0\}) = \mathbb{P}(\{1\}) = \frac{1}{2}$ sowie $\mathbb{Q}(\{0\}) = \frac{1}{3}$ und $\mathbb{Q}(\{1\}) = \frac{2}{3}$ gilt

$$\begin{aligned} KL(\mathbb{P}||\mathbb{Q}) &= \frac{1}{2}\log_2\left(\frac{3}{2}\right) + \frac{1}{2}\log_2\left(\frac{3}{4}\right) = 0.085, \\ KL(\mathbb{Q}||\mathbb{P}) &= \frac{1}{3}\log_2\left(\frac{2}{3}\right) + \frac{2}{3}\log_2\left(\frac{4}{3}\right) = 0.082, \end{aligned}$$

und damit $KL(\mathbb{P}||\mathbb{Q}) \neq KL(\mathbb{Q}||\mathbb{P})$.

Jedoch nimmt KL stets nicht-negative Werte an.

Satz 2.78

Es gilt: $KL(\mathbb{P}||\mathbb{Q}) \geq 0$ und $KL(\mathbb{P}||\mathbb{Q}) = 0$ genau dann, wenn $\mathbb{P} = \mathbb{Q}$.

Beweis.

Mit der Ungleichung von Jensen gilt

$$\begin{aligned} -KL(\mathbb{P}||\mathbb{Q}) &= -\sum_{i=1}^{n} \mathbb{P}(\{\omega_i\}) \log_2 \left(\frac{\mathbb{P}(\{\omega_i\})}{\mathbb{Q}(\{\omega_i\})} \right) \\ &\leq \log_2 \left(\sum_{i=1}^{n} \mathbb{P}(\{\omega_i\}) \frac{\mathbb{Q}(\{\omega_i\})}{\mathbb{P}(\{\omega_i\})} \right) \\ &= \log_2 \left(\sum_{i=1}^{n} \mathbb{Q}(\{\omega_i\}) \right) \\ &= 0. \end{aligned}$$

Aufgrund der konkaven Logarithmusfunktion liegt die Gleichheit genau in dem Fall vor, wenn das Argument eine Konstante ist, d.h. $\mathbb{P}(\{\omega_i\}) = k \cdot \mathbb{Q}(\{\omega_i\})$ für ein $k \in \mathbb{R}$. Mit $\sum_{i=1}^{n} \mathbb{P}(\{\omega_i\}) = \sum_{i=1}^{n} \mathbb{Q}(\{\omega_i\}) = 1$ folgt $k = 1$.

□

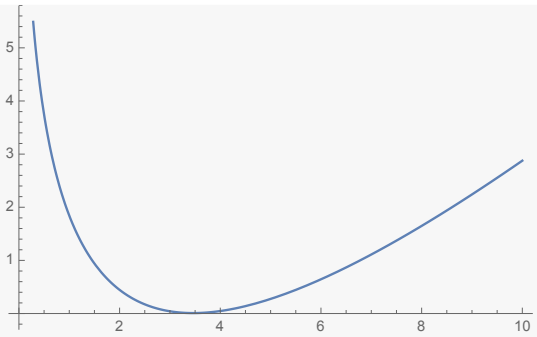
Aus Satz 2.78 folgt, dass $\mathbb{S}_{\mathbb{P},\mathbb{Q}} \geq \mathbb{S}_{\mathbb{P}}$ ist. Mit Hilfe der Kullback-Leibler Divergenz (oder der Kreuzentropie) lässt sich eine Verteilung durch eine andere schätzen. Je kleiner der Wert, desto besser ist die Schätzung. Im Idealfall ergibt sich keine Abweichung.

Beispiel 2.79

Von einer unbekannten Verteilung seien die folgenden Wahrscheinlichkeiten bekannt:

ω	$\mathbb{P}(\{\omega\})$	ω	$\mathbb{P}(\{\omega\})$
0	0.022	10	$6.7 \cdot 10^{-4}$
1	0.094	11	$1.3 \cdot 10^{-4}$
2	0.186	12	$2.0 \cdot 10^{-5}$
3	0.234	13	$2.6 \cdot 10^{-6}$
4	0.208	14	$2.7 \cdot 10^{-7}$
5	0.139	15	$2.2 \cdot 10^{-8}$
6	0.073	16	$1.5 \cdot 10^{-9}$
7	0.030	17	$7.2 \cdot 10^{-11}$
8	0.010	18	$2.5 \cdot 10^{-12}$
9	0.003	19	$5.6 \cdot 10^{-14}$

Die Verteilung soll durch eine Poissonverteilung geschätzt werden. Folgende Abbildung zeigt die Kullback-Leibler Divergenz in Abhängigkeit des Parameters λ.



Die minimale Kullback-Leibler Divergenz (0.009) ergibt sich bei einem Wert von etwa $\lambda = 3.46$ (rechnerisch lösbar!).

2.5. Grenzwertsätze

Verteilungen von Funktionen von Zufallsvariablen spielen eine große Rolle. Sind n Zufallsvariablen $X_n : \Omega \to \Psi$ gegeben, so interessiert man sich häufig für die Verteilung des arithmetischen Mittels

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

Für große n wollen wir für das arithmetische Mittel verschiedene Eigenschaften zeigen. Allerdings ist dabei stets vorausgesetzt, dass der Erwartungswert und die Varianz existieren. Insbesondere ist das standardisierte arithmetische Mittel für große n annähernd standardnormalverteilt. Diese Aussage ist für viele Anwendungen in der mathematischen Statistik von großem Nutzen.

Definition 2.80: Konvergenz in Wahrscheinlichkeit

Eine Folge $(X_n)_{n\in\mathbb{N}}$ von Zufallsvariablen konvergiert stochastisch gegen Null, wenn für beliebiges $\epsilon > 0$ gilt

$$\lim_{n\to\infty} \mathbb{P}(|X_n| > \epsilon) = 0. \tag{2.9}$$

Notation: $X_n \xrightarrow{\mathbb{P}} 0$.

Die bedeutet, dass die Wahrscheinlichkeit für das Ereignis $|X_n| > \epsilon$ für $n \to \infty$ gegen Null geht. Wegen

$$|X_n| > \epsilon = \{\omega \in \Omega;\ |X_n(\omega)| > \epsilon\} = \{\omega \in \Omega;\ |X_n(\omega)| \leq \epsilon\}^C = (|X_n| \leq \epsilon)^C$$

folgt

$$\mathbb{P}(|X_n| > \epsilon) \overset{n\to\infty}{\longrightarrow} 0 \Leftrightarrow \mathbb{P}(|X_n| \leq \epsilon) \overset{n\to\infty}{\longrightarrow} 1.$$

Damit erhalten wir $\mathbb{P}(-\epsilon < X_n < \epsilon) \overset{n\to\infty}{\longrightarrow} 1$. Ist $F_n : \Psi \to [0,1]$ die Verteilungsfunktion der X_n, dann bedeutet das

$$F_n(\epsilon) - F_n(-\epsilon) \overset{n\to\infty}{\longrightarrow} 1.$$

Mit (2.9) gilt dann auch

$$0 \overset{n\to\infty}{\longleftarrow} \mathbb{P}(X_n > \epsilon) = 1 - \mathbb{P}(X_n \leq \epsilon) = 1 - F_n(\epsilon).$$

Damit folgt zuletzt $F_n(-\epsilon) \overset{n\to\infty}{\longrightarrow} 0$. Das bedeutet, dass die Folge der Verteilungsfunktionen der Bedingung

$$\lim_{n\to\infty} F_n(x) = \begin{cases} 0, & x < 0, \\ 1, & x > 0, \end{cases} \tag{2.10}$$

genügt. Für $x \neq 0$ konvergiert F_n gegen die Einpunktverteilung. Gilt umgekehrt (2.10), ist für ein beliebiges $\epsilon > 0$

$$\begin{aligned} \lim_{n\to\infty} \mathbb{P}(X_n < -\epsilon) &\leq \lim_{n\to\infty} \mathbb{P}(X_n \leq -\epsilon) = \lim_{n\to\infty} F_n(-\epsilon) = 0, \\ \lim_{n\to\infty} \mathbb{P}(X_n > \epsilon) &\leq \lim_{n\to\infty} (1 - \mathbb{P}(X_n \leq \epsilon)) = \lim_{n\to\infty} (1 - F_n(\epsilon)) = 0. \end{aligned}$$

Aus der stochastischen Konvergenz folgt die Konvergenz der Folge der Verteilungsfunktion gegen die Einpunktverteilung in jeder Stetigkeitsstelle und umgekehrt.

Definition 2.81

Die Folge $(F_n)_{n\in\mathbb{N}}$ der Verteilungsfunktionen einer Folge $(X_n)_{n\in\mathbb{N}}$ von Zufallsvariablen heißt konvergent, falls eine Verteilungsfunktion $F : \mathbb{R} \to [0, 1]$ existiert, so dass für jede Stetigkeitsstelle $x \in \mathbb{R}$ von F die Beziehung

$$\lim_{n\to\infty} F_n(x) = F(x)$$

erfüllt ist. Die Verteilungsfunktion F heißt dann Grenzverteilungsfunktion.

Bemerkung.
(1) Stochastische Konvergenz bedeutet, dass die Grenzverteilungsfunktion die Einpunktverteilung ist.
(2) Konvergenz der Verteilungsfunktionenfolge ist schwächer als die punktweise Konvergenz von Funktionenfolgen in der Analysis.

Die stochstische Konvergenz lässt sich in verschiedenen Konstellationen formulieren:

- Eine Folge $(X_n)_{n\in\mathbb{N}}$ von Zufallsvariablen konvergiert stochastisch gegen Null genau dann, wenn die Folge $(F_n)_{n\in\mathbb{N}}$ ihrer Verteilungsfunktionen gegen die Verteilungsfunktion der Einpunktverteilung in jeder Stetigkeitsstelle dieser Funktion konvergiert.
- Ist c eine beliebige Konstante, so konvergiert $(X_n)_{n\in\mathbb{N}}$ stochastisch gegen c, wenn $(Y_n)_{n\in\mathbb{N}} = (X_n - c)_{n\in\mathbb{N}}$ stochastisch gegen Null konvergiert. Notation: $X_n \xrightarrow{\mathbb{P}} c$
- Es konvergiert $(X_n)_{n\in\mathbb{N}}$ stochastisch gegen eine Zufallsvariable X, wenn $(Y_n)_{n\in\mathbb{N}} = (X_n - X)_{n\in\mathbb{N}}$ stochastisch gegen Null konvergiert. Notation: $X_n \xrightarrow{\mathbb{P}} X$

2.5.1. Schwaches Gesetz der großen Zahlen

Bei einer großen Anzahl an unabhängigen und identisch verteilten Zufallsvariablen lässt sich bei existierendem Erwartungswert und existierender Varianz zeigen, dass sich das arithmetische Mittel mit großer Wahrscheinlichkeit in einem kleinen Intervall um den Erwartungswert realisiert.

Satz 2.82: Schwaches Gesetz der großen Zahlen

Sei $(X_n)_{n\in\mathbb{N}}$ eine Folge u.i.v. Zufallsvariablen mit $\mathbb{E}_\theta[X_n] = \mu$ und $\mathbb{V}_\theta[X_n] = \sigma^2$. Dann konvergiert die Folge $(\bar{X}_n)_{n\in\mathbb{N}}$ der arithmetischen Mittel $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ stochastisch (in Wahrscheinlichkeit) gegen μ, d. h.

$$\mathbb{P}(|\bar{X}_n - \mu| \geq c) \leq \frac{\sigma^2}{nc^2} \overset{n\to\infty}{\longrightarrow} 0,\ \bar{X}_n \overset{\mathbb{P}}{\to} \mu.$$

Beweis.

Wir betrachten für ein beliebiges aber festes $c > 0$ die Ungleichung von Tschebyschev:

$$\mathbb{P}(|\bar{X}_n - \mu| < c) \geq 1 - \frac{\sigma^2/n}{c^2} = 1 - \frac{\sigma^2}{nc^2} \overset{n\to\infty}{\longrightarrow} 1.$$

Damit gilt auch $\mathbb{P}(|\bar{X}_n - \mu| \leq c) \overset{n\to\infty}{\longrightarrow} 1$ und $\bar{X}_n \overset{\mathbb{P}}{\longrightarrow} \mu$.

□

Satz und Definition 2.83: Satz von Bernoulli

Es sei $(Y_n)_{n\in\mathbb{N}}$ eine Folge u.i.v. Zufallsvariablen mit $Y_n : \Omega \to \Psi$. Es soll mitgezählt werden, wie oft ein bestimmtes Ereignis A (mit $\mathbb{P}(A) = p$) eingetreten ist. Es seien dazu $X_n : \Omega \to \{0, 1\}$ gegeben[a] durch $X_n := I_A(Y_i(\omega))$. Für die Folge $(X_n)_{n\in\mathbb{N}}$ sei $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Dann konvergiert $\bar{X}_n$ stochastisch gegen die unbekannte Wahrscheinlichkeit p für das Eintreten des Ereignisses A. $\bar{X}_n$ heißt deshalb die relative Häufigkeit des Ereignisses A.

[a] I_A ist dabei die Indikatorfunktion, d.h. $I_A(x) = 1$, falls $x \in A$, ansonsten Null.

Beweis.

Es ist $(X_n)_{n\in\mathbb{N}}$ eine Folge u.i.v. Bernoulli-verteilter Zufallsvariablen mit $\mathbb{E}_\theta[X_n] = p$ und $\mathbb{V}_\theta[X_n] = p(1-p)$, so gilt für jedes $c > 0$ mit dem schwachen Gesetz der großen Zahlen

$$\mathbb{P}(|\bar{X}_n - p| < c) \geq 1 - \frac{p(1-p)}{nc^2} \overset{n\to\infty}{\longrightarrow} 1.$$

□

Die relative Häufigkeit kann damit zur Schätzung des Parameters p verwendet werden. Je größer n, umso „sicherer“ wird die Schätzung sein. Für gegebenes c werde eine Sicherheitswahrscheinlichkeit $1 - \alpha$ mit $0 < \alpha < 1$ festgelegt. Dann ist

$$\mathbb{P}(|\bar{X}_n - p| < c) \geq 1 - \alpha$$

genau dann, wenn

$$1 - \frac{p(1-p)}{nc^2} \geq 1 - \alpha$$

gilt. Da n der einzig freie Parameter ist, wird die Ungleichung nach n aufgelöst zu

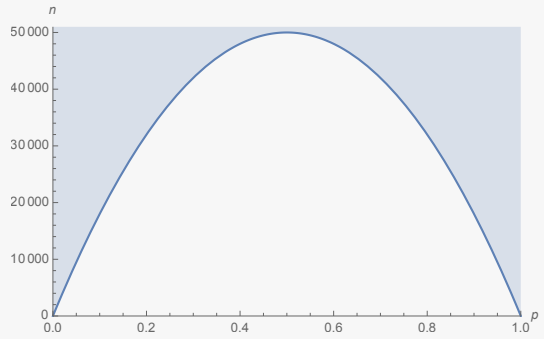
$$n \geq \frac{p(1-p)}{\alpha c^2}.$$

Beispiel 2.84

Wie groß muss n gewählt werden, damit mit einer Sicherheitswahrscheinlichkeit von 0.95 (d.h. $\alpha = 0.05$) die relative Häufigkeit von der Wahrscheinlichkeit $p = \frac{1}{7}$ um weniger als $c = \frac{1}{100}$ abweicht? Dazu muss

$$n \geq \frac{\frac{1}{7} \cdot \frac{6}{7}}{\frac{5}{100} \cdot \frac{1}{100^2}} \approx 24490$$

sein. Für $p \in [0, 1]$ zeigt die Abbildung die in der Situation notwendigen Stichprobenumfänge.



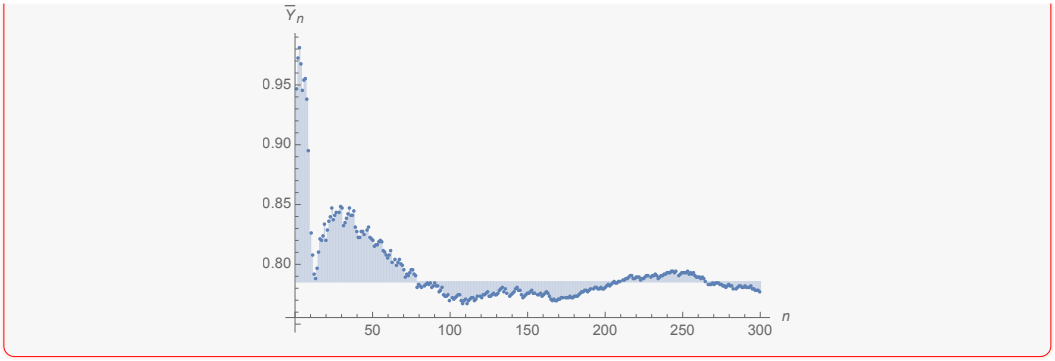
Eine weitere Anwendung des Satzes von Bernoulli ist das Berechnen eines bestimmten Integrals.

Beispiel 2.85

Es soll das Integral $\mu = \int_0^1 \sqrt{1-x^2}dx = \frac{\pi}{4} = 0.785$ numerisch berechnet werden. Sei dazu $(X_n)_{n\in\mathbb{N}}$ eine Folge unabhängiger auf $[0, 1]$ gleichverteilter Zufallsvariablen. Mit $q(x) = \sqrt{1-x^2}$ lässt sich die Zufallsvariable $Y_n = q(X_n)$ mit Erwartungswert $\mathbb{E}_\theta[Y_n] = \mu$ und Varianz $\mathbb{V}_\theta[Y_n] = \frac{2}{3} - \frac{\pi^2}{16}$ definieren (vgl. Beispiel 2.53). Dann gilt

$$\mathbb{P}(|\bar{Y}_n - \mu| < c) \geq 1 - \frac{\mathbb{V}_\theta[Y_n]}{nc^2} \overset{n\to\infty}{\longrightarrow} 1.$$

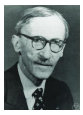
Folgende Abbildung zeigt eine Simulation der Zufallsvariablen $\bar{Y}_i$ mit $i = 1, \ldots, 300$.



2.5.2. Zentraler Grenzwertsatz

Für eine Folge $(X_n)_{n\in\mathbb{N}}$ u.i.v. Zufallsvariablen mit Erwartungswert $\mathbb{E}_\theta[X_n] = \mu$ und Varianz $\mathbb{V}_\theta[X_n] = \sigma^2$ konvergiert $\bar{X}$ stochastisch gegen den Erwartungswert μ. Selbst bei bekannter Verteilung der X_n lässt sich mit dem schwachen Gesetz der großen Zahlen nichts weiter aussagen. Um über das Verhalten von Folgen von Zufallsvariablen hinsichtlich ihrer Verteilungsfunktion Aussagen treffen zu können, benötigen wir einen Satz, der auf Lévy und Cramér[3] zurückgeht.


Cramér
1893-1985


Lévy
1886-1971

Satz 2.86

Es sei $(X_n)_{n\in\mathbb{N}}$ eine Folge von Zufallsvariablen und F_n bzw. ϕ_n seien die dazugehörigen Verteilungsfunktionen bzw. charakteristischen Funktionen. Die Folge $(F_n)_{n\in\mathbb{N}}$ strebt genau dann für $n \to \infty$ gegen die Verteilungsfunktion F, wenn in einem gewissen Werteintervall $|t| \leq \tau$ die Folge $(\phi_n)_{n\in\mathbb{N}}$ gleichmäßig gegen eine gewisse Funktion ϕ konvergiert. Die Grenzfunktion ϕ ist dann die charakteristische Funktion der Grenzverteilungsfunktion F und die Konvergenz $\phi_n \overset{n\to\infty}{\longrightarrow} \phi$ ist gleichmäßig in jedem endlichen Intervall.

Der nicht einfache Beweis findet sich z.B. in [3]. Mit dem nachfolgenden Grenzwertsatz von Lindeberg-Lévy lässt sich zeigen, dass das standardisierte arithmetische Mittel mit Wahrscheinlichkeit Eins gegen die Standardnormalverteilung strebt.


Lindeberg
1876-1932

Satz 2.87: Zentraler Grenzwertsatz

Sei $(X_n)_{n\in\mathbb{N}}$ eine Folge u.i.v. Zufallsvariablen mit Erwartungswert $\mathbb{E}_\theta[X_n] = \mu$ und Varianz $\mathbb{V}_\theta[X_n] = \sigma^2$. Dann gilt für die Verteilungsfunktion der standardisierten Zufallsvariablen des arithmetischen Mittels

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

[3] Nebenstehende Fotos von Konrad Jacobs, Erlangen, siehe http://owpdb.mfo.de/detail?photo_id=745 bzw. http://owpdb.mfo.de/detail?photo_id=2531

für alle $y \in \mathbb{R}$ der Grenzwert

$$\lim_{n\to\infty} F_n(y) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{y} e^{-\frac{y^2}{2}} dy.$$

Somit gilt

$$\mathbb{P}(Y_n \leq y) \overset{n\to\infty}{\longrightarrow} \Phi(y), \text{ d. h. } Y_n \sim \mathcal{N}(0,1) \text{ für } n \to \infty.$$

Peano
1858-1932

Taylor
1685-1731

Beweis.

Es sei in der obigen Voraussetzung Y_n geschrieben als

$$Y_n = \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu).$$

Die Zufallsvariablen $X_i - \mu$ besitzen alle dieselbe Verteilung und haben damit auch dieselbe charakteristische Funktion ϕ_X. Die charakteristische Funktion ϕ_{Y_n} von Y_n hat die Gestalt

$$\phi_{Y_n}(t) = \left(\phi_X\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n. \tag{2.11}$$

Es ist

$$\mathbb{E}[X_i - \mu] = 0 = \phi'_X(0)/i \text{ und } \mathbb{V}[X_i - \mu] = \sigma^2 = -\phi''_X(0).$$

Wir bestimmen die Taylorentwicklung um $t_0 = 0$ von ϕ_X zu

$$\begin{aligned} \phi_X(t) &= \phi_X(0) + \phi'_X(0)\cdot t + \frac{1}{2}\phi''_X(0)\cdot t^2 + R_2(t,\xi) \\ &= 1 + 0 + \frac{1}{2}\cdot(-\sigma^2)\cdot t^2 + R_2(t,\xi), \end{aligned}$$

wobei

$$R_2(t,\xi) = \frac{\phi''_X(\xi) - \phi''_X(0)}{2!}\cdot t^2, \quad \xi \text{ zwischen 0 und } t,$$

das Peano-Restglied der Taylorentwicklung ist. Wir setzen dies in die Gleichung (2.11) ein:

$$\begin{aligned} \phi_{Y_n}(t) &= \left(1 - \frac{t^2}{2n} + (\phi''_X(\xi) + \sigma^2)\frac{t^2}{2\sigma^2 n}\right)^n \\ &= \left(1 + \phi''_X(\xi)\frac{t^2}{2\sigma^2 n}\right)^n \text{ für ein } \xi \text{ zwischen } 0 \text{ und } \frac{t}{\sigma\sqrt{n}}. \end{aligned}$$

Mit $u = \phi''_X(\xi)\frac{t^2}{2\sigma^2 n}$ erhalten wir durch Logarithmieren

$$\log\left(\phi_{Y_n}(t)\right) = n\log(1+u).$$

Wegen $\xi \in \left[0, \frac{t}{\sigma\sqrt{n}}\right]$ gilt bei festen Werten t und σ^2

$$\lim_{n\to\infty} \phi''_X(\xi) = \phi''_X(0) = -\sigma^2$$

und so folgt für große n

$$|u| = \left|\phi''_X(\xi)\frac{t^2}{2\sigma^2 n}\right| < 1.$$

Das lässt die lineare Näherung $\log(1+u) = u$ zu, so dass

$$\begin{aligned}\log\left(\phi_{Y_n}(t)\right) &= n\log(1+u) = nu = n\phi''_X(\xi)\frac{t^2}{2\sigma^2 n} \\ &= \phi''_X(\xi)\frac{t^2}{2\sigma^2} \overset{n\to\infty}{\longrightarrow} -\frac{t^2}{2}\end{aligned}$$

gilt. Somit folgt in einem letzten Schritt

$$\lim_{n\to\infty} \phi_{Y_n}(t) = e^{-\frac{t^2}{2}}.$$

Das ist jedoch die charakteristische Funktion einer standardnormalverteilten Zufallsvariablen und es folgt mit Satz 2.86 die Aussage.

□

Es gibt Verallgemeinerungen des vorliegenden zentralen Grenzwertsatzes, bei denen die strikten Forderungen der Unabhängigkeit oder der identischen Verteilung aufgeweicht werden.

Teil II.

Stochastische Prozesse

3. Irrfahrten

Soll die Entwicklung der Werte einer Zufallsvariablen im zeitlichen Verlauf betrachtet werden, wird das mit stochastischen Prozessen modelliert. dazu betrachtet man einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P})$, eine Menge $I \subseteq [0, \infty)$ und für eine Menge M eine Abbildung $X : \Omega \times I \to M$. Wir einer der beiden Parameter $\omega \in \Omega$ oder $t \in I$ festgelegt, entstehen unterschiedliche Sichten auf die Funktion.

Definition 3.1: Stochastischer Prozess

Sei $I \subseteq [0, \infty)$ eine Indexmenge. Eine Familie $(X_t)_{t \in I}$ von Zufallsvariablen $X_t : \Omega \to M$ heißt stochastischer Prozess mit Zeitbereich I und Zustandsraum M.

Der Index t wird meist als Zeit interpretiert. Für $I = \{1, \ldots, k\}$ bzw. $I = \mathbb{N}$, allgemein für eine abzählbare Indexmenge, wird ein stochastischer Prozess als Prozess mit diskreter Zeit bezeichnet, ansonsten als stochastischer Prozess mit stetiger Zeit. Für jedes feste t ist X_t eine Zufallsvariable, für jedes fest gewählte $\omega \in \Omega$ wird die Abbildung $t \mapsto X_\omega(t) := X_t(\omega)$ als Pfad von ω bezeichnet und als Realisierung des stochastischen Prozesses interpretiert.

Definition 3.2: Zuwächse

Für einen stochastischen Prozess $(X_t)_{t \in I}$ heißen die Zufallsvariablen $X_t - X_s$ mit $s \leq t$ Zuwächse über (s, t). Zuwächse heißen stationär, wenn für alle $t \geq 0$ und $k \geq 0$ die Verteilung von $X_{t+k} - X_t$ nur von k abhängt. Weiter heißen Zuwächse unabhängig, wenn für jedes $n \in \mathbb{N}$ und alle $0 \leq t_0 < t_1 < \ldots < t_n, t_i \in I$, gilt $(X_{t_i} - X_{t_{i-1}})_{i=1,\ldots,n}$ ist eine Folge unabhängiger Zufallsvariablen.

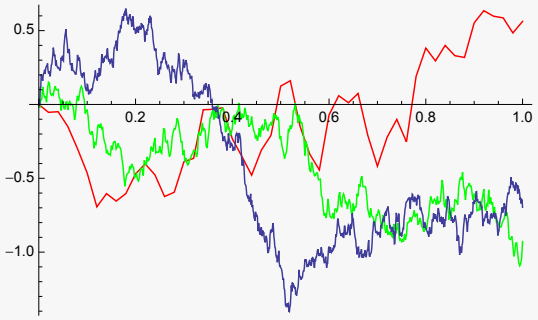
Beispiel 3.3

Ein Beispiel für einen reellwertigen stochastischen Prozess $(X_t)_{t \in [0,\infty)}$ mit stetiger Zeit ist eine Brownsche Bewegung. Dabei gilt

- $\mathbb{P}(X_0 = 0) = 1$
- $X_t \sim \mathsf{N}(0, t)$ für $t > 0$
- Der Prozess hat unabhängige und stationäre Zuwächse, insbesondere gilt $X_{t_i} - X_{t_j} \sim \mathsf{N}(0, t_i - t_j)$ für $j < i$
- der Pfad $t \mapsto X_t(\omega)$ ist stetig für jedes $\omega \in \Omega$.

Brown
1773-1858

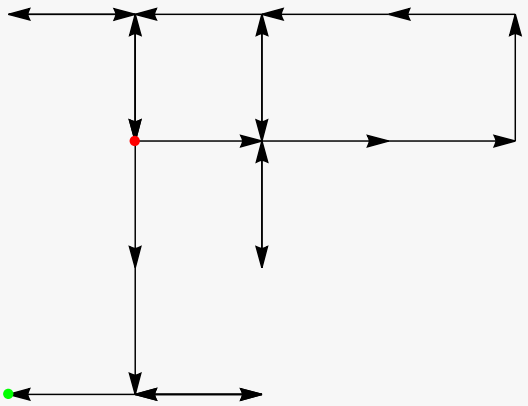
Folgende Abbildung zeigt drei Pfade einer Brownschen Bewegung mit 50, 500 und 1000 Werten für t.



Es gilt $X_{t_n} = X_{t_0} + (X_{t_1} - X_{t_0}) + \ldots + (X_{t_n} - X_{t_{n-1}}) \sim \mathsf{N}(0, t_0 + (t_1 - t_0) + \ldots + (t_n - t_{n-1}))$.

Beispiel 3.4

Beim Schach werde der König auf ein Feld des Schachbretts gestellt und danach ein (fairer) achtseitiger Würfel geworfen. Je nach dem Wurfergebnis wird der König auf eines seiner Nachbarfelder gezogen. Bei Überschreiten des Randes wird die Figur an der gegenüberliegenden Seite wieder auf das Brett gestellt. Der König befindet sich auf Position $(x, y) \in \{0, 1, \ldots, 7\}^2$ und bei jedem Zug ergibt sich als neue Position $(x', y') = (x, y) + (u, v)$ mod 8 mit $u, v \in \{-1, 0, 1\}$ und $(u, v) \neq (0, 0)$.



Der König bewegt sich rein zufällig über das Schachbrett, man spricht von einer Irrfahrt. Nach jeder Positionsänderung verändern sich die Wahrscheinlichkeiten dafür, dass sich der König nach der nächsten Positionsänderung auf einem bestimmten Feld befindet. Alle acht der aktuellen Position benachbarten Felder werden mit Wahrscheinlichkeit $1/8$ erreicht, für alle anderen ist die Wahrscheinlichkeit für das Erreichen Null.

Das letzte Beispiel zeigt, dass es für die Wahrscheinlichkeitsverteilung der nachfolgenden Position des Königs nur entscheidend ist, wo er sich aktuell befindet. Alle früheren Positionen spielen dabei keine Rolle. Das Konzept der Unabhängigkeit von Zufallsvariablen wird damit etwas gelockert.

4. Markow-Ketten

Definition 4.1: Markow-Kette

Ein stochastischer Prozess $(X_t)_{t\in\mathbb{N}_0}$ heißt Markow-Kette, falls für $n = 1, 2, \ldots$ und beliebige Parameterwerte $t_m \in \mathbb{N}_0$ mit $m = 0, 1, \ldots, n$ und $0 \leq t_0 < t_1 < \ldots < t_n$ sowie für reelle Zahlen $x, y \in \mathbb{R}$

$$\mathbb{P}(X_{t_n} = y | X_{t_{n-1}} = x, X_{t_{n-2}} = x_{n-2}, \ldots, X_{t_0} = x_0) = \mathbb{P}(X_{t_n} = y | X_{t_{n-1}} = x)$$

für alle $x_0, \ldots, x_{n-2}$ gilt.

Markow
1856-1922

Die bedingte Verteilung der Zufallsvariablen X_{t_n} unter der Bedingung, dass die Zufallsvariable $X_{t_{n-1}}$ den Wert x annimmt, ist damit von Werten von X_t zu Zeitpunkten vor t_{n-2} unabhängig.

Beispiel 4.2

Beim Roulette werden 10 Euro auf die Farbe rot gesetzt. Fällt eine der 18 roten Zahlen, gewinnt man 10 Euro, bei den 18 schwarzen und der Null ist der Einsatz verloren. Die Wahrscheinlichkeit zu gewinnen beträgt $\frac{18}{37}$, die zu verlieren $\frac{19}{37}$. Das Spiel soll beendet werden, wenn entweder alles Geld verloren oder 60 Euro Kapital vorhanden sind. Je nach Startkapital verläuft das Spiel anders. Beginnt man mit 50 Euro, so ist mit einer Wahrscheinlichkeit von $\frac{18}{37}$ das Spiel nach einer Runde für mich beendet. Mit einer Wahrscheinlichkeit von $\frac{19}{37}$ wird das Kapital bei nur noch 40 Euro sein. Dann wird weitergespielt und mit einer Wahrscheinlichkeit von $\frac{19}{37} \cdot \frac{18}{37}$ hat man 50 Euro bzw. mit einer Wahrscheinlichkeit von $\frac{19}{37} \cdot \frac{19}{37}$ nur noch 30 Euro.

Will man das Beispiel fortsetzen, wird das ganze schon deutlich mühsamer. Wir wollen das durch eine bestimmte Notation vereinfachen. Es sei p_{ij} die Wahrscheinlichkeit, mit der vom Zustand i der Zustand j erreicht wird. Entspricht also beispielsweise $i = 3$ dem Zustand 30 Euro, und $j = 4$ dem Zustand 40 Euro, so wäre $p_{34} = \frac{18}{37}$. Das bedeutet für alle Zustände $i = 0$ bis $i = 6$ im Roulette-Beispiel, dass

$$p_{ij}^{t_n} = \mathbb{P}(X_{t_n} = j | X_{t_{n-1}} = i)$$

ist. Dabei muss $\sum_{j=0}^{6} p_{ij}^{t_n} = \sum_{j=0}^{6} \mathbb{P}(X_{t_n} = j | X_{t_{n-1}} = i) = 1$ für jedes j gelten.

Definition 4.3: Stochastische Matrix

Eine quadratische Matrix $P = (p_{ij}) \in [0,1]^{n,n}$ heißt stochastische Matrix, falls

$$\sum_{j=1}^{n} p_{ij} = 1$$

für alle $i = 1, \ldots, n$ gilt.

Eine Markow-Kette heißt homogen, wenn die Wahrscheinlichkeiten $p_{ij}^{t_n}$ nicht vom Zeitpunkt t_n abhängen, d.h. $p_{ij}^{t_n} = p_{ij}$ für alle t_n. Bei einer homogenen Markow-Kette wird die dazugehörige stochastische Matrix Übergangsmatrix genannt.

Beispiel 4.4

Für das Roulette-Beispiel 4.2 ergibt sich für die Einträge der Übergangsmatrix

	0	10	20	30	40	50	60
0	1	0	0	0	0	0	0
10	$\frac{19}{37}$	0	$\frac{18}{37}$	0	0	0	0
20	0	$\frac{19}{37}$	0	$\frac{18}{37}$	0	0	0
30	0	0	$\frac{19}{37}$	0	$\frac{18}{37}$	0	0
40	0	0	0	$\frac{19}{37}$	0	$\frac{18}{37}$	0
50	0	0	0	0	$\frac{19}{37}$	0	$\frac{18}{37}$
60	0	0	0	0	0	0	1

Bei einem Startkapital von 50 Euro ergibt sich nach drei Spielrunden die Verteilung $(0, 0, 0.135, 0, 0.257, 0, 0.608)$. Mit knapp 61% Wahrscheinlichkeit hat man nach drei Spielrunden einen Gewinn gemacht. Der Erwartungswert beträgt 48.601, ist also kleiner als der Einsatz.

Die Zeilensumme einer stochastischen Matrix muss jeweils Eins ergeben. Im Beispiel wurde bereits berechnet, wie die Verteilung nach drei Spielrunden aussieht. Dabei wird als Ausgangspunkt das Startkapital genommen, welches einem der möglichen Zustände entspricht. Mit Wahrscheinlichkeit Eins ist der Zustand gegeben, der dem Startkapital entspricht. Dies kann als Einheitsvektor $\vec{e}_i$ dargestellt werden. Allgemein heißt ein Vektor $\vec{x} \in \mathbb{R}^n$ Wahrscheinlichkeitsvektor, wenn $x_i \geq 0$ und $\sum_{i=1}^{n} x_i = 1$ ist. In der i-ten Komponente steht dann die Wahrscheinlichkeit, dass der Zustand i gegeben ist, wenn die Zustände $1, \ldots, n$ eintreten können.

Satz 4.5

Die Menge der Wahrscheinlichkeitsvektoren ist konvex und abgeschlossen.

Beweis.

Sei $\lambda \in [0,1]$ und seien $\vec{x}, \vec{y} \in \mathbb{R}^n$ Wahrscheinlichkeitsvektoren. Dann gilt

$$(\lambda\vec{x} + (1-\lambda)\vec{y})^T \mathbf{1} = \sum_{i=1}^{n} (\lambda\vec{x} + (1-\lambda)\vec{y})_i = \lambda \sum_{i=1}^{n} x_i + (1-\lambda) \sum_{i=1}^{n} y_i = \lambda + (1-\lambda) = 1.$$

Beweis der Abgeschlossenheit: Eine Menge M heißt abgeschlossen, falls es ein $x \in M$ gibt, so dass für jedes $\epsilon > 0$ gilt, dass die ϵ-Umgebung von $x \in M$ nicht vollständig in M liegt, d.h. $B_\epsilon(x) = \{z \in \mathbb{R}^n;\ ||x-z|| < \epsilon\} \not\subseteq M$. Sei M die Menge der Wahrscheinlichkeitsvektoren und $x \in M$. Mit dem quadrierten Euklidischen Abstand und für $z_i = x_i + \frac{\epsilon}{n}$, $n > 1$, gilt

$$\sum_{i=1}^{n} (x_i - z_i)^2 = \sum_{i=1}^{n} \left(x_i - \left(x_i + \frac{\epsilon}{n}\right)\right)^2 = \sum_{i=1}^{n} \frac{\epsilon^2}{n^2} = \frac{\epsilon^2}{n} < \epsilon^2.$$

Damit liegt z für jedes $\epsilon > 0$ in $B_\epsilon(x)$. Aber

$$\sum_{i=1}^{n} z_i = \sum_{i=1}^{n} \left(x_i + \frac{\epsilon}{n}\right) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \frac{\epsilon}{n} = 1 + \epsilon > 1.$$

Damit ist z kein Wahrscheinlichkeitsvektor und $B_\epsilon(x) \not\subseteq M$.

□

Wird der Einheitsvektor $\vec{e}_i$ als Zeilenvektor von links an die stochastische Matrix multipliziert, ergibt sich die i-te Zeile als Ergebnis. Diese entspricht im Beispiel der Verteilung des Kapitals nach einem Spiel.

Satz 4.6

Es sei $P \in \mathbb{R}^{n,n}$ eine stochastische Matrix und $\vec{x} \in \mathbb{R}^n$ ein Wahrscheinlichkeitsvektor. Dann ist

1) $P^T\vec{x}$
2) $(P^T)^k\vec{x}$ für jedes $k \in \mathbb{N}$
3) im Falle der Existenz von $\lim_{k\to\infty} (P^T)^k$ auch $\lim_{k\to\infty} (P^T)^k\vec{x}$

ein Wahrscheinlichkeitsvektor.

Beweis.

1) Da alle Einträge von P und alle Komponenten von $\vec{x}$ nicht-negativ sind, sind auch alle Komponenten von $P^T\vec{x}$ nicht-negativ. Weiter ist $(P^T\vec{x})^T\mathbf{1} = \vec{x}^T P\mathbf{1} = \vec{x}^T\mathbf{1} = 1$.

2) Induktion nach k. Für $k = 1$ ist die Gültigkeit wegen 1) gegeben. $k \to k+1$: Die Aussage sei für k richtig, d.h. $\vec{y} = (P^T)^k\vec{x}$ ist ein Wahrscheinlichkeitsvektor. Dann folgt: $(P^T)^{k+1}\vec{x} = P^T \cdot (P^T)^k\vec{x} \overset{\text{I.V.}}{=} P^T\vec{y}$, und dies ist nach 1) wiederum ein Wahrscheinlichkeitsvektor.

3) Es existiere $\lim\limits_{k\to\infty}(P^T)^k$. Dann ist $\left((P^T)^k\vec{x}\right)_{k\in\mathbb{N}}$ eine Folge von Wahrscheinlichkeitsvektoren, deren Grenzwert wegen der Abgeschlossenheit der Menge der Wahrscheinlichkeitsvektoren wiederum ein Wahrscheinlichkeitsvektor ist.

□

Ist $\vec{x}_t$ der Wahrscheinlichkeitsvektor, der die Verteilung für die Zustände eines stochastischen Prozesses mit diskreter Zeit zu einem Zeitpunkt t angibt, so ist $\vec{x}_{t+1} = P^T\vec{x}_t$ die Verteilung für die Zustände zum Zeitpunkt $t+1$. Dies folgt aus dem Satz für die totale Wahrscheinlichkeit. Gilt $\vec{x} = P^T\vec{x}$, so heißt der Wahrscheinlichkeitsvektor stationär.

Satz 4.7

Sei $(X_t)_{t\in\mathbb{N}_0}$ eine homogene Markow-Kette mit zugehöriger stochastischer Matrix $P \in [0,1]^{n,n}$. Sei weiter $\vec{x}$ ein Wahrscheinlichkeitsvektor.
1) $\vec{x}$ ist genau dann stationär, wenn $\vec{x}$ ein Eigenvektor zum Eigenwert 1 von P^T ist.
2) 1 ist ein Eigenwert von P^T.

Beweis.

1) $\vec{y} \neq \vec{0}$ ist Eigenvektor von P^T zum Eigenwert $\lambda \in \mathbb{C}$ genau dann wenn $P^T\vec{y} = \lambda\vec{y}$ ist. Man setze $\lambda = 1$. Sind alle Komponenten von $\vec{y}$ nicht-negativ, so lässt sich durch $\vec{x} = \vec{y}/\sum\limits_{i=1}^{n} y_i$ ein stationärer Wahrscheinlichkeitsvektor erzeugen.
2) Es gilt $P\mathbf{1} = (p_1^T\mathbf{1} \dots p_n^T\mathbf{1})^T = \mathbf{1}$. Damit ist $\mathbf{1}$ Eigenvektor von P zum Eigenwert 1. Jeder Eigenwert von P ist wegen $\det(P - \lambda E_n) = \det((P - \lambda E_n)^T) = \det(P^T - \lambda E_n)$ auch Eigenwert von P^T.

□

Der Eigenwert 1 einer stochastischen Matrix spielt eine wichtige Rolle. Zunächst ist damit gezeigt, dass es wenigstens einen Eigenwert λ_{max} von P gibt, dessen Betrag mindestens Eins ist, $|\lambda_{max}| \geq 1$. Sei nun $\vec{v} \neq \vec{0}$ ein Eigenvektor von P zum Eigenwert λ. Dann gilt aber

auch

$$\begin{aligned}
|\lambda| \cdot \max_{i=1,\ldots,n}\{|v_i|\} &= \max_{i=1,\ldots,n}\{|\lambda v_i|\} = \max_{i=1,\ldots,n}\{|(Pv)_i|\} \\
&= \max_{i=1,\ldots,n}\left\{\left|\sum_{j=1}^{n} p_{ij}v_j\right|\right\} \\
&\leq \max_{i=1,\ldots,n}\left\{\sum_{j=1}^{n}|p_{ij}v_j|\right\} = \max_{i=1,\ldots,n}\left\{\sum_{j=1}^{n}|p_{ij}|\cdot|v_j|\right\} \\
&\leq \max_{i=1,\ldots,n}\left\{\sum_{j=1}^{n}|p_{ij}|\cdot\max_{k=1,\ldots,n}\{|v_k|\}\right\} \\
&= \max_{i=1,\ldots,n}\left\{\sum_{j=1}^{n}|p_{ij}|\right\}\cdot\max_{i=1,\ldots,n}\{|v_i|\} \\
&= 1\cdot\max_{i=1,\ldots,n}\{|v_i|\}
\end{aligned}$$

Insgesamt gilt somit

$$1 \text{ ist Eigenwert von } P^T, |\lambda| \leq 1 \text{ und } |\lambda_{max}| \geq 1 \Rightarrow |\lambda_{max}| = 1.$$

Es soll nun um die Frage gehen, wann der Grenzwert $\lim_{k\to\infty}(P^T)^k$ existiert. Dazu betrachten wir die (komplexe) Jordansche Normalform von P^T. Es sei

$$\chi_{P^T}(\lambda) = \prod_{i=1}^{k}(\lambda_i - \lambda)^{r_i}$$

das (komplexe) charakteristische Polynom von P^T, d.h. λ_i ist ein r_i-facher Eigenwert von P^T und $\lambda_i \neq \lambda_j$ für $i \neq j$. Die Jordansche Normalform ist eine Matrix J in Tridiagonalgestalt, die durch eine Basistransformation aus P^T entsteht,

$$J = Q^{-1}P^TQ = \begin{pmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_k \end{pmatrix}$$

und aus so genannten Jordan-Blöcken der Gestalt

$$J_i = \boxed{\begin{pmatrix} \lambda_i & q_{i1} & & & \\ & \lambda_i & q_{i2} & & \\ & & \ddots & \ddots & \\ & & & \lambda_i & q_{i,r_i-1} \\ & & & & \lambda_i \end{pmatrix}}$$

zusammengesetzt ist. Dabei gibt es so viele Blöcke wie es verschiedene Eigenwerte von P^T gibt und jeder Block hat die Größe, die der algebraischen Vielfachheit des zugehörigen Eigenwerts entspricht. Es ist $q_{ij} \in \{0, 1\}$. Nur wenn die algebraische Vielfachheit des

Eigenwerts seiner geometrischen entspricht, sind die q_{ij} allesamt Null, ansonsten gibt es Einträge $q_{ij} = 1$. Weiter gilt

$$(P^T)^k = (QJQ^{-1})^k = QJ^kQ^{-1}.$$

Damit kann anhand der Jordanschen Normalform entschieden werden, ob der Grenzwert $\lim_{k\to\infty}(P^T)^k$ existiert. Dazu betrachten wir einen einzelnen Jordan-Block J_i, da bei der Berechnung von J^k jeder Jordan-Block sich nur auf sich und auf keinen der anderen Jordan-Blöcke auswirkt. Innerhalb eines Jordan-Blocks interessiert uns die Konstellation

$$\begin{bmatrix} \lambda_i & q \\ & \lambda_i \end{bmatrix}$$

und wir untersuchen, was beim k-fachen Produkt von J passiert. Zunächst entsteht dann der Block

$$\begin{bmatrix} \lambda_i^k & kq\lambda_i^{k-1} \\ & \lambda_i^k \end{bmatrix}$$

Wir unterscheiden fünf Fälle:

- $\lambda_i = 0$: $\begin{bmatrix} 0 & q \\ & 0 \end{bmatrix} \overset{k\to\infty}{\to} \begin{bmatrix} 0 & 0 \\ & 0 \end{bmatrix}$
- $|\lambda_i| = 1, q = 0$: $\begin{bmatrix} \lambda_i^k & 0 \\ & \lambda_i^k \end{bmatrix}$, für $\lambda_i \neq 1$ divergent, für $\lambda_i = 1$ konvergent
- $|\lambda_i| = 1, q = 1$: $\begin{bmatrix} \lambda_i^k & k\lambda_i^{k-1} \\ & \lambda_i^k \end{bmatrix}$, für $\lambda_i \neq 1$ divergent, für $\lambda_i = 1$ bestimmt divergent
- $|\lambda_i| < 1, q = 0$: $\begin{bmatrix} \lambda_i^k & 0 \\ & \lambda_i^k \end{bmatrix} \overset{k\to\infty}{\to} \begin{bmatrix} 0 & 0 \\ & 0 \end{bmatrix}$
- $|\lambda_i| < 1, q = 1$: $\begin{bmatrix} \lambda_i^k & k\lambda_i^{k-1} \\ & \lambda_i^k \end{bmatrix} \overset{k\to\infty}{\to} \begin{bmatrix} 0 & 0 \\ & 0 \end{bmatrix}$

Zu ungünstigen Situationen kommt es, wenn ein Eigenwert mit Betrag 1 ungleich 1 vorkommt, da dann J^k divergiert und es keine Grenzmatrix gibt. Ist $\lambda = 1$, stimmen aber die geometrische und algebraische Vielfachheit nicht überein, so „springt“das System immer zwischen verschiedenen Zuständen hin und her.

Satz 4.8

Es sei P eine stochastische Matrix. Genau dann wenn $\lambda = 1$ der einzige Eigenwert von P^T mit $|\lambda| = 1$ und die algebraische gleich der geometrischen Vielfachheit für $\lambda = 1$ ist, existiert der Grenzwert $\lim_{k\to\infty}(P^T)^k$. Ist in dieser Situation $\lambda = 1$ ein einfacher Eigenwert, so ist der Grenzwert dadurch gekennzeichnet, dass alle Spalten der Grenzmatrix gleich sind.

Beweis.

Der erste Teil wurde oben gezeigt. Zum zweiten Teil ist zu sagen, dass für den Fall der Einfachheit des Eigenwerts 1 die Grenzmatrix $\lim_{k\to\infty}(P^T)^k$ nur eine 1 auf der Diagonalen stehen hat. Alle anderen Einträge der Matrix sind Null.

□

Beispiel 4.9

Setzen wir das Roulettespiel aus Beispiel 4.2 fort, so ergeben sich für die stochastische Matrix P die Eigenwerte $\lambda_1 = 1$ (doppelt), $\lambda_2 = 0.8657, \lambda_3 = -0.8657, \lambda_4 = 0.4998$ und $\lambda_5 = 0$. P^T ist diagonalisierbar. Da λ_1 einziger Eigenwert mit $|\lambda| = 1$ ist, existiert der Grenzwert $\lim_{k\to\infty}(P^T)^k$ und es ist

$$G^T = \lim_{k\to\infty}(P^T)^k = \begin{pmatrix} 1.00 & 0.86 & 0.70 & 0.54 & 0.37 & 0.19 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.14 & 0.30 & 0.46 & 0.63 & 0.81 & 1.00 \end{pmatrix}$$

Die Grenzmatrix zeigt die Verteilung der Zustände je nach Wahl eines Startzustandes. Steigt man im Beispiel mit 50 Euro in das Spiel ein, wird mit Wahrscheinlichkeit 0.19 alles verloren und mit Wahrscheinlichkeit 0.81 endet das Spiel mit 60 Euro. Der erwartete Gewinn dabei ist

$$(10 \cdot 0.81 - 50 \cdot 0.19) \text{ Euro} = -1.40 \text{ Euro}.$$

In keinem der Fälle ist der erwartete Gewinn positiv. Diese Spielstrategie lohnt sich also nicht.

Teil III.

Induktive Statistik

5. Entscheidungen

Wissenschaftliche Untersuchungen dienen zur Überprüfung wissenschaftlicher Hypothesen. Mit Hilfe einer Stichprobe sollen Erkenntnise gewonnen und möglichst verallgemeinert werden, um letztlich die Hypothese beurteilen zu können. In der Statistik werden Aussagen über Sachverhalte einer Grundgesamtheit getätigt. Da meist nicht die vollständige Grundgesamtheit erfassbar ist, unterliegen diese Aussagen einer gewissen Unsicherheit. Sie basieren dann auf einer vorliegenden Stichprobe aus der Grundgesamtheit. Eine Aussage über die unbekannte Verteilung einer Zufallsvariablen X oder einen unbekannten Parameter auf Grundlage einer gewonnenen Stichprobe wird statistische Entscheidung genannt. Dazu muss es eine Möglichkeit geben, mittels jeder möglichen Beobachtung eine Entscheidung treffen zu können. Es werden drei Klassen von Aufgaben zur Entscheidungsfindung unterschieden: Schätzprobleme, Bereichsschätzprobleme und Testprobleme. Um die Begriffe zu erläutern, betrachten wir folgendes Beispiel.

Beispiel 5.1: Schätzprobleme

Ein bestimmtes Smartphone soll gekauft werden. Dazu werden die Preise von zehn Anbietern verglichen. Die Preise unterliegen oftmals sogar stündlichen Schwankungen, die von vielen Einflussfaktoren getragen und so nur schwer zu erfassen sind. Obwohl einer dieser vielen Einflussfaktoren sicherlich der Blick auf den Preis der Konkurrenz ist, sei die Unabhängigkeit der Preisbildung zwischen den Anbietern angenommen. Die Preise werden zufallsabhängig modelliert über Zufallsvariablen $X_i : \Omega \to \mathbb{R}$ für $i \in \{1, \dots, 10\}$. Dazugehörig werde eine Verteilungsannahme $\mathbb{P}_\theta$, $\theta \in \Theta$, für den Preis dieses Smartphones getroffen, zu deren Parameter der erwartete Preis μ gehöre. Dabei wird festgelegt, aus welcher Wertemenge M der Parameter μ stammen soll. Für die zehn Anbieter gebe es folgende Realisierungen:

$$x_1 = 320,\ x_2 = 370,\ x_3 = 310,\ x_4 = 390,\ x_5 = 350,$$
$$x_6 = 350,\ x_7 = 340,\ x_8 = 355,\ x_9 = 335 \text{ und } x_{10} = 360$$

Sind wir daran interessiert, Aussagen über den erwarteten Preis μ zu treffen, liegt ein Schätzproblem vor. Wir benötigen eine Funktion, die aus den Realisierungen einen Wert für μ liefert. Die beiden Schätzfunktionen

$$T_1 : \mathbb{R}^{10} \to \mathbb{R},\ (x_1, \dots, x_{10}) \mapsto T_1(x_1, \dots, x_{10}) := \frac{x_1 + x_{10}}{2} \text{ oder}$$
$$T_2 : \mathbb{R}^{10} \to \mathbb{R},\ (x_1, \dots, x_{10}) \mapsto T_2(x_1, \dots, x_{10}) := \frac{1}{10} \sum_{i=1}^{10} x_i$$

könnten dafür geeignet sein. Hierbei erhalten wir die beiden Schätzwerte

$$\begin{aligned}\mu_1 &:= T_1(x_1,\ldots,x_{10}) = 340,\\ \mu_2 &:= T_2(x_1,\ldots,x_{10}) = 348.\end{aligned}$$

Es stellt sich die Frage, welche der beiden Schätzfunktionen einen „besseren“ Schätzwert liefert. In der Untersuchung von Schätzproblemen, der Schätztheorie, wird diese Frage näher untersucht.

Oftmals interessiert nicht nur ein einzelner Schätzwert $\mu_0 \in M$ für den unbekannten Erwartungswert μ, sondern ein Bereich $M_0 \subseteq M$, der den unbekannten Erwartungswert μ mit einer vorgegebenen Wahrscheinlichkeit überdeckt. Zumeist wird ein solcher Bereich in Form eines Intervalls $[\mu_u, \mu_o] \subset M$ angegeben. Derartige Intervalle basieren auf der zugrunde liegenden Verteilungsannahme. Wird im Beispiel die Normalverteilungsannahme $\mathbb{P}_\theta = \mathcal{N}(\mu, \sigma^2)$ mit $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+$ getroffen, so lässt sich für die Schätzfunktion T_2 ein Intervall, in dem mit 90%-iger Wahrscheinlichkeit der Erwartungswert μ liegt durch

$$[w_u, w_o] = \left[348 - 1.83 \cdot \frac{23.36}{\sqrt{10}}, 348 + 1.83 \cdot \frac{23.36}{\sqrt{10}}\right] = [334.46, 361.54]$$

angeben. Das Problem wird Bereichsschätzung genannt.

Ist zu entscheiden, in welchem von zwei vorgegebenen disjunkten Bereichen der unbekannte Parameter μ liegt, handelt es sich um ein Testproblem. So könnte gefragt sein, ob das Smartphone einen erwarteten Preis hat, der kleiner oder gleich einem vorgegeben Wert μ_0 oder größer ist, $\mu \leq \mu_0$ bzw. $\mu > \mu_0$. Wiederum kann mit einer Verteilungsannahme eine Entscheidung getroffen werden. Dabei gibt es keine 100%-ige Sicherheit, es kann eine der beiden Alternativen angenommen werden, obwohl die andere richtig ist. Dazu gibt man sich eine Irrtumswahrscheinlichkeit vor, die erste Alternative abzulehnen, obwohl sie richtig ist. Für $\mathbb{P}_\theta = \mathcal{N}(\mu, \sigma^2)$, $\mu_0 = 330$ und die Irrtumswahrschenlichkeit 0.05 würde hier die Entscheidung zugunsten der zweiten Alternative fallen. Auch hier stellt sich die Frage nach einer bestmöglichen Modellierung von Testproblemen. Dies wird in der Testtheorie untersucht.

Es seien $(\Omega, \mathcal{F})$ und $(\Psi, \mathcal{G})$ Messräume[1]. Zu $(\Omega, \mathcal{F})$ sei $W_{\theta\in\Theta}$ eine Familie von Wahrscheinlichkeitsmaßen mit Parameterraum Θ. Ist $X : \Omega \to \Psi$ eine Zufallsvariable, so heißt

$$(\Psi, \mathcal{G}, \mathbb{P}_{X,W})$$

statistischer Raum.

5.1. Schätztheorie

Die Grundannahme der Mathematischen Statistik, die laut [13] dadurch gegeben ist, die „Beobachtungen als Realisierungen von Zufallsgrößen aufzufassen und damit (zu) unterstellen, dass sich der Vorgang durch eine Wahrscheinlichkeitsverteilung beschreiben lässt“, liefert eine Schnittstelle zwischen einer daten- und modellgetriebenen Untersuchung

[1] Ein Messraum besteht aus einer Grundmenge mit einer σ-Algebra über dieser Menge.

in der Stochastik. Eine Beobachtung, z.B. eine Messung bei der Durchführung eines physikalischen Experiments, ist demnach dem Zufall unterworfen. Zufallseinflüsse werden durch einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P})$, der nicht näher spezifiziert wird, modelliert.

5.1.1. Datenmatrix

Für jede Datenerhebung gibt es eine Gesamtheit aller Information tragenden Objekte. Die Gesamtheit der Objekte wird durch Identifikationsmerkmale, d.h. sachliche, räumliche oder zeitliche Kriterien eindeutig festgelegt. Die Anzahl Objekte ist in realen Situationen oftmals so groß, dass nicht für jedes einzelne Objekt Daten erhoben werden können. Dann wird eine Teilmenge herangezogen, die repräsentativ für die Gesamtheit aller Objekte ist. Ein einzelnes Objekt der Untersuchung muss sich von anderen Objekten ebenfalls durch Identifikationsmerkmale eindeutig unterscheiden. Ein einzelnes Objekt, das die für eine Untersuchung interessierende Information trägt, heißt Merkmalsträger $g_{i\cdot}$, $i \in I$, wobei I eine beliebige Indexmenge sei. Der Punkt im Index deute an, dass es zu jedem Merkmalsträger Ausprägungen zu verschiedenen Eigenschaften geben kann. Wir fassen alle durch festgelegte Identifikationsmerkmale zu einer virtuellen Gesamtheit zusammengebrachten Merkmalsträger konkret in der Grundgesamtheit $G := \{g_{i\cdot};\ i \in I\}$ zusammen. Wie die einzelnen Merkmalsträger erfasst werden, spielt dabei keine Rolle. Sie werden nun als Einheit in der Grundgesamtheit betrachtet. Werden die Merkmalsträger in der Grundgesamtheit G zusammengefasst, so nehmen wir an, dass die Merkmalsträger gleiche jedem von ihnen gegebene Eigenschaften haben, die empirisch beobachtet oder gemessen werden können.

Die Beobachtungsdaten werden nach bestimmten Gesichtspunkten entsprechend einer vorgegebenen Fragestellung gesammelt. Eine zentrale Aufgabe besteht in der Modellierung der durch die Fragestellung gegebenen Aufgabe. Die Modellierung umfasst dabei nicht nur die Festlegung von zu erhebenden Merkmalen, der Träger der Merkmale und die Überlegung, auf welche Weise die Informationen abgegriffen und quantifiziert werden, sondern auch je nach geplanter Weiterverarbeitung der Daten die Festlegung des zugrunde liegenden stochastischen Modells. Dies erfordert auch einen gewissen Mut, sich auf ein bestimmtes Modell festzulegen. Das hat aber die Konsequenz, dass die Ergebnisse nachfolgender Datenanalysen bereits aufgrund der Modellierung in eine bestimmte Richtung gelenkt werden können. Für jedes Beobachtungsdatum sind drei Sichten zu unterscheiden: Welcher Wertemenge entstammt das Datum, welche Eigenschaft repräsentiert das Datum und welchem Objekt ist das Datum zugeordnet? Ein Datum ohne Zufallseinfluss kann damit als Wert einer Abbildung von der Menge der Objekte in die vorgesehene Wertemenge, den Merkmalsraum M, angesehen werden, $e : G \to M$. Die unbekannte Abbildung $e \in \mathbb{E}$ entstammt einem Funktionenraum $\mathbb{E}$ und steht dabei für die festgelegte gemeinsame Eigenschaft der Objekte.

Da nicht immer alle Merkmalsträger einer Grundgesamtheit untersucht werden können, muss eine Auswahl getroffen werden. Es wird also lediglich von einer Teilmenge der Grundgesamtheit eine Datenerhebung vorgenommen (eine so genannte Teilerhebung). Eine solche Teilmenge $S \subseteq G$ heißt Stichprobenmenge der Grundgesamtheit G. $S = \{s_{1\cdot}, \ldots, s_{n\cdot}\} \subseteq G$ mit $n \in \mathbb{N}$ sei als endlich vorausgesetzt. n heißt der Stichprobenumfang. Seien nun $\mathbb{E} = \{e_1, \ldots, e_k\}$, $e_j : S \to M_j$ $(j = 1, \ldots, k)$ und $M = \bigcup_{j=1}^{k} M_j$. Eine zufallsunabhängi-

ge Beobachtung kann nun als Abbildung $\psi_S : \mathbb{E} \to M^n$, $e \mapsto (x_1, \ldots, x_n)^T := \psi_S(e) := (e(s_{1\cdot}), \ldots, e(s_{n\cdot}))^T$ beschrieben werden. Wird die Grundgesamtheit herangezogen, wird dann anstelle von ψ_S entsprechend ψ_G geschrieben mit $\psi_G(e) = (e(g_{1\cdot}), e(g_{2\cdot}), \ldots)^T$. Der Vektor $\mathbf{x} := (x_1, \ldots, x_n)^T$ heißt Stichprobe vom Umfang n.

Die Abbildung e, die eine Eigenschaft der Merkmalsträger repräsentiert, wird deswegen auch als Merkmal bezeichnet. Die Vektoren von k zufalls(un)abhängigen Beobachtungen mit Stichprobenumfang n werden in der Datenmatrix

$$\begin{array}{l} \\ s_{1\cdot} = g_{i_1\cdot} \\ s_{2\cdot} = g_{i_2\cdot} \\ \vdots \\ s_{n\cdot} = g_{i_n\cdot} \end{array} \begin{array}{c} \begin{array}{cccc} e_1 & e_2 & \ldots & e_k \end{array} \\ \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1k} \\ x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix} \end{array} = (x_{ij}) =: X$$

zusammengefasst. Jedes x_{ij} heißt Merkmalswert bzw. Datum des Merkmalsträgers $s_{i\cdot}$ für das Merkmal e_j. Liegt eine Datenmatrix vor, sind eine Vielzahl verschiedener Fragestellungen möglich. Eine Auswahl daraus soll in den weiteren Kapiteln betrachtet werden. Abkürzend schreiben wir meist X_j und meinen dabei sowohl das Merkmal e_j als auch den Vektor $x_{\cdot j}$ der Merkmalswerte aller Merkamlsträger für das Merkmal e_j.

5.1.2. Skalenarten

Es gibt verschiedene Arten von Merkmalen:

- Merkmale mit Nominalskala: Einzelne Merkmalswerte können lediglich hinsichtlich ihrer Gleichheit oder Ungleichheit unterschieden werden. Als Merkmalsraum verwenden wir endliche Mengen, deren Elemente Kategorien genannt werden. Wir sprechen hier von qualitativen Merkmalen.
- Merkmale mit Ordnialskala: Die Merkmalswerte können zudem in eine Reihenfolge „$\preceq$“ gebracht werden. Es sind keine Abstände zwischen den einzelnen Merkmalswerten interpretierbar.
- Merkmale mit Kardinalskala: Es können zusätzlich Abstände und Verhältnisse zwischen Merkmalswerten gebildet und interpretiert werden. Wir werden für solche Merkmale als Merkmalsraum die reellen Zahlen benutzen, $M = \mathbb{R}$. Sie sind die in den Ingenieurwissenschaften am häufigsten anzutreffende Merkmalsart.

Jedes Merkmal mit Kardinalskala kann als Merkmal mit Ordinalskala und jedes Merkmal mit Ordinalskala als Merkmal mit Nominalskala aufgefasst werden. Dies ist jeweils mit einem Informationsverlust verbunden. Nicht-qualitative Merkmale nennen wir auch quantitative Merkmale.

Es bezeichne

$$B := \{c \in M;\ c = x_i \text{ für ein } i = 1, \ldots, n\} \tag{5.1}$$

die Menge der beobachteten Ausprägungen $c \in M$ einer Stichprobe $\mathbf{x}$ vom Umfang n des Merkmals X. Liegt eine stochastische Modellierung der Datenmatrix zugrunde, sprechen wir auch von den Realisierungen anstelle der beobachteten Ausprägungen. Wir werden beide Begriffe synonym verwenden.

5.1.3. Punktschätzungen

Definition 5.2: Schätzfunktion und Schätzwert

Seien $X_1, \ldots, X_n$ u.i.v. Zufallsvariablen mit $X_i : \Omega \to M$, $X_i \sim \mathbb{P}_{\theta \in \Theta}$ und $(x_1, \ldots, x_n)^T \in M^n$ eine dazugehörige Stichprobe. Sei weiter $\gamma : \Theta \to \Gamma$ eine Abbildung vom Parameterraum Θ in Γ. Dann heißt eine Zufallsvariable $T : M^n \to \Gamma$ Schätzfunktion für γ und $T(x_1, \ldots, x_n)$ heißt Schätzwert für $\gamma(\theta)$.

Von einer Schätzfunktion wird erwartet, die Abbildung γ möglichst gut wiederzugeben.

Definition 5.3: Erwartungstreue

Eine Schätzfunktion $T : M^n \to \Gamma$ heißt erwartungstreu, wenn für alle $\theta \in \Theta$ gilt:

$$\mathbb{E}_\theta[T] = \gamma(\theta).$$

Der Begriff der Erwartungstreue wurde von Gauß eingeführt.

Gauß
1777-1855

Erläuterung

Es werde angenommen, dass die Daten unabhängig normalverteilt mit den Parametern μ und σ^2 sind, $X_i \sim N(\mu, \sigma^2)$ für alle $i = 1, \ldots, n$. Es soll der Parameter μ geschätzt werden. Es ist $\Theta = \mathbb{R} \times \mathbb{R}^+$, $\theta = (\mu, \sigma^2)$ und

$$\gamma_1 : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}(\mu, \sigma^2) \mapsto \gamma_1(\mu, \sigma^2) := \mu.$$

Eine Möglichkeit besteht darin, den Mittelwert aus der ersten und letzten Realisierung zu bestimmen. Eine zweite Möglichkeit liefert die Berechnung des arithmetischen Mittels.

$$\begin{aligned} T_1 : \mathbb{R}^n \to \mathbb{R} \quad (x_1, \ldots, x_n)^T \mapsto g_1(x_1, \ldots, x_n) &:= \tfrac{1}{2}(x_1 + x_n) \\ T_2 : \mathbb{R}^n \to \mathbb{R} \quad (x_1, \ldots, x_n)^T \mapsto g_2(x_1, \ldots, x_n) &:= \tfrac{1}{n} \sum_{i=1}^{n} x_i. \end{aligned}$$

Welcher Schätzwert bzw. welche Schätzfunktion ist im Sinne der Mathematischen Statistik „besser“? Die Mathematische Statistik liefert dafür verschiedene Gütekriterien. Eines davon ist die eben definierte Erwartungstreue. So sind beide Schätzfunktionen T_1 und T_2 erwartungstreue Schätzfunktionen für γ_1, denn es gilt

$$\begin{aligned} \mathbb{E}_\theta[T_1] &= \mathbb{E}_\theta\left[\frac{1}{2}(X_1 + X_n)\right] = \frac{1}{2}(\mu + \mu) = \mu = \gamma_1(\mu, \sigma^2), \\ \mathbb{E}_\theta[T_2] &= \mathbb{E}_\theta\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu = \gamma_1(\mu, \sigma^2). \end{aligned}$$

Hinsichtlich der Erwartungstreue sind beide Schätzfunktionen gleichermaßen geeignet.

Ist eine Schätzfunktion nicht erwartungstreu, wird sie verzerrt genannt. Die Differenz zwi-

schen $\mathbb{E}_\theta[T]$ und $\gamma(\theta)$ heißt Bias:

$$\text{Bias}_\theta(T, \gamma) = \mathbb{E}_\theta[T] - \gamma(\theta).$$

Die erwartete quadratische Abweichung von T zu $\gamma(\theta)$ dient ebenfalls als Maß der Beurteilung einer Schätzfunktion.

Definition 5.4: MSE

Es sei $T : M^n \to \Gamma$ eine Schätzfunktion für $\gamma(\theta)$. Dann heißt

$$\text{MSE}_\theta(T, \gamma) = \mathbb{E}_\theta[(T - \gamma(\theta))^2]$$

der Mean Square Error (MSE) der Schätzfunktion T.

Bias und MSE hängen über

$$\begin{aligned}
\text{MSE}_\theta(T, \gamma) &= \mathbb{E}_\theta[(T - \gamma(\theta))^2] && (5.2)\\
&= \mathbb{E}_\theta[T^2] - 2\gamma(\theta)\mathbb{E}_\theta[T] + \gamma(\theta)^2 && (5.3)\\
&= \mathbb{E}_\theta[T^2] - \mathbb{E}_\theta[T]^2 + \mathbb{E}_\theta[T]^2 - 2\gamma(\theta)\mathbb{E}_\theta[T] + \gamma(\theta)^2 && (5.4)\\
&= \mathbb{V}_\theta[T] + \text{Bias}_\theta(T, \gamma)^2 && (5.5)
\end{aligned}$$

zusammen. Sind T_1 und T_2 zwei unterschiedliche Schätzfunktionen für $\gamma(\theta)$, so heißt T_1 MSE-besser als T_2, wenn

$$\text{MSE}_\theta(T_1, \gamma) \leq \text{MSE}_\theta(T_2, \gamma) \text{ für alle } \theta \in \Theta$$

gilt und es mindestens ein $\theta^* \in \Theta$ mit $\text{MSE}_{\theta^*}(T_1, \gamma) < \text{MSE}_\theta(T_2, \gamma)$ gibt.

Beispiel 5.5

Es seien $X_1, \ldots, X_n$ u.i.v. Zufallsvariablen mit Erwartungswert μ und Varianz σ^2. Eine mögliche Schätzfunktion für die Varianz σ^2 $(\gamma(\theta) = \sigma^2)$ lautet mit dem arithmetischen Mittel $\bar{x} = T_2(x_1, \ldots, x_n)$

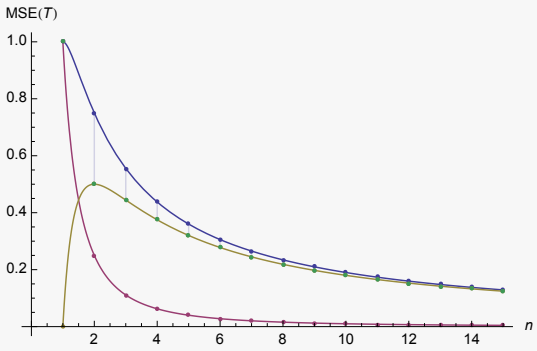
$$T_3(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Diese Schätzfunktion ist nicht erwartungstreu, denn

$$\begin{aligned}
\mathbb{E}_\theta[T_3] &= \mathbb{E}_\theta\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_\theta[X_i^2 - 2X_i\bar{X} + \bar{X}^2] \\
&= \frac{1}{n}\left(\mathbb{E}_\theta\left[\sum_{i=1}^{n}X_i^2\right] - n\mu^2 - n\mathbb{E}_\theta[\bar{X}^2] + n\mu^2\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}\mathbb{V}_\theta[X_i] - n\mathbb{V}_\theta[\bar{X}]\right) \\
&= \sigma^2\left(\frac{n-1}{n}\right)
\end{aligned}$$

Mit $S^2 = \frac{n}{n-1}T_3$ ergibt sich eine erwartungstreue Schätzfunktion für σ^2. Für den Mean Square Error von T_3 gilt bei Annahme normalverteilter X_i:

$$\mathsf{MSE}_{(\mu,\sigma^2)}(T_3,\sigma^2) = \mathbb{V}_{(\mu,\sigma^2)}[T_3] + \mathsf{Bias}_{(\mu,\sigma^2)}(T_3,\sigma^2)^2 = \frac{2(n-1)}{n^2}\sigma^4 + \frac{1}{n^2}\sigma^4.$$



Für wachsendes n stimmt der MSE nahezu mit der Varianz der Schätzfunktion überein und der Bias (zum Quadrat) ist fast Null. Die Schätzfunktion wird dann asymptotisch erwartungstreu genannt.

Für eine konstante Schätzfunktion $T(X) = \theta_0$, für die der Mean Square Error $\mathsf{MSE}_{\theta_0}(T,\gamma)$ gleich Null ist, gibt es keinen MSE-besten Schätzer im Sinne der Definition von MSE-besser. Daher wird oft die Klasse der zugelassenen Schätzfunktionen eingeschränkt. Ein weiteres Gütekriterium ist die Effizienz einer Schätzfunktion. Dabei gilt eine Schätzfunktion mit kleinerer Varianz als besser geeignet.

Definition 5.6

Eine erwartungstreue Schätzfunktion T des Parameters θ heißt effizient, wenn T unter allen erwartungstreuen Schätzfunktionen für θ die kleinste Varianz hat.

Erläuterung

Man betrachte noch einmal die beiden erwartungstreuen Schätzfunktionen für den Erwartungswert μ.

$$\begin{aligned}\mathbb{V}_\theta[T_1] &= \mathbb{V}_\theta\left[\frac{1}{2}(X_1+X_n)\right] = \frac{1}{4}\left(\sigma^2+\sigma^2\right) = \frac{\sigma^2}{2},\\ \mathbb{V}_\theta[T_2] &= \mathbb{V}_\theta\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}n\cdot\sigma^2 = \frac{\sigma^2}{n} \overset{n>2}{<} \mathbb{V}_\theta[g_1].\end{aligned}$$

T_2 bleibt als Kandidat für eine effiziente Schätzfunktion für μ im Rennen.

Ein effizienter Schätzer ist damit ein MSE-bester Schätzer. Es stellt sich die Frage, auf welche Weise sinnvolle Schätzfunktionen konstruiert werden können.

Maximum Likelihood-Schätzung


Fisher
1890-1962

Die von Fisher vorgeschlagene Maximum-Likelihood Schätzung zeichnet sich dadurch aus, dass verschiedene Gütekriterien erfüllt werden. Sei F die Verteilungsfunktion einer Zufallsvariablen X mit zugehöriger diskreter oder stetiger Dichte. Hängt die Verteilungsfunktion von k unbekannten Parametern $\theta_1,\ldots,\theta_k$ ab, die mittels einer Stichprobe $(x_1,\ldots,x_n)$ vom Umfang n geschätzt werden sollen, betrachtet man die gemeinsame Wahrscheinlichkeitsdichtefunktion $f_{X_1\ldots X_n}(x_1,\ldots,x_n;\theta_1,\ldots,\theta_k)$ in Abhängigkeit der Parameter. Sind die Parameter bekannt, lässt sich der Wert der Dichte anhand einer Stichprobe bestimmen. Ist umgekehrt die Stichprobe bekannt, lässt sich für jede beliebige Parameterkonstellation der Wert der Dichte ermitteln. Die letzte Sichtweise auf die Dichte führt zur Likelihood-Funktion.

Definition 5.7: Likelihood-Funktion

Es seien $X_1,\ldots,X_n$ u.i.v. mit zugehöriger diskreter oder stetiger Dichte in Abhängigkeit eines Parameter-Tupels $(\theta_1,\ldots,\theta_k)\in\Theta$ eines Parameterraums Θ. Dann heißt $L_{X_1\ldots X_n}:\Theta\to\mathbb{R}_0^+$

$$\begin{aligned}L_{X_1\ldots X_n}(\theta_1,\ldots,\theta_k;x_1,\ldots,x_n) &= f_{X_1\ldots X_n}(x_1,\ldots,x_n;\theta_1,\ldots,\theta_k)\\ &= \prod_{i=1}^{n} f_{X_i}(x_i;\theta_1,\ldots,\theta_k)\end{aligned}$$

Likelihood-Funktion. Die Loglikelihood-Funktion $l_{X_1\ldots X_n}$ entsteht aus der Likelihood-Funktion durch Logarithmieren,

$$\begin{aligned}l_{X_1\ldots X_n}(\theta_1,\ldots,\theta_k;x_1,\ldots,x_n) &= \log\left(L_{X_1\ldots X_n}(x_1,\ldots,x_n;\theta_1,\ldots,\theta_k)\right)\\ &= \sum_{i=1}^{n}\log\left(f_{X_i}(x_i;\theta_1,\ldots,\theta_k)\right).\end{aligned}$$

Auf Basis einer Stichprobe lässt sich mit der Likelihood-Funktion eine Schätzfunktion für $\gamma(\theta_1,\ldots,\theta_k)$ bestimmen, indem diejenige Kombination $(\hat\theta_1,\ldots,\hat\theta_k)$ in Abhängigkeit der Stich-

probe bestimmt wird, bei der die Dichte den größten Wert annimmt. Die Dichte wird bei gegebener Stichprobe für die bestimmten Parameterwerte maximiert.

Definition 5.8: Maximum Likelihood-Schätzung

Ein Schätzwert $(\hat{\theta}_1, \ldots, \hat{\theta}_k) = T(x_1, \ldots, x_n)$ einer Schätzfunktion T heißt Maximum Likelihood-Schätzwert für $\gamma(\theta_1, \ldots, \theta_k)$, wenn für alle $(\theta_1, \ldots, \theta_k) \in \Theta$ die Bedingung

$$L_{X_1,\ldots,X_n}(\hat{\theta}_1, \ldots, \hat{\theta}_k; x_1, \ldots, x_n) \geq L_{X_1,\ldots,X_n}(\theta_1, \ldots, \theta_k; x_1, \ldots, x_n)$$

bzw.

$$l_{X_1,\ldots,X_n}(\hat{\theta}_1, \ldots, \hat{\theta}_k; x_1, \ldots, x_n) \geq l_{X_1,\ldots,X_n}(\theta_1, \ldots, \theta_k; x_1, \ldots, x_n)$$

gilt. T ist dann die Maximum Likelihood-Schätzfunktion für γ.

Da das Logarithmieren eine streng monoton wachsende Abbildung ist, wird der Maximalpunkt für l und L im selben Punkt angenommen. Es ist somit das System

$$\frac{\partial L}{\partial \theta_i} = 0 \text{ bzw. } \frac{\partial l}{\partial \theta_i} = 0 \text{ für alle } i = 1, \ldots, n$$

zu lösen und zu überprüfen, ob es sich um eine Maximalstelle handelt.

Beispiel 5.9

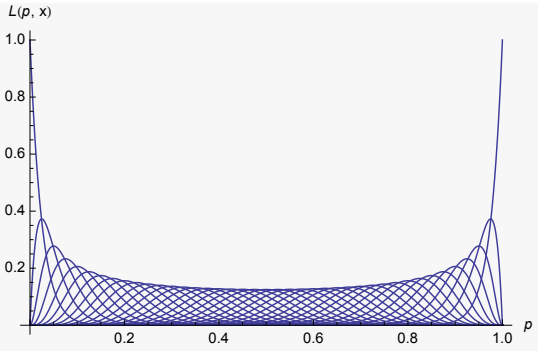
Es seien $X_i \sim Ber(p)$ u.i.v mit Parameter $p \in (0,1)$ und $i = 1, \ldots, n$. Dann lautet mit $x = \sum_{i=1}^{n} x_i$ die Likelihood-Funktion

$$L_{X_1 \ldots X_n}(p; x_1, \ldots, x_n) = p^x (1-p)^{n-x}.$$

Zur Bestimmung des Maximum Likelihood-Schätzwerts betrachte man aufgrund der einfacheren Handhabung die Loglikelihood-Funktion, leite diese nach p ab und setze sie Null. Für $0 < p < 1$ gilt dann

$$\begin{aligned} \frac{dl_{X_1 \ldots X_n}(p; x_1, \ldots, x_n)}{dp} &= \frac{d}{dp}\left(x \log(p) + (n-x)\log(1-p)\right) \\ &= \frac{x}{p} - \frac{n-x}{1-p} \stackrel{!}{=} 0. \end{aligned}$$

Durch Auflösen nach p ergibt sich $p = \frac{x}{n}$. Die zweite Ableitung nach p an dieser Stelle ist $-\frac{x}{p^2} - \frac{n-x}{(1-p)^2} \stackrel{p=x/n}{=} -\frac{n^2}{x} - \frac{n^2}{n-x} < 0$ ($x \neq n$ und $x \neq 0$). Also liegt ein Maximum vor.



Für $x = 0$ bzw. $x = n$ ergeben sich Maximalstellen am Rand der Likelihood-Funktion, d.h. bei $p = 0$ und $p = 1$.

Beispiel 5.10

Es seien $X_i \sim \mathcal{N}(\mu, \sigma^2)$ u.i.v. mit Parametern $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+$ und $i = 1, \ldots, n$. Die Loglikelihood-Funktion lautet

$$\begin{aligned} l_{X_1 \ldots X_n}(\mu, \sigma^2; x_1, \ldots, x_n) &= \sum_{i=1}^{n} \log\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right) \\ &= -n \log\left(\sqrt{2\pi\sigma^2}\right) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2. \end{aligned}$$

Mit $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ gilt:

$$\begin{aligned} \frac{\partial l_{X_1 \ldots X_n}}{\partial \mu} &= \frac{1}{2\sigma^2} \cdot 2\sum_{i=1}^{n}(x_i - \mu) = \frac{n}{\sigma^2}(\bar{x} - \mu) \stackrel{!}{=} 0 \Leftrightarrow \mu = \bar{x} \\ \frac{\partial l_{X_1 \ldots X_n}}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu)^2 \stackrel{!}{=} 0 \stackrel{\mu=\bar{x}}{\Leftrightarrow} \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 =: \hat{\sigma}^2. \end{aligned}$$

Die Hessematrix an der Stelle $\left(\bar{x}, \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$,

$$H_l\left(\bar{x}, \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right) = \begin{pmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & -\frac{n}{2(\hat{\sigma}^2)^2} \end{pmatrix},$$

ist negativ definit und damit liegt eine Maximalstelle vor. Die Maximum Likelihood-Schätzfunktion für die Varianz einer Normalverteilung ist die aus Beispiel 5.5 stammende nicht erwartungstreue Schätzfunktion.

Beispiel 5.11

Für die Exponentialverteilung mit Parameter λ gilt mit $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$$l_{X_1 \dots X_n}(\lambda; x_1, \dots, x_n) = n \log(\lambda) - \sum_{i=1}^{n} \lambda x_i = n \log(\lambda) - \lambda n \bar{x}.$$

Der Maximum Likelihood-Schätzwert ergibt sich mit

$$\frac{\partial l_{X_1 \dots X_n}}{\partial \lambda} = \frac{n}{\lambda} - n\bar{x} \stackrel{!}{=} 0 \Leftrightarrow \lambda = \frac{1}{\bar{x}} \text{ und } \frac{\partial^2 l_{X_1 \dots X_n}}{\partial \lambda^2} = -\frac{n}{\lambda^2} < 0$$

zu

$$\lambda = \frac{1}{\bar{x}}.$$

Regressionsanalyse

Zum Ende des Abschnitts wollen wir uns ein wichtiges Anwendungsbeispiel mit Hilfe der Maximum Likelihood-Schätzung ansehen: die Regressionsanalyse. Dabei geht es darum, dass ein kausaler Zusammenhang zwischen einem abhängigen Merkmal Y und unabhängigen Merkmalen $X_1, \dots X_m$ in der Form

$$y = f_\beta(x_1, \dots, x_m)$$

modelliert wird. Durch Messungen werden sowohl Werte y_i als auch Werte x_{iu} mit $i = 1, \dots, n$ und $u = 1, \dots, m$ beschafft, um das Modell aufzustellen. Wir gehen davon aus, dass die Funktion f einen linearen Zusammenhang der Form

$$f_\beta(x_1, \dots, x_m) = \sum_{j=1}^{k} \beta_j g_j(x_1, \dots, x_m)$$

beschreibt. Hierbei sind g_j modellierte reelle Funktionen der Eingangsgrößen $X_1, \dots, X_m$. Bei den Messungen stellt sich meist heraus, dass kein exakter Zusammenhang in der dargestellten Form möglich ist. Deswegen wird in das Modell ein additiver Term eingefügt, der diese Fehler modelliert,

$$y = f_\beta(x_1, \dots, x_m) + \epsilon.$$

Eine gängige Annahme ist, dass die Fehler zufälliger Natur sind. Es wird davon ausgegangen, dass jeder einzelne Messfehler $E_i \sim \mathcal{N}(0, \sigma^2)$, $i = 1 \dots, n$, normalverteilt mit Erwartungswert Null und gleicher Varianz σ^2 ist, was jedoch in jedem Fall überprüft werden muss. Werden die X_u als deterministische Merkmale modelliert, lassen sich die $Y_i = f_\beta(x_{i1}, \dots, x_{im}) + E_i \sim \mathcal{N}(f_\beta(x_1, \dots, x_m), \sigma^2)$ ebenfalls als normalverteilte Zufallsvariablen mit gleicher Varianz darstellen. Für die Normalverteilungen sind der Erwartungswert und die Varianz zu schätzen. Das bedeutet, dass die Parameter β_j und σ^2 bestimmt werden müssen. Unter der Annahme der Unabhängigkeit der Fehler E_i lässt sich mit

$E_i = Y_i - f_\beta(x_{i1}, \ldots, x_{im})$ die Loglikelihood-Funktion

$$\begin{aligned} l_{E_1 \ldots E_n}(\beta_1, \ldots, \beta_k, \sigma^2; y_1, \ldots, y_n, X) &= \log\left(\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\epsilon_i}{\sigma}\right)^2}\right) \\ &= \log\left(\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_i - \sum_{j=1}^{k} \beta_j g_j(x_{i1}, \ldots, x_{im})}{\sigma}\right)^2}\right) \\ &= -n \log\left(\sigma\sqrt{2\pi}\right) \\ &\quad -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{k} \beta_j g_j(x_{i1}, \ldots, x_{im})\right)^2 \end{aligned}$$

aufstellen. Seien nun $\vec{y} = (y_1, \ldots, y_n)^T$, $\vec{\beta} = (\beta_1, \ldots, \beta_k)^T$ und $X = (\tilde{x}_{ij}) \in \mathbb{R}^{n,k}$, $\tilde{x}_{ij} = g_j(x_{i1}, \ldots, x_{im})$ und $i = 1, \ldots, n$, $j = 1 \ldots, k$. X wird dabei die Designmatrix genannt. Dann gilt

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{k} \vec{\beta}_j g_j(x_{i1}, \ldots, x_{im})\right)^2 = (\vec{y} - X\vec{\beta})^T(\vec{y} - X\vec{\beta}). \tag{5.6}$$

Zur Bestimmung des Maximum Likelihood-Schätzwerts sind die Nullstellen der partiellen Ableitungen nach den Parametern zu bestimmen:

$$\begin{aligned} \nabla_{\vec{\beta}} l_{E_1 \ldots E_n}(\vec{\beta}, \sigma^2) &= -\frac{1}{2\sigma^2} \cdot (-2X^T\vec{y} + 2X^TX\vec{\beta}) \stackrel{!}{=} \vec{0} \\ &\quad \Leftrightarrow \vec{\beta} = (X^TX)^{-1}X^T\vec{y} \text{ und } X^TX \text{ invertierbar,} \\ \frac{\partial l_{E_1 \ldots E_n}}{\partial \sigma^2}(\vec{\beta}, \sigma^2) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\vec{y} - X\vec{\beta})^T(\vec{y} - X\vec{\beta}) \stackrel{!}{=} 0 \\ &\quad \Leftrightarrow \sigma^2 = \frac{1}{n}(\vec{y} - X\vec{\beta})^T(\vec{y} - X\vec{\beta}) \end{aligned}$$

Für die Hessematrix an den Nullstellen des Gradienten gilt

$$H_{l_{E_1 \ldots E_n}} = -\begin{pmatrix} \frac{1}{\sigma^2}X^TX & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}.$$

Diese Matrix ist im Falle der Invertierbarkeit von X^TX negativ definit und somit liegt ein Maximum vor.

Kleinst-Quadrate-Schätzung

Der Maximum Likelihood-Schätzung bei der Regressionsanalyse kann auch geometrisch gedeutet werden. Der Ausdruck (5.6) stellt den quadrierten Euklidischen Abstand des durch das Modell geschätzten Vektors $X\vec{\beta}$ von dem gemessenen $\vec{y}$ dar. Wird $\vec{\beta}$ so bestimmt, dass dieser Abstand minimal wird, landen wir exakt beim Maximum Likelihood-Schätzwert für $\vec{\beta}$ von oben. Das Vorgehen wird als Kleinst-Quadrate-Schätzung bezeichnet. Dabei gilt

$$(y - X\vec{\beta})^T X\vec{\beta} = \vec{y}^T X(X^TX)^{-1}X^T\vec{y} - \vec{y}^T X(X^TX)^{-1}X^T\vec{y} = 0.$$

$\vec{y}$ wird orthogonal auf den von den Spalten von X aufgespannten Vektorraum auf den Vektor $X\vec{\beta}$ projiziert. Besitzt X (und damit X^TX) nicht vollen Rang oder sind die Spalten von X nahezu linear abhängig (es liegt ein schlecht konditioniertes Problem vor), ist die Kleinst-Quadrate-Lösung entweder nicht eindeutig bestimmt oder sie macht keinen Sinn. Abhilfe kann hier schaffen, das Problem so zu formulieren, dass es wieder eine eindeutige und gut konditionierte Lösung gibt. Eine Möglichkeit besteht darin, die Matrix X mit einem $\lambda \neq 0$ zur Matrix

$$\begin{pmatrix} X \\ \lambda I \end{pmatrix} \in \mathbb{R}^{n+k,k}$$

und $\vec{y}$ zu $\vec{y} = (y_1, \ldots, y_n, 0, \ldots, 0)^T \in \mathbb{R}^{n+k}$ zu erweitern. Die Matrix besitzt vollen Rang k und ist durch entsprechende Wahl von λ gut konditioniert. Nun betrachtet man das Problem

$$\min_{\vec{\beta} \in \mathbb{R}^n} \left\{ \left[\begin{pmatrix} \vec{y} \\ \vec{0} \end{pmatrix} - \begin{pmatrix} X \\ \lambda I \end{pmatrix} \vec{\beta} \right]^T \left[\begin{pmatrix} \vec{y} \\ \vec{0} \end{pmatrix} - \begin{pmatrix} X \\ \lambda I \end{pmatrix} \vec{\beta} \right] \right\}$$

für ein frei wählbares aber gegebenes λ. Als Lösung des Problems ergibt sich

$$\begin{aligned} \vec{\beta} &= \left[\begin{pmatrix} X \\ \lambda I \end{pmatrix}^T \begin{pmatrix} X \\ \lambda I \end{pmatrix} \right]^{-1} \begin{pmatrix} X \\ \lambda I \end{pmatrix}^T \begin{pmatrix} \vec{y} \\ \vec{0} \end{pmatrix} \\ &= (X^TX + \lambda^2 I)^{-1} X^T \vec{y} \end{aligned}$$

Diese Vorgehensweise wird als Regularisierung bezeichnet.

Beispiel 5.12

Es seien die Messwerte $(x_{11}, y_1) = (1, 1)$, $(x_{21}, y_2) = (2, 0)$ und $(x_{31}, y_3) = (1, 0)$ gegeben. Es soll das Modell $y = \beta_1 x_1$ untersucht werden. β_1 lässt sich mit $X = (1, 2, 1)^T$ berechnen zu

$$\beta_1 = (X^TX)^{-1}X^T\vec{y} = \frac{1}{6}$$

Sei eine weitere Eingangsgröße durch die Messwerte $x_{12} = 2$, $x_{22} = 4$ und $x_{32} = 2$ bekannt. Als Modell werde nun $y = \beta_1 x_1 + \beta_2 x_2$ angesetzt. Die Designmatrix

$$X = \begin{pmatrix} 1 & 2 \\ 2 & 4 \\ 1 & 2 \end{pmatrix}$$

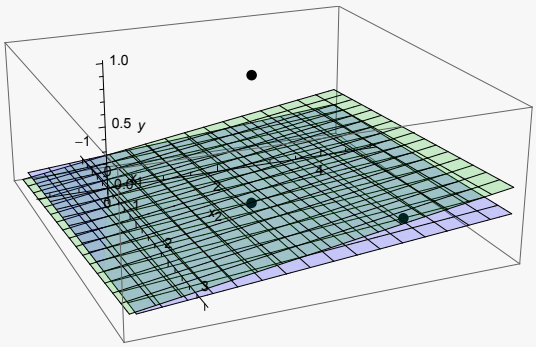
besitzt nur Rang 1<2, deshalb ist keine eindeutige Lösung bestimmbar. Es ergibt sich $\beta_1 = \frac{1}{6} - 2\beta_2$ für ein frei wählbares β_2, etwa $\beta_2 = \frac{1}{15}$ und $\beta_1 = \frac{1}{30}$. Deshalb werde das regularisierte Problem betrachtet. Für $\vec{\beta}$ errechnet sich

$$\vec{\beta} = \frac{1}{\lambda^2 + 30} \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

Für die Eigenwerte von $\tilde{X}^T\tilde{X}$ der erweiterten Designmatrix

$$\tilde{X} = \begin{pmatrix} 1 & 2 \\ 2 & 4 \\ 1 & 2 \\ \lambda & 0 \\ 0 & \lambda \end{pmatrix}$$

ergibt sich $\mu_1 = \lambda^2$ und $\mu_2 = 30 + \lambda^2$ und somit für die Kondition $\kappa = \frac{30+\lambda^2}{\lambda^2}$. Für kleine λ wird das Problem schlecht konditioniert sein.



5.1.4. Bereichsschätzungen

Bei einer Punktschätzung ist im stetigen Fall die Wahrscheinlichkeit, den oder die wahren Parameter zu treffen, gleich Null. Es gibt keinerlei Genauigkeitsaussage hinsichtlich der Schätzung. Deshalb versucht man auf Basis einer Stichprobe $(x_1, \ldots, x_n)^T$ u.i.v. Zufallsvariablen $X_1, \ldots, X_n$, bei der Schätzung eines einzelnen unbekannten Parameters θ ein Intervall $[I_u, I_o]$ anzugeben, in dem mit einer vorgegebenen Wahrscheinlichkeit von mindestens $1 - \alpha$ $(0 < \alpha < 1)$ der gesuchte unbekannte Parameter seinen Wert annimmt, d.h.

$$\mathbb{P}(I_u(X_1, \ldots, X_n) \leq \theta \leq I_o(X_1, \ldots, X_n)) \geq 1 - \alpha.$$

Die Wahrscheinlichkeit $1 - \alpha$ heißt das Konfidenzniveau. Dabei ist nicht gesagt, dass das Intervall den wahren Wert des Parameters tatsächlich überdeckt. Er wird nur mit vorgegebenem Konfidenzniveau überdeckt. α stellt dabei die Irrtumswahrscheinlichkeit dar und muss vorab festgelegt werden.

Seien nun $X_i \sim \mathcal{N}(\mu, \sigma^2)$ für $i = 1, \ldots, n$ und σ^2 bekannt. Als Schätzfunktion für $\gamma(\mu, \sigma^2) = \mu$ werde das arithmetische Mittel $\bar{X}$ verwendet. Bei Gültigkeit der Annahme gilt $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$ und weiter

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

Somit gilt

$$\begin{aligned}\mathbb{P}\left(\left|\sqrt{n}\frac{\bar{X}-\mu}{\sigma}\right| \leq \Phi_{1-\frac{\alpha}{2}}\right) &= 1-\alpha \\ \mathbb{P}\left(-\frac{\sigma}{\sqrt{n}}\Phi_{1-\frac{\alpha}{2}} \leq \bar{X}-\mu \leq \frac{\sigma}{\sqrt{n}}\Phi_{1-\frac{\alpha}{2}}\right) &= 1-\alpha \\ \mathbb{P}\left(\bar{X}-\frac{\sigma}{\sqrt{n}}\Phi_{1-\frac{\alpha}{2}} \leq \mu \leq \bar{X}+\frac{\sigma}{\sqrt{n}}\Phi_{1-\frac{\alpha}{2}}\right) &= 1-\alpha\end{aligned}$$

Damit lässt sich für μ das Intervall

$$[I_u(X_1,\ldots,X_n), I_o(X_1,\ldots,X_n)] = \left[\bar{X}-\frac{\sigma}{\sqrt{n}}\Phi_{1-\frac{\alpha}{2}}, \bar{X}+\frac{\sigma}{\sqrt{n}}\Phi_{1-\frac{\alpha}{2}}\right]$$

angeben. Ein solches Intervall wird Konfidenzintervall genannt. Die Länge des Intervalls,

$$L = 2 \cdot \frac{\sigma}{\sqrt{n}}\Phi_{1-\frac{\alpha}{2}}$$

hängt vom Stichprobenumfang n und dem gewählten α ab. Soll bei gleichbleibender Irrtumswahrscheinlichkeit α die Länge des Intervalls fest vorgegeben werden, kann damit der mindestens notwendige Stichprobenumfang bestimmt werden zu

$$n \geq \left(\frac{2\Phi_{1-\frac{\alpha}{2}}\sigma}{L}\right)^2.$$

Beispiel 5.13

Aus einem Sortiment von Schrauben werden zehn entnommen und deren Länge in Millimetern gemessen. Es ergibt sich die Stichprobe $(10, 8, 9, 10, 11, 11, 9, 12, 8, 12)^T$. Vom Hersteller der Schrauben wird eine Varianz von $\sigma^2 = 4$ vorgegeben. Unter der Annahme einer normalverteilten Grundgesamtheit soll ein Konfidenzintervall zur Irrtumswahrscheinlichkeit von $\alpha = 0.05$ für den unbekannten Erwartungswert μ bestimmt werden. Es ist $\Phi_{0.975} = 1.960$. Das arithmetische Mittel der Stichprobe ist $\bar{x} = 10$. Mit $\Phi_{0.975} \cdot \frac{\sigma}{\sqrt{n}} = 1.240$ ergibt sich das Konfidenzintervall $[8.760, 11.240]$.

Das Konfidenzintervall besitzt die Länge $L = 2.480$. Um die Länge auf $L = 1$ zu verkürzen, müsste ceteris paribus der Stichprobenumfang auf $n = \left\lceil \left(\frac{2 \cdot 1.960 \cdot 2}{1}\right)^2 \right\rceil = 62$ erhöht werden.

Bereichsschätzungen können auf Basis jeder gegebenen Verteilung vorgenommen werden. Ist etwa S^2 eine Schätzfunktion für die Varianz gemäß der in Abschnitt 7.1 beschriebenen empirischen Varianz, so gilt mit den entsprechenden Quantilen der χ^2-Verteilung

$$\mathbb{P}\left(\frac{(n-1)S^2}{\chi^2_{n-1}(1-\alpha/2)} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1}(\alpha/2)}\right) = 1-\alpha.$$

5.2. Testtheorie

Im eingangs eingeführten Beispiel (5.1) liegt eine Verteilungsannahme zugrunde. Es wird ein parametrischer statistischer Raum $(\Psi, \mathcal{G}, \mathbb{P}_{\mathbb{X}, \mathcal{W}_{\vartheta \in \not\subseteq}})$ betrachtet. Doch nicht immer kann eine Verteilungsannahme getroffen werden. Auch für diesen Fall lassen sich Testprobleme formulieren. Allgemein ist eine Entscheidung zwischen zwei Aussagealternativen hinsichtlich der Grundgesamtheit zu treffen. Die erste Aussagealternative wird als Nullhypothese H_0, die zweite als Alternativhypothese H_1 bezeichnet. Beide Alternativen müssen sich gegenseitig ausschließen. Die Entscheidung für eine der beiden Alternativen wird auf Basis einer Stichprobe getroffen. Dazu wird eine Entscheidungsfunktion definiert, welche eine Regel zur Entscheidungsfindung angibt.

5.2.1. Parametrische Tests

Zunächst fassen wir die ersten Überlegungen zu Testproblemen auf Basis parametrisierter Verteilungsannahmen zusammen.

Definition 5.14

Sei $\left(\Psi, \mathcal{G}, \mathbb{P}_{X, W_{\vartheta \in \Theta}}\right)$ ein statistischer Raum, $\gamma : \Theta \to \Pi$ eine Abbildung und $\{\Pi_0, \Pi_1\}$ eine Partition von Π. Die Formulierung

$$H_0 : \gamma(\vartheta) \in \Pi_0, \quad \gamma(\vartheta) \in \Pi_1,$$

heißt parametrisches Testproblem.

Ist eine Stichprobe gegeben, soll eine Entscheidung über die Annahme oder Ablehnung von H_0 herbeigeführt werden. Dazu werden n u.i.v. Zufallsvariablen X_i betrachtet und eine Testfunktion $T(X_1, \ldots, X_n)$ erzeugt, deren Verteilung von den X_i abhängt. Der Wertebereich von T wird in zwei Bereiche unterteilt: einen kritischen Bereich K und einen Annahmebereich $\bar{K}$. Ist auf Basis einer Stichprobe $(x_1, \ldots, x_n)^T$ der Wert $T(x_1, \ldots, x_n)$ der Prüfgröße in K, $T(x_1, \ldots, x_n) \in K$, so wird H_0 abgelehnt, ansonsten d.h. $T(x_1, \ldots, x_n) \in \bar{K}$ sagt man, H_0 kann nicht abgelehnt werden. Letzteres ist aber nicht als Argument gegen H_1 zu verstehen. Es treten vier Situationen auf:

- H_0 ist wahr und wird nicht abgelehnt (richtige Entscheidung),
- H_0 ist wahr und wird abgelehnt (Fehler 1. Art),
- H_0 ist nicht wahr und wird nicht abgelehnt (Fehler 2. Art),
- H_0 ist nicht wahr und wird abgelehnt (richtige Entscheidung).

Zur Konstruktion von Tests wird ein so genanntes Signifikanzniveau α mit $0 < \alpha < 1$ festgelegt. Der Test heißt dann Signifikanztest zum Niveau α und wird so konstruiert, dass

$$\mathbb{P}_\theta(T(X_1, \ldots, X_n) \in K) \leq \alpha \text{ für alle } \theta \in \Theta \text{ und } \gamma(\theta) \in \Pi_0$$

gilt. Wird H_0 abgelehnt, gilt H_1 als statistisch signifikant mit einer Irrtumswahrscheinlichkeit von höchstens α. Der Fehler 1. Art ist vorgegeben, der Fehler 2. Art,

$$\mathbb{P}_\theta(T(X_1, \ldots, X_n) \in \bar{K}) \text{ für alle } \theta \in \Theta \text{ und } \gamma(\theta) \in \Pi_1,$$

soll dabei möglichst klein sein. Die Funktion

$$G(\theta) = \mathbb{P}_\theta(T(X_1, \ldots, X_n) \in K)$$

in Abhängigkeit von θ heißt Gütefunktion des Tests, die Funktion

$$o(\theta) = 1 - G(\theta) = \mathbb{P}_\theta(T(X_1, \ldots, X_n) \in \bar{K})$$

die Operationscharakteristik des Tests.

Ist $g(\theta) \geq \alpha$ für alle $\theta \in \Theta$ und $\gamma(\theta) \in \Pi_1$, so heißt der Test unverfälscht. Bei $\Pi_0 = \{\theta_0\}$ und $\Pi_1 = \Pi \setminus \Pi_0$ spricht man von einem zweiseitigen Test, bei $\Pi_0 = \{\pi \in \Pi_0;\ \pi \geq \theta_0\}$ oder $\Pi_0 = \{\pi \in \Pi_0;\ \pi \leq \theta_0\}$ und $\Pi_1 = \Pi \setminus \Pi_0$ von einem einseitigen Test. Neben der Angabe des kritischen Bereichs wird auch der so genannte p-value benutzt. Dabei gilt für den einseitigen Test

$$p = \mathbb{P}_{\theta_0}(T(X_1, \ldots, X_n) > t) \text{ bzw. } p = \mathbb{P}_{\theta_0}(T(X_1, \ldots, X_n) < t)$$

und für den zweiseitigen Test

$$p = \mathbb{P}_{\theta_0}(|T(X_1, \ldots, X_n)| > t).$$

Ist $p \leq \alpha$, so wird H_0 abgelehnt, ansonsten kann H_0 nicht abgelehnt werden.

Beispiel 5.15: einfacher Gauß-Test

Es seien $X_i \sim \mathcal{N}(\mu, \sigma^2)$ u.i.v. mit bekannter Varianz σ^2 und unbekanntem Erwartungswert μ. Es sei $H_0 : \mu = \mu_0$, also $\gamma(\mu, \sigma^2) = \mu$ und damit $H_1 : \mu \neq \mu_0$. Für das Signifikanzniveau wird oft $\alpha = 0.05$ festgelegt. Dies ist eine willkürliche Wahl. Durch das arithmetische Mittel mit anschließender Standardisierung lässt sich eine unter H_0 entsprechend verteilte Testgröße

$$T(X_1, \ldots, X_n) = \sqrt{n}\frac{\bar{X} - \mu_0}{\sigma} \overset{H_0}{\sim} \mathcal{N}(0, 1)$$

bilden. Liegt der Wert $t = T(x_1, \ldots, x_n)$ nicht in der Nähe von Null, so trifft die Nullhypothese nicht zu. Der Bereich $K = \mathbb{R} \setminus [-k, k]$ umfasst unter der Nullhypothese eine Wahrscheinlichkeit von α derart, dass

$$\mathbb{P}_{\mu_0}(|T(X_1, \ldots, X_n)| > k) = \alpha$$

ist. Mit $k = \phi_{1-\frac{\alpha}{2}}$ gilt $K = (-\infty, -\phi_{1-\frac{\alpha}{2}}) \cup (\phi_{1-\frac{\alpha}{2}}, \infty)$. Gilt $|t| > \phi_{1-\frac{\alpha}{2}}$, so wird die Nullhypothese abgelehnt.

Analog zu den Intervallschätzungen lässt sich eine Vielzahl an Paramtertests auf Basis gegebener Verteilungen erzeugen. Ist etwa die Varianz beim Gauß-Test unbekannt und muss geschätzt werden, ergibt sich eine sogenannte t-Verteilung für die Testfunktion T. Anstelle der Quantile der Normalverteilung werden entsprechend Quantile der t-Verteilung betrachtet.

Teil IV.

Deskriptive Statistik

6. Merkmale mit Nominalskala

Erläuterung

Es gibt manchmal Wertemengen zu Eigenschaften, die sich nicht in das Konzept der Menge der reellen Zahlen einbetten lassen. Mit Konzept ist dabei gemeint, wie sich die erhobenen Daten weiterverarbeiten lassen. Dennoch ist es sinnvoll und notwendig, auch solche Daten durch Kenngrößen und Visualisierungen zu beschreiben. Ebenso ist es häufig von großem Interesse, ob es eine Beziehung zwischen zwei oder mehreren solchen Eigenschaften gibt.

Die Analyse gemeinsam auftretender Werte (Items) hat durch die rasante Entwicklung des Internets enorm an Bedeutung gewonnen. Ausgehend von vereinzelten Marketing-basierten Verhaltensanalysen von Kunden anfang der 1990er Jahre, der so genannten Warenkorbanalyse, sind heute derartige Ansätze weit verbreitet. Ein oft anzutreffender Text auf diversen Webportalen ist etwa

> „Kunden, die dieses Produkt gekauft haben, kauften auch . . . “

Die Analyse des Nutzerverhaltens (Click-Verhalten) auf einzelnen Webseiten oder die Untersuchung von Texten auf Zusammenhänge zwischen vorkommenden Worten und einzelnen Schlüsselbegriffen sind Beispiele für die Untersuchung gemeinsam auftretender Werte. Auch in den Ingenieurwissenschaften finden entsprechende Verfahren Anwendung.

Beispiel 6.1: Diagnostik von Störungen

Oft ist es schwierig, die Ursache für eine Störung an technischen Anlagen zu finden. Diagnosesysteme nutzen physikalische Modelle und Erfahrungen, um die Ursachenfindung zu unterstützen. Physikalische Eigenschaften der technischen Anlage und konkrete Messungen einzelner Größen werden zu so genannten Assoziationsregelmodellen verarbeitet. Die Wirkung (Störung) wird dabei auf Basis gewisser Kriterien durch Ursachen (Werte physikalischer Größen) erklärt.

Grundlegend für eine Vielzahl an Verfahren zur Zusammenhangsanalyse ist der so genannte Item-Support. Das heißt nichts anderes als dass wir das Vorkommen von (Merkmals)Werten abzählen müssen.

6.1. Absolute und relative Häufigkeiten

Allen Skalenarten ist gemein, dass bei einer Stichprobe die beobachteten Ausprägungen zusammengefasst und ihre Anzahl ausgezählt werden kann. Insbesondere bei qualitativen

Merkmalen ist eine sinnvolle und wichtige Beschreibung über das Auszählen der Realisierungen, das Bestimmen von Häufigkeiten, möglich. Wir betrachten zunächst ein qualitatives Merkmal X mit dem endlichen Merkmalsraum $M = \{c_1, \ldots, c_m\}$. Bei qualitativen Merkmalen steht keine Ordnung und keine sinnvolle Abstandsdefinition zweier Merkmalsausprägungen zur Verfügung. Wir unterscheiden zwei Begriffe.

Auszählen von Häufigkeiten

Qualitative Merkmale können in einer Häufigkeitstabelle erfasst werden. Dabei werden für insgesamt n Merkmalsträger $s_1, \ldots, s_n$ einer Stichprobe S und $m = |M|$ Kategorien $c_1, \ldots, c_m$ eines qualitativen Merkmals X entweder absolute Häufigkeiten in Form der Anzahl der Fälle n_l in der Kategorie c_l oder relative Häufigkeiten in Form der Proportion h_l in der Kategorie c_l erfasst. Um die Anzahl auszuzählen, benötigen wir eine Indikatorfunktion

$$I_X^L : M \to \{0,1\}, \quad x \mapsto I_X^L(x) := \begin{cases} 1, & x \in L \text{ (hier: } L = \{c_l\}), \\ 0, & \text{sonst.} \end{cases} \tag{6.1}$$

Die Anzahl der Fälle und die Proportion einer Stichprobe ergeben sich dann über

$$aH : M^n \to \mathbb{N}_0^m, \ \mathbf{x} \mapsto aH(\mathbf{x}) := \left(\sum_{i=1}^n I_X^{\{c_1\}}(x_i), \ldots, \sum_{i=1}^n I_X^{\{c_m\}}(x_i) \right)^T,$$
$$rH : M^n \to \mathbb{N}_0^m, \ \mathbf{x} \mapsto rH(\mathbf{x}) := \left(\frac{1}{n} \sum_{i=1}^n I_X^{\{c_1\}}(x_i), \ldots, \frac{1}{n} \sum_{i=1}^n I_X^{\{c_m\}}(x_i) \right)^T.$$

Der Wert der Abbildung $n : M \to \mathbb{N}_0$, $c_l \mapsto n_l := n(c_l) := aH(x_1, \ldots, x_n)_l$ heißt absolute Häufigkeit der Kategorie c_l und entsprechend heißt der Wert der Abbildung $h : M \to \mathbb{N}_0$, $c_l \mapsto h_l := h(c_l) := rH(x_1, \ldots, x_n)_l$ relative Häufigkeit der Kategorie c_l.

Beispiel 6.2: Passagierdaten vom Untergang der Titanic

Wir betrachten Daten von den 2201 Passagieren zum Untergang der Titanic. Es werden die Klasse, das Alter und Geschlecht sowie das Überleben der jeweiligen Person erfasst. Hier ein Ausschnitt der Datentabelle:

Datentabelle „Titanic“

Class	Age	Sex	Survived
1st	Adult	Male	Yes
1st	Adult	Male	Yes
1st	Adult	Male	Yes
1st	Adult	Male	Yes
⋮	⋮	⋮	⋮
Crew	Adult	Female	Yes
Crew	Adult	Female	No
Crew	Adult	Female	No
Crew	Adult	Female	No

Bei allen vier Merkmalen handelt es sich um qualitative Merkmale und wir können die Fälle auszählen:

Class

Kategorie	n_l	h_l
1st	325	$325/2201$
2nd	285	$285/2201$
3rd	706	$706/2201$
Crew	885	$885/2201$
gesamt	2201	1

Age

Kategorie	n_l	h_l
Adult	2092	$2092/2201$
Child	109	$109/2201$
gesamt	2201	1

Sex

Kategorie	n_l	h_l
Female	470	$470/2201$
Male	1731	$1731/2201$
gesamt	2201	1

Survived

Kategorie	n_l	h_l
Yes	1490	$1490/2201$
No	711	$711/2201$
gesamt	2201	1

Einen einfachen Datensatz können wir noch direkt abzählen. Doch schon beim Titanic-Datensatz mit 2201 Merkmalsträgern ist eine Abzählung offenbar zu aufwändig. Wir nutzen stattdessen R[1].

```
setwd("/Daten/Datensatz")
titanic<-read.table("Titanic.txt",sep="\t",head=TRUE)
attach(titanic)
head(titanic)
str(titanic)
```

In der ersten Zeile setzen wir das Arbeitsverzeichnis. Der Titanic-Datensatz wird aus der entsprechenden Textdatei eingelesen, die Spalten sind durch einen Tabulator getrennt und es existiert eine Kopfzeile. Die einzelnen Variablen werden durch die dritte Zeile zugänglich, d.h. das Merkmal Age kann nun entweder indirekt über „titanic$Age“ oder direkt angesprochen werden. Die vierte Zeile führt zu folgender Ausgabe.

```
  Class   Age  Sex Survived
1 First Adult Male      Yes
2 First Adult Male      Yes
3 First Adult Male      Yes
4 First Adult Male      Yes
```

[1] https://www.r-project.org, ein interessantes Buch zur Anwendung von R ist [5]

```
5 First Adult Male      Yes
6 First Adult Male      Yes
```

Der str-Befehl liefert Informationen über die Datenstruktur eines Objekts.

```
'data.frame':   2201 obs. of  4 variables:
 Class    : Factor w/ 4 levels "Crew","First",..:
                     2 2 2 2 2 2 2 2 2 2 ...
 Age      : Factor w/ 2 levels "Adult","Child":
                     1 1 1 1 1 1 1 1 1 1 ...
 Sex      : Factor w/ 2 levels "Female","Male":
                     2 2 2 2 2 2 2 2 2 2 ...
 Survived : Factor w/ 2 levels "No","Yes":
                     2 2 2 2 2 2 2 2 2 2 ...
```

Um die Häufigkeiten eines Merkmals zu bestimmen, können wir den table-Befehl nutzen.

```
table(Class)
prop.table(table(Class))
```

```
Class
  Crew   First  Second   Third
   885     325     285     706

Class
     Crew      First     Second      Third
0.4020900 0.1476602 0.1294866 0.3207633
```

Meist treten in den Anwendungen mehrere Merkmale gleichzeitig auf. Jedoch ist es sinnvoll, Merkmale auch einzeln zu untersuchen. Das erfolgt nicht zuletzt deshalb, weil es für eine Vielzahl von Verfahren in der angewandten Statistik Annahmen hinsichtlich der Beschaffenheit einzelner Merkmale - wie etwa Verteilungsannahmen - gibt. Weiter können so Untersuchungen hinsichtlich Ausreißern (z.B. Messfehler oder Eingabefehler) durchgeführt werden.

6.2. Modus und Informationsentropie

Bevor wir eine Item-Analyse beginnen können, benötigen wir noch etwas Vorwissen für die Untersuchung von Merkmalen mit Nominalskala.

Lageparameter

Bei einer Realisierung eines qualitativen Merkmals gibt es die Möglichkeit, die Ausprägungen auszuzählen. Interessant dabei ist, welche Ausprägung am häufigsten auftritt. Diese Kenngröße legen wir fest in

Definition 6.3: Modus

Sei X ein qualitatives Merkmal mit $m = |M|$ Kategorien $c_1, \ldots, c_m$ und repräsentiere das Tupel $(n_1, \ldots, n_m)$ die absoluten Häufigkeiten der Kategorien $c_1, \ldots, c_m$ einer Stichprobe von X. Die Abbildung

$$\begin{aligned} md: \quad & \mathbb{N}^m \to \mathcal{P}(\{c_1, \ldots, c_m\}), \\ & (n_1, \ldots, n_m) \mapsto Mod := md(n_1, \ldots, n_m) := \{c_i;\ n_i \geq n_l\ \forall l \in \{1, \ldots, m\}\} \end{aligned}$$

liefert die am häufigsten beobachteten Kategorien und jede Kategorie $c_l \in Mod$ heißt Modus des Merkmals X.

Beispiel 6.4

Der Modus des Merkmals Class aus dem Titanic-Beispiel ist die Kategorie „Crew“. Der Modus des Merkmals Age ist die Kategorie „Adult“.

In R kann der Modus über folgende Funktion bestimmt werden.

Bestimmung des Modus mit R

```
modus<-function(X)
{       a<-table(X, exclude=NULL)
        c<-which(a==max(a))
        return(c)
}
modus(Class)
```

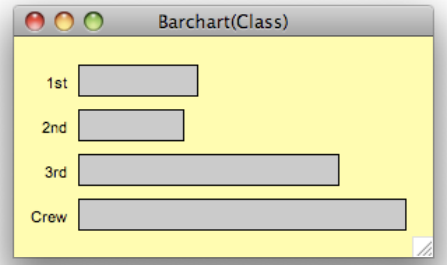
Wir erzeugen zunächst eine Tabelle der absoluten Häufigkeiten des Merkmals, wobei auch fehlende Werte mit aufgenommen werden. Über den which-Befehl werden diejenigen Indizes und Kategoriebezeichnungen gespeichert, für welche die Häufigkeit am größten ist.

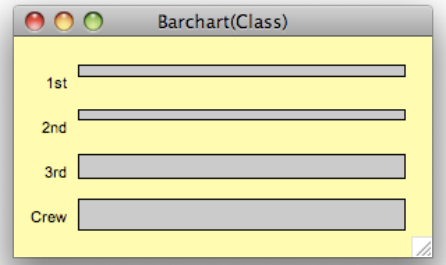
```
Crew
   1
```

Die erste Kategorie der Crew ist der Modus des Merkmals Class im Titanic-Beispiel.

Barchart und Spineplot

Barcharts und Spineplots werden verwendet, um die absoluten oder relativen Häufigkeiten von Kategorien darzustellen. Für jede Kategorie wird eine Rechtecksfläche (Balken) erzeugt. Dabei ist die Häufigkeit proportional zur dargestellten Fläche. Um die Häufigkeiten vergleichen zu können, wird beim Barchart die gleiche Breite gewählt. Die Höhe der Balken ist damit maßgebend für die Häufigkeiten.






```
dat=read.table("Titanic.txt",sep="\t",head=TRUE)
attach(dat)
barplot(table(Class))
mosaicplot(table(Class))
```

Das linke Bild zeigt einen Barchart der Kategorien des Merkmals Class der Titanic-Daten. In einem Barchart lässt sich oft ohne zusätzlichen Aufwand der Modus erkennen. In unserem Fall ist es die Kategorie Crew.

Beim Spineplot ist dagegen, wie im rechten Bild (trotz der Beschriftung mit Barchart) zu sehen, die Höhe der Balken gleich und die Breite der Balken ist damit proportional zu den Häufigkeiten. Oftmals interessiert uns ein gemeinsames Auftreten von Merkmalswerten unterschiedlicher Merkmale. Wenn wir wissen wollen, wie viele Frauen unter den Überlebenden waren, so können wir dies zunächst mit Hilfe einer Kreuztabelle (siehe Abschnitt 6.3) untersuchen.

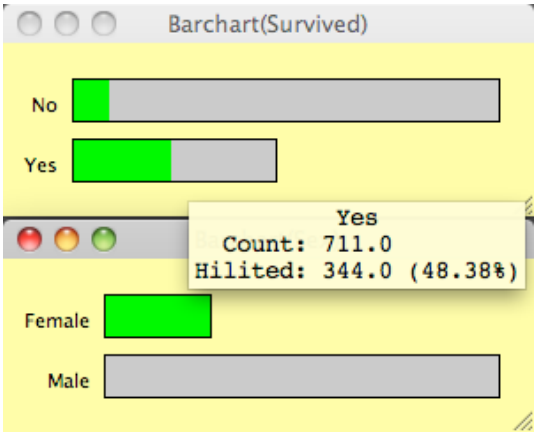
```
ftable(Sex,Survived)
```

Wir nutzen den ftable-Befehl, da er in der Darstellung Vorteile hat, wie wir noch sehen werden.

```
        Survived   No  Yes
Sex
Female             126  344
```

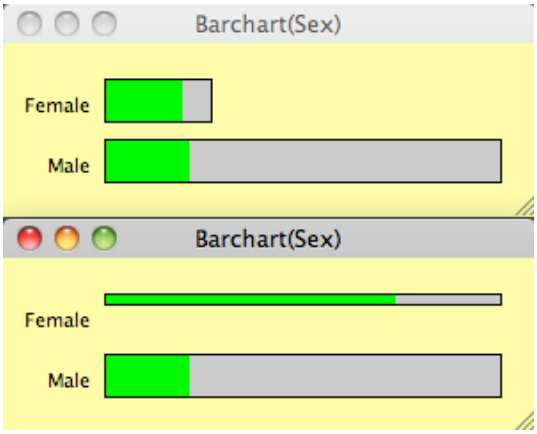
```
Male              1364   367
```

Wir sehen, dass 344 Frauen überlebt haben. Doch die Frage nach dem Anteil Frauen unter den Überlebenden können wir nicht direkt beantworten. Dazu müssen wir den Anteil der Frauen unter allen Überlebenden betrachten, also $\frac{344}{344+367} = 48\%$. Hier gibt es mit Hilfe interaktiver Techniken die Möglichkeit, schneller ans Ziel zu gelangen. Durch gelinktes Higlighting in Verbindung mit einer Abfrage können wir die Antwort sehen.



Durch Highlighting wird der Anteil der Häufigkeiten in jeder Kategorie andersfarbig dargestellt, der beispielsweise durch die Kategorie Yes des Merkmals Survived repräsentiert wird.

Der Spineplot hat Vorteile gegenüber dem Barchart, wenn wir Anteile in den Häufigkeiten vergleichen wollen. Wir fragen uns, ob anteilig mehr Männer oder Frauen überlebt haben.

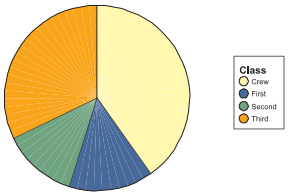


Wir sehen den entscheidenden Vorteil bei der Anwendung des gelinkten Highlighting, da wir sofort erkennen können, dass anteilig mehr Frauen überlebt haben als Männer. Der Barchart gibt die absoluten Zahlen wider, die bei beiden Kategorien annähernd gleich sind.

Pie charts

Ein Pie chart ist eine Graphik, bei der die Fläche eines Kreises entsprechend der Häufigkeiten der Kategorien eines qualitativen Merkmales in Segmente aufgeteilt wird.

Nebenstehendes Bild zeigt einen Pie chart des Merkmals Class. Die Intention des Pie charts entspricht dem des Barcharts. Pie charts sind jedoch eine weniger vorteilhafte Art und Weise, um Informationen über Häufigkeiten darzustellen. Das Auge ist gut darin, lineare Maße abzuschätzen und schlecht darin, relative Flächen zu beurteilen. Ist in dem Pie chart die grüne (Second) oder die blaue (First) Fläche größer?



Streuparameter und Konzentrationsmessung

Die Häufigkeiten eines qualitativen Merkmals können verwendet werden, um zu untersuchen, wie gleichmäßig sich die Daten auf die einzelnen Kategorien verteilen. Entfallen sämtliche Ausprägungen auf einen Wert, so liegt keine Streuung in den Daten, d.h. wir nehmen sicher an, dass bei einer weiteren Realisierung derselbe Wert auftritt. Ist dies nicht der Fall, sind die Daten nicht konzentriert und es liegt eine Unsicherheit bzgl. der Vorhersage der Merkmalsausprägung einer neuen Realisierung vor. Wir können uns somit überlegen, dass die Menge an Information, die in einer neuen Realisierung liegt, als nichtnegative Funktion der relativen Häufigkeiten beschrieben werden kann. Was fordern wir von einer solchen Funktion $s : [0, 1] \to [0, \infty]$? Zunächst soll $s(0) = \infty$ und $s(1) = 0$ gelten. Die Information zweier Ereignisse, die sich gegenseitig nicht beeinflussen, soll sich zudem aufaddieren, d.h. $s(ab) = s(a) + s(b)$. Eine stetige Funktion, die diese Eigenschaften erfüllt, ist die Logarithmus-Funktion. Zur Bestimmung der Streuung bei qualitativen Merkmalen kann das Konzept der Informationsentropie verwendet werden. Die Informationsentropie beschreibt die Gleichmäßigkeit der Häufigkeiten aufgrund der mittleren zu erwartenden Menge an Information (genannt Entropie) und ist ein Erwartungswert.

Definition 6.5: Streuung qualitativer Merkmale

Sei X ein qualitatives Merkmal mit $m = |M|$ Kategorien und seien $h_l, l = 1, \ldots, m$, die relativen Häufigkeiten der Kategorien bei einer gegebenen Stichprobe mit $\sum_{l=1}^{m} h_l = 1$. Es gelte $0 \cdot \ln 0 := 0$. Die Informationsentropie V der Verteilung ist dann definiert als Funktion $V : [0, 1]^m \to [0, 1]$ mit

$$(h_1, \ldots, h_m) \mapsto V(h_1, \ldots, h_m) := -\frac{1}{\ln(m)} \sum_{l=1}^{m} h_l \ln(h_l).$$

Der Wertebereich für V soll zwischen 0 und 1 liegen. Um das zu zeigen, müssen wir ein Minimierungsproblem lösen.

Satz und Definition 6.6: Allgemeine Minimierungsprobleme

Ein allgemeines Minimierungsproblem (MP) ist von der Form

$$\begin{array}{ll} \min\limits_{\mathbf{x}\in\Gamma} & \{f(\mathbf{x})\} \\ \text{unter} & g_i(\mathbf{x}) \leq 0 \quad \forall i = 1,\ldots,m, \\ & h_j(\mathbf{x}) = 0 \quad \forall j = 1,\ldots,l, \end{array} \tag{6.2}$$

mit $f : \Gamma \to \mathbb{R}$, $g_i : \Gamma \to \mathbb{R}$, $h_j : \Gamma \to \mathbb{R}$ und $\Gamma \subseteq \mathbb{R}^n$.

Oftmals lassen sich Minimierungsprobleme mit Nebenbedingungen (die (Un-)gleichungen mit g_i und h_j) nicht direkt lösen. Eine Möglichkeit besteht in der Umformulierung zu einem so genannten Lagrange-Problem.

Satz und Definition 6.7: Lagrange-Probleme

Gegeben sei ein Minimierungsproblem (MP). Dann heißt das Problem

$$\begin{array}{ll} \max\limits_{\mathbf{u}\in\mathbb{R}^m,\, \mathbf{v}\in\mathbb{R}^l} & \left\{ \Theta(\mathbf{u},\mathbf{v}) := \inf\limits_{\mathbf{x}\in\Gamma} \{ f(\mathbf{x}) + \sum\limits_{i=1}^{m} u_i g_i(\mathbf{x}) + \sum\limits_{j=1}^{l} v_j h_j(\mathbf{x}) \} \right\} \\ \text{unter} & u_i \geq 0 \quad \forall i = 1,\ldots,m, \\ & \mathbf{v} \in \mathbb{R}^l, \end{array} \tag{6.3}$$

mit der Lagrange-Funktion $\Theta : \mathbb{R}^m \times \mathbb{R}^l \to \mathbb{R}$ das Lagrange-Duale Problem (DP) zu (MP).

In manchen Fällen gibt es Kandidaten für Optimalpunkte, d.h. Lösungen von (MP) und (DP). Dazu müssen sie folgende Bedingungen erfüllen.

Satz und Definition 6.8: KKT-Bedingungen

Ein Minimierungsproblem liege in der Formulierung (DP) gemäß (6.3) vor. Ein $\mathbf{x} \in \Gamma$ erfüllt die KKT-Bedingungen, falls f und $\mathbf{g} = (g_1,\ldots,g_m)$ differenzierbar sind und u_i, v_j existieren mit

$$\begin{aligned} \nabla f(\mathbf{x})^T + \mathbf{u}^T J_g(\mathbf{x}) + \mathbf{v}^T J_h(\mathbf{x}) &= \mathbf{0}, \\ \mathbf{u}^T \mathbf{g}(\mathbf{x}) &= \mathbf{0}, \\ \mathbf{g}(\mathbf{x}) &\leq \mathbf{0}, \\ \mathbf{h}(\mathbf{x}) &= \mathbf{0}, \\ \mathbf{u} &\geq \mathbf{0}. \end{aligned} \tag{6.4}$$

Sind f und diejenigen g_i mit $g_i(\mathbf{x}) = 0$ konvex bei $\mathbf{x}$ und sind für $v_j \neq 0$ die h_j affin linear, dann ist $\mathbf{x}$ ein Optimalpunkt für (MP) und $(\mathbf{u},\mathbf{v})$ einer für (DP).

Nun können wir uns den Wertebereich von V überlegen.

Satz 6.9

Die Funktion ist sinnvoll definiert, d.h. es gilt: $0 \leq V(h_1, \ldots, h_m) \leq 1$.

Beweis.

Es ist $h_l \ln(h_l) \leq 0$ für jedes l und damit ist V stets positiv. V ist eine durch die Nebenbedingung $\sum_{l=1}^{m} h_l = 1$ im Definitionsbereich eingeschränkte Funktion und nimmt das Maximum dort an, wo $h_l = \frac{1}{m}$ für alle $l \in \{1, \ldots, m\}$ gilt. Denn betrachten wir die dazugehörige zu minimierende Lagrange-Funktion

$$\Theta(v) := \frac{1}{\ln m} \sum_{l=1}^{m} h_l \ln h_l + v \left(\sum_{l=1}^{m} h_l - 1 \right),$$

und bilden die partiellen Ableitungen nach h_j, $\frac{\partial \Theta}{\partial h_j} = \frac{1}{\ln m}(\ln h_j + 1) + v$, bzw. v, $\frac{\partial \Theta}{\partial v} = \left(\sum_{l=1}^{m} h_l - 1 \right)$, so erhalten wir durch Nullsetzen der Gradienten

$$\begin{aligned} \frac{\partial L}{\partial h_j} = 0 \quad &\Leftrightarrow \quad h_j = e^{-u \ln m - 1} \text{ bzw. durch Einsetzen} \\ \frac{\partial \Theta}{\partial u} = 0 \quad &\Leftrightarrow \quad u = 1 - \frac{1}{\ln m}. \end{aligned}$$

Damit ergibt sich $h_j = \frac{1}{m}$. Weil die Hessematrix $H(h_1, \ldots, h_m)$ positiv definit ist, folgt, dass es sich um eine Minimalstelle handelt. Wir untersuchen die beiden Extremfälle. Sei zunächst $h_i = 1$ für ein $i \in \{1, \ldots, m\}$ und $h_j = 0$ für alle $j \neq i$. Dann ist

$$-\frac{1}{\ln(m)} \sum_{l=1}^{m} h_l \ln(h_l) = -\frac{1}{\ln(m)} 1 \cdot \ln(1) = 0.$$

Sei andererseits $h_l = \frac{1}{m}$ für alle $l \in \{1, \ldots, m\}$. Dann haben wir

$$-\frac{1}{\ln(m)} \sum_{l=1}^{m} h_l \ln(h_l) = -\frac{1}{\ln(m)} m \cdot \frac{1}{m} \ln\left(\frac{1}{m}\right) = -\frac{\ln(m^{-1})}{\ln(m)} = 1.$$

□

Beispiel 6.10: Titanic: Class und Age

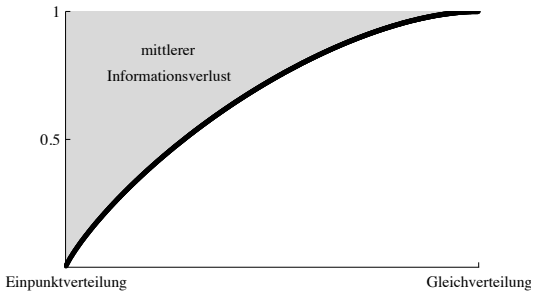
Wir berechnen die Informationsentropien V_1, V_2 für die beiden qualitativen Merkmale Class und Age. Es ist

$$\begin{aligned} V_2 &= -\frac{1}{\log(4)}\left(\frac{885}{2201}\log\left(\frac{885}{2201}\right)+\ldots+\frac{706}{2201}\log\left(\frac{706}{2201}\right)\right) = 0.922 \\ V_1 &= -\frac{1}{\log(2)}\left(\frac{2092}{2201}\log\left(\frac{2092}{2201}\right)+\frac{109}{2201}\log\left(\frac{109}{2201}\right)\right) = 0.284 \end{aligned}$$

Für das Merkmal Klasse wird 92% der möglichen Information geliefert, es liegt viel Information in diesem Merkmal. Hingegen erhalten wir bei Age nur 28% Information, d.h. wir haben einen mittleren Informationsverlust von 72%. Es liegt wenig Information im diesem Merkmal.

```
informationsentropie<-function(X){
k<-prop.table(table(X))
inf<-sum(k*log(k))*(-1/log(length(k)))
print(inf)
}
informationsentropie(CarHire)
```

Bemerkung.
Die Informationsentropie darf nicht als linear interpretiert werden. Die folgende Abbildung verdeutlicht diesen Umstand durch die schematische Entwicklung der Informationsentropie von der Einpunkt- zur Gleichverteilung von relativen Häufigkeiten. Bei einer Einpunktverteilung enthalten die Daten keinerlei Information, während bei einer Gleichverteilung maximale Information in den Daten vorhanden ist.



6.3. Assoziationen

Die Häufigkeitstabellen einzelner qualitativer Merkmale können wir in einer Matrix, der so genannten Kontingenztabelle, zusammenbringen. Sie ist die datenbasierte Form der Kontigenztabelle für Wahrscheinlichkeiten gemäß Tabelle 2.2. Seien k qualitative Merkmale $X_1, \ldots, X_k$ gegeben. Die Kategorien des ersten Merkmals werden in die erste Spalte der

Matrix eingetragen. Die Kategorien des zweiten Merkmals in der ersten Zeile. Mit dem dritten Merkmal findet eine neuerliche Unterteilung der Spalten statt. In die zweite Spalte werden nämlich für jede Kategorie des ersten Merkmals sämtliche Kategorien des dritten Merkmals eingetragen. Das vierte Merkmal unterteilt die erste Zeile mit dem zweiten Merkmal usw. Sei m_j die Anzahl der Kategorien des Merkmals X_j. Dann ergibt sich eine $\prod_{l=2j-1}^{k} m_l \times \prod_{l=2j}^{k} m_l$-Zellen Matrix mit $1 \leq j \leq k$, in deren Zellen die jeweils entsprechend der Kategorien vorhandenen Fälle einzutragen sind.

Beispiel 6.11

Wir erstellen eine Kontingenztabelle mit absoluten Häufigkeiten für die drei Merkmale Class, Age und Sex (in dieser Reihenfolge).

Kontingenztabelle

Class	Sex	Age		
		Adult	Child	Summe
Crew	Female	23	0	23
	Male	862	0	862
1st	Female	144	1	145
	Male	175	5	180
2nd	Female	93	13	106
	Male	168	11	179
3rd	Female	165	31	196
	Male	462	48	510
Summe		2092	109	2201

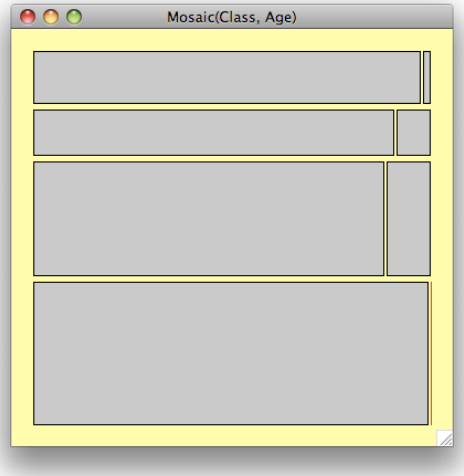
Die letzte Zeile und die letzte Spalte wurden hinzugefügt, um Zeilen- bzw. Spaltensummen zu erfassen. Genauso gut hätten wir die Kontingenztabelle mittels der relativen Häufigkeiten der einzelnen Fälle befüllen können.

Konteingenztabellen bilden häufig die Grundlage für verschiedene Tests.

Mosaicplot

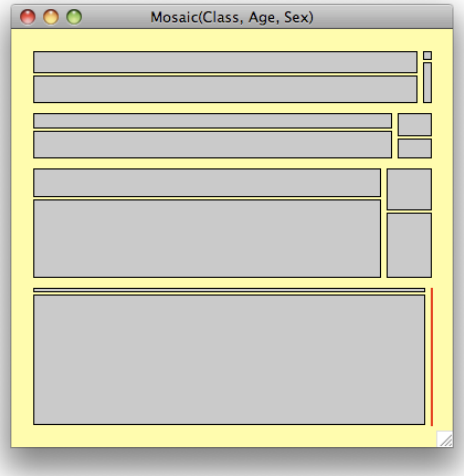
Der Prozess der Erstellung des Mosaicplots ist grundsätzlich gleich zum Aufbau einer Kontingenztabelle, wie wir es in Abschnitt 6.3 bei der Erstellung einer Kontingenztabelle gesehen haben.

Wir beginnen mit einem Spineplot für das erste Merkmal. Das zweite Merkmal wird hinzugefügt, indem die Flächen jeder Kategorie vertikal entsprechend der Proportionen der Kategorien des zweiten Merkmals unterteilt werden. Im Beispiel auf der rechten Seite beginnen wir mit einem Spineplot für das Merkmal Class und fügen das Merkmal Age mit seinen zwei Kategorien hinzu. Die Fläche oben rechts repräsentiert hier den Anteil der Kinder der ersten Klasse, die Fläche oben links den Anteil der Erwachsenen der ersten Klasse. Unten rechts ist durch den roten Strich zu erkennen, dass keine Kinder Mitglied der Crew der Titanic waren.




```
mosaicplot(Class,Age)
```

Das Merkmal Sex gibt das Geschlecht jedes Merkmalsträgers an. Erweitern wir den Mosaicplot um das Merkmal, werden die Flächen wiederum horizontal unterteilt. Die oberen Flächen repräsentieren weibliche Personen, die unteren Flächen männliche. Es zeigt sich, dass es in der ersten Klasse lediglich ein Mädchen unter sechs Kindern gab. Lässt sich nun die Frage beantworten, ob auch auf der Titanic das Prinzip „Frauen und Kinder zuerst galt“ oder ob überproportional viele Personen der ersten Klasse überlebt haben? An dieser Stelle soll auf die interaktiven Möglichkeiten der Exploration der Daten verwiesen werden! Die Vorgehensweise des Hinzufügens von Merkmalen kann nun beliebig mit weiteren qualitativen Merkmalen fortgesetzt werden.

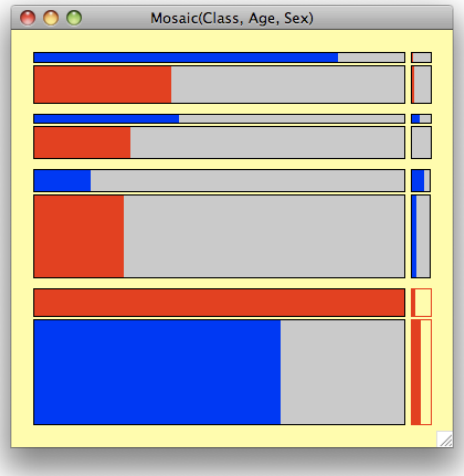


Anhand des Mosaicplots können wir Aussagen über die Abhängigkeit der einzelnen Merkmale treffen. Um das zu untersuchen, müssen sämtliche Reihenfolgen der betrachteten Merkmale gebildet und davon jeweils ein Mosaicplot erstellt werden. Wir betrachten die Räume zwischen den Flächen. Lassen sich diese in jedem Mosaicplot ohne „übermäßiges Abknicken“ durchlaufen, d.h. es liegt eine gleichmäßige Partitionierung vor, kann von der Unabhängigkeit der Merkmale ausgegangen werden (mutual independence). Andernfalls muss eine Abhängigkeit angenommen werden, die weiter untersucht werden kann.

Bei drei Merkmalen können dann die partial independence, die conditional independence, die no three-way interaction und die three-way interaction erkannt werden. Bei der partial independence ist ein Merkmal von den beiden anderen unabhängig, was sich anhand

der Gleichmäßigkeit der beiden Plots zeigt, in denen das bestimmte Merkmal mittig angeordnet ist. Die conditional independence bedeutet, dass zwei Merkmale unter gegebenem dritten unabhängig voneinander sind. Das zeigt sich anhand einer gleichmäßigen Partitionierung innerhalb der Kategorien des gegebenen ersten Merkmals. Allerdings ist die Partitionierung in den beiden entsprechenden Mosaicplots nicht gleichförmig. Die beiden letzten Formen der Unabhängigkeit sind sowohl aus Sicht der Interpretation als auch anhand der Mosaicplots schwer zu erkennen. In unserem Beispiel interpretieren wir anhand der sechs Mosaicplots, dass bei gegebenem Merkmal Class die beiden anderen Merkmale nahezu unabhängig sind, also eine conditional independence vorliegt. Wir stellen die Hypothese auf, dass sich ein Modell aus den drei Merkmalen, und so genannten Interaktionstermen zwischen Class und Age bzw. Class und Sex zusammensetzt.

Es besteht die Möglichkeit, einen Mosaicplot nicht in der Intention wie bisher, sondern unter Annahme des Unabhängigkeitsmodells die erwarteten Werte darzustellen. Ist eine beobachtete Häufigkeit einer „Zelle“ größer als die erwartete, wird die Proportion des Überschusses blau dargestellt. Ist dagegen die beobachtete Häufigkeit kleiner als die erwartete, wird die Proportion des Fehlers rot gezeichnet. In unserem Beispiel werden in der Zelle oben links deutlich mehr erwachsene Frauen in der ersten Klasse erwartet als tatsächlich auf dem Schiff waren. Auch leere Zellen, wie wir sie bei Kindern in der Crew haben, werden nun entsprechend der erwarteten Werte dargestellt. Dieser Mosaicplot ist durchgängig gleichmäßig partitioniert. Wären die beobachteten gleich den erwarteten Werten, ergäbe sich genau dieser Mosaicplot.



Die Darstellung in einer Kontingenztabelle bietet die Möglichkeit herauszufinden, ob es Assoziationen zwischen verschiedenen Merkmalen gibt. Ein derartiges Assoziationsmaß sollte auf das Intervall $[-1, 1]$ beschränkt sein, wobei der Wert 0 die Unabhängigkeit beider Variablen, -1 einen perfekt negativen und 1 einen perfekt positiven Zusammenhang charakterisieren sollte. Werte zwischen 0 und ± 1 deuten dann auf einen mehr oder weniger starken, jedoch keinen perfekten Zusammenhang hin. Die Richtung des Zusammenhangs (negativ, positiv) ist nur bei mindestens ordinal skalierten Variablen sinnvoll interpretierbar. Einige Assoziationsmaße für nominal skalierte Variablen verwenden daher nur das Intervall $[0, 1]$ und werden auch als richtungslose Assoziationsmaße bezeichnet. Wir betrachten zunächst zwei qualitative Merkmale und folgende Kontingenztafel:

		Merkmal X_2			
		$\tilde{c}_1$	$\dots$	$\tilde{c}_{m_2}$	
Merkmal X_1	c_1	n_{11}	$\dots$	n_{1m_2}	$n_{1\cdot}$
	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	c_{m_1}	$n_{m_1 1}$	$\dots$	$n_{m_1 m_2}$	$n_{m_1 \cdot}$
		$n_{\cdot 1}$	$\dots$	$n_{\cdot m_2}$	n

Die Grundlage eines Assoziationsmaßes für zwei qualitative Merkmale stellt die χ^2-Größe dar. Seien in einer Kontingenztabelle bestehend aus zwei qualitativen Merkmalen X_1 und X_2 für $i = 1, \ldots, m_1$ und $j = 1, \ldots, m_2$ n_{ij} die absoluten Häufigkeiten für die Anzahl Fälle in Kategorie c_i des Merkmals X_1 und in Kategorie $\tilde{c}_j$ des Merkmals X_2 bzw. h_{ij} die relativen Häufigkeiten für die Proportion in Kategorie c_i des Merkmals X_1 und in Kategorie $\tilde{c}_j$ des Merkmals X_2. Die Zeilensumme der i-ten Zeile wird mit $n_{i\cdot} = n_{i1} + \ldots + n_{im_2}$ für alle $i = 1, \ldots, m_1$ und die Spaltensumme der j-ten Spalte wird mit $n_{\cdot j} = n_{1j} + \ldots + n_{m_1 j}$ für alle $j = 1, \ldots, m_2$ bezeichnet. Zeilen- und Spaltensummen werden als Randsummen bezeichnet. Entsprechendes gilt für die relativen Häufigkeiten und die Zeilen- und Spaltensummen, sie heißen Randhäufigkeiten.

Ein wichtiger Begriff ist die bedingte Häufigkeitsverteilung. Wir bilden die bedingte Häufigkeit für eine Merkmalsausprägung eines Merkmals unter der Bedingung, dass für ein anderes Merkmal eine feste Merkmalsausprägung eintritt. Dann heißt

$$h_{c_i|\tilde{c}_j} = \frac{n_{ij}}{n_{\cdot j}} = \frac{h_{ij}}{h_{\cdot j}} \tag{6.5}$$

für $n_{\cdot j} > 0$ und $i = 1, \ldots, m_1$ bedingte Häufigkeit von $X_1 = c_i$ unter der Bedingung $X_2 = \tilde{c}_j$. Weiter heißt

$$h_{\tilde{c}_j|c_i} = \frac{n_{ij}}{n_{i\cdot}} = \frac{h_{ij}}{h_{i\cdot}} \tag{6.6}$$

für $n_{i\cdot} > 0$ und $j = 1, \ldots, m_2$ bedingte Häufigkeit von $X_2 = \tilde{c}_j$ unter der Bedingung $X_1 = c_i$. Somit kann eine bedingte Häufigkeitsverteilung für ein Merkmal unter der Bedingung einer festgelegten Merkmalsausprägung eines anderen Merkmals bestimmt werden.

Beispiel 6.12

Die bedingte Häufigkeit von CarHire=No unter der Bedingung Main.Purpose=2 ist $\frac{2}{3}$. Die bedingte Häufigkeit von Main.Purpose=2 unter der Bedingung CarHire=No ist $\frac{2}{22}$. Die bedingte Häufigkeitsverteilung für CarHire unter der Bedingung Main.Purpose=2 lautet $h_{\text{No}|2} = \frac{2}{3}, h_{\text{Yes}|2} = \frac{1}{3}$. Die bedingten relativen Häufigkeiten summieren sich wieder zu 1.

Es gilt

$$\begin{aligned} 1 &= \sum_{i=1}^{m_1} h_{c_i|\tilde{c}_j} = \sum_{i=1}^{m_1} \frac{n_{ij}}{n_{\cdot j}} = \frac{n_{\cdot j}}{n_{\cdot j}} \\ &= \sum_{j=1}^{m_2} h_{\tilde{c}_j|c_i} = \sum_{j=1}^{m_2} \frac{n_{ij}}{n_{i\cdot}} = \frac{n_{i\cdot}}{n_{i\cdot}}. \end{aligned}$$

Wir definieren nun die χ^2-Größe, die uns die Grundlage für verschiedene Assoziationsmaße liefert. Hierzu nutzen wir die Randhäufigkeiten. Nehmen wir an, dass die beiden Merkmale empirisch unabhängig voneinander sind, so lassen sich die bedingten Häufigkeiten dafür, dass $X_1 = c_i$ und $X_2 = \tilde{c}_j$ gilt, multiplizieren und wir erhalten dann eine erwartete Häufigkeit (absolut)

$$\nu_{ij} = n \cdot \frac{n_{i\cdot} n_{\cdot j}}{n^2} = \frac{n_{i\cdot} n_{\cdot j}}{n} \tag{6.7}$$

für $X_1 = c_i$ und $X_2 = \tilde{c}_j$. Hieraus lässt sich eine Testgröße bestimmen, indem die normierten quadrierten Abstände der tatsächlichen Häufigkeiten zu den erwarteten Häufigkeiten aufsummiert werden.

Definition 6.13: χ^2-Größe

Die χ^2-Größe zweier qualitativer Merkmale X_1 und X_2 mit m_1 bzw. m_2 Kategorien und n Realisierungen lautet

$$\chi^2 := \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \frac{(n_{ij} - \nu_{ij})^2}{\nu_{ij}}, \tag{6.8}$$

wobei ν_{ij} die erwarteten Häufigkeiten gemäß Gleichung (6.7) sind.

Bemerkung.
Wir schließen den Fall einer Randhäufigkeit von Null aus, da dann die jeweilige Kategorie nicht berücksichtigt werden müsste und sie weggelassen wird.

Ferner hat die χ^2-Größe ihren Namen nicht zu Unrecht. Unter Normalverteilungsannahme ist die Größe χ^2-verteilt und es kann ein entsprechender einseitiger Test $H_0 : \chi^2 = 0$ gegen $H_1 : \chi^2 \neq 0$ auf Unabhängigkeit durchgeführt werden (vgl. [2], S. 337f.

Bei empirischer Unabhängigkeit der beiden Merkmale stimmen die beobachteten mit den erwarteten Häufigkeiten überein. Dann ist der χ^2-Wert Null. Auf der anderen Seite ist der χ^2-Wert nach oben beschränkt. Allgemein gilt zunächst

$$\begin{aligned}
\sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \frac{(n_{ij} - \nu_{ij})^2}{\nu_{ij}} &= \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \frac{n_{ij}^2 - 2n_{ij}\nu_{ij} + \nu_{ij}^2}{\nu_{ij}} \\
&= \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \left(\frac{n_{ij}^2}{\nu_{ij}} - 2n_{ij} + \nu_{ij} \right) = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \left(\frac{n_{ij}^2}{\nu_{ij}} \right) - n \\
&= n \left(\sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \left(\frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} \right) - 1 \right).
\end{aligned} \tag{6.9}$$

Um den χ^2-Wert nach oben abzuschätzen, müssen wir den Fall bestimmen, bei welchem die Merkmale vollkommen abhängig sind. Zwei Merkmale sind vollkommen abhängig, wenn in einer Kontingenztabelle

- für $m_1 > m_2$ in jeder Zeile die Häufigkeiten in genau einem Feld konzentriert sind, $h_{\tilde{c}_j | c_i} = \frac{n_{ij}}{n_{i\cdot}}$ ist 1 für genau ein j ansonsten 0,
- für $m_1 = m_2$ in jeder Zeile und Spalte die Häufigkeiten in genau einem Feld konzentriert sind, $h_{\tilde{c}_j | c_i} = h_{c_i | \tilde{c}_j} = \frac{n_{ij}}{n_{i\cdot}} = \frac{n_{ij}}{n_{\cdot j}}$ ist 1 für genau ein j bzw. i ansonsten 0, bzw.
- für $m_1 < m_2$ in jeder Spalte die Häufigkeiten in genau einem Feld konzentriert sind, $h_{c_i | \tilde{c}_j} = \frac{n_{ij}}{n_{\cdot j}}$ ist 1 für genau ein i ansonsten 0.

Die Gleichung (6.9) vereinfacht sich zu

$$\chi^2 = n \cdot (\min\{m_1, m_2\} - 1).$$

Insgesamt haben wir gezeigt, dass

$$0 \leq \chi^2 \leq n \cdot (\min\{m_1, m_2\} - 1).$$

gilt und damit ist auch zu sehen, dass in Abhängigkeit von der Anzahl der Realisierungen n der χ^2-Wert unbeschränkt ansteigt. Das ist jedoch im Hinblick auf ein Assoziationsmaß problematisch, da stets zuerst die obere Schranke bestimmt werden muss. Deswegen wird der χ^2 selbst nicht als Maß verwendet, sondern daraus werden Maße konstruiert. Wir definieren

Definition 6.14: Kontingenzkoeffizient nach Pearson

Der Kontingenzkoeffizient C nach Pearson ist definiert als

$$C := \sqrt{\frac{\chi^2}{n + \chi^2}}. \tag{6.10}$$

Setzen wir die untere bzw. obere Schranke für den χ^2-Wert ein, so ist der Kontingenzkoeffizient C beschränkt durch

$$0 \leq C \leq \sqrt{\frac{n \cdot (\min\{m_1, m_2\} - 1)}{n + n \cdot (\min\{m_1, m_2\} - 1)}} = \sqrt{\frac{\min\{m_1, m_2\} - 1}{\min\{m_1, m_2\}}} < 1.$$

Ein noch vorhandenes Problem besteht darin, dass der Koeffizient von der Anzahl der Kategorien abhängt. Auch das Problem können wir durch eine Normierung beseitigen.

Definition 6.15: Korrigierter Kontingenzkoeffizient nach Pearson

Der korrigierte Kontingenzkoeffizient nach Pearson ist definiert als

$$C_* := C\sqrt{\frac{\min\{m_1, m_2\}}{\min\{m_1, m_2\} - 1}}. \tag{6.11}$$

Es ist offensichtlich, dass $0 \leq C_* \leq 1$ gilt. Dabei liegt völlige empirische Unabhängigkeit für $C_* = 0$ und völlige empirische Abhängigkeit für $C_* = 1$ vor. Der korrigierte Kontingenkoeffizient C_* ist ungerichtet. Der korrigierte Kontingenzkoeffizient beschreibt also die Stärke des Zusammenhangs zweier Merkmale auf Basis eines beliebigen Skalenniveaus.

Beispiel 6.16

		Sex		
		Female	Male	
Survived	No	126 (*318.2*)	1364 (*1171.8*)	1490
	Yes	344 (*151.8*)	367 (*559.2*)	711
		470	1731	2201

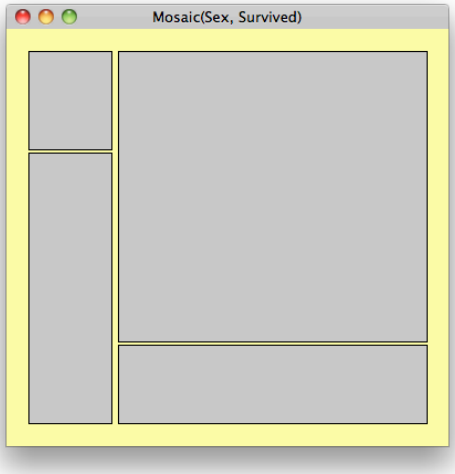
Für die beiden qualitativen Merkmale Sex und Survived ergibt sich die aufgeführte Kontingenztabelle aus den Beobachtungswerten, in Klammern stehen die erwarteten Werte. Aus $\chi^2 = 456.9$ ergeben sich die Kontingenzkoeffizienten $C = 0.415$ und $C_* = 0.586$. Das lässt auf eine vorhandene empirische Abhängigkeit schließen.

Auch in R gibt es die Möglichkeit, den Pearsonschen-Kontingenzkoeffizienten zu bestimmen. Dazu benötigen wir die Library vcd und den Befehl assocstats.

```
library(vcd)
assocstats(ftable(Sex, Survived))
```

```
                    X^2  df  P(> X^2)
Likelihood Ratio  434.47   1         0
Pearson           456.87   1         0
Phi-Coefficient   : 0.456
Contingency Coeff.: 0.415
Cramers V         : 0.456
```

Wir können $\chi = 456.87$ (Pearson) und $C = 0.415$ (Contingency Coeff.) ablesen und damit $C_* = 0.586$ (siehe Beispiel 6.16) bestimmen. Dies können wir auch graphisch erkennen, indem wir uns den Mosaicplot der beiden Merkmale ansehen. Unabhängigkeit würde bedeuten, dass die Fläche des Quadrats so in Rechtecksflächen aufgeteilt würde, dass jede Rechtecksfläche das Produkt der dazugehörigen Randhäufigkeiten darstellt. Es ergäben sich durchgezogene Linien innerhalb des Quadrats, da sich in jeder Seitenlänge die entsprechende Randhäufigkeit wiederspiegelte. Je weiter ein Mosaicplot davon abweicht, desto weniger ist diese Unabhängigkeitsannahme gegeben.



Das Beispiel zeigt deutlich die Abweichung von der Unabhängigkeitsannahme.

Bemerkung.
(I) Es sei noch einmal betont, dass eine Assoziation nichts über kausale Zusammenhänge zwischen Merkmalen aussagt.
(II) Bedingte Häufigkeiten und damit auch Kontingenzkoeffizienten lassen sich genauso für mehr als zwei Merkmale bestimmen. Für den korrigierten Kontingenzkoeffizienten wird entsprechend die minimale Kategorienanzahl aller Merkmale bestimmt.

Beispiel 6.17

Wir betrachten Daten zum Unglück der Titanic im Jahre 1912. Von 2201 Personen überlebten 711 den Untergang des Schiffs. Wir haben die qualitativen Merkmale Class, Age, Sex und Survived zur Verfügung. Stellen wir die drei ersten Merkmale (etwa in dieser Reihenfolge) in einer Kontingenztabelle dar, so erhalten wir $\chi^2 = 523.740$, $C = 0.438$ und $C_* = 0.620$. Das ergibt einen relativ hohen Wert, der auf eine ausgeprägte empirische Abhängigkeit schließen lässt.

7. Merkmale mit Kardinalskala

Die Grundannahme der Mathematischen Statistik, die laut [13] dadurch gegeben ist, die „Beobachtungen als Realisierungen von Zufallsgrößen aufzufassen und damit (zu) unterstellen, dass sich der Vorgang durch eine Wahrscheinlichkeitsverteilung beschreiben lässt“, liefert eine Schnittstelle zwischen einer daten- und modellgetriebenen Untersuchung in der Stochastik. Eine Beobachtung, z.B. eine Messung bei der Durchführung eines physikalischen Experiments, ist demnach dem Zufall unterworfen. Zufallseinflüsse werden durch einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P})$, der nicht näher spezifiziert wird, modelliert.

7.1. Kenngrößen für ein Merkmal

Sei $\mathbf{x} = (x_1, \ldots, x_n)^T$ eine Stichprobe eines kardinalen Merkmals X. Die drei zumeist benutzten Kenngrößen zur Beschreibung kardinaler Merkmale sind der empirische Mittelwert $\bar{x}$, die empirische Varianz s^2 und die empirische Standardabweichung s,

$$\bar{x} : \mathbb{R}^n \to \mathbb{R}, \qquad \mathbf{x} \mapsto \bar{x} := \bar{x}(\mathbf{x}) := \frac{1}{n} \sum_{i=1}^{n} x_i, \tag{7.1}$$

$$s^2 : \mathbb{R}^n \to \mathbb{R}, \qquad \mathbf{x} \mapsto s^2 := s^2(\mathbf{x}) := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2, \tag{7.2}$$

$$s : \mathbb{R}^n \to \mathbb{R}, \qquad \mathbf{x} \mapsto s := s(\mathbf{x}) := \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}. \tag{7.3}$$

Die drei Funktionen stellen die empirischen Varianten von Schätzfunktionen dar. Der durch Einsetzen der Stichprobenelemente erhaltene Wert ist der Schätzwert. Wir übernehmen somit die Konzepte aus der Wahrscheinlichkeitstheorie. Sei G eine Menge von Schätzfunktionen und $g \in G$. Dann heißt g konsistent, wenn $\mathbb{V}_\theta[g] \overset{n\to\infty}{\longrightarrow} 0$.
Die drei Kenngrößen sind demnach erhaltene Schätzwerte der angegebenen Schätzfunktionen. Der empirische Mittelwert ist ein Lageparameter, während die empirische Varianz und die Standardabweichung Streuparameter sind.

Beispiel 7.1

Sei $\mathbf{x} = (5, 10, 4, 13, 8)^T$. Dann gilt

$$\begin{aligned} \bar{x} &= \frac{1}{5}(5+10+4+13+8) = 8, \\ s^2 &= \frac{1}{4}\left((5-8)^2+(10-8)^2+(4-8)^2+(13-8)^2+(8-8)^2\right) = \frac{54}{4} = 13.5, \\ s &= \sqrt{13.5} = 3.67. \end{aligned}$$

Beispiel 7.2: Wisconsin Brustkrebs-Daten

Bei einer Fine-needle aspiration biopsy (FNA)[a] werden durch eine feine Nadel dem Körper Zellen entnommen und unter dem Mikroskop untersucht. Im vorliegenden Datensatz[b] wurden jedem der 569 Patienten 10-40 Zellen entnommen und deren Zellkerne untersucht. Dabei wurden 10 Merkmale erfasst: radius, texture, peri, area, smooth, comp, scav, ncav, symt und fracd. Über alle Kerne wurden der empirische Mittelwert, der größte Wert und die empirische Standardabweichung gebildet, so dass für jeden Patienten diese 30 Merkmale zur Verfügung stehen. Darüber hinaus wird erfasst, ob das Gewebe bös- oder gutartig ist. Für das Merkmal radius.mv erhalten wir beispielsweise

$$\begin{aligned} \bar{x} &= 14.13, \\ s^2 &= 12.42, \\ s &= 3.52. \end{aligned}$$

[a] http://en.wikipedia.org/wiki/Fine-needle_aspiration
[b] siehe https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

Äquivarianz

Bei Merkmalen spielt möglicherweise die Messskala eine wichtige Rolle. So können bei einer Messung von Temperaturen die Messwerte in Grad Celsius, Kelvin oder Fahrenheit angegeben werden. Ein Lage- oder Streuparameter sollte einen Wert unabhängig von der gewählten Messskala liefern. Schätzfunktionen für Lageparameter $m : \mathbb{R}^n \to \mathbb{R}$ werden als äquivariant bezeichnet, wenn für beliebige $a, b \in \mathbb{R}$

$$m(a \cdot x_1 + b, \ldots, a \cdot x_n + b) = a \cdot m(x_1, \ldots, x_n) + b$$

gilt, während Schätzfunktionen für Streuparameter $s : \mathbb{R}^n \to \mathbb{R}$ als äquivariant bezeichnet werden, wenn für beliebige $a, b \in \mathbb{R}$

$$s(a \cdot x_1 + b, \ldots, a \cdot x_n + b) = |a| \cdot s(x_1, \ldots, x_n)$$

gilt.

Satz 7.3

Die Schätzfunktionen für den empirischen Mittelwert und die empirische Standardabweichung sind äquivariant.

Beweis.

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}(a\cdot x_i+b) &= a\cdot\frac{1}{n}\sum_{i=1}^{n}x_i+\frac{1}{n}\cdot n\cdot b = a\cdot\frac{1}{n}\sum_{i=1}^{n}x_i+b \\ \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(a\cdot x_i+b-(a\cdot\bar{x}+b))^2} &= |a|\cdot\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2} \end{aligned}$$

□

Robustheit

Manchmal finden sich bei Messungen Werte, die sich deutlich von den anderen unterscheiden. Sie werden Ausreißer genannt. Kenngrößen sollten möglichst wenig durch einzelne Ausreißer beeinflusst werden. Solche Kenngrößen werden als robust bezeichnet. Um den Einfluss einzelner Beobachtungen bestimmen zu können, gibt es die Möglichkeit der Betrachtung des so genannten Sensitivitätsdiagramms. Dabei wird für jede Realisierung x_j eine skalierte Differenz SC zwischen der Kenngröße m_n der gesamten Stichprobe und der Kenngröße $m_{n(j)}$ der um den einen Merkmalswert reduzierten Stichprobe ermittelt,

$$SC(x_j, m) := k\cdot|m_n - m_{n(j)}|. \tag{7.4}$$

Die Skalierung mit k hängt von der jeweils untersuchten Kenngröße ab. Betrachten wir den Graphen

$$\{(x_j, SC(x_j,m));\ j=1,\ldots,n\} \text{ bzw. } \{(j, SC(x_j,m));\ j=1,\ldots,n\},$$

so werden grundsätzlich besonders große Werte als Ausreißer identifiziert und interpretiert. Mit $\mathbf{1} := (1,\ldots,1)^T$ werde der Vektor (mit entsprechender Dimension) bezeichnet, dessen jede Komponente 1 ist, mit $\tilde{\mathbf{1}} := (1,\ldots,1,0,1,\ldots,1)^T$ der Vektor, dessen j-te Komponente 0, alle anderen Komponenten 1 sind. Es ist $\bar{x} = \frac{1}{n}\mathbf{1}^T\mathbf{x}$. Sei

$$\bar{x}_j := \frac{1}{n-1}\sum_{\substack{i=1\\ i\neq j}}^{n} x_i = \frac{1}{n-1}\mathbf{x}^T\tilde{\mathbf{1}}. \tag{7.5}$$

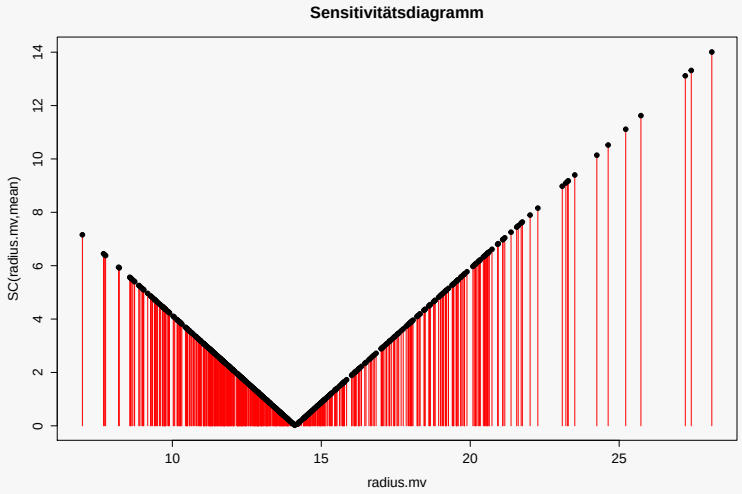
Mit $k = n - 1$ erhalten wir zunächst für den empirischen Mittelwert

$$\begin{aligned} SC(x_j, \bar{x}) &= (n-1) \cdot |\bar{x} - \bar{x}_j| \\ &= (n-1) \left| \bar{x} - \left(\frac{n}{n-1} \left(\bar{x} - \frac{x_j}{n} \right) \right) \right| \\ &= (n-1) \left| \frac{n-1-n}{n-1} \bar{x} - \frac{x_j}{n-1} \right| \\ &= |x_j - \bar{x}|. \end{aligned}$$

SC ist damit ein Maß für die absolute Abweichung von x_j zum empirischen Mittelwert der anderen Realisierungen.

Beispiel 7.4

Für das Merkaml radius.mv und den empirischen Mittelwert erhalten wir folgendes Sensitivitätsdiagramm:



Es ist linear in x_j und es ergeben sich Werte nahe 0 im Bereich der empirischen Mittelwerte (um 14.1).

Wir untersuchen noch die skalierte Differenz für die empirische Varianz. Zur einfacheren Berechnung führen wir noch zwei Bezeichnungen ein. Mit

$$\mathbf{x} - \bar{x} \cdot \mathbf{1} = \mathbf{x} - \frac{1}{n} \mathbf{1}\mathbf{1}^T \mathbf{x} = \underbrace{I - \frac{1}{n} \mathbf{1}\mathbf{1}^T}_{H} \mathbf{x}$$

sei $H := I - \frac{1}{n} \mathbf{1}\mathbf{1}^T$. Entsprechend sei $\tilde{H} := \tilde{I} - \frac{1}{n-1} \tilde{\mathbf{1}}\tilde{\mathbf{1}}^T$, wobei $\tilde{I} = I - \mathbf{e}_j \mathbf{e}_j^T$ eine modifizierte Einheitsmatrix sei, deren j, j-te Komponente 0 ist. Damit lässt sich die empirische Varianz schreiben als

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1} \mathbf{x}^T H^T H \mathbf{x} \overset{(7.18)}{=} \frac{1}{n-1} \mathbf{x}^T H \mathbf{x}.$$

Analog zu (7.5) ist $s_j^2 = \frac{1}{n-2}\mathbf{x}^T\tilde{H}\mathbf{x}$. Setzen wir $k = n-1$, so bekommen wir

$$\begin{aligned}
SC(x_j, s^2) &= (n-1)\cdot|s^2 - s_j^2| \\
&= \left|\mathbf{x}^T H\mathbf{x} - \frac{n-1}{n-2}\mathbf{x}^T\tilde{H}\mathbf{x}\right| \\
&= \left|\mathbf{x}^T\left(H - \frac{n-1}{n-2}\tilde{H}\right)\mathbf{x}\right| \\
&= \left|\mathbf{x}^T\left(\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) - \frac{n-1}{n-2}\left(\tilde{I} - \frac{1}{n-1}\tilde{\mathbf{1}}\tilde{\mathbf{1}}^T\right)\right)\mathbf{x}\right| \\
&= \left|\mathbf{x}^T\mathbf{x} - \frac{1}{n}\mathbf{x}^T\mathbf{1}\mathbf{1}^T\mathbf{x} - \frac{n-1}{n-2}\mathbf{x}^T\tilde{I}\mathbf{x} + \frac{1}{n-2}\mathbf{x}^T\tilde{\mathbf{1}}\tilde{\mathbf{1}}^T\mathbf{x}\right| \\
&= \left|\frac{-1}{n-2}\mathbf{x}^T\tilde{I}\mathbf{x} + x_j^2 - \frac{1}{n}\underbrace{(\mathbf{x}^T\tilde{\mathbf{1}} + \mathbf{x}^T\mathbf{e}_j)(\mathbf{x}^T\tilde{\mathbf{1}} + \mathbf{x}^T\mathbf{e}_j)^T}_{\mathbf{x}^T\tilde{\mathbf{1}}\mathbf{e}_j^T\mathbf{x} + \mathbf{x}^T\mathbf{e}_j\mathbf{e}_j^T\mathbf{x} + \mathbf{x}^T\tilde{\mathbf{1}}\tilde{\mathbf{1}}^T\mathbf{x} + \mathbf{x}^T\mathbf{e}_j\tilde{\mathbf{1}}^T\mathbf{x}} + \frac{1}{n-2}\mathbf{x}^T\tilde{\mathbf{1}}\tilde{\mathbf{1}}^T\mathbf{x}\right| \\
\overset{x_j=\mathbf{e}_j^T\mathbf{x}}{=} &\left|\frac{n-1}{n}x_j^2 - \frac{2}{n}\mathbf{x}^T\tilde{\mathbf{1}}\cdot x_j + \frac{-1}{n-2}\mathbf{x}^T\tilde{I}\mathbf{x} + \frac{2}{n(n-2)}\mathbf{x}^T\tilde{\mathbf{1}}\tilde{\mathbf{1}}^T\mathbf{x}\right| \qquad (7.6) \\
&= \left|\frac{n}{n-1}(x_j - \bar{x})^2 - s_j^2\right|
\end{aligned}$$

Wir untersuchen, für welchen Wert von x_j die skalierte Differenz gleich 0 ist. Da SC quadratisch in x_j ist, erhalten wir allgemein die beiden folgenden Lösungen:

$$\begin{aligned}
SC(x_j, s^2) &= 0 \\
\Leftrightarrow x_j &= \frac{n}{2(n-1)}\cdot\left(\frac{2}{n}\mathbf{x}^T\tilde{\mathbf{1}} \pm \left(\frac{4}{n^2}(\mathbf{x}^T\tilde{\mathbf{1}})^2\right.\right. \\
&\quad \left.\left. -\frac{4(n-1)}{n}\left(-\frac{1}{n-2}\mathbf{x}^T\tilde{I}\mathbf{x} + \frac{2}{n(n-2)}\mathbf{x}^T\tilde{\mathbf{1}}\tilde{\mathbf{1}}^T\mathbf{x}\right)\right)^{\frac{1}{2}}\right) \\
&= \frac{\mathbf{x}^T\tilde{\mathbf{1}}}{n-1} \pm \left((\mathbf{x}^T\tilde{\mathbf{1}})^2\cdot\frac{-n}{(n-1)^2(n-2)} + \mathbf{x}^T\tilde{I}\mathbf{x}\cdot\frac{n}{(n-1)(n-2)}\right)^{\frac{1}{2}} \\
&= \frac{\mathbf{x}^T\tilde{\mathbf{1}}}{n-1} \pm \sqrt{\frac{n}{n-1}}\cdot s_j. \qquad (7.7)
\end{aligned}$$

Bei der Untersuchung des empirischen Mittelwerts hat sich ergeben, dass Werte von x_j nahe des empirischen Mittelwerts $\bar{x}_j$ eine kleine Differenz ergeben. Setzen wir $x_j = \frac{\mathbf{x}^T\tilde{\mathbf{1}}}{n-1}$ an, ergibt sich als Differenz

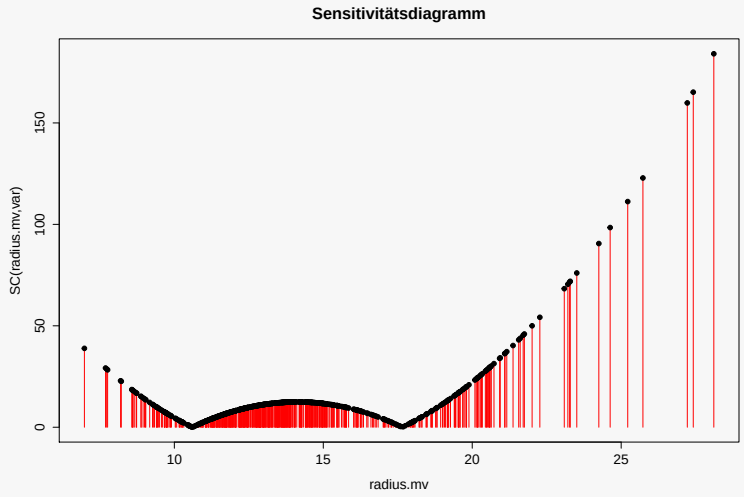
$$\begin{aligned}
SC\left(\frac{\mathbf{x}^T\tilde{\mathbf{1}}}{n-1}, s^2\right) &= \left|\frac{(\mathbf{x}^T\tilde{\mathbf{1}})^2}{n(n-1)} - \frac{2(\mathbf{x}^T\tilde{\mathbf{1}})^2}{n(n-1)} + \frac{-1}{n-2}\mathbf{x}^T\tilde{I}\mathbf{x} + \frac{2}{n(n-2)}\mathbf{x}^T\tilde{\mathbf{1}}\tilde{\mathbf{1}}^T\mathbf{x}\right| \\
&= \left|\frac{1}{n-2}\left(-\mathbf{x}^T\tilde{I}\mathbf{x} + \frac{1}{n-1}(\mathbf{x}^T\tilde{\mathbf{1}})^2\right)\right| \\
&= s_j^2. \qquad (7.8)
\end{aligned}$$

Damit verringert sich die Varianz in diesem Fall zu

$$s^2 = s_j^2 - \frac{1}{n-1}s_j^2 = \frac{n-2}{n-1}s_j^2.$$

Innerhalb des Intervalls $[\bar{x}_j - s_j, \bar{x}_j + s_j]$ verringert sich die Varianz, danach vergrößert sie sich.

Beispiel 7.5




```
x1=radius.mv
mean(x1)
var(x1)
sd(x1)
u=c()
for(i in 1:l)
{   u=append(u,l*abs(mean(v1)-mean(v1[-i])))
}
plot(x1,u,pch=19,type="h",col=2)
```

Im letzten Beispiel ändert sich die empirische Varianz durch Weglassen des größten beobachteten Merkmalswertes relativ um 2.7 Prozent. Um extreme Werte bei Kenngrößen zu berücksichtigen, können die größten oder kleinsten α-Prozent der Werte bei der Berechnung eines Lageparameters weggelassen werden. Dazu müssen wir zunächst die Merkmalswerte einer Stichprobe sortieren. Wir betrachten die Permutation

$$T : \mathbb{R}^n \to \mathbb{R}^n, \quad \mathbf{x} = (x_1, \ldots, x_n)^T \mapsto \left(x_{(1)}, \ldots, x_{(n)}\right)^T := T(x_1, \ldots, x_n)$$
$$\text{mit } x_{(i)} \leq x_{(j)} \text{ für alle } i < j.$$

T heißt Ordnungsstatistik auf $\mathbb{R}^n$. Die Abbildung

$$T_i : \mathbb{R}^n \to \mathbb{R}, \ \mathbf{x} = (x_1, \ldots, x_n) \mapsto T_i(x_1, \ldots, x_n) := T(x_1, \ldots, x_n)_i = x_{(i)}$$

heißt i-te Ordnungsstatistik auf $\mathbb{R}^n$. Das α-Quantil $c_{(\alpha)}$ wird dann für $0 \leq \alpha \leq 1$ durch das

Element mit dem kumulierten Gewicht von mindestens $n \cdot \alpha$ bestimmt durch

$$c_{(\alpha)} : \mathbb{R}^n \to \mathbb{R},\ c_{(\alpha)} := c_{(\alpha)}(\mathbf{x}) := \begin{cases} T_{\lceil n\cdot\alpha \rceil}(\mathbf{x}) = x_{(\lceil n\cdot\alpha \rceil)}, & 0 < \alpha \leq 1, \\ x_{(1)}, & \alpha = 0. \end{cases} \tag{7.9}$$

Das getrimmte Mittel ist nun für $\frac{1}{n} < \alpha < \frac{1}{2}$ definiert durch

$$\bar{x}_\alpha : \mathbb{R}^n \to \mathbb{R},\ \mathbf{x} \mapsto \bar{x}_\alpha := \bar{x}_\alpha(\mathbf{x}) := \frac{1}{n - 2\lfloor \alpha \cdot n \rfloor + 1} \sum_{i=\lfloor \alpha\cdot n \rfloor}^{\lceil (1-\alpha)\cdot n \rceil} x_{(i)} \tag{7.10}$$

Ein weiterer wichtiger robuster Lageparameter ist der Median $\tilde{x}$,

$$\text{med} : \mathbb{R}^n \to \mathbb{R},\ \mathbf{x} \mapsto \tilde{x} := \text{med}(\mathbf{x}) := c_{(0.5)}. \tag{7.11}$$

Der Median ist demnach das 50-Prozent Quantil. Auf Basis der Quantile lassen sich auch weitere Streuparameter bestimmen wie etwa der Interquartilsabstand IQR oder der Median der absoluten Abweichung vom Median MAD,

$$IQR : \mathbb{R}^n \to \mathbb{R},\ \mathbf{x} \mapsto \text{IQR} := IQR(\mathbf{x}) := c_{(0.75)} - c_{(0.25)}, \tag{7.12}$$

$$MAD : \mathbb{R}^n \to \mathbb{R},\ \mathbf{x} \mapsto \text{MAD} := MAD(\mathbf{x}) := \text{med}(|x_1 - \tilde{x}|, \ldots, |x_n - \tilde{x}|). \tag{7.13}$$

Der untere Whisker errechnet sich über $W_u := c_{(0.25)} - 1.5 \cdot \text{IQR}$, der obere Whisker über $W_o := c_{(0.75)} + 1.5 \cdot \text{IQR}$. Mit Hilfe der Ausreißer-Regel von Tukey lassen sich ebenfalls mögliche Ausreißer entdecken. Dabei werden solche x_i gesucht, für die

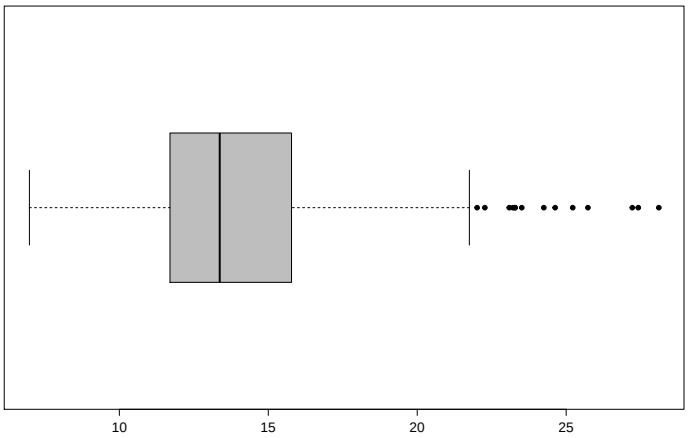
$$x_i < W_u \text{ bzw. } x_i > W_o$$

gilt. Mit Hilfe des Medians, der Whisker und der für die Whisker verwendeten Quantile lässt sich ein Merkmal übersichtlich graphisch darstellen.

```
fivenum(x1)
```

Boxplot

Ein Boxplot hilft bei der Entdeckung potentieller Ausreißer. Dabei wird nicht jeder beobachtete Merkmalswert einzeln gezeichnet, sondern die Werte werden zu Boxen zusammengefasst und der Median wird gekennzeichnet. In der Abbildung ist das als Querstrich zu erkennen. Sämtliche Werte, die sich zwischen oberem und unterem Quartil $(c_{(0.75)}, c_{(0.25)})$ befinden, werden in der inneren Box um den Median zusammengefasst. Die schwarzen Linien umfassen alle Werte vom oberen Quartil bis zum oberen Whisker bzw. vom unteren Quartil zum unteren Whisker. Diejenigen Werte, die noch nicht erfasst sind, werden außerhalb einzeln dargestellt. Sie lassen sich als Ausreißer interpretieren. Die obere Graphik zeigt einen Boxplot für das Merkmal aus dem Beispiel 7.2 und deutet mehrere mögliche Ausreißer hin.




```
boxplot(radius.mv, horizontal=TRUE, col="gray",pch=19)
```

Histogramm

Wir unterteilen den Bereich zwischen dem kleinsten und dem größten Wert der Stichprobe in m nicht notwendigerweise gleichbreite Intervalle $[b_i, b_{i+1}[$, $i = 1, \ldots, m$, mit $b_1 \leq \min\{x_1, \ldots, x_n\}$ und $b_{m+1} > \max\{x_1, \ldots, x_n\}$. Jedes Intervall $[b_i, b_{i+1}[$ enthält eine bestimmte Anzahl n_i an Merkmalswerten. Um die Anzahl auszuzählen, benötigen wir für eine beliebige Menge L eine Indikatorfunktion gemäß (6.1). Es ist $n_i = \sum_{j=1}^{n} I^{[b_i, b_{i+1}[}(x_j)$.
In einem Histogramm werden die Intervalle auf einer Achse angetragen. Jedem Intervall $[b_i, b_{i+1}[$ wird entsprechend der relativen Häufigkeit $h_i = \frac{n_i}{n}$ der enthaltenen Merkmalswerte eine Höhe k_i so zugeordnet, dass für die Fläche $(b_{i+1} - b_i) \cdot k_i = h_i$ gilt. Fassen wir ein Histogramm als Graph der (reellen) Funktion

$$h : \mathbb{R} \to \mathbb{R},\ b \mapsto h(b) := \begin{cases} \frac{h_i}{b_{i+1} - b_i}, & b \in [b_i, b_{i+1}[, \\ 0, & \text{sonst,} \end{cases}$$

auf, so gilt

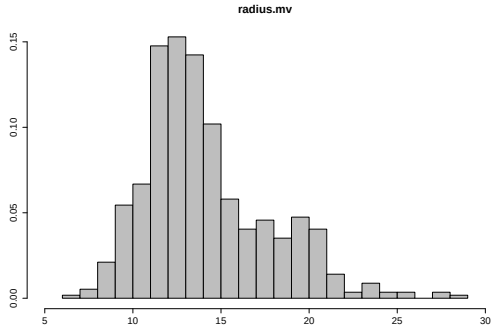
$$\int_{-\infty}^{\infty} h(x)dx = \sum_{i=1}^{m} (b_{i+1} - b_i) \cdot \frac{h_i}{b_{i+1} - b_i} = \sum_{i=1}^{m} \frac{n_i}{n} = 1.$$

h ist stets nicht-negativ und damit wird h zu einer Dichtefunktion. Wir können ein Histogramm als Schätzer der Dichte interpretieren.

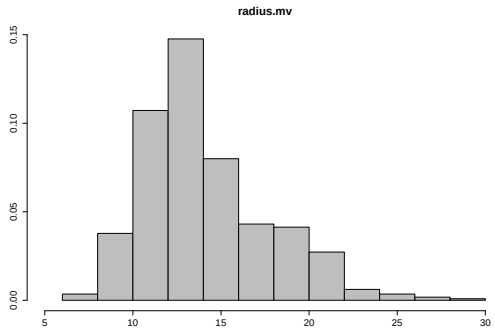
Es gibt die Faustregel nach Freedman und Diaconis, dass die Breite der Intervalle (Binbreite) eines Histogramms gleich $\frac{2 \cdot \text{IQR}}{\sqrt[3]{n}}$ sein soll. Dies ist aber lediglich eine Richtschnur.

```
hist(radius.mv, freq=FALSE, right=FALSE, xlab="", ylab="",
     main="radius.mv", col="gray", xlim=c(5,30), breaks="FD")
hist(radius.mv, freq=FALSE, right=FALSE, xlab="", ylab="",
     main="radius.mv", col="gray", xlim=c(5,30))
hist(radius.mv, freq=FALSE, right=FALSE, xlab="", ylab="",
     main="radius.mv", col="gray", xlim=c(5,30), breaks=seq(4,32,4))
```

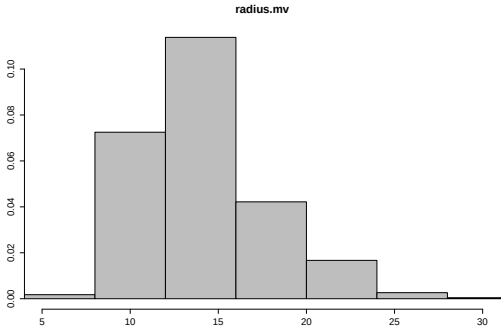
Die folgende Abbildung zeigt ein Histogramm für das Merkmal radius.mv aus Beispiel 7.2.



Ein weiterer wichtiger Aspekt ist die Festlegung der linken Intervallgrenze b_1. Standardmäßig wird der kleinste Merkmalswert genommen. Dies kann jedoch zu Fehlinterpretationen führen. Zur Exploration sollten mehrere Intervallzahlen und verschiedene Startwerte für b_1 gewählt werden.



Es erweist sich als Vorteil, wenn die Intervallbreite identisch gewählt wird. Dennoch ist bei der Interpretation und der Wahl der Intervallbreite Vorsicht geboten. Oft wird ein Ergebnis suggeriert, das lediglich auf einer unglücklichen Wahl der Intervallbreite zurückzuführen ist. Dies tritt insbesondere bei der Untersuchung von Lücken in Erscheinung.



Eine Merkregel ist, dass eine große Zahl an Intervallen eine gleichmäßige Verteilung erzeugt, bei einer geringen Intervallzahl Details verschluckt werden.
Eine weitere Interpretationsmöglichkeit besteht in der Beurteilung der Schiefe der Verteilung. Vergleichen wir den empirischen Mittelwert mit dem Median, so können wir Aussagen bezüglich der Schiefe der empirischen Verteilung treffen, wie folgende Übersicht zeigt.

Schiefe der Verteilung

Vergleich	Interpretation
$\bar{x} = \tilde{x}$	Symmetrische empirische Verteilung
$\bar{x} > \tilde{x}$	Rechtsschiefe empirische Verteilung
$\bar{x} < \tilde{x}$	Linksschiefe empirische Verteilung

Hier liegt eine rechtsschiefe empirische Verteilung vor, was sich durch Rechnung belegen lässt: $\bar{x} = 14.13 > 13.37 = \tilde{x}$.

Kern-Dichteschätzung

Wir haben für das Histogramm die Eigenschaft einer Dichte nachgewiesen. Allerdings sind für die Höhe des Histogramms lediglich Ausprägungen im jeweiligen Intervall entscheidend. Um das zu vermeiden, kann ein so genanntes gleitendes Histogramm verwendet werden. Für ein beliebiges $b \in \mathbb{R}$ approximieren wir die Dichte mittels $\tilde{h} : \mathbb{R} \to \mathbb{R}$,

$$\tilde{h}(b) := \frac{\frac{1}{n} \cdot \sum_{i=1}^{n} I^{[b-h,b+h[}(x_i)}{2h}.$$

Seien $x_1, \ldots, x_n$ die Merkmalswerte. Wir notieren die Funktion $\tilde{h}$ in der Form

$$\tilde{h}(b) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{b - x_i}{h}\right), \quad K\left(\frac{b - x_i}{h}\right) := \begin{cases} \frac{1}{2}, & x_i - h < b \leq x_i + h, \\ 0, & \text{sonst.} \end{cases}$$

Dann gilt

$$\int_{-\infty}^{\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{b - x_i}{h}\right) db = \frac{1}{nh} \sum_{i=1}^{n} \int_{-\infty}^{\infty} K(x) \cdot h dx = \frac{1}{nh} \cdot nh = 1.$$

Die bei der Integration durchgeführte Substitution $x := \frac{b-x_i}{h}$ liefert eine Funktion

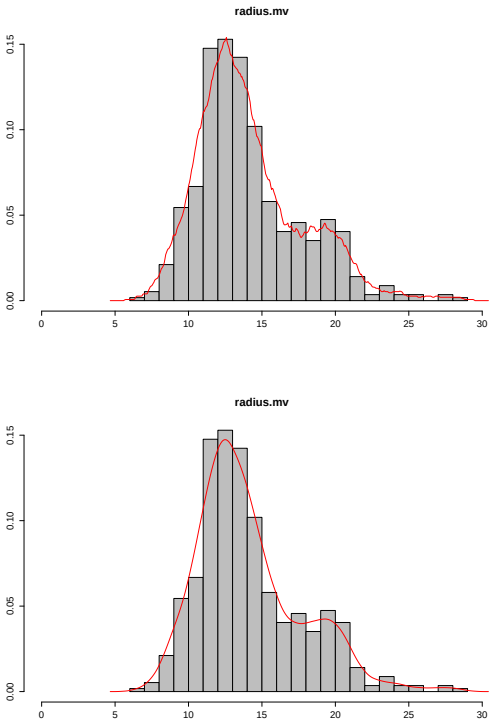
$$K(x) = \begin{cases} \frac{1}{2}, & -1 < x \leq 1, \\ 0, & \text{sonst,} \end{cases}$$

die zwischen -1 und 1 nur positive Werte annimmt und den Integralwert 1 hat. Wir nennen eine solche Funktion eine Kernfunktion und die Dichte $\tilde{h}(b)$ einen Kerndichteschätzer. Die hier benutzte Rechteckskernfunktion führt zu einer nichtstetigen Dichtefunktion.

Durch eine andere Wahl der Kernfunktion können stetige Dichten erzeugt werden. Hierzu müssen die Kernfunktionen stetig sein. Die additive Überlagerung führt dann zu einer stetigen Funktion. Als Beispiel sei die Bisquare-Kernfunktion

$$K(x) := \begin{cases} \frac{15}{16}(1-x^2)^2, & -1 < x \leq 1, \\ 0, & \text{sonst,} \end{cases}$$

vorgestellt. In der Abbildung sind die Rechtecks- und Bisquare-Kernfunktion zu sehen. Betrachten wir wieder die Daten aus Beispiel 7.2 und schätzen die Dichte von radius.mv mit Hilfe der Rechtecks- bzw. der Bisquare-Kernfunktion. Problematisch bei der Kerndichteschätzung ist ähnlich zu den Histogrammen die Wahl der Bandbreite h. Es gibt verschiedene Ansätze zur optimalen Bandbreitenwahl, jedoch sind diese jeweils nicht unproblematisch. Je größer die Bandbreite, desto glatter wird die geschätzte Dichtefunktion.




```
hist(radius.mv, freq=FALSE, right=FALSE, xlab="", ylab="",
     main="radius.mv", col="gray", xlim=c(0,30), breaks="FD")
lines(density(radius.mv, kernel=c("rectangular")), col=2)
lines(density(radius.mv, kernel=c("biweight")), col=2)
```

7.2. Kenngrößen für mehrere Merkmale

Die Lageparameter lassen sich grundsätzlich auf den zwei- oder mehrdimensionalen Fall übertragen[1]. Seien k Merkmale $X_1, \ldots, X_k$ und eine Datenmatrix $X \in \mathbb{R}^{n,k}$ gegeben.

[1] vgl. auch [8]

Dann werden der Schwerpunkt $\bar{\mathbf{x}}$ und der Medianpunkt $\tilde{\mathbf{x}}$ definiert als

$$\bar{\mathbf{x}} : \mathbb{R}^{n,k} \to \mathbb{R}^k, \qquad X \mapsto \bar{\mathbf{x}} := \bar{\mathbf{x}}(X) := \frac{1}{n}\mathbf{1}^T X = (\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_k), \tag{7.14}$$

$$\tilde{\mathbf{x}} : \mathbb{R}^{n,k} \to \mathbb{R}^k, \qquad X \mapsto \tilde{\mathbf{x}} := \tilde{\mathbf{x}}(X) := (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_k). \tag{7.15}$$

Für weitere Betrachtungen sind die Vektoren als Zeilenvektoren geschrieben. Nun lässt sich die Datenmatrix X mit Hilfe des Schwerpunkts zentrieren,

$$\begin{pmatrix} x_{11} - \bar{\mathbf{x}}_1 & \ldots & x_{1k} - \bar{\mathbf{x}}_k \\ x_{21} - \bar{\mathbf{x}}_1 & \ldots & x_{2k} - \bar{\mathbf{x}}_k \\ \vdots & & \vdots \\ x_{n1} - \bar{\mathbf{x}}_1 & \ldots & x_{nk} - \bar{\mathbf{x}}_k \end{pmatrix} = X - \mathbf{1}\bar{\mathbf{x}} = X - \mathbf{1}\frac{1}{n}\mathbf{1}^T X = \underbrace{\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)}_{H} X.$$

$H \in \mathbb{R}^{n,n}$ wird Zentrierungsmatrix genannt und ist identisch zur Matrix H nach Beispiel 7.4.

Definition 7.6: Orthogonalprojektion

Eine lineare Abbildung $\mathbb{R}^n \to \mathbb{R}^n$, die durch die Matrix $M \in \mathbb{R}^{n,n}$ repräsentiert wird, heißt Orthogonalprojektion auf den Unterraum $U \subseteq \mathbb{R}^n$, falls

$$M\mathbf{x} \in U \qquad \text{für alle } \mathbf{x} \in \mathbb{R}^n, \tag{7.16}$$

$$(\mathbf{x} - M\mathbf{x})^T\mathbf{u} = 0 \qquad \text{für alle } \mathbf{u} \in U,\ \mathbf{x} \in \mathbb{R}^n. \tag{7.17}$$

Satz 7.7

Die Zentrierungsmatrix H ist eine Orthogonalprojektion auf den Unterraum

$$U = \{\mathbf{x} \in \mathbb{R}^n;\ \mathbf{1}^T\mathbf{x} = 0\}.$$

Beweis.

Zunächst gilt

$$\begin{aligned} H^T &= (I - \tfrac{1}{n}\mathbf{1}\mathbf{1}^T)^T = (I - \tfrac{1}{n}\mathbf{1}\mathbf{1}^T) = H, \\ H^T H &= H^2 = (I - \tfrac{1}{n}\mathbf{1}\mathbf{1}^T)(I - \tfrac{1}{n}\mathbf{1}\mathbf{1}^T) = I - \tfrac{2}{n}\mathbf{1}\mathbf{1}^T + \tfrac{1}{n^2}n\mathbf{1}\mathbf{1}^T = H. \end{aligned}$$

Damit ist H erstens symmetrisch. Zweitens ist H idempotent und so eine Projektion. Um den Unterraum U bestimmen zu können, untersuchen wir die Eigenwerte und Eigenräume von H.

Wir behaupten, dass H die Eigenwerte 0 und 1 besitzt. Das gilt wegen $(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{1} = \mathbf{0}$ und $(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T)(1, -1, 0, \ldots, 0)^T = (1, -1, 0, \ldots, 0)^T$. Wir bestimmen für $\lambda = 0$ bzw.

$\lambda = 1$ die Lösungen der Gleichungen $H\mathbf{v} = \lambda\mathbf{v}$. Zunächst für $\lambda = 0$:

$$H\mathbf{v} = \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\mathbf{v} = \mathbf{v} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{v} \stackrel{!}{=} \mathbf{0} \Leftrightarrow v_i = \bar{\mathbf{v}} \text{ für alle } i = 1, \ldots, n, v_i \neq 0.$$

Damit ist $\lambda = 0$ Eigenwert von H zum eindimensionalen Eigenraum $E_0 = \{\mathbf{v} \in \mathbb{R}^n;\ v_i = v_j \text{ für alle } i, j\}$. Weiter ist für $\lambda = 1$:

$$H\mathbf{v} = \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\mathbf{v} = \mathbf{v} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{v} \stackrel{!}{=} \mathbf{v} \Leftrightarrow \mathbf{1}\mathbf{1}^T\mathbf{v} = \mathbf{0} \text{ und } \mathbf{v} \neq \mathbf{0}.$$

Damit ist $\lambda = 1$ Eigenwert von H zum $n-1$-dimensionalen Eigenraum $E_1 = \{\mathbf{v} \in \mathbb{R}^n;\ \mathbf{1}^T\mathbf{v} = 0\}$.

Da zudem für beliebiges $\mathbf{x} \in \mathbb{R}^n$ und $\mathbf{v} \in E_1$ dann $(\mathbf{x} - H\mathbf{x})^T\mathbf{v} = \left(\frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{x}\right)^T\mathbf{v} = \frac{1}{n}\mathbf{x}^T\mathbf{1}\mathbf{1}^T\mathbf{v} = 0$ ist, wird H zu einer Orthogonalprojektion.textbf

□

Mit den Überlegungen lässt sich die empirische Varianz eines Merkmals schreiben als

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}||H\mathbf{x}||^2 = \frac{1}{n-1}\mathbf{x}^T H\mathbf{x}. \tag{7.18}$$

Dabei ist $||.||$ die euklidische Norm, d.h. für $\mathbf{x} \in \mathbb{R}^n$ ist

$$||\mathbf{x}|| = \sqrt{\mathbf{x}^T\mathbf{x}} = \sqrt{x_1^2 + \ldots + x_n^2}.$$

Assoziationsmaße für zwei Merkmale

Sind zwei Merkmale gegeben, kann die empirische Varianz jeweils als Streuparameter herangezogen werden. Darüber hinaus interessiert aber auch ein gemeinsames Streuverhalten und damit verbunden auch ein gemeinsames Verhalten der Merkmale. Wir suchen eine Kenngröße, die uns etwas über den Zusammenhang der beiden Merkmale aussagen kann.

Wir bestimmen ein Assoziationsmaß für zwei quantitative Merkmale X_1 und X_2 mit jeweils n Realisierungen. Für ein quantitatives Merkmal lassen sich jeweils die empirischen Varianzen $s^2_{X_1}$ bzw. $s^2_{X_2}$ bestimmen. Seien $\bar{x}_1$ und $\bar{x}_2$ die beiden empirischen Mittelwerte und x_{ij} mit $i \in \{1, \ldots, n\}$ und $j \in \{1, 2\}$ die Realisierungen. Für das Produkt der Abstände zu den Mittelwerten gilt

$$\begin{aligned}
(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) > 0 \quad &\Leftrightarrow \quad \begin{cases} \text{(I)} & (x_{i1} - \bar{x}_1) > 0 \wedge (x_{i2} - \bar{x}_2) > 0 \text{ oder} \\ \text{(III)} & (x_{i1} - \bar{x}_1) < 0 \wedge (x_{i2} - \bar{x}_2) < 0 \end{cases} \\
(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) < 0 \quad &\Leftrightarrow \quad \begin{cases} \text{(II)} & (x_{i1} - \bar{x}_1) > 0 \wedge (x_{i2} - \bar{x}_2) < 0 \text{ oder} \\ \text{(IV)} & (x_{i1} - \bar{x}_1) < 0 \wedge (x_{i2} - \bar{x}_2) > 0 \end{cases}
\end{aligned} \tag{7.19}$$

Dieser Sachverhalt kann graphisch dargestellt werden durch die zweidimensionale Punktmenge $\{(x_{i1}, x_{i2}) | 1 \leq i \leq n\} \subset \mathbb{R}^2$. Wir unterteilen das x-y-Koordinatensystem durch

die Geraden $y = \bar{x}_2$ und $x = \bar{x}_1$ in vier Bereiche (I) bis (IV). Jede Realisierung (x_{i1}, x_{i2}) liegt je nach Zusammenhang in (7.19) im entsprechenden Bereich. Werden nun sämtliche Abstandsprodukte addiert und ergibt sich ein deutlich positiver Wert, so müssen die Realisierungen überwiegend in den Bereichen (I) und (III) liegen. Ist die Summe deutlich negativ, müssen die Realisierungen überwiegend in den Bereichen (II) und (IV) liegen. Dementsprechend liegt ein positiver bzw. negativer Zusammenhang zwischen X_1 und X_2 vor. Wir halten fest

Definition 7.8: Empirische Kovarianz

Seien X_1 und X_2 quantitative Merkmale mit jeweils n Realisierungen und den arithmetischen Mittelwerten $\bar{x}_1$ bzw. $\bar{x}_2$. Dann heißt der Schätzwert der Schätzfunktion $s^2_{X_1X_2} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$,

$$(\mathbf{x}_1, \mathbf{x}_2) \mapsto s^2_{X_1X_2} := s^2_{X_1X_2}(\mathbf{x}_1, \mathbf{x}_2) := \frac{1}{n-1}\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)$$

empirische Kovarianz zwischen den Merkmalen X_1 und X_2.

Bemerkung.
(I) Für $X_1 = X_2$ erhalten wir die empirische Varianz s^2 von X_1.
(II) Offensichtlich gilt $s^2_{X_1X_2} = s^2_{X_2X_1}$.

Fassen wir für k Merkmale X_1 bis X_k die paarweisen empirischen Kovarianzen in einer Matrix zusammen, erhalten wir die symmetrische empirische Varianz-Kovarianzmatrix $S_X = \left(s^2_{X_iX_j}\right) \in \mathbb{R}^{k,k}$. Sie lässt sich bestimmen über

$$S_X = \frac{1}{n-1}(HX)^T HX = \frac{1}{n-1}X^T H^T HX = \frac{1}{n-1}X^T HX.$$

Doch offen bleibt die Frage, was ein deutlich positiver Wert der Kovarianz zweier Merkmale ist. Denn die Summe hängt stark von den Realisierungen ab. Werden nämlich sämtliche Realisierungen von X_1 mit einem Faktor multipliziert, erhöht sich auch die Kovarianz um diesen Faktor. Entsprechendes gilt für das zweite Merkmal und für beide zusammen. Mit dem Faktor u für das erste und dem Faktor v für das zweite Merkmal ergibt sich

$$\frac{1}{n-1}\sum_{i=1}^{n}(u \cdot x_{i1} - u \cdot \bar{x}_1)(v \cdot x_{i2} - v \cdot \bar{x}_2) = \frac{u \cdot v}{n-1}\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)$$

Diesen Nachteil können wir durch eine Normierung mit dem Produkt der empirischen Standardabweichungen beider Merkmale ausgleichen. Damit kann für je zwei kardinal skalierte Merkmale ein vergleichbarer Wert für den Zusammenhang bestimmt werden. Wir erhalten folgende Definition.

Definition 7.9: Empirischer Korrelationskoeffizient nach Bravais-Pearson

Für zwei Merkmale X_1 und X_2 mit jeweils n Realisierungen und den empirischen Mittelwerten $\bar{x}_1$ bzw. $\bar{x}_2$ heißt der Schätzwert der Schätzfunktion $\rho_{X_1X_2} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$,

$$\rho_{X_1X_2} := \rho_{X_1X_2}(\mathbf{x}_1, \mathbf{x}_2) := \frac{s^2_{X_1X_2}}{s_{X_1} s_{X_2}} = \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2} \cdot \sqrt{\sum_{i=1}^{n}(x_{i2} - \bar{x}_2)^2}} \tag{7.20}$$

der empirische Korrelationskoeffizient der Merkmale X_1 und X_2.

Bemerkung.
Es gilt: $-1 \leq \rho_{X_1X_2} \leq 1$, da mit $\mathbf{y}_1 = \mathbf{x}_1 - \bar{x}_1\mathbf{1} = H\mathbf{x}_1$ und $\mathbf{y}_2 = \mathbf{x}_2 - \bar{x}_2\mathbf{1} = H\mathbf{x}_2$

$$\rho_{X_1X_2} = \frac{\mathbf{y}_1^T\mathbf{y}_2}{||\mathbf{y}_1|| \cdot ||\mathbf{y}_2||} = \cos(\phi) \in [-1, 1] \tag{7.21}$$

für einen Winkel ϕ ist.

Alle Korrelationen in einer Matrix zusammengefasst erhalten wir die empirische Korrelationsmatrix $R_X = (\rho_{X_iX_j}) \in \mathbb{R}^{k,k}$. Auch sie lässt sich in Matrixschreibweise darstellen. Dazu überlegen wir uns, dass mit (7.21) gilt:

$$\rho_{X_iX_j} = \frac{\mathbf{y}_i^T\mathbf{y}_j}{||\mathbf{y}_i|| \cdot ||\mathbf{y}_j||} = \frac{1}{n-1} \cdot \frac{\mathbf{x}_i^T H \mathbf{x}_j}{s_{X_i} s_{X_j}} = \frac{1}{n-1} \cdot \left(\frac{\mathbf{x}_i}{s_{X_i}}\right)^T H \left(\frac{\mathbf{x}_j}{s_{X_j}}\right).$$

Somit ist $R_X = D_X^{-\frac{1}{2}} S_X D_X^{-\frac{1}{2}}$ mit $D_X^{-\frac{1}{2}} = \text{diag}\left((s^2_{X_i})^{-\frac{1}{2}}\right)$.

Beispiel 7.10

Gegeben sei die Datenmatrix

$$X = \begin{pmatrix} 22 & 5 \\ 25 & 10 \\ 21 & 4 \\ 28 & 13 \\ 24 & 8 \end{pmatrix} \in \mathbb{R}^{5,2} \text{ mit } \bar{\mathbf{x}} = (24, 8) \text{ und } \mathbf{s}^2 = (7.5, 13.5).$$

Wir erhalten als empirische Kovarianz $s^2_{X_1X_2} = \frac{40}{4} = 10$ und damit $\rho_{X_1X_2} = \frac{10}{\sqrt{7.5}\sqrt{13.5}} = 0.994$. Es liegt ein fast perfekter positiver Zusammenhang vor.

Beispiel 7.11

Für die Merkmale radius.mv und peri.mv erhalten wir als empirische Kovarianz $s^2_{X_1X_2} = 85.45$ und damit $\rho_{X_1X_2} = 0.998$. Es liegt ein nahezu perfekter positiver Zusammenhang vor.

In R erfolgt die Berechnung über die Befehle cov und cor.

```
cov(x1,x2)
cor(x1,x2)
```

Robustheit

Auch bei mehreren Merkmalen ist darauf zu achten, mögliche Ausreißer zu identifizieren. Ausreißer müssen hier zusätzlich im Zusammenhang aller Merkmale gesehen werden. Um eine solche Betrachtung durchzuführen, erweitern wir die Idee des Sensitivitätsdiagramms auf mehrere Merkmale. In Erweiterung zu (7.4) legen wir die Differenz SC fest über

$$SC(\mathbf{x}_{i\cdot}, \mathbf{m}) := n \cdot ||\mathbf{m}_n - \mathbf{m}_{n(i)}||. \tag{7.22}$$

Ist **m** etwa der Schwerpunkt, so kann als Norm z.B. die euklidische benutzt werden. Ist **m** die empirische Varianz-Kovarianzmatrix, so muss mit einer Matrixnorm gearbeitet werden. Als natürliche Erweiterung der euklidischen Norm kann die Frobeniusnorm einer Matrix $A = (a_{ij}) \in \mathbb{R}^{n,k}$

$$||A|| = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{k} a_{ij}^2}.$$

benutzt werden. Zur Exploration von einflussreichen Datenpunkten untersuchen wir wiederum einen Graphen der Form

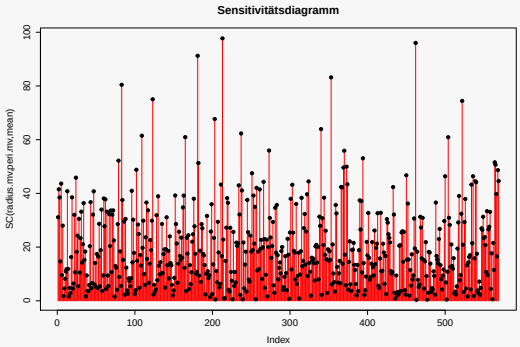
$$\{(x_{i1}, x_{i2}, SC(\mathbf{x}_{i\cdot}, \mathbf{m})); \ i = 1, \ldots, n\}$$

oder

$$\{(i, SC(\mathbf{x}_{i\cdot}, \mathbf{m})); \ i = 1, \ldots, n\}.$$

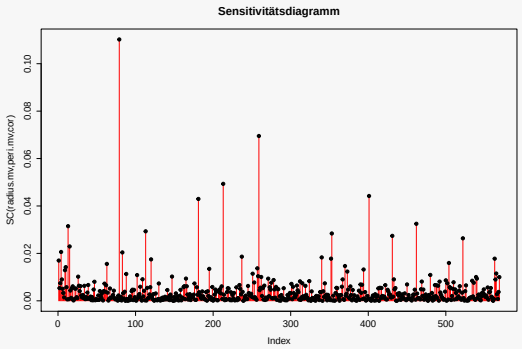
Beispiel 7.12

Gegeben seien die Merkmale radius.mv und peri.mv. Das Sensitivitätsdiagramm zum Schwerpunkt



zeigt ein paar Kandidaten für Ausreißer. Betrachten wir das Sensitivitätsdiagramm zur

empirischen Varianz-Kovarianzmatrix,

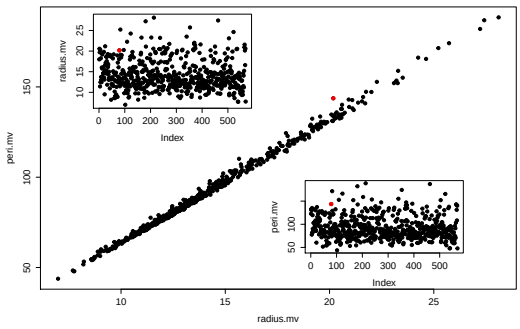


so ist ein Wert besonders auffällig, der auch zu den Kandidaten im ersten Diagramm zählt.

```
u=c()
for(i in 1:l)
{
        u=append(u,l*abs(sqrt(sum(
            (c(mean(radius.mv),mean(peri.mv))
            -c(mean(radius.mv[-i]),mean(peri.mv[-i])))^2)
            )))
}
plot(u,pch=19,main="Sensitivitaetsdiagramm",type="h",
    ylab="SC(radius.mv,peri.mv,mean)",xlab="Index",col=2)
points(u,pch=19)
u=c()
for(i in 1:l)
{
        u=append(u,l*abs( sqrt(2*(cor(radius.mv,peri.mv)
                -cor(radius.mv[-i],peri.mv[-i]))^2) ))
}
plot(u,pch=19,main="Sensitivitaetsdiagramm",type="h",
    ylab="SC(radius.mvmperi.mv,mean)",xlab="Index",col=2)
points(u,pch=19)
```

Scatterplot

Im letzten Beispiel haben wir einen Merkmalsträger identifiziert, der Einfluss auf die empirische Varianz-Kovarianzmatrix und den empirischen Mittelwert nimmt. Um ihn besser beurteilen zu können, betrachten wir einen so genannten Scatterplot der beiden Merkmale radius.mv und peri.mv.



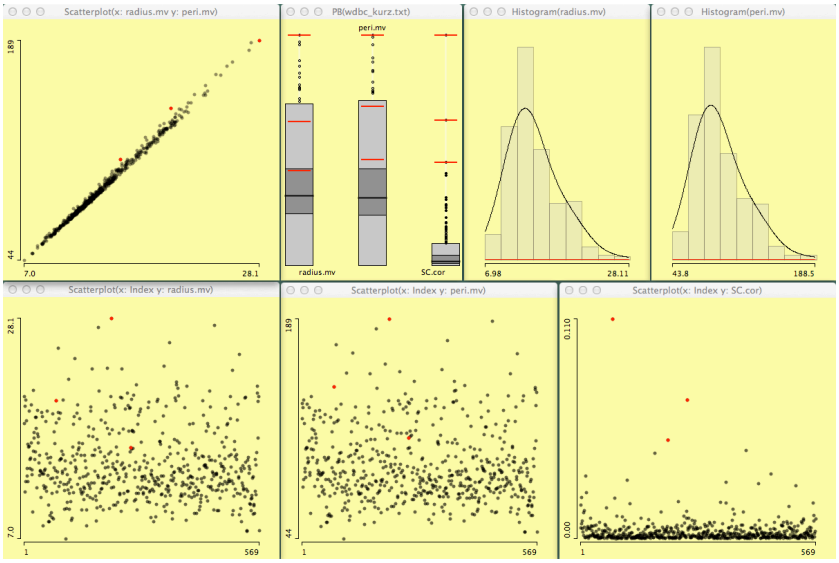

```
plot(radius.mv, peri.mv, pch=19)
```

Dabei werden die Merkmalswertepaare $\{(x_{i1}, x_{i2});\ i = 1, \ldots, n\}$ in ein Koordinatensystem angetragen. Grundsätzlich lässt sich hier der positiv lineare Zusammenhang erkennen. Es ist gedanklich möglich, eine Gerade durch die Daten zu legen. Die Modellierung und Interpretation eines solchen Zusammenhangs erfolgt im Kapitel **??** über die Regressionsanalyse. Der mögliche Ausreißer ist rot gekennzeichnet. Klein sind Scatterplots Index gegen das jeweilige Merkmal gezeichnet. In keinem der drei Scatterplots würde der rote Merkmalsträger als Ausreißer identifiziert werden.

Interaktivität in statistischer Graphik


Tukey
1915-2000

Die bisherigen graphischen Darstellungen haben wir separat voneinander betrachtet. Es wäre aber sicher interessant, die beiden erkannten „Ausreißer“ im Sensitivitätsdiagramm aus Beispiel 7.12 mit den einzelnen Boxplots oder dem Scatterplot zu vergleichen. Dies ist mit Hilfe interaktiver Elemente in statistischer Software möglich, wie es von Tukey in [11] erstmals gefordert wurde. Besonders wichtig dabei ist die Technik des gelinkten Highlighting. Dabei werden alle Graphiken eines Datensatzes verbunden, so dass sich Änderungen, Auswahlen oder Anpassungen in einer Graphik oder dem zugrunde liegenden Datensatz auf alle Graphiken übertragen werden (Linking). Eine auf den Merkmalsträger bezogene Teilauswahl der Daten (Selektion) wird durch eine entsprechende farbliche Darstellung der ausgewählten Objekte (Highlighting) illustriert. Durch die Kombination von Linking, Selektion und Highlighting wird eine Selektion in einer Graphik auf alle anderen Graphiken exakt übertragen. Ein weiterer Aspekt ist die Versorgung mit zusätzlichen Informationen auf Wunsch (Abfrage). Im Beispiel selektieren wir die beiden „Ausreißer“ und erfahren, wo sich diese in den Boxplots bzw im Scatterplot wiederfinden.



Selektieren wir die drei größten Werte im Sensitivitätsdiagramm für die empirische Korrelation, können wir durch Highlicghting in allen anderen Plots die entsprechenden Merkmalsträger betrachten. Die beiden größten Werte werden nicht als univariate (für ein Merkmal) Ausreißer erkannt. Lediglich der dritte Wert ist der größte Werte für beide Merkmale.

Ein Histogramm, in dem für ein Intervall keine Werte vorliegen, wird ein Rechteck der Höhe 0 gezeichnet. Das sollte durch einen roten Querstrich gekennzeichnet werden, es erfolgt eine Warnung, eine durch das Tool automatisch erzeugte Darstellung eines besonderen Sachverhalts.

Monotone Zusammenhänge

Der Einfluss eines Merkmalsträgers auf die Kovarianz bzw. Korrelation kann beachtlich groß sein. Die Schätzwerte für den linearen Zusammenhang sind also sehr ausreißeranfällig. Wir können stattdessen bestimmen, wie gut eine monotone Funktion den Zusammenhang beschreiben kann. Vorausgesetzt wird lediglich eine monotone Folge von Werten, auf denen eine Metrik definiert werden kann. Da diese Überlegung auch für komparative Merkmale gilt, betrachten wir kurz Eigenschaften komparativer Merkmale.

Realisierungen eines komparativen Merkmals können der Größe nach geordnet und in eine Rangfolge gebracht werden. Damit eröffnet sich eine erste Beschreibungsmöglichkeit für die Daten, indem die relativen Häufigkeiten aufsummiert werden. Sei B gemäß (5.1) die Menge der beobachteten Ausprägungen, eine Teilmenge des Merkmalsraums M. Wir erhalten

Definition 7.13: Empirische Verteilungsfunktion

Sei X ein komparatives Merkmal mit Merkmalsraum M und m beobachteten Ausprägungen $c_1, \ldots, c_m \in B \subseteq M$, wobei $c_i \preceq c_j$ für $i < j$ gelte. Seien weiter durch $h(c_i) = h_i,\ i = 1, \ldots, m$, die relativen Häufigkeiten einer beobachteten Anzahl von

Realisierungen gegeben. Dann heißt die Abbildung $H : M \to [0,1]$ mit

$$H(c) := \sum_{c_i \preceq c} h(c_i)$$

empirische Verteilungsfunktion des Merkmals X.

Bemerkung.
(I) Die empirische Verteilungsfunktion ist eine rechtsseitig stetige Treppenfunktion.
(II) Definition 7.13 schreibt keineswegs $m = |M|$ vor und damit auch nicht $c = c_l$ für ein $l \in \{1, \ldots, m\}$.

Jede beobachtete Merkmalsausprägung eines Merkmals X erhält einen so genannten Rang $r : B \to \mathbb{R}$

$$c_l \mapsto r(c_l) := \begin{cases} \sum_{i=1}^{l-1} n_i + \frac{n_l(n_l+1)}{2n_l} = n \cdot H(c_{l-1}) + \frac{n_l+1}{2}, & l \neq 1, \\ \frac{n_1+1}{2}, & l = 1, \end{cases}$$

wobei $n_l = n(c_l)$ die absoluten Häufigkeiten und $H(c_l)$ die empirische Verteilungsfunktion des Merkmals X ist. Der letzte Summand ist notwendig, um so genannte Bindungen zu berücksichtigen. Eine Bindung entsteht, wenn mehr als ein Merkmalsträger den gleichen Merkmalswert eines Merkmals X besitzt. Damit erhalten die sortierten Merkmalsträger einer Stichprobenmenge S den Rang

$$r_X : S \to \mathbb{R},\ s_{(i)\cdot} \mapsto R_X(s_{(i)\cdot}) := r\left(x_{(i)}\right).$$

Der „Rangdurchschnitt“ $\bar{r}_X$ eines Merkmals X mit n Realisierungen und m beobachteten Merkmalsausprägungen ist $\bar{r}_X = \frac{n+1}{2}$.

$$\begin{aligned} \bar{r}_X &= \frac{1}{n}\sum_{i=1}^{n} r_X(\tilde{\omega}_{i\cdot}) = \frac{1}{n}\sum_{i=1}^{n} r(X(\tilde{\omega}_{i\cdot})) \\ &= \frac{1}{n}\sum_{l=1}^{m} n_l r(c_l) = \frac{1}{n}\sum_{l=1}^{m} n_l \left(\sum_{i=1}^{l-1} n_i + \frac{n_l+1}{2}\right) \\ &= \frac{1}{n}\sum_{l=1}^{m} \frac{1}{2}\left(n_l^2 + n_l + 2\sum_{i=1}^{l-1} n_l n_i\right) \\ &= \frac{1}{2n}\left(\sum_{l=1}^{m} n_l^2 + \sum_{l=1}^{m} n_l + 2\sum_{l=1}^{m}\sum_{i=1}^{l-1} n_l n_i\right) \\ &= \frac{1}{2n}\left(\sum_{l=1}^{m} n_l + \left(\sum_{l=1}^{m} n_l\right)^2\right) \\ &= \frac{1}{2n}\left(n + n^2\right) \\ &= \frac{n+1}{2} \end{aligned} \tag{7.23}$$

Die gemittelten quadrierten Abstände der Ränge vom Rangdurchschnitt ergeben die Rang-

varianz und diese errechnet sich über

$$\sigma^2_{r_X} = \frac{1}{n-1}\sum_{i=1}^{n}(r_X(s_{(i)\cdot}) - \bar{r}_X)^2.$$

Mit diesen Voraussetzungen können wir ein Assoziationsmaß festlegen.

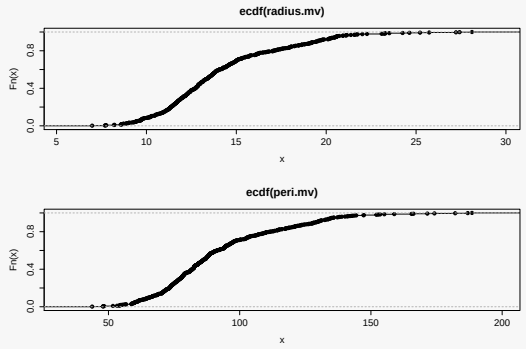
Definition 7.14: Korrelationskoeffizient nach Spearman

Der Spearmansche Rangkorrelationskoeffizient zwischen zwei komparativen Merkmalen X_1 und X_2 ist gegeben durch

$$\rho_{X_1,X_2} := \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{r_{X_1}(s_{(i)\cdot}) - \bar{r}_{X_1}}{\sqrt{\sigma^2_{r_{X_1}}}} \cdot \frac{r_{X_2}(s_{(i)\cdot}) - \bar{r}_{X_2}}{\sqrt{\sigma^2_{r_{X_2}}}}\right). \tag{7.24}$$

Beispiel 7.15

Die empirischen Verteilungsfunktionen bei den Merkmalen radius.mv und peri.mv bei den Brustkrebsdaten sehen folgendermaßen aus:



Der Rangdurchschnitt ist $\bar{R}_{X_1} = \bar{R}_{X_2} = 285$. Es gibt bei beiden Merkmalen einzelne Bindungen. Die Rangvarianzen ergeben sich zu $\sigma^2_{R_{X_1}} = 27027.37$, $\sigma^2_{R_{X_2}} = 27027.45$ und der Spearmansche Rangkorrelationskoeffizient ist $\rho_{X_1,X_2} = 0.998$. Es besteht eine starke monotone Beziehung zwischen radius.mv und peri.mv.

```
rank(radius.mv)
rank(peri.mv)
var(rank(radius.mv))
var(rank(peri.mv))
cor(rank(radius.mv),rank(peri.mv))
cor(radius.mv,peri.mv,method = c("spearman"))
plot(ecdf(radius.mv))
plot(ecdf(peri.mv))
```

Verallgemeinerte Varianz

Eine Verallgemeinerung der Varianz auf die Situation mit mehr als einem Merkmal führt auf zwei Ansätze: Die Totalvariation s_t^2 und die verallgemeinerte empirische Varianz s_v^2.

$$s_t^2 : \mathbb{R}^{n,k} \to \mathbb{R}, X \mapsto s_t^2(X) := \text{sp}(S) = \sum_{i=1}^{k} s_{X_i X_i}^2, \tag{7.25}$$

$$s_v^2 : \mathbb{R}^{n,k} \to \mathbb{R}, X \mapsto s_v^2(X) := \det(S). \tag{7.26}$$

Beide Werte verdichten die vorhandene Information sehr stark. Die Idee der Totalvarianz hat den Vorteil, dass sie sich bei einer orthogonalen Transformation nicht ändert. Für einen Punkt $\mathbf{x} = (x_1, \ldots, x_k)^T \in \mathbb{R}^k$ gilt $\mathbf{x} = (x_1, \ldots, x_k)^T = \sum_{i=1}^{k} x_i \mathbf{e}_i$ mit der kanonischen Basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$. Der Punkt soll mit Hilfe einer anderen Orthonormalbasis $\{\mathbf{q}_1, \ldots, \mathbf{q}_k\}$ dargestellt werden. Sei Q die Matrix, deren Spalten die Basisvektoren $\mathbf{q}_1, \ldots, \mathbf{q}_k$ sind. Es folgt

$$\sum_{i=1}^{k} x_i \mathbf{e}_i = \sum_{i=1}^{k} y_i \mathbf{q}_i \Leftrightarrow I^{k,k}\mathbf{x} = Q\mathbf{y} \Leftrightarrow \mathbf{y} = Q^T\mathbf{x}.$$

Ist $X \in \mathbb{R}^{n,k}$ eine Datenmatrix, so entsprechen die Zeilen den Objektpositionen (Punkten) im $\mathbb{R}^k$. Für das j-te Objekt gilt $\sum_{i=1}^{k} x_{ji}\mathbf{e}_i = \sum_{i=1}^{k} y_{ji}\mathbf{q}_i$ genau dann, wenn $I^{k,k}\mathbf{x}_{j\cdot}^T = Q\mathbf{y}_{j\cdot}^T$, somit $\mathbf{y}_{j\cdot} = \mathbf{x}_{j\cdot}Q$. Stellen wir alle Zeilen von X in der neuen Basis dar, so erhalten wir $Y = XQ$.

Wie ändert sich die empirische Varianz-Kovarianzmatrix bei einer orthogonalen Transformation der Daten? Mit $S_X = \frac{1}{n-1}X^T H X$ gilt

$$S_Y = \frac{1}{n-1}(XQ)^T H (XQ) = Q^T \frac{1}{n-1} X^T H X Q = Q^T S_X Q.$$

Satz und Definition 7.16: Spur einer Matrix

Sei $A = (a_{ij}) \in \mathbb{R}^{k,k}$. Die Spur von A ist die Summe der Diagonalelemente von A, $\text{sp}(A) := \sum_{i=1}^{k} a_{ii}$. Mit $B = (b_{ij}) \in \mathbb{R}^{k,k}$ gilt $\text{sp}(AB) = \text{sp}(BA)$.

Beweis.

$$\text{sp}(AB) = \sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}b_{ji} = \sum_{j=1}^{k}\sum_{i=1}^{k} b_{ji}a_{ij} = \text{sp}(BA).$$

□

Mit Satz 7.16 folgt

$$\text{sp}(S_Y) = \text{sp}(Q^T S_X Q) = \text{sp}(QQ^T S_X) = \text{sp}(S_X). \tag{7.27}$$

Dies muss z.B. bei einer Hauptkomponentenanalyse benutzt werden. Dafür berücksichtigt die Kenngröße keinerlei Information über Kovarianzen zwischen den Merkmalen. Zwar geschieht dies bei der verallgemeinerten empirischen Varianz, dennoch kann der einzelne Wert keine unterschiedlichen Strukturen in den Kovarianzen erklären.

Standardisierung

Durch Standardisierung lassen sich Daten unterschiedlicher Skalierung miteinander vergleichen. Eine durch die Skalierung bedingte ungleiche Einflussnahme auf datenanalytische Modelle wird dadurch vermieden. Dividieren wir die zentrierten Daten eines Merkmals X durch die Standardabweichung (hier mit Index X), $w_i := \frac{x_i - \bar{x}}{s_X}$, ergibt sich ein neues Merkmal $W = HXD_X^{-\frac{1}{2}}$ mit empirischem Mittelwert von 0 und einer empirischen Varianz von 1:

$$\begin{aligned}
\bar{w} &= \frac{1}{n}\sum_{i=1}^{n} w_i = \frac{1}{n}\sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_X} = 0, \\
s_W^2 &= \frac{1}{n-1}\sum_{i=1}^{n} w_i^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_X}\right)^2 = \frac{1}{s_X^2} \cdot \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{s_X^2}{s_X^2} = 1.
\end{aligned}$$

Deren Varianz-Kovarianzmatrix ist gemäß (7.21) zugleich ihre Korrelationsmatrix:

$$\begin{aligned}
S_W &= \frac{1}{n-1}(HXD_X^{-\frac{1}{2}})^T H(HXD_X^{-\frac{1}{2}}) = D_X^{-\frac{1}{2}} S_X D_X^{-\frac{1}{2}} = R_X. \\
R_W &= S_W = R_X.
\end{aligned}$$

Die aus X durch

$$\mathbf{z} := \frac{H\mathbf{x}}{||H\mathbf{x}||\sqrt{n-1}} \tag{7.28}$$

mit $z_i = \frac{x_i - \bar{x}}{s_X\sqrt{n-1}}$ gewonnenen Daten heißen kleine standardisierte Daten. Um eine Datenmatrix X so zu standardisieren, ist

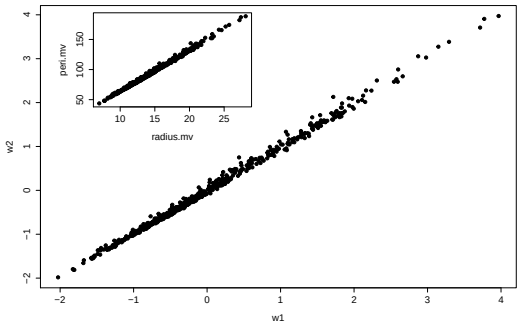
$$Z = \frac{1}{\sqrt{n-1}} HXD_X^{-\frac{1}{2}} = \frac{1}{\sqrt{n-1}} W$$

zu berechnen. Hierbei gilt

$$\begin{aligned}
S_Z &= \frac{1}{n-1}\frac{1}{\sqrt{n-1}}(HXD_X^{-\frac{1}{2}})^T H \cdot \frac{1}{\sqrt{n-1}}(HXD_X^{-\frac{1}{2}}) = D_X^{-\frac{1}{2}} S_X D_X^{-\frac{1}{2}} = \frac{1}{n-1} R_X, \\
R_Z &= \mathrm{diag}(\sqrt{n-1})\frac{1}{n-1} R_X \,\mathrm{diag}(\sqrt{n-1}) = R_X, \\
Z^T Z &= \frac{1}{n-1} D_X^{-\frac{1}{2}} X^T H^T H X D_X^{-\frac{1}{2}} = R_X.
\end{aligned}$$

```
w1=(radius.mv-mean(radius.mv))/sd(radius.mv)
w2=(peri.mv-mean(peri.mv))/sd(peri.mv)
mean(w1)
var(w1)
cov(cbind(w1,w2))
cor(cbind(w1,w2))
plot(w1,w2,pch=19)
```

Die Abbildung zeigt den Scatterplot der standardisierten Daten, oben eingeklinkt ist der Scatterpelot der Originaldaten zu sehen.



Literatur

[1] Georg Cantor. »Ueber eine elementare Frage der Mannigfaltigkeitslehre.« In: *Jahresbericht der Deutschen Mathematiker-Vereinigung* 1 (1890).

[2] Gerd Fischer, Matthias Lehner und Angela Puchert. *Einführung in die Stochastik - Die grundlegenden Fakten mit zahlreichen Erläuterungen, Beispielen und Übungsaufgaben*. Berlin Heidelberg New York: Springer-Verlag, 2015. ISBN: 978-3-658-07903-1.

[3] Marek Fisz. *Wahrscheinlichkeitsrechnung und mathematische Statistik: mit 40 Tab*. Bd. 40. Hochschulbücher für Mathematik. Berlin: Dt. Verl. d. Wiss, 1989. ISBN: 3326000790.

[4] Otto Forster. *Analysis 1*. Vieweg Verlag, Friedrich und Sohn Verlagsgesellschaft mbH, 2001. ISBN: 3528572248.

[5] R. Kabacoff. *R in Action: Data Analysis and Graphics With R*. Manning Pubs Co Series. Manning Publications Company, 2011. ISBN: 9781935182399. URL: http://books.google.de/books?id=qWpWRwAACAAJ.

[6] S. Schäffler. *Mathematik der Information*. Springer Spektrum, 2015. ISBN: 9783662463819.

[7] S. Schäffler und D. Meintrup. *Stochastik*. Berlin Heidelberg: Springer Verlag, 2005. ISBN: 3-540-21676-6.

[8] R. Schlittgen. *Multivariate Statistik*. Lehr- und Handbücher der Statistik : Fachgebiet Biometrie. Oldenbourg Wissensch.Vlg, 2009. ISBN: 9783486585957. URL: http://books.google.de/books?id=Qn2tPAAACAAJ.

[9] Dietrich Stoyan. *Stochastik fuer Ingenieure und Naturwissenschaftler: eine Einführung in die Wahrscheinlichkeitstheorie und die Mathematische Statistik*. Berlin: Akademie-Verl, 1993. ISBN: 3055016033. URL: http://d-nb.info/931107628/04.

[10] Helge Toutenburg, Angela Dörfler und Nina Quitzau. *Induktive Statistik: eine Einführung mit SPSS für Windows*. 3., überarb. Aufl. Springer-Lehrbuch. Berlin: Springer, 2005. ISBN: 3540242937 (kart.) URL: http://www.gbv.de/dms/hebis-darmstadt/toc/128074957.pdf.

[11] J.W. Tukey. *Exploratory Data Analysis*. New York: Addison-Wesley, 1977. ISBN: 0-201-07616-0.

[12] G. Walz. *Lexikon der Statistik*. Elsevier GmbH, 2004. ISBN: 3-8274-1423-7.

[13] H. Witting. *Mathematische Statistik I*. Teubner, 1985.

Index