

Skriptum zu MAS3/GST3 WS2014/2015

© Stephan Dreiseitl

Inhaltsverzeichnis

1	Einleitung	3
2	Formalisierung der Wahrscheinlichkeitsrechnung	7
2.1	Begriffsbildung	7
2.2	Axiomatisierung	10
3	Elementare Zähltheorie	16
3.1	Geordnete Proben mit Wiederholung	17
3.2	Geordnete Proben ohne Wiederholung	18
3.3	Permutationen ohne Wiederholung	19
3.4	Permutationen mit Wiederholung	19
3.5	Ungeordnete Proben ohne Wiederholung	21
3.6	Ungeordnete Proben mit Wiederholung	22
4	Bedingte Wahrscheinlichkeit	24
5	Zufallsvariable	31
5.1	Eindimensionale Zufallsvariable	31
5.2	Mehrdimensionale Zufallsvariable	36
6	Verteilungen von diskreten Zufallsvariablen	37
6.1	Gleichverteilung	37
6.2	Bernoulliverteilung	37
6.3	Binomialverteilung	38
6.4	Hypergeometrische Verteilung	40
6.5	Poissonverteilung	42
6.6	Gemeinsame Verteilungen	45
6.7	Unabhängigkeit von Zufallsvariablen	49
7	Erwartungswert und Varianz	52
7.1	Erwartungswert	53
7.2	Varianz	58
7.3	Kovarianz und Korrelationskoeffizient	62
8	Stetige Verteilungen	69
8.1	Unterschiede zu diskreten Verteilungen	69
8.2	Die Normalverteilung	74

9	Einführung in die induktive Statistik	83
9.1	Deskriptive Statistik	84
9.2	Mathematische Grundbegriffe	86
9.3	Stichprobenmittel und Stichprobenvarianz	88
9.4	Kovarianz und Korrelationskoeffizient einer Stichprobe	92
10	Maximum Likelihood Schätzungen	95
10.1	Anwendung: Lineare Regression	98
10.2	Anwendung: Logistische Regression	103
11	Einführung in die Theorie der Konfidenzintervalle	107
12	Einführung in die Testtheorie	116

Kapitel 1

Einleitung

Die Wahrscheinlichkeitsrechnung nimmt in der Mathematik eine herausragende Stellung ein, da sich in kaum einem anderen Bereich mit derart geringem (formal-)technischen Aufwand korrekte und einleuchtende Ergebnisse erzielen lassen. Ebenso wie in der Geometrie ist das theoretische Gerüst der Wahrscheinlichkeitsrechnung aus der Formalisierung von Anwendungen gewachsen. Obwohl der Begriff der Wahrscheinlichkeit anhand einfacher Beispiele leicht zu verstehen ist, ergaben sich bei der Einbettung in die moderne Mathematik erhebliche Schwierigkeiten. Zudem verleiten einfache Fragestellungen oft dazu, sich von seiner Intuition auf Irrwege leiten zu lassen; nicht selten liefert ein formal korrektes Durchrechnen unerwartete (richtige) Lösungen.

Auf Grund dieser Diskrepanz zwischen Intuition und formalem Apparat zählte *David Hilbert*, einer der berühmtesten Mathematiker aller Zeiten, im Jahr 1900 die Formalisierung der W.Rechnung¹ zu den wichtigsten ungeklärten Problemen der Mathematik. Erst 1933 gelang es *A.N. Kolmogorow* mit Hilfe von Maßtheorie und Mengenlehre, eine solche axiomatische Grundlage zu spezifizieren.

Die Probleme, die hier angesprochen werden, werden uns in dieser Vorlesung nicht beschäftigen. Sie werden hier nur erwähnt um anzudeuten, dass sich hinter den recht oberflächlichen Anwendungen, wie wir sie praktizieren werden, ein mächtiger formaler Apparat befindet. Dieser formale Apparat ist auch notwendig, denn wir werden bald sehen, dass man sich selten so leicht vom Hausverstand zu einer falschen Vorgangsweise (und zu einem falschen Ergebnis) verleiten lassen kann wie bei der W.Rechnung. Das geschieht vermutlich deshalb, weil praktisch alle Probleme direkt aus der Realität gegriffen sind, und bei solchen Aufgabenstellungen lässt sich der Hausverstand nur schwer umgehen.

Der W.Rechnung kann man sich, wie den meisten Bereichen der Mathematik, von zwei Seiten nähern: Einerseits, indem man bei der Axiomatisierung beginnend einfache Zusammenhänge aus den Grundgesetzen ableitet, oder aber, indem man aufgrund von Beispielen diejenigen Sachverhalte herausfiltert, die zu einer Axiomatisierung führen können. Da das Verständnis für die Methoden der W.Rechnung durch den zweiten Ansatz sicherlich vergrößert wird, wird er in dieser Vorlesung angewandt. In diesem Sinn beginnen wir mit einigen klassischen Beispielen aus der

¹Anmerkung: Im Folgenden wird zur Abkürzung statt Wahrscheinlichkeit nur mehr W. geschrieben, auch in zusammengesetzten Worten wie W.Rechnung.

W.Rechnung, wobei nicht unerwähnt bleiben soll, welche berühmte Mathematiker diese an sich einfachen Problemstellungen nicht korrekt lösen konnten. Die hier angeführten Beispiele sind [Beaumont, 1986; Blom, 1989; Freudenthal, 1973; Ross, 1987; Székely, 1990] entnommen. Der Aufbau dieser Vorlesung folgt großteils den Skripten [Dür, 1988] und [Weiss, 1992], aus denen ebenfalls etliche Beispiele entnommen sind. Weitere Beispiele stammen aus [Larson and Marx, 2006] und den frei verfügbaren Skripten [Grinstead and Snell, 1997; Bialas, 2005].

Beispiel 1.1 Bekanntlich haben bei einem idealen Würfel jeder der Zahlen 1 bis 6 die gleiche Chance (nämlich $\frac{1}{6}$), geworfen zu werden. Wenn die Augen zweier Würfel zusammengezählt werden, liegt das Ergebnis zwischen 2 und 12. Natürlich ist dabei das Ergebnis 2 weniger wahrscheinlich als etwa die Ergebnisse 9 oder 10, weil sich 2 nur als 1+1, 9 aber als 3+6 oder 4+5 schreiben lässt (und 10 als 4+6 oder 5+5). Beim Würfeln hat man allerdings herausgefunden, dass die Summe 9 häufiger auftritt als 10. Warum?

Wichtige Einsicht (etwa von *Gerolamo Cardano* in “De Ludo Aleae” (1663) oder *Galileo Galilei* in “Consideratione sopra il Giuoco dei Dadi” (1718), nicht aber von *Gottfried Leibniz*, der sich hier irrte): Beim Werfen mit zwei Würfeln müssen beide Würfel getrennt betrachtet werden. Das Ergebnis des Würfels ist ein Paar (n_1, n_2) , wobei n_1 die Augenzahl des ersten, und n_2 die Augenzahl des zweiten Würfels angibt. Für jede Zahl zwischen 1 und 6 liegt die W., dass n_1 gleich dieser Zahl ist, bei $\frac{1}{6}$; dasselbe gilt für n_2 . Bei getrennten Ereignissen, wie es hier der Fall ist (der Wurf des einen Würfel beeinflusst nicht das Ergebnis des anderen), errechnet sich die W. des Paares der Ereignisse als das Produkt der W. der einzelnen Ereignisse. So ist für jedes (geordnete!) Paar (n_1, n_2) von Augenzahlen die W. des Auftretens dieses Paares genau $\frac{1}{36}$. Die W., dass die Summe der beiden Würfel 9 beträgt, ergibt sich als die Summe der W. aller Paare (n_1, n_2) mit $n_1 + n_2 = 9$. Dies sind die Paare $(3, 6)$, $(6, 3)$, $(4, 5)$ und $(5, 4)$. Somit ist die W., mit zwei Würfeln 9 zu erhalten, gleich $\frac{4}{36}$. Für das Ergebnis 10 gibt es allerdings nur die drei Kombinationen $(4, 6)$, $(6, 4)$ und $(5, 5)$, woraus sich eine W. von $\frac{3}{36}$ ergibt.

Zu merken: In diesem Beispiel lassen sich bereits zwei wichtige Grundregeln der W.Rechnung erkennen.

- Die W. für das Auftreten eines Ereignisses, das sich aus zwei unabhängigen Teilereignissen zusammensetzt, ist das Produkt der W. der Teilereignisse.
- Wenn ein Ereignis durch mehrere disjunkte Möglichkeiten zustande kommen kann, so ist die W. dieser Ereignisse die Summe der W. dieser Möglichkeiten. \square

Beispiel 1.2 *Chevalier de Méré*, ein bekannter französischer Glücksspieler, traf den berühmten *Blaise Pascal* und fragte ihn nach einer Erklärung für das folgende, seiner Meinung nach paradoxe Phänomen: Er habe bemerkt, dass die W., beim viermaligen Werfen eines Würfels mindestens eine Sechs zu erhalten, größer als $\frac{1}{2}$ ist. Dies decke sich auch mit seiner mathematischen Interpretation des Problems. Wenn man jetzt statt mit einem mit zwei Würfeln wirft, so habe er entdeckt, dass die W., beim 24-maligen Würfeln mindestens eine Doppelsechs zu erhalten, kleiner als $\frac{1}{2}$ ist. Dies decke sich allerdings *nicht* mit seiner mathematischen Interpretation, denn schließlich sei $4 \cdot \frac{1}{6} = \frac{2}{3} = 24 \cdot \frac{1}{36}$, somit beides größer als $\frac{1}{2}$.

Wichtige Einsicht (die Pascal mit *Pierre de Fermat* teilte. Später löste auch *Isaac Newton* unabhängig ein ähnliches Problem): De Méré's mathematische Interpretation des Sachverhaltes ist falsch. Wenn sich W. so einfach aufsummieren ließen, könnte man daraus schließen, dass bei sechsmaligem Würfeln *sicher* (also mit W. 1) mindestens eine Sechs geworfen wird—ebenso wie mindestens eine Eins, eine Zwei usw.

Die korrekte Betrachtungsweise dieses Problems ist die folgende: Wie vorher wird das Ergebnis des viermaligen Würfels als Tupel (n_1, n_2, n_3, n_4) interpretiert. Die W., dass n_1 keine Sechs ist beträgt $\frac{5}{6}$; ebenso für n_2, n_3, n_4 . Das Ereignis "mindestens eine Sechs" ist das Gegenteil des Ereignisses "keine Sechs". Da die einzelnen Würfel unabhängig voneinander geworfen werden, ist die W. von "keine Sechs bei viermaligem Würfeln" $\frac{5}{6} \cdot \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{5}{6}$, die W. des Komplementäreignisses "mindestens eine Sechs bei viermaligem Würfeln" ist $1 - (\frac{5}{6})^4 \approx 0.516$. Hier stimmte also die falsche Interpretation de Méré's mit der mathematisch korrekten in dem Sinn überein, dass beide Ergebnisse größer als $\frac{1}{2}$ sind.

Beim Werfen mit zwei Würfeln ist die W., keine Doppelsechs zu erhalten, gleich $\frac{35}{36}$. Damit ist die W. von "keine Doppelsechs bei 24-maligem Würfeln" gleich $(\frac{35}{36})^{24}$, und die W. von "mindestens eine Doppelsechs bei 24-maligem Würfeln" gleich $1 - (\frac{35}{36})^{24} \approx 0.491$.

Zu merken: Wie auch im vorigen Beispiel haben wir hier die W. für das 24-malige unabhängige Würfeln als das Produkt der einzelnen W. erhalten. "Mindestens einmal" ist das Gegenteil von "nie"; da das sichere Ereignis W. 1 hat, ergibt sich für die W. p und \bar{p} zweier komplementärer Ereignisse die Beziehung $\bar{p} = 1 - p$. Wir werden in Abschnitt 2.2 sehen, wie sich diese Rechenregeln leicht mit Hilfe der Mengenlehre formal korrekt schreiben lassen. \square

Beispiel 1.3 Die Lösung des folgenden Aufteilungsproblems wird von manchen als die "Geburtsstunde der W.Rechnung" angesehen. Das Problem soll ebenfalls von de Méré an Pascal gestellt worden sein.

Das Aufteilungsproblem ergibt sich aus einem Spiel, das zwei Personen solange spielen, bis einer von ihnen eine gewisse Anzahl von Runden (etwa sechs) gewonnen hat. Um den Reiz des Spiels zu erhöhen, wird um einen Geldbetrag gespielt. Bei jeder Runde hat jeder der beiden Spieler die gleiche Chance, die Runde zu gewinnen. Wenn jetzt das Spiel aufgrund externer Umstände *vor* Erreichen der erforderlichen Rundenzahl abgebrochen werden muss (wenn etwa Spieler A fünf, Spieler B aber erst drei Runden gewonnen hat), wie soll dann der eingesetzte Geldbetrag aufgeteilt werden?

Wichtige Einsicht (Das Problem wurde zum ersten Mal von *Fra Lucca Paccioli* in seinem Werk "Summa de arithmetica, geometria, proportioni et proportionalità" erwähnt und dort nicht korrekt gelöst; auch *Niccolo Tartaglia* gab wenig später eine falsche Lösung an, bevor ebenfalls wieder Pascal und Fermat 1654 unabhängig voneinander zur selben korrekten Lösung kamen): Pascal und Fermat erkannten, dass die richtige Antwort nicht die bereits gewonnenen Runden zu berücksichtigen hat, sondern die Chancen der Spieler auf den Gesamtsieg (und damit als Anwendung der W.Rechnung zu betrachten ist). Fermat formulierte die Lösung für die oben gegebenen Zahlen in etwa so: Wenn das Spiel weitergespielt werden kann, steht nach weiteren drei Runden sicher ein Sieger fest. Da dies nicht möglich ist, werden diese drei Runden fiktiv ausgespielt. Da die W., dass ein Spieler eine Runde gewinnt,

gerade $\frac{1}{2}$ ist, ergeben sich folgende Möglichkeiten für die Sieger dieser drei Runden, wobei jede Kombination gleich wahrscheinlich ist (dabei bedeutet etwa A-A-B, dass Spieler A Runde 7 und Runde 8 und Spieler B Runde 9 gewinnt):

A-A-A	A-A-B	A-B-A	A-B-B
B-A-A	B-A-B	B-B-A	B-B-B

In nur einem dieser acht Fälle (nämlich bei B-B-B) gewinnt Spieler B das gesamte Spiel, sonst immer Spieler A. Daher ist der Gewinn im Verhältnis 7:1 zu teilen.

Zu merken: Die W.Rechnung kann auch in Bereichen eingesetzt werden, denen dies auf den ersten Blick nicht anzusehen ist. \square

Außerdem kommt es in der W.Rechnung oft vor, dass verschiedene Möglichkeiten abgezählt werden müssen. Dies war im vorigen Beispiel noch sehr einfach möglich. Die Lehre von richtigen Zählen wird *elementare Zähltheorie* genannt (ein Teilgebiet der *Kombinatorik*) und in dieser Vorlesung in Kapitel 3 behandelt.

Kapitel 2

Formalisierung der Wahrscheinlichkeitsrechnung

In diesem Kapitel werden die in der Einleitung gewonnenen Einsichten zu einem axiomatischen Gerüst der W.Rechnung ausgebaut. Wir beschränken uns dabei auf die *diskrete W.Rechnung*, bei der die Menge aller möglichen Ergebnisse eines Experiments endlich oder abzählbar unendlich ist. Dies erlaubt es uns, mit Summen zu arbeiten und auf Integrale zu verzichten, die im allgemeinen Fall notwendig sind. Mit dieser Einschränkung ergeben sich auch die in der Einleitung angesprochenen Schwierigkeiten bei der Axiomatisierung nicht.

2.1 Begriffsbildung

Experimente Die W.Rechnung beschäftigt sich mit der Durchführung von *Zufallsexperimenten*, also Versuchen, die sowohl *wiederholbar* als auch *zufällig* sind; die *Ausgänge* der Experimente sind also nicht vorhersehbar. Dabei wird jedem Experiment ein *Ereignisraum* Ω zugeordnet. Dieses Ω ist die Menge aller möglichen Ausgänge (Ergebnisse, Realisierungen) des Experiments. Das Finden des Ereignisraumes Ω stellt den wichtigen ersten Schritt zur Formalisierung der W.Rechnung dar. Da wir später speziell auch an der Größe des Ereignisraums interessiert sind, wird diese in den folgenden Beispielen mit angegeben.

Beispiel 2.1 Experiment: Einmaliges Werfen eines Würfels.

$$\Omega = \{1, 2, 3, 4, 5, 6\}, \quad |\Omega| = 6. \quad \square$$

Beispiel 2.2 Experiment: Einmaliges Werfen einer Münze.

$$\Omega = \{\text{KOPF}, \text{ZAHL}\}, \quad |\Omega| = 2. \quad \square$$

Beispiel 2.3 Experiment: Einmaliges Werfen zweier Würfeln (wie schon in Beispiel 1.1 besprochen).

$$\begin{aligned} \Omega &= \{(1, 1), (1, 2), \dots, (1, 6), (2, 1), (2, 2), \dots, (6, 6)\} = \\ &= \{(n_1, n_2) \mid 1 \leq n_i \leq 6 \text{ für } i = 1, 2\}, \quad |\Omega| = 36. \quad \square \end{aligned}$$

Beispiel 2.4 Viermaliges Werfen eines Würfels (aus Beispiel 1.2)

$$\begin{aligned}\Omega &= \{(1, 1, 1, 1), (1, 1, 1, 2), \dots, (2, 1, 1, 1), (2, 2, 2, 2), \dots, (6, 6, 6, 6)\} = \\ &= \{(n_1, n_2, n_3, n_4) \mid 1 \leq n_i \leq 6 \text{ für } i = 1, 2, 3, 4\}, \quad |\Omega| = 6^4.\end{aligned}$$

Wir sehen, dass bei der Abstraktion kein Unterschied zwischen dem gleichzeitigen Werfen von n Würfeln und dem n -maligen Werfen eines Würfels gemacht wird. \square

Beispiel 2.5 24-maliges Werfen zweier Würfel (wieder aus Beispiel 1.2)

$$\Omega = \Omega_1^{24} = \underbrace{\Omega_1 \times \Omega_1 \times \dots \times \Omega_1}_{24\text{-mal}}$$

wobei Ω_1 der Ereignisraum des einmaligen Werfens mit zwei Würfeln ist. Dies ist der Ereignisraum, der in Beispiel 2.3 angegeben ist (und dort mit Ω bezeichnet wird). \square

Ereignisse Eine beliebige Teilmenge A von Ω bezeichnet man als *Ereignis*. Dabei ist $\bar{A} = \Omega \setminus A$ das *Komplementärereignis* von A , Ω das *sichere Ereignis*, und die leere Menge \emptyset das *unmögliche Ereignis*. Zwei disjunkte Ereignisse nennt man *unvereinbar*.

Jedes Element $\omega \in \Omega$ bezeichnet man als *Elementarereignis*. Man sagt, dass ein Ereignis $A \subseteq \Omega$ eingetreten ist, wenn das Ergebnis des Experiments ein Elementarereignis $\omega \in A$ ist.

Beispiel 2.6 Experiment: Werfen eines Würfels. Sei $A =$ “die Augenzahl ist gerade”. Dieses Ereignis tritt bei $\omega = 2$, $\omega = 4$ oder $\omega = 6$ ein. \square

Beispiel 2.7 Experiment: Werfen zweier Würfel. $A =$ “die Augensumme beider Würfel ist 10”. Dieses Ereignis tritt bei $\omega = (6, 4)$, $\omega = (4, 6)$ oder $\omega = (5, 5)$ ein. \square

Wahrscheinlichkeit Der umgangssprachliche Begriff der *Wahrscheinlichkeit* wird in der Mathematik mit Hilfe von Axiomen fixiert. Dabei sind zwei Dinge zu beachten:

- Da es sich bei Axiomen um mathematische Sätze handelt, die nicht bewiesen werden, sondern als wahr angenommen werden, muss große Sorgfalt darauf gelegt werden, dass sie nicht im Widerspruch zum Rest der Mathematik stehen. Auf diese Frage gehen wir hier aber nicht näher ein.
- Die Richtigkeit der Axiome soll einsichtig sein. Das trifft speziell auf unsere Behandlung der W.Rechnung zu, bei der die Theorie nur unterstützend zur Anwendung benötigt wird.

Von der Anwendungsseite kommend möchte man also die W. als eine Funktion P definieren, die jedem Ereignis $A \subseteq \Omega$ eines Experiments eine Zahl zwischen 0 und 1 zuordnet. Diese Zahl sagt aus, wie wahrscheinlich das Auftreten von A ist.

Es gibt mehrere Möglichkeiten, die Funktion P festzulegen:

- Sie kann aus a-priori Überlegungen folgen (wie etwa aus der Geometrie eines Würfels oder einer Münze);

- sie kann subjektive Einschätzungen von nicht-deterministischen Ereignissen ausdrücken (Barcelona schlägt Bayern München in der Champions League);
- oder etwa aus der relativen Häufigkeit des Auftretens eines Ereignisses in einer Versuchsreihe abgelesen werden (morgen ist das Wetter mit 80%-iger W. schön).

Wir werden im Folgenden sehen, wie man aus den Eigenschaften der relativen Häufigkeit zu einer mathematischen Spezifizierung von W . kommen kann. Zuerst benötigen wir folgende Definition.

Definition 2.1 (Relative Häufigkeit)

Ein Experiment mit Ereignisraum Ω werde m -mal identisch durchgeführt. Dabei bezeichne m_A die absolute Häufigkeit des Auftretens von A in der Versuchsreihe. Dann ist die *relative Häufigkeit* (auch *empirische Wahrscheinlichkeit*) h eines Ereignisses A definiert als

$$h(A) := \frac{m_A}{m},$$

also als Verhältnis der Anzahl des Auftretens von A in der Versuchsreihe zur Länge der Versuchsreihe.

Beispiel 2.8 Eine Münze wird 100 mal geworfen. Das Ereignis KOPF erscheint dabei 46 mal. Damit ist $h(\text{KOPF}) = 0.46$. Wenn die Versuchsreihe auf 1000-maliges Werfen erweitert wird und KOPF 487 mal erscheint, ergibt sich eine relative Häufigkeit von $h(\text{KOPF}) = 0.487$. \square

Beispiel 2.9 Ein Würfel wird 100 mal geworfen. Das Ereignis $A =$ “die Augenzahl ist ungerade” tritt 52 mal ein, womit $h(A) = 0.52$ ist. Wenn bei 1000-maligem Werfen $h(A) = 0.56$ und bei 10000-maligem Werfen $h(A) = 0.55$ ist, wird man sich überlegen, ob der Würfel homogen ist, oder ob einige Ereignisse wahrscheinlicher als andere sind. Es ist Aufgabe der *Testtheorie*, eines eigenen Teilgebiets der Statistik, Fragestellungen dieser Art zu untersuchen. Wir werden Fragestellungen dieser Art in Kapitel 12 kurz betrachten. \square

Beispiel 2.10 In einem vereinfachten Text habe das Experiment “Lesen eines Buchstabens” 27 verschiedene Ausgänge, nämlich die Buchstaben a bis z und das Leerzeichen, das hier mit “-” bezeichnet wird. Für dieses Experiment wird der L^AT_EX-Quelltext dieses Skriptums verwendet, bei dem alle Großbuchstaben in Kleinbuchstaben umgewandelt wurden. Die relative Häufigkeit der einzelnen Buchstaben ist graphisch in Abbildung 2.1 zu sehen. Gegenüber Text in normaler Sprache sind dabei einzelne Buchstaben überrepräsentiert, wie etwa x und q, die als Variablennamen bzw. in L^AT_EX-Befehlen häufig vorkommen.

Dieses und davon abgeleitete Beispiele sind aus einem Buch von David MacKay [MacKay, 2003] entnommen. \square

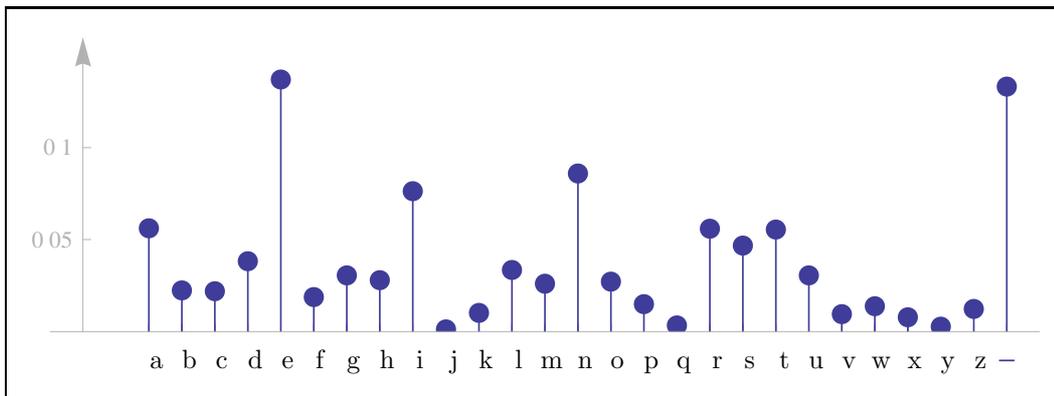


Abbildung 2.1: Graphische Repräsentation der relativen Häufigkeit der Buchstaben a bis z und des Leerzeichens “-” in diesem Skriptum.

Aus der Mengenlehre ergibt sich zudem noch die Forderung, dass die W. eines Ereignisses A gleich der Summe der W. der Teilmengen von A ist (wobei diese Teilmengen disjunkt sein müssen). Für endliches A ist diese Forderung klar einsichtig, sie soll aber auch für unendliche Mengen gelten.

Beispiel 2.11 Ein homogener Würfel wird 100 mal geworfen. Das Ereignis $A =$ “die Augenzahl ist gerade” setzt sich aus den drei disjunkten Elementarereignissen $\omega_2 =$ “Augenzahl 2”, $\omega_4 =$ “Augenzahl 4” und $\omega_6 =$ “Augenzahl 6” zusammen. Aus der Versuchsreihe ergeben sich die absoluten Häufigkeiten $m_{\omega_2} = 17$, $m_{\omega_4} = 14$ und $m_{\omega_6} = 18$. Damit ist dann

$$h(A) = \frac{m_A}{m} = \frac{m_{\omega_2} + m_{\omega_4} + m_{\omega_6}}{m} = \frac{49}{100}. \quad \square$$

2.2 Axiomatisierung

Die wichtigsten Begriffe aus Abschnitt 2.1 waren *Ereignisraum*, *Ereignis* und *Wahrscheinlichkeit*. In den Beispielen konnten wir bereits einige elementare Eigenschaften dieser Konzepte beobachten. Die Gesetzmäßigkeiten, die sich aus den Überlegungen in Abschnitt 2.1 ergeben, lassen sich wie folgt festlegen. Dabei bezeichnet $\text{Pot}(\Omega)$ die *Potenzmenge* (also die Menge aller Teilmengen) von Ω .

Definition 2.2 (Mathematischer Wahrscheinlichkeitsbegriff)

Sei Ω eine abzählbare Menge. Eine Abbildung $P : \text{Pot}(\Omega) \rightarrow [0, 1]$ heißt *Wahrscheinlichkeit* (auch *Wahrscheinlichkeitsmaß*, *Wahrscheinlichkeitsverteilung*) in Ω , wenn folgende beiden Bedingungen erfüllt sind:

$$P(\Omega) = 1 \quad (\text{normiert})$$

$$P\left(\bigcup_{i \geq 1} A_i\right) = \sum_{i \geq 1} P(A_i) \quad (\sigma\text{-additiv})$$

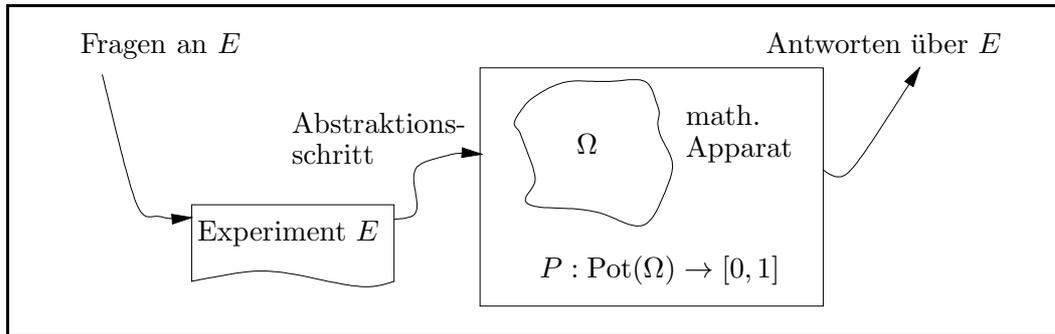


Abbildung 2.2: Abstraktionsschritt unter Verwendung des mathematischen Apparats.

für jede Folge A_1, A_2, \dots von paarweise disjunkten Teilmengen von Ω (also $A_i \cap A_j = \emptyset$ für $i \neq j$). Die drei charakterisierenden Elemente dieser Definition werden zu einem Tripel $(\Omega, \text{Pot}(\Omega), P)$ zusammengefasst und als *Wahrscheinlichkeitsraum* bezeichnet. Für abzählbares Ω nennt man $(\Omega, \text{Pot}(\Omega), P)$ einen *diskreten W.Raum*.

Der mathematische W.Begriff gibt also nicht an, wie sich W. berechnen lassen, sondern nur, welche Gesetzmäßigkeiten eine W. zu erfüllen hat. In diesem Sinn wird man durch das Umgehen mit dem abstrahierten mathematischen W.Begriff dazu gezwungen, formal korrekt über Fragen nachzudenken, die sich aus der Beobachtung von Experimenten ergeben. Dies ist in Abbildung 2.2 dargestellt.

Als erstes werden wir nachprüfen, dass die empirische W. h aus Definition 2.1 auch wirklich eine W. im Sinne der obigen Definition ist.

Satz 2.1 Die empirische W. h erfüllt die Bedingungen (normiert) und (σ -additiv) aus Definition 2.2.

Beweis h ist normiert:

$$h(\Omega) = \frac{m_\Omega}{m} = \frac{m}{m} = 1.$$

h ist σ -additiv: Für jede Folge A_1, A_2, \dots von paarweise disjunkten Teilmengen von Ω gilt

$$h\left(\bigcup_{i \geq 1} A_i\right) = \frac{m_{\bigcup_{i \geq 1} A_i}}{m} = \frac{\sum_{i \geq 1} m_{A_i}}{m} = \sum_{i \geq 1} \frac{m_{A_i}}{m} = \sum_{i \geq 1} h(A_i). \quad \square$$

Aus den zwei Axiomen (normiert) und (σ -additiv) einer W. lassen sich mit Hilfe der Mengenlehre sofort folgende Schlüsse ziehen.

Satz 2.2 Für die in Definition 2.2 eingeführten $W.$ gilt

- (1) $P(\emptyset) = 0$
- (2) $P(\bar{A}) = 1 - P(A)$
- (3) Aus $A \subseteq B$ folgt $P(A) \leq P(B)$
- (4) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Beweis Etwa der Aussagen (1) und (3):

- (1) Da Ω und \emptyset disjunkt sind, gilt $P(\Omega \cup \emptyset) = P(\Omega) + P(\emptyset)$. Wegen $P(\Omega \cup \emptyset) = P(\Omega)$ und $P(\Omega) = 1$ ergibt sich daraus $P(\emptyset) = 0$.
- (3) Für $A \subseteq B$ kann man $B = A \cup (B \setminus A)$ schreiben, wobei A und $B \setminus A$ disjunkt sind. Da $P(B \setminus A) \geq 0$ ist (wie für jede Menge), gilt

$$P(B) = P(A) + P(B \setminus A) \geq P(A). \quad \square$$

Durch die Modellierung von Ereignissen als Mengen lassen sich $W.$ über elementare Mengenoperationen ausdrücken und ausrechnen. Wir betrachten dazu einige Beispiele; diese Beispiele lassen sich besonders einfach über Visualisierungen der Mengen mit sogenannten *Venn-Diagrammen* lösen.

Beispiel 2.12 In einer Gruppe von 100 Studenten gibt es 30 Fußballer und 15 Ultimate Frisbee Spieler. Es stellt sich heraus, dass 10 Studenten sowohl die eine, als auch die andere der beiden genannten Sportarten ausüben. Wieviele der Studenten sind weder Fußballer noch Ultimate Spieler?

Lösung: Mit den Abkürzungen F für Fußballer und U für Ultimate Spieler wissen wir aus der Angabe, dass $|F \cap U| = 10$ ist. Damit ergibt sich durch einfaches Aufzeichnen, dass 20 Studenten *nur* Fußballer sind, und 5 Studenten *nur* Ultimate Spieler. Somit sind $10 + 20 + 5 = 35$ Studenten in diesen Sportarten aktiv, und 65 Studenten betreiben keine der beiden Sportarten. \square

Beispiel 2.13 Eine Studentin verabredet sich mit zwei notorisch unzuverlässigen Freundinnen in einem Cafe. Die $W.$, dass die erste bzw. zweite Freundin erscheint, seien $P(A) = 0.3$ bzw. $P(B) = 0.5$. Die $W.$, dass zumindest eine von beiden auftaucht, sei 0.7. Was ist die $W.$, dass beide ins Cafe kommen? Was sind die $W.$, dass nur jede allein kommt?

Lösung: Mit der Angabe gilt $P(A \cup B) = 0.7$, gesucht ist $P(A \cap B)$. Durch Umformen von Aussage (4) aus Satz 2.2, oder einfaches Aufzeichnen wie in Abbildung 2.3, ergibt sich

$$P(A \cap B) = P(A) + P(B) - P(A \cup B) = 0.3 + 0.5 - 0.7 = 0.1.$$

Die $W.$, dass nur die erste Freundin, nicht aber die zweite auftaucht, ist

$$P(A \setminus B) = P(A) - P(A \cap B) = 0.3 - 0.1 = 0.2;$$

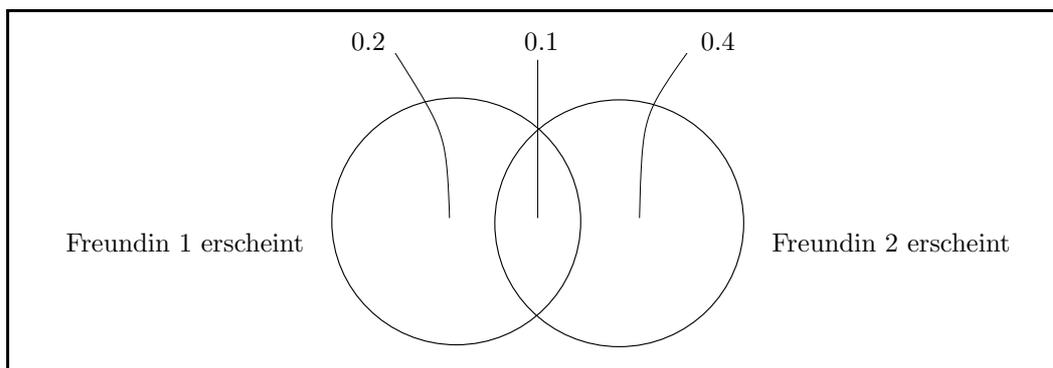


Abbildung 2.3: Venn-Diagramm zur Veranschaulichung der disjunkten Ereignisse und ihrer W. aus Beispiel 2.13.

analog ergibt sich für die andere Freundin

$$P(B \setminus A) = P(B) - P(A \cap B) = 0.5 - 0.1 = 0.4. \quad \square$$

Durch den Zusammenhang mit der Mengenlehre wird es vielleicht schon aufgefallen sein, dass die W. eines Ereignisses A nur von der W. der Elementarereignisse von A abzuhängen scheint. Dass das auch tatsächlich so ist, sehen wir nach der folgenden Definition.

Definition 2.3 (Wahrscheinlichkeitsfunktion)

Die Funktion

$$f : \Omega \rightarrow [0, 1] \\ \omega \mapsto f(\omega) := P(\{\omega\}),$$

die einem Elementarereignis seine W. zuweist, heißt *Wahrscheinlichkeitsfunktion* von P .

Folgender Satz drückt aus, dass man die W. eines Ereignisses durch die W. der darin enthaltenen Elementarereignisse berechnen kann.

Satz 2.3 Ein W.Maß P ist durch seine W.Funktion eindeutig bestimmt.

Beweis Jedes $A \subseteq \Omega$ lässt sich schreiben als $A = \bigcup_{\omega \in A} \{\omega\}$. Damit ist

$$P(A) = P\left(\bigcup_{\omega \in A} \{\omega\}\right) = \sum_{\omega \in A} P(\{\omega\}) = \sum_{\omega \in A} f(\omega). \quad \square$$

Aus der Bedingung (normiert) in Definition 2.2 folgt sofort, dass die aufsummierten W.funktionen aller Elementarereignisse 1 ergeben müssen. Für beliebige

W.funktion f gilt also

$$\sum_{\omega \in \Omega} f(\omega) = 1.$$

Sobald man also die W.funktion einer W. kennt, kann man die W. jedes Ereignisses als Summe der W. der Elementarereignisse ausrechnen. Im Folgenden werden wir uns näher mit einer Verteilung beschäftigen, die eine speziell einfache W.funktion aufweist: bei dieser ist jedes Elementarereignis gleich wahrscheinlich. Dazu müssen wir allerdings annehmen, dass der Ereignisraum Ω endlich ist.

Definition 2.4 (Gleichverteilung)

Ein endlicher Ereignisraum Ω mit W.Maß P heißt *Laplace-Raum*, wenn

$$f(\omega) = P(\{\omega\}) = \frac{1}{|\Omega|}.$$

Das W.Maß P eines Laplace-Raums Ω heißt *Gleichverteilung* in Ω .

In einem Laplace-Raum gilt immer

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

wie aus dem Beweis von Satz 2.3 unmittelbar folgt. Mit Hilfe einiger Beispiele kann man sich auch von der Plausibilität (und nicht nur der mathematischen Richtigkeit) sofort überzeugen.

Beispiel 2.14 Ein homogener Würfel wird dreimal geworfen. Dabei sei A das Ereignis "die Summe der Augenzahlen beträgt 11". Formal kann dieses Ereignis als

$$A = \{(n_1, n_2, n_3) \mid n_1 + n_2 + n_3 = 11\}$$

im Laplace-Raum $\Omega = \{(n_1, n_2, n_3) \mid 1 \leq n_i \leq 6 \text{ für } i = 1, 2, 3\}$ geschrieben werden. Die W. jedes Elementarereignisses $\omega = (n_1, n_2, n_3)$ ist $\frac{1}{6^3}$, somit ist

$$P(A) = \frac{|A|}{|\Omega|} = \frac{27}{6^3} = \frac{1}{8}. \quad \square$$

Dieses Ergebnis setzt aber Kenntnis von $|A| = 27$ voraus. Das nächste Kapitel wird zeigen, wie wir die Kardinalität von Mengen ohne Aufschreiben und Abzählen jedes einzelnen Elements berechnen können. Zum jetzigen Zeitpunkt beschränken wir uns darauf, die *ungeordneten* Möglichkeiten zum Erreichen der Augensumme 11 zu notieren. Es sind dies 6-4-1, 6-3-2, 5-5-1, 5-4-2, 5-3-3 und 4-4-3. Durch Abzählen der möglichen Umordnungen dieser Kombinationen erhält man die Lösung 27.

Beispiel 2.15 Beim Toto werden Tipps auf den Ausgang von 12 Fußballspielen abgegeben. Jedes Spiel kann entweder mit einem Sieg der Heim- oder Gastmannschaft bzw. mit einem Unentschieden enden. Wenn jeder dieser Ausgänge für jedes Spiel gleich wahrscheinlich ist, wie groß ist dann die Chance, einen Zwölfer zu erlangen, also *alle* Tipps richtig zu haben? Wie groß ist die W., *keinen* der Tipps richtig zu haben?

Lösung: Der Ereignisraum des Totospiels besteht aus 12-Tupeln, bei denen jeder der Komponenten entweder 1, 2 oder x ist. Damit ist $\Omega = \{1, 2, x\}^{12}$ und $|\Omega| = 3^{12}$. Wenn jede einzelne Tippkombination wie angenommen gleich wahrscheinlich ist, ergibt sich für die W. jedes einzelnen Tipps $P(\{\omega\}) = \frac{1}{3^{12}}$. Das Ergebnis $A = \text{“12 richtige”}$ ist ein Elementarereignis (wird nur von einem $\omega \in \Omega$ realisiert) und hat daher die W.

$$P(A) = \frac{|A|}{|\Omega|} = \frac{1}{3^{12}}.$$

Das Ereignis $B = \text{“kein einziges richtiges Resultat”}$ ist *kein* Elementarereignis, da es auf viele verschiedene Möglichkeiten realisiert werden kann. Da es für jedes der 12 Spiele zwei Möglichkeiten gibt, falsch zu tippen, besteht B aus 2^{12} Elementarereignissen. Somit ist

$$P(B) = \frac{|B|}{|\Omega|} = \frac{2^{12}}{3^{12}} = \left(\frac{2}{3}\right)^{12} \approx 0.0077. \quad \square$$

Die beiden vorangegangenen Beispiele haben gezeigt, dass sich viele alltägliche Probleme durch Laplace-Räume beschreiben lassen. Dabei benötigen wir zum Berechnen der W. Methoden, wie man effizient die Kardinalität von Mengen berechnen kann. Mit diesem Thema beschäftigen wir uns im nächsten Kapitel.

Kapitel 3

Elementare Zähltheorie

Das Ziel dieses Kapitels ist es, Denkansätze zum Abzählen von Mengen in geordnete Bahnen zu lenken. Dabei ist die Schwierigkeit meist nicht das Anwenden einer Methode auf ein Problem, sondern vielmehr herauszufinden, welche Methode auf ein gegebenes Problem anwendbar ist. Wir werden bald herausfinden, dass dieser Vorgang durch das Erkennen von *Ähnlichkeiten von Problemen* erleichtert wird. So gibt es einige Standardprobleme, deren Lösungen recht offensichtlich sind. Die Kunst der elementaren Zähltheorie besteht darin, neue Probleme auf die schon bekannten Standardprobleme zu reduzieren, für die man die Lösungen bereits kennt. Dieser Vorgang wird in Abbildung 3.1 schematisch dargestellt. Deshalb werden wir Standardprobleme häufig über ihre Charakteristik beschrieben und nicht so oft mit den offiziellen Bezeichnungen, die wenig aussagekräftig sind; zudem sind diese Bezeichnungen nicht einheitlich.

In dieser Vorlesung beschäftigen wir uns mit *geordneten* und *ungeordneten Proben* sowie *Permutationen*, alle drei sowohl *mit* als auch *ohne Wiederholungen*. Damit ergeben sich sechs mögliche Kombinationen, die wie folgt über die Analogie des Ziehens von durchnummerierten Bällen aus einer Urne charakterisiert werden können: Geordnete Proben unterscheiden sich von ungeordneten Proben dadurch, dass es bei ersteren auf die genaue Reihenfolge der gezogenen Kugeln ankommt, bei zweiteren aber nicht. So ist etwa die (zeitlich in dieser Reihenfolge gezogene) Folge von Lottozahlen “3 – 21 – 9 – 5 – 43 – 17” eine *geordnete Probe*, die Präsentation “3 – 5 – 9 – 17 – 21 – 43” als Ergebnis der Ziehung eine *ungeordnete Probe*. Die Terminologie ist hier potentiell verwirrend, da die zweite (als ungeordnet bezeichnete) Probe ja das Result des Ordnen der ersten (als geordnet bezeichneten) Probe ist. Die Bezeichnung “geordnet/ungeordnet” bezieht sich hier aber darauf, ob es auf genau die Reihenfolge ankommt, in der die Probe gezogen wird: Dies ist bei der ersten Probe der Fall (“3 – 21 – 9 – 5 – 43 – 17” ist eine andere Sequenz gezogener Bälle als “3 – 17 – 5 – 43 – 9 – 21”), bei der zweiten aber nicht (“3 – 5 – 9 – 17 – 21 – 43” sind die gleichen Gewinnzahlen wie “3 – 17 – 5 – 43 – 9 – 21”).

Beide diese Beispiele sind Proben *ohne Wiederholung*, da ja jede Lottozahl maximal einmal gezogen werden kann. Die zugehörigen Proben *mit Wiederholung* ergeben sich dadurch, dass die Bälle nach jedem Ziehen wieder in die Urne zurückgegeben werden, und damit auch mehrmals gezogen werden können bzw. dass gleiche Bälle öfters (wiederholt) in der Urne sind.

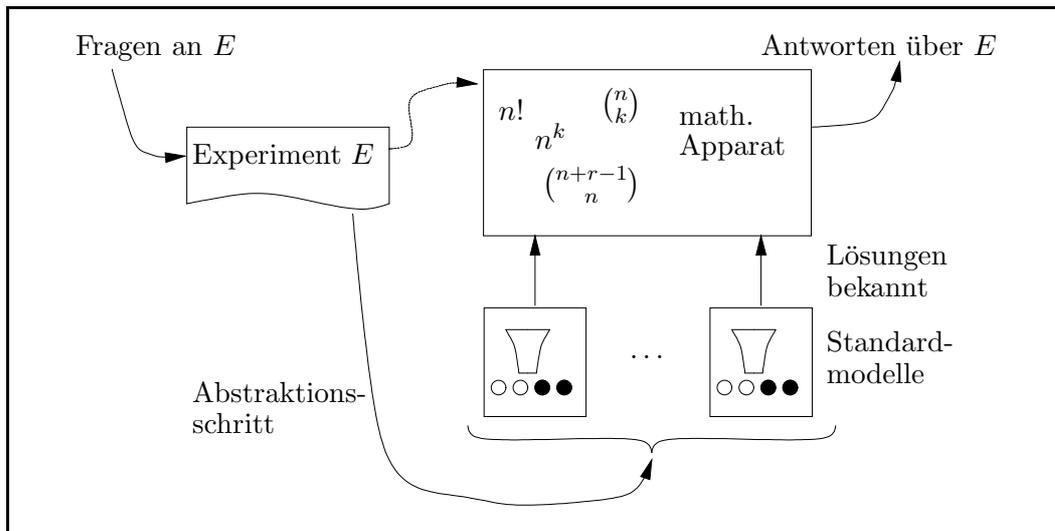


Abbildung 3.1: Lösen eines Problems durch Reduktion auf ein Standardmodell.

Von *Permutationen* spricht man, wenn man die möglichen Anordnungen von Objekten abzählen möchte. Dies entspricht für den Fall von Permutationen *ohne Wiederholung* dem Urnenmodell, aus dem *alle Bälle* (unterscheidbar) gezogen werden. Im Gegensatz dazu befinden sich bei *Permutationen mit Wiederholung* auch mehrere gleiche Bälle (untereinander ununterscheidbar) in der Urne.

Nach dieser Erläuterung der Terminologie beginnen wir mit dem einfachsten Zählproblem, dessen Lösung wir unbewusst schon einige Male verwendet haben.

3.1 Geordnete Proben mit Wiederholung

Beispiel 3.1 Ein Würfel wird viermal geworfen. Wieviele mögliche Elementarereignisse gibt es, was ist also $|\Omega|$?

Lösung: Beim ersten bis vierten Wurf kann jeweils eine von sechs Augenzahlen erscheinen. Somit ergeben sich $6 \cdot 6 \cdot 6 \cdot 6 = 1296$ Elementarereignisse. \square

Das Beispiel 2.15 (Fußball-Toto) behandelt eine ähnliche Art von Problemen. Ihre Lösung wird im folgenden Satz zusammengefasst.

Satz 3.1 (geordnete Proben mit Wiederholung) Aus einer Menge G mit $|G| = n$ können n^k k -Tupel gebildet werden.

Standardmodell (geordnete Proben mit Wiederholung) Aus einer Urne mit n verschiedenen Kugeln wird k mal mit Zurücklegen gezogen. Bei diesem Experiment ergeben sich n^k verschiedene Sequenzen von gezogenen Kugeln.

Beispiel 3.2 Es gibt 10^5 fünfstellige Telefonnummern, wenn jede Ziffernkombination eine gültige Nummer ist. \square

Beispiel 3.3 Es gibt $2^8 = 256$ binäre Zahlen der Länge 8, und $2^{16} = 65536$ der Länge 16. \square

Eine einfache Erweiterung ergibt sich, wenn man in Satz 3.1 nicht nur eine, sondern mehrere Mengen zulässt.

Satz 3.2 Die Mengen G_1, G_2, \dots, G_r enthalten jeweils n_1, n_2, \dots, n_r Elemente. Dann enthält $G_1 \times G_2 \times \dots \times G_r$ genau $n_1 \cdot n_2 \cdot \dots \cdot n_r$ Elemente.

Beispiel 3.4 Bei einer Auswahl von fünf Vor-, sieben Haupt- und drei Nachspeisen lassen sich $5 \cdot 7 \cdot 3 = 105$ verschiedene Menus zusammenstellen. \square

Das folgende Beispiel illustriert, dass die meisten Schwierigkeiten sich ergeben, wenn aus der Problemstellung nicht sofort der Lösungsweg ersichtlich ist.

Beispiel 3.5 Wieviele Möglichkeiten gibt es, zehn verschiedene Münzen auf zwei Personen aufzuteilen? \square

3.2 Geordnete Proben ohne Wiederholung

Beispiel 3.6 Wieviele fünfstellige Telefonnummern gibt es, bei denen jede Ziffer verschieden ist?

Lösung: Die erste Stelle kann frei aus $\{0, \dots, 9\}$ gewählt werden. Für die zweite bleiben nur noch neun Ziffern, für die dritte acht, die vierte sieben und die fünfte sechs. Damit ergeben sich $10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 = 30240$ Möglichkeiten. \square

Beispiel 3.7 Aus sechs Läufern soll eine Staffel von vier Läufern gebildet werden. Dabei gibt es sechs Möglichkeiten für den ersten, fünf für den zweiten, vier für den dritten und schließlich drei für den Schlussläufer. Somit ergeben sich $6 \cdot 5 \cdot 4 \cdot 3 = 360$ Möglichkeiten. \square

Zu beachten ist, dass im letzten Beispiel die Reihenfolge der Läufer eine Rolle spielt; so sind etwa 1-4-3-5 und 4-1-3-5 verschiedene Staffeln. Wir werden im Abschnitt 3.4 sehen, wie man korrekt abzählt, wenn die Reihenfolge nicht relevant ist.

Satz 3.3 (geordnete Proben ohne Wiederholung) Aus einer Menge mit n Elementen können $n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot (n-k+1)$ k -Tupel gebildet werden, bei denen jede Komponente verschieden ist.

Standardmodell (geordnete Proben ohne Wiederholung) Aus einer Urne mit n verschiedenen Kugeln wird k mal ohne Zurücklegen gezogen. Hier ergeben sich $n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot (n-k+1)$ verschiedene Sequenzen gezogener Kugeln.

Wenn in Satz 3.3 $n = k$ ist, erhält man die im folgenden Abschnitt präsentierte Vereinfachung.

3.3 Permutationen ohne Wiederholung

Beispiel 3.8 Das Kartenspiel *Schnapsen* spielt man mit 20 Karten. Von einem gemischten Stapel werden der Reihe nach alle Karten aufgedeckt. Wieviele verschiedene gemischte Stapel (und damit aufgedeckte Sequenzen) gibt es?

Lösung: Wie aus dem letzten Abschnitt ersichtlich gibt es genau $20 \cdot 19 \cdot \dots \cdot 1$ Möglichkeiten. \square

Definition 3.1 (Fakultät)

Die Zahl $n!$ (sprich n Fakultät) ist definiert als

$$n! = n \cdot (n - 1) \cdot \dots \cdot 1.$$

Außerdem setzt man $0! = 1$.

Mit Einführung der Fakultätsschreibweise kann $n \cdot (n - 1) \cdot \dots \cdot (n - k + 1)$ in Abschnitt 3.2 auch als $\frac{n!}{(n-k)!}$ geschrieben werden.

Satz 3.4 (Permutationen ohne Wiederholung) Es gibt $n!$ verschiedene Möglichkeiten, eine Menge mit n Elementen anzuordnen.

Standardmodell (Permutationen ohne Wiederholung) Aus einer Urne mit n verschiedenen Kugeln werden alle n Kugeln gezogen (ohne Zurücklegen). Dabei erhält man $n!$ verschiedene Sequenzen gezogener Kugeln.

Beispiel 3.9 Gegeben sei eine Menge M mit n Elementen. Dann gibt es genau $n!$ Bijektionen $M \rightarrow M$, wie man aus dem folgenden Argument sehen kann: Man nummeriere die Elemente von M beliebig durch; dann gibt es für das erste Element n mögliche Bildelemente, für das zweite $n - 1$ usw., bis für das letzte Element nur mehr eine Möglichkeit bleibt. \square

Bis jetzt haben wir nur Mengen mit verschiedenen Elementen betrachtet, wie es nach Definition des Mengenbegriffs auch notwendig ist: In einer Menge kann bekanntlicherweise jedes Element nur einmal enthalten sein. Im Folgenden behandeln wir auch Mengen, bei denen einzelne Elemente bezüglich einer Eigenschaft ununterscheidbar sind. Um dem strengeren mathematischen Mengenbegriff gerecht zu werden, müssen wir dann annehmen, dass sie sich bezüglich einer anderen, nicht weiter spezifizierten Eigenschaft unterscheiden.

3.4 Permutationen mit Wiederholung

Beispiel 3.10 In Beispiel 2.14 haben wir angegeben, dass es 27 Möglichkeiten gibt, mit drei Würfeln die Augensumme 11 zu würfeln. Wie kommt man auf diese Zahl?

Lösung: Ohne die verschiedenen Anordnungen der einzelnen Würfel zu unterscheiden, kommt man auf die sechs Kombinationen 6-4-1, 6-3-2, 5-5-1, 5-4-2, 5-3-3 und 4-4-3. Da in 6-4-1 alle Augen verschieden sind, gibt es $3! = 6$ Möglichkeiten, mit Hilfe dieser Augen auf die Summe 11 zu kommen. Bei 6-3-2 und 5-4-2 verhält es sich gleich. Bei 5-5-1, 5-3-3 und 4-4-3 ergeben sich jeweils nur mehr drei Möglichkeiten. Die anderen drei fallen weg, weil zwei der Würfel ununterscheidbare Augenzahlen zeigen. Somit ergeben sich insgesamt $6 + 6 + 6 + 3 + 3 + 3 = 27$ Kombinationen, die alle in Summe 11 ergeben.

Die wichtige Einsicht aus diesem Beispiel ist die folgende: 6-4-1 lässt sich auf $3!$ Arten anordnen, 5-5-1 nur mehr auf $\frac{3!}{2!}$ Arten. In dieses erweiterte Modell gepresst könnte man auch sagen, dass sich 5-5-5 nur mehr auf $\frac{3!}{3!} = 1$ Art anordnen lässt. Es werden also alle möglichen Kombinationen durch die ununterscheidbaren Kombinationen dividiert. \square

Beispiel 3.11 Auf einer Kette werden 15 farbige Perlen aufgefädelt, von denen fünf rot, fünf schwarz, und fünf blau sind. Auf wieviele Arten können die 15 Perlen angeordnet werden?

Lösung: 15 unterscheidbare Perlen können auf $15!$ Arten angeordnet werden. Die fünf schwarzen Perlen können untereinander nicht unterschieden werden; damit fallen $5!$ Kombinationen weg. Das gleiche gilt für die $5!$ Kombinationen von roten und die $5!$ Kombinationen von blauen Perlen, die untereinander nicht unterscheidbar sind. Weil diese Ununterscheidbarkeit in *jeder* der $15!$ möglichen Anordnungen gilt, wird durch das "Ausmaß der Ununterscheidbarkeit" *dividiert*. Somit ergeben sich $\frac{15!}{5!5!5!} = 756756$ optisch unterscheidbare Anordnungen. \square

Wir verzichten auf den Beweis, dass hier dividiert (und nicht etwa subtrahiert) werden muss. Das folgende Beispiel illustriert vielleicht besser als ein Beweis, dass dies auch tatsächlich zutrifft.

Beispiel 3.12 Zwei Paare von ein-eigenen Zwillingen werden für ein Foto nebeneinander plziert. Wieviele optisch nicht unterscheidbare Anordnungen gibt es?

Lösung: Vier Personen können auf $4! = 24$ Arten angeordnet werden. Weil jeweils zwei Personen ununterscheidbar sind, ergeben sich die folgenden $\frac{4!}{2!2!} = 6$ Anordnungen:

$$\begin{array}{lll} \text{A-A-B-B} & \text{A-B-A-B} & \text{A-B-B-A} \\ \text{B-A-A-B} & \text{B-A-B-A} & \text{B-B-A-A} \end{array}$$

Man kann sich überzeugen, dass es keine 20 Anordnungen gibt. Auf diese Zahl kommt man, wenn man fälschlicherweise die ununterscheidbaren Kombinationen von allen möglichen abzieht, also $4! - 2! - 2! = 20$ rechnet. \square

Satz 3.5 (Permutationen mit Wiederholung) Eine Menge mit n Elementen, von denen k_1, \dots, k_r Elemente jeweils untereinander ununterscheidbar sind, kann auf $\frac{n!}{k_1! \cdots k_r!}$ Möglichkeiten angeordnet werden.

Standardmodell (Permutationen mit Wiederholung) In einer Urne befinden sich n Kugeln, wovon jeweils k_i ($i = 1, \dots, r$) Kugeln die gleiche Farbe haben. Alle n Kugeln werden ohne Zurücklegen aus der Urne gezogen. Dann gibt es $\frac{n!}{k_1! \cdots k_r!}$ farblich verschiedene Sequenzen von (in jeder Farbe ununterscheidbaren) gezogenen Kugeln.

3.5 Ungeordnete Proben ohne Wiederholung

Beispiel 3.13 Beim österreichischen Lotto “6 aus 45” können $45 \cdot 44 \cdots 40$ verschiedene *geordnete* Sequenzen von sechs Zahlen gezogen werden. Jeweils $6!$ dieser Sequenzen bestehen aus den gleichen Zahlen, liefern also die gleiche *ungeordnete* Sequenz. Mit den Überlegungen aus Abschnitt 3.4 wissen wir, dass die Anzahl verschiedener Kombinationen von Gewinnzahlen (als ungeordnete Sequenz) $\frac{45 \cdot 44 \cdots 40}{6!} = 8\,145\,060$ beträgt. \square

Beispiel 3.14 Aus einer Gruppe von 50 Studierenden soll eine Gruppe von drei Vertretern ausgewählt werden. Wenn wir annehmen, dass es keine Unterschiede zwischen den Vertretern gibt, so werden die $50 \cdot 49 \cdot 48$ Möglichkeiten, drei *unterschiedliche* Vertreter zu wählen, noch durch $3!$ dividiert, wodurch sich $\frac{50 \cdot 49 \cdot 48}{3!} = 19\,600$ Möglichkeiten ergeben. \square

In beiden Beispielen wurde der Übergang von einer geordneten zu einer ungeordneten Probe dadurch erreicht, dass durch die Fakultät des Umfangs der Probe dividiert wird. Da dieser Umfang auch in den Zähler des Bruches einfließt, ergibt sich eine besonders einfache Formel.

Definition 3.2 (Binomialkoeffizient)

Die Zahl $\binom{n}{k}$ (sprich *n über k*) ist definiert als

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

Die Anzahl der verschiedenen Lottozahlen aus Beispiel 3.13 ist somit $\binom{45}{6}$, die der verschiedenen Gruppen von Studentenvertretern aus Beispiel 3.14 $\binom{50}{3}$.

Satz 3.6 (Ungeordnete Proben ohne Wiederholung) Aus einer Menge von n Elementen können k Elemente auf $\binom{n}{k}$ Arten ausgewählt werden, wenn es auf die Reihenfolge der k Elemente nicht ankommt.

Standardmodell (Ungeordnete Proben ohne Wiederholung) Aus einer Urne mit n verschiedenen Kugeln werden mit einem Griff k Kugeln gezogen—es gibt also keine Reihenfolge unter diesen k Kugeln. Dann gibt es $\binom{n}{k}$ verschiedene Resultate dieses Ziehens.

Ein anderes ebenfalls sehr gebräuchliches Modell für ungeordnete Proben ist die Anzahl der Teilmengen einer Menge. So gibt es $\binom{n}{k}$ k -elementige Teilmengen einer Menge mit n Elementen.

Beispiel 3.15 Gegeben sei eine Menge M mit vier Elementen. Es gibt $\binom{4}{0} = 1$ null-elementige Teilmenge (die leere Menge), $\binom{4}{1} = 4$ ein-elementige, $\binom{4}{2} = 6$ zwei-elementige, $\binom{4}{3} = 4$ drei-elementige und $\binom{4}{4} = 1$ vier-elementige Teilmengen von M . \square

Wie aus diesem Beispiel ersichtlich, hat eine vier-elementige Menge $2^4 = 16$ Teilmengen. Diese Tatsache lässt sich wie folgt verallgemeinern.

Satz 3.7 Eine Menge mit n Elementen hat 2^n Teilmengen, oder anders ausgedrückt:

$$\sum_{k=0}^n \binom{n}{k} = 2^n.$$

Zum Abschluss dieses Kapitels wird noch folgende Abzählungsregel erwähnt, die allerdings etwas weniger allgemein anwendbar ist als die bisher präsentierten. Hier soll vor allem die besondere Art hervorgehoben werden, wie das Zählproblem mit der geeigneten Darstellung auf ein einfacheres schon bekanntes Problem reduziert werden kann.

3.6 Ungeordnete Proben mit Wiederholung

Beispiel 3.16 An einem schönen Herbsttag sollen 10 frisch geerntete Äpfel auf die vier Kinder aufgeteilt werden, die beim Pflücken geholfen haben. Auf wieviele Arten ist dies möglich, wenn manche Kinder auch keine Äpfel bekommen können (etwa, weil sie keine mögen)?

Lösung: Es gibt $\binom{10+4-1}{10} = \binom{13}{10} = 286$ solcher Aufteilungen. Die genaue Begründung folgt unmittelbar aus dem Beweis des folgenden Satzes. \square

Satz 3.8 (Ungeordnete Proben mit Wiederholung) Es können $\binom{n+r-1}{n}$ r -Tupel α gebildet werden, für die $\sum_{i=1}^r \alpha_i = n$ gilt. Man beachte, dass $\alpha_i = 0$ erlaubt ist.

Standardmodell (Ungeordnete Proben mit Wiederholung) Eine Menge von n ununterscheidbaren Kugeln soll auf r Urnen aufgeteilt werden, wobei manche Urnen auch leer bleiben können. Dies ist auf $\binom{n+r-1}{n}$ Arten möglich.

Beweis (in der Sprache des Standardmodells) Gegeben seien r Urnen und n Kugeln, wobei Urne i genau α_i Kugeln enthält ($1 \leq i \leq r$). Es gilt also $\sum_{i=1}^r \alpha_i = n$. Die Richtigkeit des Satzes folgt am einleuchtendsten, wenn man sich Folgendes

überlegt: Wir bilden ein Tupel der Länge n , dessen Komponenten alle 1 sind. Jetzt fügen wir als "Trennstellen" noch $r - 1$ Nullen ein, wodurch ein Tupel der Länge $n + r - 1$ entsteht. Der Wert von α_i sei gegeben durch die Anzahl der zwischen der $(i - 1)$ ten und i ten Null stehenden Einser. So kodiert etwa für die Aufteilung von 5 Kugeln auf 3 Urnen (also $n = 5, r = 3$) das Tupel $(0, 1, 1, 0, 1, 1, 1)$ die Aufteilung $\alpha_1 = 0, \alpha_2 = 2, \alpha_3 = 3$, oder in der Sprache des obigen Satzes, das Tupel $(0, 2, 3)$.

Man kann sich überlegen, dass durch beliebige Wahl der $r - 1$ Trennstellen im Tupel alle möglichen Aufteilungen der Kugeln auf die Urnen beschrieben werden können. Da es $\binom{n+r-1}{r-1}$ Möglichkeiten gibt, diese Trennstellen zu setzen, gibt es auch genau so viele mögliche Aufteilungen von n Kugeln auf r Urnen.

In der Erklärung wird als Lösung $\binom{n+r-1}{r-1}$ angegeben; dies ist wegen

$$\binom{n+r-1}{r-1} = \frac{(n+r-1)!}{(r-1)!n!} = \binom{n+r-1}{n}$$

das gleiche wie der Binomialkoeffizient in Satz 3.8. Die Gleichheit dieser Ausdrücke kann man sich auch dadurch überlegen, dass durch die Wahl der $r - 1$ Trennstellen die Lage der n Einser eindeutig bestimmt ist, und umgekehrt. \square

Beispiel 3.17 (Fortsetzung von Beispiel 3.16) Die 10 Äpfel entsprechen den n Kugeln im Standardmodell, und die vier Kinder den r Urnen. Mit der Formel erhält man sofort die Lösung von $\binom{13}{10}$ möglichen Aufteilungen. \square

Beispiel 3.18 Wie wir aus Abschnitt 3.1 wissen, gibt es 10^5 fünfstellige Telefonnummern; aus Abschnitt 3.2 wissen wir; dass es $\frac{10!}{5!}$ Telefonnummern gibt, bei denen jede Ziffer verschieden ist. Wieviele fünfstellige Telefonnummern aber gibt es, die aufsteigend sortiert sind? So sind etwa 12268 und 15689 aufsteigend sortiert, nicht aber 12253.

Lösung: Als erstes muss man überlegen, wie diese Aufgabenstellung auf das Standardproblem der ungeordneten Proben mit Wiederholung abgebildet werden kann. Dies ist etwa folgendermaßen möglich: Es sollen $n = 5$ Kugeln (die die Stellen repräsentieren) auf $r = 10$ Urnen (die die Ziffern repräsentieren) aufgeteilt werden. Die Anzahl der Kugeln pro Urne gibt an, wie oft diese Ziffer in der aufsteigenden Telefonnummer vorkommt. Die Urnenbelegung $(0, 0, 2, 0, 0, 1, 1, 1, 0, 0)$ ergibt somit die aufsteigende Telefonnummer 22567. Man kann erkennen, dass jede solche Urnenbelegung genau eine aufsteigende Telefonnummer repräsentiert. Mit Satz 3.8 gibt es genau $\binom{10+5-1}{5} = 2002$ solcher Telefonnummern. Die Repräsentation der Anforderung "aufsteigend sortiert" erfolgt somit implizit über die Reihenfolge der Urnen. \square

Beispiel 3.19 Eine genügend oft stetig differenzierbare Funktion $f : \mathbb{R}^r \rightarrow \mathbb{R}$ soll n -mal partiell abgeleitet werden. Eine partielle Ableitung n -ter Ordnung ist definiert als

$$\frac{\partial^n f}{\partial x_1^{\alpha_1} \cdots \partial x_r^{\alpha_r}} \quad \text{wobei } \alpha_1 + \cdots + \alpha_r = n,$$

d.h., f wird nach allen Koordinatenrichtungen zusammengezählt n -mal abgeleitet. Wieviele partielle Ableitungen n -ter Ordnung gibt es?

Lösung: Für eine Funktion $f : \mathbb{R}^r \rightarrow \mathbb{R}$ gibt $\binom{n+r-1}{n}$ partielle Ableitungen n -ter Ordnung, wie aus Satz 3.8 hervorgeht. \square

Kapitel 4

Bedingte Wahrscheinlichkeit

Bis jetzt haben wir immer den relativ einfachen Fall betrachtet, dass ein Ereignis E des Ereignisraums Ω in r paarweise disjunkte Teilereignisse A_i aufgespalten werden kann. Bei Laplace-Experimenten ergibt sich dann die W. $P(E)$ als

$$P(E) = \frac{|E|}{|\Omega|} = \frac{|A_1| + \dots + |A_r|}{|\Omega|}.$$

Im Folgenden werden wir sehen, dass in vielen Anwendungen der W.Rechnung die paarweise Disjunktheit der Teilereignisse A_i nicht gegeben ist.

Beispiel 4.1 In der Bevölkerung einer Kleinstadt mit 10170 Einwohnern gibt es 5280 Frauen, von denen 23 farbenblind sind. Von den 4890 Männern sind nur 10 farbenblind. Wie groß ist die W., dass eine zufällig ausgewählte Person dieser Stadt farbenblind ist? Wie ändert sich die W., wenn die ausgewählte Person eine Frau ist?

Lösung: Die W., eine der 33 farbenblinden Personen zufällig auszuwählen, beträgt $\frac{33}{10170} = 0.0032$. Unter den Frauen beträgt ist diese W. $\frac{23}{5280} = 0.0043$. \square

Beispiel 4.2 In einer Urne befinden sich ein weißer und zwei rote Bälle. Einer der Bälle wird zufällig gezogen. Wenn es der weiße ist, wird er zusammen mit einem weiteren weißen wieder zurückgelegt; wenn es einer der roten ist, wird er zusammen mit zwei weiteren roten wieder zurückgelegt. Anschließend wird noch ein Ball gezogen. Was ist die W., dass der zweite gezogene Ball rot ist (wenn das Resultat der ersten Ziehung nicht bekannt ist)?

Lösung: Wir bezeichnen mit R_1 das Ereignis "rot bei erster Ziehung", mit R_2 "rot bei zweiter Ziehung" und mit W_1 "weiß bei erster Ziehung". Die W. von R_2 setzt sich aus den W. der disjunkten Ereignisse $R_1 \cap R_2$ und $W_1 \cap R_2$ zusammen, kann also durch

$$P(R_2) = P(R_1 \cap R_2) + P(W_1 \cap R_2)$$

berechnet werden. Dabei ist $P(R_1 \cap R_2) = \frac{2}{3} \cdot \frac{4}{5}$, da $P(R_1) = \frac{2}{3}$ ist, und dann für die zweite Ziehung vier von fünf Bällen in der Urne rot sind. Mit einem analogen Argument kann man nachrechnen, dass $P(W_1 \cap R_2) = \frac{1}{3} \cdot \frac{2}{4}$ ist. Damit ergibt sich für R_2 die W. $P(R_2) = \frac{2}{3} \cdot \frac{4}{5} + \frac{1}{3} \cdot \frac{2}{4} = \frac{7}{10}$. Ohne weitere Rechnung erhält man sofort die W. von "weiß bei zweiter Ziehung" als $1 - \frac{7}{10}$. \square

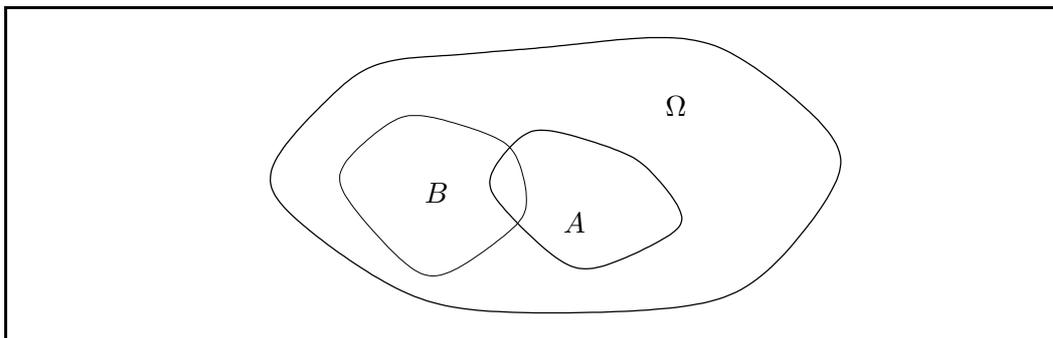


Abbildung 4.1: Illustration zum Verständnis der bedingten W.

Im letzten Beispiel haben wir W. wie $P(R_1 \cap R_2)$ “händisch”, also ohne formalen Apparat ausgearbeitet. Wir verwenden die Überlegungen, die zur Lösung des obigen Beispiels geführt haben, für folgende Definition.

Definition 4.1 (Bedingte Wahrscheinlichkeit)

Gegeben seien zwei Ereignisse A und B , wobei $P(A) > 0$ gefordert ist. Die W.

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

heißt *bedingte Wahrscheinlichkeit von B unter A* und gibt an, wie sich die W. von B verändert, wenn schon bekannt ist, dass das Ereignis A eingetreten ist. Durch Umformen erhält man einen Ausdruck für die W. des gemeinsamen Eintretens zweier Ereignisse:

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

Die letzte Gleichheit ergibt sich aus daraus, dass $P(A \cap B) = P(B \cap A)$ ist.

Der Begriff der bedingten W. lässt sich wie in Abbildung 4.1 gezeigt veranschaulichen. Entspricht die “normale” W. von B dem Verhältnis der Flächen von B und Ω (wie es bei Laplace-Experimenten der Fall ist), so entspricht die bedingte W. von B unter A dem Verhältnis der Fläche von $B \cap A$ zu A . Mit dem neuen Begriff können die letzten beiden Beispiele wie folgt geschrieben werden.

Beispiel 4.3 (nochmals Beispiel 4.1) Sei A = “Frau ausgewählt” und B = “farbenblinde Person ausgewählt”. Dann ist

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{23}{10170}}{\frac{5280}{10170}} = \frac{23}{5280}. \quad \square$$

Beispiel 4.4 (nochmals Beispiel 4.2) Mit gleichen Bezeichnungen wie oben ist

$$P(R_2) = P(R_1 \cap R_2) + P(W_1 \cap R_2)$$

$$\begin{aligned}
&= P(R_2|R_1)P(R_1) + P(R_2|W_1)P(W_1) = \\
&= \frac{4}{5} \cdot \frac{2}{3} + \frac{2}{4} \cdot \frac{1}{3} = \frac{7}{10}. \quad \square
\end{aligned}$$

Einige der elementarsten Begriffe und Sätze der W.Rechnung leiten sich von Begriff der bedingten W. ab. Die folgende Definition nimmt dabei eine zentrale Rolle ein.

Definition 4.2 (Unabhängigkeit von Ereignissen)

Zwei Ereignisse A und B heißen *unabhängig*, wenn

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

gilt.

Mit der Definition der bedingten W. folgt sofort, dass A und B genau dann unabhängig sind, wenn $P(B|A) = P(B)$, wenn also Kenntnis von Eintreten von A keine Auswirkung auf die W. von B hat.

In allen Beispielen vor diesem Kapitel sind nur unabhängige Ereignisse vorgekommen; so stellen die Ergebnisse der einzelnen Wiederholungen eines Experiments (z.B. Münzen oder Würfel werfen) unabhängige Ereignisse dar.

Beispiel 4.5 Die W. des Ereignisses A = “Doppelsechs bei zweimaligem Werfen eines Würfels” ist leicht aus der W. des Ereignisses B = “Sechs bei einmaligem Werfen eines Würfels” zu errechnen:

$$P(A) = P(B) \cdot P(B) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}. \quad \square$$

Beispiel 4.6 Bei einem Kartenspiel sind die Ereignisse A = “Ass ziehen” und B = “Herz ziehen” unabhängig. Dies ist nicht nur anschaulich klar, sondern kann auch mit Definition 4.2 nachgerechnet werden. So ist $P(A) = \frac{4}{52}$, $P(B) = \frac{13}{52}$ und $P(A \cap B)$ (W., die Herz-Ass zu ziehen) ist

$$P(A \cap B) = \frac{1}{52} = \frac{4}{52} \cdot \frac{13}{52} = P(A) \cdot P(B). \quad \square$$

In den folgenden Beispielen soll der Unterschied zwischen *disjunkten Ereignissen* A und B (also mit $P(A \cap B) = \emptyset$) und *unabhängigen Ereignissen* (also mit $P(A \cap B) = P(A) \cdot P(B)$) nochmals betont werden.

Beispiel 4.7 Wir betrachten bei einem Kartenspiel die beiden Ereignisse A = “Herz ziehen” und B = “schwarze Karte ziehen”. Da es keine schwarzen Herzkarten gibt, sind A und B disjunkt. Es ist $P(B) = \frac{1}{2}$, aber $P(B|A) = 0$, somit sind A und B abhängig.

Disjunkte Ereignisse können also nicht unabhängig sein: Wenn $P(A \cap B) = 0$ ist, dann ist $P(A \cap B) \neq P(A) \cdot P(B)$, außer natürlich im uninteressanten Fall, dass A oder B unmögliche Ereignisse sind.

Die Umkehrung dieser Aussage gilt allerdings nicht: Wenn Ereignisse *nicht* disjunkt sind, können sie sowohl abhängig oder auch unabhängig sein. Mit den Zahlen aus Beispiel 2.13, nämlich $P(A) = 0.3$, $P(B) = 0.5$, und $P(A \cap B) = 0.1$ sind A und B nicht disjunkt, aber auch nicht unabhängig. Mit etwas anderen Zahlen, nämlich $P(A) = 0.3$, $P(B) = 0.8$, und $P(A \cap B) = 0.24$ sind die beiden Ereignisse ebenfalls wieder nicht disjunkt, aber unabhängig. \square

Beispiel 4.8 Eine neues Sicherheitssystem zum Passagier-Check am Flughafen bestehe aus drei unabhängigen Komponenten, die hintereinander passiert werden müssen. Die Ereignisse A , B und C seien, dass die jeweiligen Komponenten ausfallen; diese Ereignisse seien unabhängig voneinander. Die W. dieser Ereignisse seien $P(A) = 0.01$, $P(B) = 0.02$, und $P(C) = 0.03$. Die W., dass das ganze System fehlerfrei funktioniert, ist somit

$$P(\bar{A}) \cdot P(\bar{B}) \cdot P(\bar{C}) = 0.99 \cdot 0.98 \cdot 0.97 = 0.941.$$

Die W., dass genau *eine* Komponente ausfällt, ist

$$P(A) \cdot P(\bar{B}) \cdot P(\bar{C}) + P(\bar{A}) \cdot P(B) \cdot P(\bar{C}) + P(\bar{A}) \cdot P(\bar{B}) \cdot P(C) = \\ 0.01 \cdot 0.98 \cdot 0.97 + 0.99 \cdot 0.02 \cdot 0.97 + 0.99 \cdot 0.98 \cdot 0.03 = 0.058.$$

An dieser letzten Berechnung kann man schön die Multiplikation der W. für unabhängige Ereignisse und die Addition der W. für disjunkte Ereignisse ablesen. \square

Satz 4.1 (Multiplikationsregel für bedingte Wahrscheinlichkeit) Für n beliebige Ereignisse A_1, \dots, A_n gilt

$$P(A_1 \cap A_2 \dots \cap A_n) = P(A_1) \cdot P(A_2 | A_1) \cdot P(A_3 | A_1 \cap A_2) \dots \\ \dots P(A_n | A_1 \cap \dots \cap A_{n-1}).$$

Beweis Nach Definition der bedingten W. kann die rechte Seite im Satz auch als

$$P(A_1) \cdot \frac{P(A_2 \cap A_1)}{P(A_1)} \cdot \frac{P(A_3 \cap A_2 \cap A_1)}{P(A_2 \cap A_1)} \dots \frac{P(A_1 \cap \dots \cap A_n)}{P(A_1 \cap \dots \cap A_{n-1})}$$

geschrieben werden. Wie man sehen kann, kürzt sich bis auf den Faktor $P(A_1 \cap \dots \cap A_n)$ alles weg, womit der Satz gezeigt ist. \square

Folgendes Beispiel illustriert anschaulich, dass sich hinter den manchmal etwas umständlichen Formulierungen der W.Rechnung einfache Zusammenhänge verbergen.

Beispiel 4.9 Aus einem Kartenspiel mit 52 Karten wird viermal ohne Zurücklegen gezogen. Was ist die W., dass alle vier gezogenen Karten Herz sind?

Lösung: Ohne theoretische Überlegungen und Anwendung von Satz 4.1 kommt man zum korrekten Resultat von $\frac{13}{52} \frac{12}{51} \frac{11}{50} \frac{10}{49}$, da beim ersten Ziehen 13 der 52 Karten

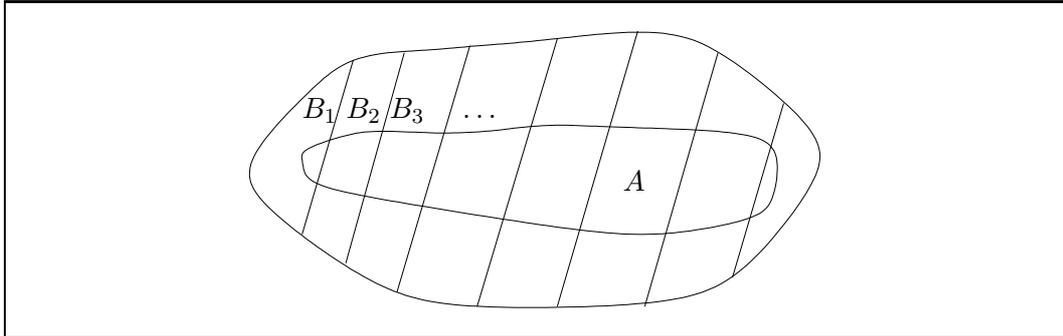


Abbildung 4.2: Illustration zum Satz von der totalen W.

Herz sind, beim zweiten Ziehen (wenn schon eine Herz gezogen wurde) nur mehr 12 der 51, usw.

Im theoretischen Modell von Satz 4.1 lässt sich dies wie folgt formulieren: Sei A_i für $i = 1, 2, 3, 4$ das Ereignis "Herz beim i -ten Ziehen". Gesucht ist dann $P(A_1 \cap A_2 \cap A_3 \cap A_4)$, also die W. von "Herz bei jedem Ziehen". Daraus ergibt sich wie zuerst

$$\begin{aligned} P\left(\bigcap_{i=1}^4 A_i\right) &= P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_2 \cap A_1) \cdot P(A_4|A_3 \cap A_2 \cap A_1) \\ &= \frac{13}{52} \cdot \frac{12}{51} \cdot \frac{11}{50} \cdot \frac{10}{49} = 0.0026. \end{aligned} \quad \square$$

Satz 4.2 (Satz von der totalen Wahrscheinlichkeit) Gegeben sei eine disjunkte Zerlegung B_1, B_2, \dots des Ereignisraums Ω (d.h., $B_i \cap B_j = \emptyset$ für $i \neq j$ und $\bigcup_{i=1}^{\infty} B_i = \Omega$) mit $P(B_i) > 0$ für alle i . Dann lässt sich die W. jedes Ereignisses A schreiben als

$$P(A) = \sum_{i=1}^{\infty} P(A|B_i)P(B_i).$$

Beweis Da $\sum_{i=1}^{\infty} P(A|B_i)P(B_i) = \sum_{i=1}^{\infty} P(A \cap B_i)$ ist, besagt der Satz nichts weiter, als dass sich die W. eines Ereignisses als Summe disjunkter Teilereignisse $A \cap B_i$ schreiben lässt. Dies ist mit der Skizze in Abbildung 4.2 anschaulich klar.

In Formel ausgedrückt lässt sich diese Tatsache argumentieren durch: $A \subseteq \Omega$, daher ist $A \cap \Omega = A$. Mit $\Omega = \bigcup_{i=1}^{\infty} B_i$ ist $A = A \cap \bigcup_{i=1}^{\infty} B_i$ und die gesuchte Gleichheit ergibt sich als

$$\begin{aligned} P(A) &= P\left(A \cap \bigcup_{i=1}^{\infty} B_i\right) = P\left(\bigcup_{i=1}^{\infty} A \cap B_i\right) = \sum_{i=1}^{\infty} P(A \cap B_i) \\ &= \sum_{i=1}^{\infty} P(A|B_i)P(B_i). \end{aligned}$$

Entscheidend ist hier, dass alle Mengen $A \cap B_i$ paarweise disjunkt sind. □

Beispiel 4.10 In einer Fabrik werden Waren auf drei Maschinen gefertigt; dabei produziert Maschine 1 40% der Waren, Maschine 2 35%, und Maschine 3 25%. Die Ausschussware (fehlerhafte Stücke), die die einzelnen Maschinen produzieren, verteilt sich wie folgt: Maschine 1 2%, Maschine 2 4%, Maschine 3 5%. Was ist die W., dass eine zufällig gezogene Ware defekt ist?

Lösung: Sei B_i für $i = 1, 2, 3$ das Ereignis “Ware wurde von Maschine i produziert”. Dann erfüllen die B_i die Bedingungen aus Satz 4.2: Sie sind disjunkt und ergeben zusammen den vollständigen Ereignisraum. Wir sind an der W. des Ereignisses $A =$ “Ware ist defekt” interessiert. Die Prozentzahlen der Ausschusswaren liefern gerade die Werte für $P(A|B_i)$ für $i = 1, 2, 3$. Damit ergibt sich unmittelbar aus Satz 4.2 die Lösung:

$$\begin{aligned} P(A) &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3) \\ &= 0.02 \cdot 0.4 + 0.04 \cdot 0.35 + 0.05 \cdot 0.25 = 0.0345 \end{aligned} \quad \square$$

Mit Hilfe eines einfachen Tricks (der aber fundamentale Wichtigkeit hat) kann man $P(A|B)$ aus $P(B|A)$ ausrechnen, wie der folgende Satz besagt.

Satz 4.3 (Satz von Bayes) Die bedingte W. von A unter B ergibt sich als

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

aus der bedingten W. von B unter A , wenn $P(B) > 0$ gilt.

Beweis Da $P(A \cap B) = P(B \cap A)$ ist, kann man

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

über die bedingte W. ausdrücken. Damit ergibt sich sofort

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

Meist wird der Satz von Bayes in Kombination mit dem Satz von der totalen W. angegeben. Dann ergibt sich

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_{i=1}^{\infty} P(A|B_i)P(B_i)},$$

wobei hier einfach B durch B_k und $P(A)$ wie in Satz 4.2 ersetzt wurde.

Beispiel 4.11 (Fortsetzung von Beispiel 4.10) Meist ist man an einer Umkehrung der Fragestellung von Beispiel 4.10 interessiert: Nicht die W., dass eine Ware defekt ist, sondern die W., dass eine defekte Ware von einer bestimmten Maschine produziert wird soll ermittelt werden. Dies ist mit der Kombination der Sätze von Bayes

und der totalen W. einfach möglich. So ergibt sich für die W., dass eine defekte Ware von Maschine 1 produziert wurde

$$P(B_1|A) = \frac{P(A|B_1)P(B_1)}{\sum_{i=1}^{\infty} P(A|B_i)P(B_i)} = \frac{0.02 \cdot 0.4}{0.0345} = 0.2319.$$

Analog dazu erhält man $P(B_2|A) = 0.4058$ und $P(B_3|A) = 0.3623$. □

Beispiel 4.12 Zur Identifizierung einer Infektionskrankheit wird ein Bluttest administriert, der bei 99% der infizierten Personen die korrekte Diagnose liefert. Dieser Test diagnostiziert allerdings auch 1% der nichtinfizierten Personen fälschlicherweise als Krankheitsträger.

Wenn bekannt ist, dass 0.5% der Bevölkerung infiziert ist, was ist dann die W., dass eine Person tatsächlich krank ist, wenn ihr Testresultat positiv ist?

Lösung: Es sei A das Ereignis "Person ist infiziert" und B das Ereignis "Test ist positiv". Gesucht ist somit die bedingte W. von A unter B . Dies ist mit dem Satz von Bayes unter Verwendung von \bar{A} als Komplement des Ereignisses A

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})} \\ &= \frac{\frac{99}{100} \frac{5}{1000}}{\frac{99}{100} \frac{5}{1000} + \frac{1}{100} \frac{995}{1000}} = 0.3322. \end{aligned}$$

Die Gleichheit $P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A})$ gilt wegen des Satzes von der totalen W., da A und \bar{A} eine disjunkte Zerlegung des Ereignisraumes sind. □

Kapitel 5

Zufallsvariable

Bis jetzt haben wir uns bei der Abstraktion eines Zufallsexperiments zu einem mathematischen Modell auf den Ereignisraum Ω und die W. Verteilung P auf Ω , also auf den W. Raum $(\Omega, \text{Pot}(\Omega), P)$ konzentriert. Tatsächlich sind aber bei einigen Experimenten nicht die Ausgänge der Experimente direkt interessant, sondern eine davon abgeleitete Größe—in manchen Fällen auch mehrere solche Größen.

5.1 Eindimensionale Zufallsvariable

Beispiel 5.1 Ein dreimaliges Werfen eines Würfels wird durch den Ereignisraum $\Omega = \{(n_1, n_2, n_3) \mid 1 \leq n_i \leq 6 \text{ für } i = 1, 2, 3\}$ beschreiben. Wenn man die Summe der Augenzahlen beobachten möchte, interessiert nur die Funktion

$$\begin{aligned} Z : \Omega &\rightarrow \mathbb{N}_0 \\ (n_1, n_2, n_3) &\mapsto Z(n_1, n_2, n_3) := n_1 + n_2 + n_3. \end{aligned} \quad \square$$

Beispiel 5.2 Beim zehnmaligen Werfen einer Münze ist der Ereignisraum $\Omega = \{(b_1, \dots, b_{10}) \mid b_i \in \{\text{KOPF}, \text{ZAHL}\} \text{ für } i = 1, \dots, 10\}$. Wenn nur die Anzahl des Ausganges KOPF beobachtet werden soll, so kann man sich auf den Wert der Funktion

$$\begin{aligned} Z : \Omega &\rightarrow \mathbb{N}_0 \\ (b_1, \dots, b_{10}) &\mapsto Z(b_1, \dots, b_{10}) := |\{j \mid b_j = \text{KOPF für } j = 1, \dots, 10\}| \end{aligned}$$

beschränken. □

In der mathematischen Abstraktion von Abbildung 2.2 entspricht das der Einführung eines weiteren Abstraktionsniveaus, wie in Abbildung 5.1 gezeigt. Die für den Beobachter interessanten Fragen über Ausgänge eines Experiments werden nicht mehr direkt durch Auswertung von Ω beantwortet, sondern erst nach einer weiteren Übertragung auf eine Zahlenmenge.

Mit den Eigenschaften dieser Übertragung werden wir uns in diesem Kapitel beschäftigen. Dafür benötigen wir als erstes eine genaue Definition des oben erwähnten zweiten Abstraktionsschrittes.

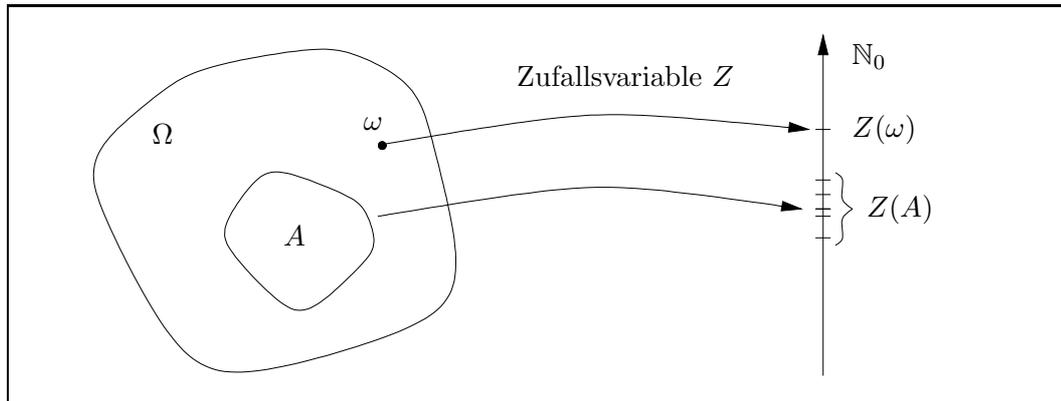


Abbildung 5.1: Übertragung eines Ereignisraumes Ω auf die natürlichen Zahlen.

Definition 5.1 (Zufallsvariable)

Sei $(\Omega, \text{Pot}(\Omega), P)$ ein diskreter W.Raum. Eine *Zufallsvariable* (kurz ZV) Z ist eine Funktion

$$Z : \Omega \rightarrow \mathbb{N}_0.$$

Zufallsvariable werden allgemein mit Großbuchstaben wie X, Y, Z bezeichnet. Ein W.Maß P auf Ω induziert eindeutig ein W.Maß P_Z auf \mathbb{N}_0 , das *Verteilung von Z* genannt wird.

Zufallsvariable übertragen also den diskreten Ereignisraum Ω auf die natürlichen Zahlen, wie in Abbildung 5.1 dargestellt. Dabei ist die Wahl der natürlichen Zahlen als Bildbereich von Z nicht verpflichtend; jede andere diskrete Menge ist ebenso geeignet. Erweiterungen auf den stetigen Fall (wenn der Bildbereich von Z \mathbb{R} und nicht \mathbb{N}_0 ist) werden im Kapitel 8 behandelt. Zufallsvariable bieten die Möglichkeit, Charakteristika einzelner Experimente herauszufiltern und so ein gemeinsames Verhalten verschiedener Experimente zu betrachten. Was das genau bedeutet, werden wir in Kapitel 6 sehen.

Durch eine ZV soll ein W.Maß P auf Ω in geeigneter Weise auf ein W.Maß P_Z auf \mathbb{N}_0 übertragen werden. Dies ist auf sehr einfache Weise möglich. Um zu sehen, dass P_Z eindeutig und vollständig durch P gegeben ist, führen wir zuerst folgende Schreibweisen für eine Zahl $k \in \mathbb{N}_0$ und eine Teilmenge $A \in \text{Pot}(\mathbb{N}_0)$, also $A \subseteq \mathbb{N}_0$ ein:

$$\begin{aligned} \{Z = k\} &:= \{\omega \in \Omega \mid Z(\omega) = k\} \\ \{Z \in A\} &:= \{\omega \in \Omega \mid Z(\omega) \in A\} \end{aligned}$$

Beim genauen Betrachten dieser Schreibweisen sollte klar werden, dass Z keine Bijektion zwischen Ω und \mathbb{N}_0 sein muss: Es ist meist der Fall, dass mehrere $\omega \in \Omega$ durch Z auf ein $k \in \mathbb{N}_0$ abgebildet werden.

Definition 5.2 (Verteilung einer Zufallsvariablen)

Für ein beliebiges $A \subseteq \mathbb{N}_0$ ist die *Verteilung* P_Z einer ZV $Z : \Omega \rightarrow \mathbb{N}_0$ definiert durch

$$P_Z(A) := P(\{Z \in A\}) = \sum_{k \in A} P(\{Z = k\}) = \sum_{\substack{\omega \in \Omega \\ Z(\omega) \in A}} P(\omega)$$

Die Schreibweise $P(\{Z \in A\})$ ist natürlich formal korrekt und weist darauf hin, dass die W. von ZV eigentlich auf dem Definitionsbereich dieser ZV definiert ist. Wir werden aber im Folgenden eine einfachere Notation von ZV verwenden.

Definition 5.3 (Notation von Zufallsvariablen)

Für zwei ZV $X, Y : \Omega \rightarrow \mathbb{N}_0$ definieren wir abkürzend

$$P_X(k) := P(X = k) := P(\{X = k\}),$$

lassen also die Mengenklammern weg. Weiters sei

$$P(X = k \wedge Y = j) := P(X = k, Y = j) := P(\{X = k\} \cap \{Y = j\}).$$

Die letzte Definition des gemeinsamen Auftretens von Werten zweier ZV lässt sich auf beliebig viele ZV verallgemeinern; ebenso kann der Boolesche Junktoren \vee durch die Mengenvereinigung definiert werden. Wir werden diese Notation speziell in Abschnitt 6.6 benötigen.

Man kann nachprüfen, dass die oben definierte Verteilung von ZV wirklich ein W.Maß ist.

Satz 5.1 Die durch das W.Maß P induzierte Verteilung P_Z von Z ist ein W.Maß im Sinn von Definition 2.2, erfüllt also die Bedingungen (normiert) und (σ -additiv).

Beweis als Übung. □

In Analogie zu Definition 2.3 gibt es für die W. von einzelnen natürlichen Zahlen (den zu Elementarereignissen in Ω vergleichbaren Objekten) einen eigenen Begriff.

Definition 5.4 (W.Funktion einer Zufallsvariable)

Für eine ZV $Z : \Omega \rightarrow \mathbb{N}_0$ nennt man die Funktion

$$\begin{aligned} f_Z : \mathbb{N}_0 &\rightarrow [0, 1] \\ k &\mapsto f_Z(k) := P(Z = k) \end{aligned}$$

die *Wahrscheinlichkeitsfunktion* von Z .

Mit der Notation aus Definition 5.3 ist klar, dass $f_Z(k) = P_Z(k)$ ist. Wir werden im Folgenden beide Schreibweisen verwenden. Mit der Definition einer Wahrscheinlichkeitsfunktion, die nur auf Elementen des Wertebereichs einer ZV definiert ist (und nicht auch auf Mengen) vermeidet man die Mehrdeutigkeiten, die die Notation mit dem Symbol P bedingt. Bei den diskreten Verteilungen, die hier angesprochen werden, stellt dies noch keinen Vorteil dar, da P_Z und f_Z identische Werte liefern. Wir werden aber bei den in Kapitel 8 diskutierten *stetigen Verteilungen* sehen, dass dann P_Z und f_Z *unterschiedliche* Funktionen sind. Es ist sinnvoll, auf diese dann auftretende Unterscheidung bereits bei diskreten ZV hinzuweisen.

Mit obigen Definitionen kann man nun eine Verteilung P_Z auf \mathbb{N}_0 berechnen, wenn nur der W.Raum $(\Omega, \text{Pot}(\Omega), P)$ und die ZV Z bekannt sind. Wir abstrahieren also $(\Omega, \text{Pot}(\Omega), P)$ zu einem neuen W.Raum $(\mathbb{N}_0, \text{Pot}(\mathbb{N}_0), P_Z)$.

Als abschließende Bemerkung zu dieser theoretischen Einführung soll speziell erwähnt werden, dass man nur am Zufalls-Verhalten einer ZV Z interessiert ist. Diese Information ist vollständig durch P_Z gegeben, sodass man sich nur mehr für P_Z und nicht mehr für Z selbst interessiert.

Es folgen einige elementare Beispiele, um den Begriff der ZV vertraut zu machen.

Beispiel 5.3 (vergleiche Beispiel 5.1) Ein Würfel wird dreimal geworfen. Was ist die Verteilung der Augensummen der drei Würfe, was also ist $P_Z(k)$ für $k = 3, \dots, 18$?

Lösung: Wir betrachten die drei Schritte, die sich in den theoretischen Überlegungen ergeben haben, für dieses Beispiel sehr ausführlich. Mit einiger Übung können diese auch kürzer behandelt werden.

- (1) (Erstellen des Modells) Der Ereignisraum Ω ist

$$\Omega = \{(n_1, n_2, n_3) \mid 1 \leq n_i \leq 6 \text{ für } i = 1, 2, 3\},$$

die W. jedes Elementarereignisses $\omega = (n_1, n_2, n_3)$ ist $\frac{1}{6^3}$. Das W.Maß auf Ω ist somit die Gleichverteilung, und für beliebiges $A \subseteq \Omega$ ist $P(A) = |A|/|\Omega|$.

- (2) (Definition der ZV) Wenn wir uns nur für die Summe der Augenzahlen interessieren, genügt es, die Werte der ZV

$$\begin{aligned} Z : \Omega &\rightarrow \mathbb{N}_0 \\ (n_1, n_2, n_3) &\mapsto Z(n_1, n_2, n_3) := n_1 + n_2 + n_3 \end{aligned}$$

zu betrachten.

- (3) (Bestimmung der Verteilung von Z) Die Verteilung P_Z ist durch ihre W.Funktion f_Z eindeutig bestimmt. Diese W.Funktion gibt an, wie wahrscheinlich einzelne Augensummen k sind. Dabei sei A_k das Ereignis "die Augensumme ist k ", also $A_k = \{(n_1, n_2, n_3) \mid n_1 + n_2 + n_3 = k\}$. Die W.Funktion ist damit nach Definition 5.4

$$f_Z(k) = P(Z = k) = P(A_k) = \frac{|A_k|}{|\Omega|},$$

Zur Angabe von $f_Z(k)$ muss man also nur die Mächtigkeit dieser Mengen abzählen. Dabei kann man wie in Beispiel 3.10 vorgehen und kommt so zu folgendem Ergebnis (die W.Funktion ist nur für k zwischen 3 und 18 ungleich null):

k	$f_Z(k)$	k	$f_Z(k)$	k	$f_Z(k)$	k	$f_Z(k)$
3	1/216	7	15/216	11	27/216	15	10/216
4	3/216	8	21/216	12	25/216	16	6/216
5	6/216	9	25/216	13	21/216	17	3/216
6	10/216	10	27/216	14	15/216	18	1/216

□

Beispiel 5.4 Eine Münze wird n -mal geworfen. Unter einem *Run* versteht man eine Folge von gleichen Ausgängen eines Experiments; so enthält etwa für $n = 10$ die Folge ZAHL KOPF ZAHL KOPF KOPF ZAHL ZAHL ZAHL ZAHL ZAHL fünf Runs: drei der Länge eins und jeweils einen der Länge zwei und fünf. Wir interessieren uns für die Anzahl der Runs: Wie wahrscheinlich ist es, dass bei n -maligem Werfen der Münze im Ergebnis gerade k Runs auftreten?

Lösung: Gefragt ist also die Verteilung der ZV, die die Anzahl der Runs in einer Folge von Münzwürfen angibt. Der Ereignisraum ist

$$\Omega = \{(b_1, \dots, b_n) \mid b_i \in \{\text{KOPF}, \text{ZAHL}\} \text{ für } i = 1, \dots, n\}.$$

Da jede Folge (b_1, \dots, b_n) gleich wahrscheinlich ist, ist das W.Maß auf Ω die Gleichverteilung. Dabei ist $|\Omega| = 2^n$.

Die ZV Z ist gegeben durch

$$Z : \Omega \rightarrow \mathbb{N}_0$$

$$(b_1, \dots, b_n) \mapsto Z(b_1, \dots, b_n) := \text{Anzahl der Runs in } (b_1, \dots, b_n).$$

Gefragt ist P_Z , also die Verteilung von Z . Diese lässt sich für $k = 1, \dots, n$ schreiben als

$$P_Z(k) = P(\{(b_1, \dots, b_n) \mid \text{es gibt } k \text{ Runs in } b_1, \dots, b_n\}).$$

Die Problemstellung lässt sich wie im letzten Beispiel darauf reduzieren, die Mächtigkeit einer Menge zu berechnen (und dann noch durch die Mächtigkeit von Ω zu dividieren). Hier müssen wir die Menge aller Folgen (b_1, \dots, b_n) mit k Runs abzählen.

Man verwendet dazu am besten die Tatsache, dass k Runs in b_1, \dots, b_n genau $k - 1$ Vorzeichenwechsel entsprechen. Wir zählen dann die Menge aller Folgen mit $k - 1$ Vorzeichenwechseln folgendermaßen ab: Das erste Folgenglied b_1 kann auf zwei Arten festgelegt werden (entweder KOPF oder ZAHL). Unter den verbleibenden $n - 1$ Stellen müssen dann $k - 1$ Vorzeichenwechsel sein. Dies entspricht einem Standardmodell aus Abschnitt 3.5: Es sollen $k - 1$ Elemente aus einer $(n - 1)$ -elementigen Menge ausgewählt werden. Dafür gibt es $\binom{n-1}{k-1}$ Möglichkeiten. Zusammenfassend ergibt sich dann

$$P_Z(k) = \frac{|\{(b_1, \dots, b_n) \mid \text{es gibt } k \text{ Runs in } b_1, \dots, b_n\}|}{|\Omega|} = \frac{2 \binom{n-1}{k-1}}{2^n} = \frac{\binom{n-1}{k-1}}{2^{n-1}}. \quad \square$$

Wie wir in den letzten beiden Beispielen gesehen haben, lässt sich die Lösung völlig losgelöst vom ursprünglichen Experiment über die Verteilung von P_Z beschreiben. Im Zuge dieser Abstraktion von P in Ω zu P_Z in \mathbb{N}_0 können verschiedene Experimente über die ihnen gemeinsame Verteilung charakterisiert werden. Die gebräuchlichsten besitzen eigene Namen; einige werden im nächsten Kapitel genauer betrachtet.

5.2 Mehrdimensionale Zufallsvariable

Wenn bei einem Zufallsexperiment nicht nur eine sondern mehrere Messzahlen des Experiments betrachtet werden sollen, so lassen sich diese Beobachtungen über mehrere ZV beschreiben, die als Komponenten eines ZV-Vektors betrachtet werden. Solche mehrdimensionale ZV spielen für gemeinsame Verteilungen und den Begriff der Unabhängigkeit von ZV eine große Rolle; dies wird in den Abschnitten 6.6 und 6.7 näher erläutert.

Beispiel 5.5 Bei zufällig ausgewählten österreichischen Familien werden die Anzahl der Söhne und Töchter in der Familie gezählt. Der Ereignisraum Ω ist die Menge aller österreichischen Familien; für eine zufällig bestimmte Familie ω gibt $Z_1 : \Omega \rightarrow \mathbb{N}_0$ die Anzahl der Söhne und $Z_2 : \Omega \rightarrow \mathbb{N}_0$ die Anzahl der Töchter an. Beide ZV können als Komponenten einer ZV $Z = (Z_1, Z_2) : \Omega \rightarrow \mathbb{N}_0^2$ aufgefasst werden.

Unter 10000 Familien seien die folgenden (fiktiven) Zahlen von Söhnen und Töchtern beobachtet worden:

		Söhne				
		0	1	2	3	4
Töchter	0	3817	1611	375	127	65
	1	1657	939	199	96	8
	2	435	228	134	3	2
	3	115	89	7	4	2
	4	78	7	1	0	1

Aus dieser Tabelle können z.B. folgende Schlüsse gezogen werden, wobei die W. relative Häufigkeiten sind: $P(Z_1 = 0) = 0.6102$ durch Aufsummieren der ersten Spalte; $P(Z_2 = 3) = 0.0217$ durch Aufsummieren der dritten Zeile; und $P(Z_1 = 0 \wedge Z_2 = 3) = 115$ durch direktes Ablesen aus der Tabelle. \square

Beispiel 5.6 Wir betrachten einen Würfel, der zweimal geworfen wird. Auf dem Ereignisraum $\Omega = \{(n_1, n_2) \mid n_i \in \{1, \dots, 6\}\}$ seien zwei ZV X und Y definiert. Dabei gebe für ein Elementarereignis (n_1, n_2) die ZV $X : \Omega \rightarrow \mathbb{N}_0$ die Augensumme $n_1 + n_2$ an, und $Y : \Omega \rightarrow \mathbb{N}_0$ den Wert $\max\{n_1, n_2\}$. Die W. des gemeinsamen Auftretens von Werten von X und Y kann folgender Tabelle entnommen werden:

		X										
		2	3	4	5	6	7	8	9	10	11	12
Y	1	$\frac{1}{36}$	0	0	0	0	0	0	0	0	0	0
	2	0	$\frac{2}{36}$	$\frac{1}{36}$	0	0	0	0	0	0	0	0
	3	0	0	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	0	0	0	0	0	0
	4	0	0	0	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	0	0	0	0
	5	0	0	0	0	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	0	0
	6	0	0	0	0	0	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

\square

Kapitel 6

Verteilungen von diskreten Zufallsvariablen

Zufallsvariable erlauben das Abstrahieren (und gegebenenfalls Vereinfachen) von W.Räumen $(\Omega, \text{Pot}(\Omega), P)$ zu $(\mathbb{N}_0, \text{Pot}(\mathbb{N}_0), P_Z)$; die Verteilung P_Z wird dann losgelöst von den unterliegenden W.Räumen $(\Omega, \text{Pot}(\Omega), P)$ betrachtet. Dies erlaubt uns etwa, die W. des k -fachen Eintretens eines Experiment-Ausgangs beim n -fachen Wiederholen zu bestimmen, und zwar unabhängig davon, ob es sich bei dem Experiment etwa um Würfeln, Münzwurfen oder ein anderes Zufallsexperiment handelt.

Durch die Vielzahl von unterliegenden W.Räumen und Beobachtungen der Ausgänge ergibt sich eine große Menge von interessanten Verteilungen; wir werden hier nur einige der einfachsten behandeln.

6.1 Gleichverteilung

Wie sich mit einer kurzen Überlegung (siehe unten) erkennen lässt, ist \mathbb{N}_0 kein Laplace-Raum, es gibt also auf den natürlichen Zahlen keine Gleichverteilung. Sie wird trotzdem hier erwähnt, weil es auf jeder *endlichen* Teilmenge $A \subseteq \mathbb{N}_0$ eine Gleichverteilung gibt; für jedes $k \in A$ ist die W.Funktion gegeben als

$$f_Z(k) = \frac{1}{|A|}.$$

Die Unmöglichkeit einer Gleichverteilung auf ganz \mathbb{N}_0 lässt sich folgendermaßen begründen: Für eine Gleichverteilung müsste für jedes $k \in \mathbb{N}_0$ gelten, dass $f_Z(k) = c$ ist (für ein fixes, noch zu bestimmendes c). Wegen der Normiertheits-Bedingung $P(\Omega) = 1$ aus Definition 2.2 muss andererseits $\sum_{k \geq 0} f_Z(k) = \sum_{k \geq 0} c = 1$ gelten; dies ist für kein $c \in \mathbb{R}$ erfüllt. Somit kann es keine Gleichverteilung auf \mathbb{N}_0 geben.

6.2 Bernoulliverteilung

Ein Zufallsexperiment habe zwei mögliche Ausgänge, die durch die Elementarereignisse ω_1 bzw. ω_2 beschrieben werden. Dabei gelte $P(\omega_1) = 1 - p$ und $P(\omega_2) = p$. Eine Zufallsvariable

$$Z : \Omega \rightarrow \mathbb{N}_0$$

$$\omega_1 \mapsto 0$$

$$\omega_2 \mapsto 1$$

weise diesen Elementarereignissen natürliche Zahlen zu; es gilt $P_Z(0) = 1 - p$ und $P_Z(1) = p$.

Definition 6.1 (Bernoulliverteilung)

Eine Verteilung P_Z mit W.Funktion $f_Z : \mathbb{N}_0 \rightarrow [0, 1]$, für die

$$f_Z(k) = \begin{cases} 1 - p & \text{für } k = 0 \\ p & \text{für } k = 1 \\ 0 & \text{sonst} \end{cases}$$

gilt, heißt *Bernoulliverteilung*.

Die Bernoulliverteilung ist für sich allein genommen recht uninteressant; wenn aber Experimente mit Bernoulliverteilung n mal unabhängig durchgeführt werden, ist die Anzahl des Auftretens jedes der beiden Ereignisse binomialverteilt. Dies werden wir im nächsten Abschnitt genauer sehen.

6.3 Binomialverteilung

Beispiel 6.1 Ein Zufallsexperiment mit den beiden möglichen Ausgängen A und B wird n mal unabhängig wiederholt. Dabei ist $P(A) = p$ und $P(B) = 1 - p$. Was ist die W., dass unter den n Versuchen genau k mal das Ereignis A eintritt?

Lösung: Wir formulieren zuerst das mathematische Modell. Der Ereignisraum ist

$$\Omega = \{(b_1, \dots, b_n) \mid b_i \in \{A, B\} \text{ für } i = 1, \dots, n\}.$$

Die ZV Z ist gegeben durch

$$Z : \Omega \rightarrow \mathbb{N}_0 \\ (b_1, \dots, b_n) \mapsto Z(b_1, \dots, b_n) := |\{j \mid b_j = A \text{ für } j = 1, \dots, n\}|,$$

also durch die Anzahl der Experimente, bei denen das Ereignis A eingetreten ist. Gesucht ist die Verteilung P_Z von Z . Im Gegensatz zu den Beispielen 5.3 und 5.4 ist das W.Maß auf Ω *nicht* die Gleichverteilung, da ja nicht jede Folge von Ausgängen gleich wahrscheinlich ist.

Es ist zu erkennen, dass $P_Z(0) = (1 - p)^n$ und $P_Z(n) = p^n$ ist, da dann A gerade 0 oder n mal eintritt. Man könnte versucht sein, auch $P_Z(1) = p(1 - p)^{n-1}$ und allgemein $P_Z(k) = p^k(1 - p)^{n-k}$ als richtig anzunehmen. Dies ist aber noch nicht vollständig korrekt. Die Zahl $p(1 - p)^{n-1}$ etwa gibt an, wie wahrscheinlich *eine* Folge von Ausgängen B, B, A, \dots, B ist, bei denen A nur einmal auftritt. Insgesamt gibt es aber n solcher Folgen, da A ja an jeder der n Stellen eintreten kann.

Ebenso ist die W. *einer* Versuchsreihe, bei der k mal das Ereignis A und $(n - k)$ mal das Ereignis B eintritt, gleich $p^k(1 - p)^{n-k}$. Wir interessieren uns aber für die W.

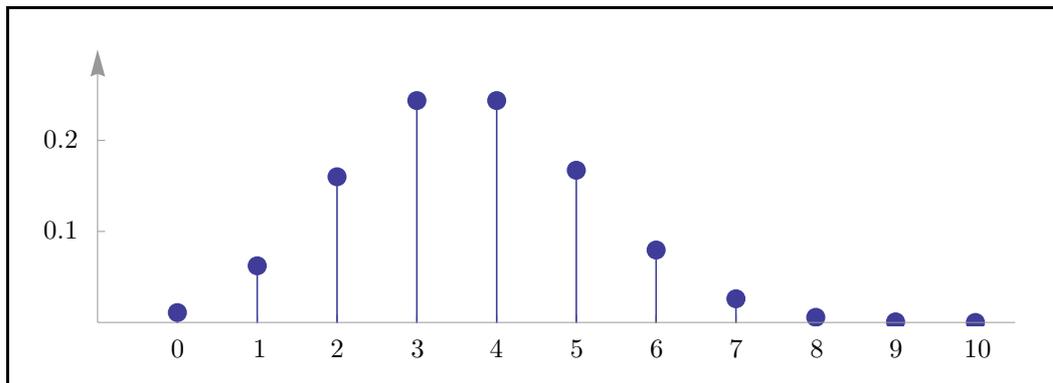


Abbildung 6.1: Verteilung der binomialverteilten Zufallsvariable aus Beispiel 6.2.

aller Versuchsreihen, bei denen k mal das Ereignis A eintritt. Dies entspricht einem Standardmodell aus Abschnitt 3.5: Aus n möglichen Stellen sollen die k Stellen ausgewählt werden, an denen A auftritt. Für diese Auswahl gibt es $\binom{n}{k}$ Möglichkeiten. Somit ist die gesuchte Verteilung von Z gegeben durch

$$P_Z(k) = \binom{n}{k} p^k (1-p)^{n-k}. \quad \square$$

Definition 6.2 (Binomialverteilung)

Eine Verteilung P_Z mit W.Funktion $f_Z : \mathbb{N}_0 \rightarrow [0, 1]$, für die

$$f_Z(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{für } k = 0, \dots, n \\ 0 & \text{sonst} \end{cases}$$

gilt, heißt *Binomialverteilung* mit den Parametern n und p . Für die Binomialverteilung mit diesen Parametern schreiben wir auch $\text{Bin}(n, p)$.

Beispiel 6.2 Aus einer Urne mit vier roten und sieben schwarzen Kugeln wird zehnmal mit Zurücklegen gezogen. Was ist die W., dabei genau k rote Kugeln zu ziehen?

Lösung: Es sei $Z : \Omega \rightarrow \mathbb{N}_0$ die ZV, die einer Versuchsreihe die Anzahl der gezogenen roten Kugeln zuweist. Weil die gezogenen Kugeln wieder zurückgelegt werden, sind alle zehn Versuche gleich. Bei jedem einzelnen Versuch ist die W., eine rote Kugel zu ziehen, gleich $\frac{4}{11}$. Damit lässt sich die Argumentation von Beispiel 6.1 sofort anwenden (wo wir ein viel allgemeineres Problem gelöst haben). Wir sehen, dass Z $\text{Bin}(10, \frac{4}{11})$ -verteilt ist. Demnach ist

$$P_Z(k) = \binom{10}{k} \left(\frac{4}{11}\right)^k \left(\frac{7}{11}\right)^{10-k}.$$

Es ist etwa $P_Z(0) = 0.0109$, $P_Z(1) = 0.0622$ und $P_Z(2) = 0.16$. Eine graphische Repräsentation von $P_Z(k)$ ist in Abbildung 6.1 zu sehen. \square

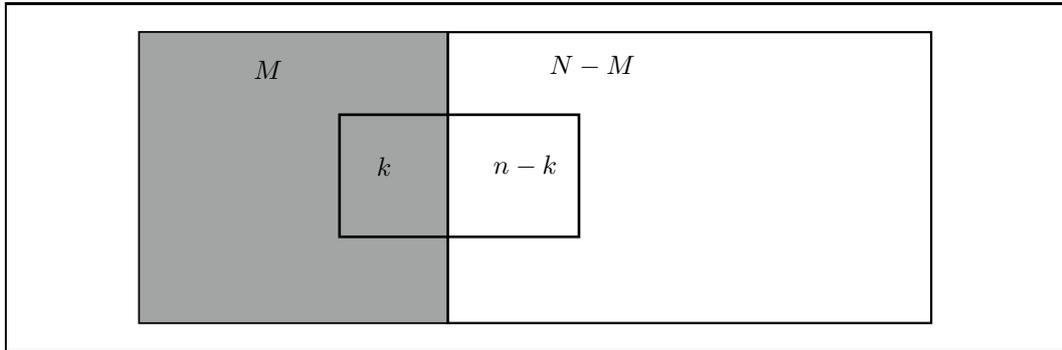


Abbildung 6.2: Illustration zur hypergeometrischen Verteilung: M von N Kugeln haben eine besondere Eigenschaft; aus allen N Kugeln werden n gezogen. Gesucht ist die W., genau k der M besonderen Kugeln zu ziehen.

Beispiel 6.3 Eine Münze wird viermal geworfen. Was ist die W., dass mindestens zweimal KOPF erscheint?

Lösung: Der geeignete Ereignisraum ist $\Omega = \{\text{KOPF, ZAHL}\}^4$. Die ZV $Z : \Omega \rightarrow \mathbb{N}_0$ misst die Anzahl des Auftretens von KOPF in der Versuchsreihe. Das Münzwürfen wird viermal unabhängig wiederholt, somit ist Z $\text{Bin}(4, \frac{1}{2})$ -verteilt. Die gesuchte W. ist dann

$$P_Z(\{k \geq 2\}) = P_Z(\{2, 3, 4\}) = \sum_{k=2}^4 \binom{4}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{4-k} = \frac{6}{16} + \frac{4}{16} + \frac{1}{16} = \frac{11}{16}. \quad \square$$

6.4 Hypergeometrische Verteilung

Beispiel 6.4 In einer Urne befinden sich N Kugeln, von denen M eine besondere Eigenschaft haben, die sie von den anderen $N - M$ Kugeln unterscheiden. Aus dieser Urne werden ohne Zurücklegen (oder, was äquivalent dazu ist, mit einem Griff) n Kugeln gezogen. Was ist die W., dass sich unter den n gezogenen genau k der M Kugeln mit der besonderen Eigenschaft befinden? Diese Situation ist graphisch in Abbildung 6.2 dargestellt.

Lösung: Der Lösungsraum Ω dieses Zufallsexperiments besteht aus allen n -elementigen Teilmengen der N -elementigen Menge aller Kugeln in der Urne. Mit Satz 3.6 gibt es genau $\binom{N}{n}$ solcher Teilmengen. Da jede der so gezogenen Teilmengen gleich wahrscheinlich ist, ist dieser W.Raum ein Laplace-Raum. Um die gewünschte W. zu erhalten, muss man noch ausrechnen, wieviele der $\binom{N}{n}$ n -elementigen Teilmengen die gewünschte Eigenschaft haben, genau k der M besonderen Kugeln zu enthalten.

Eine ähnliche Überlegung zeigt dass es $\binom{M}{k}$ Möglichkeiten gibt, k der M besonderen Kugeln zu ziehen. Wenn man sich allerdings für Anzahl der Möglichkeiten interessiert, unter n gezogenen Kugeln k der besonderen zu erhalten so muss man berücksichtigen dass dies auch bedeutet, unter den restlichen $n - k$ gezogenen *keine*

der besonderen Kugeln mehr zu haben. Man muss daher auch noch die Möglichkeiten berücksichtigen muss, aus den $N - M$ "normalen" Kugeln $n - k$ zu ziehen. Wie schon oben argumentiert, ist dies auf $\binom{N-M}{n-k}$ Arten möglich.

Zusammenfassend ergibt sich also, dass von allen $\binom{N}{n}$ Möglichkeiten, n von N Kugeln zu ziehen, genau $\binom{M}{k} \cdot \binom{N-M}{n-k}$ die gewünschte Eigenschaft haben. Da alle Möglichkeiten des Ziehens von n aus N Kugeln gleich wahrscheinlich sind, erhält man als gesuchte W. den Quotienten $\binom{M}{k} \cdot \binom{N-M}{n-k} / \binom{N}{n}$. \square

Das Ergebnis des obigen Beispiels lässt sich formal als Definition der hypergeometrischen Verteilung verwenden.

Definition 6.3 (Hypergeometrische Verteilung)

Eine Verteilung P_Z mit W.Funktion $f_Z : \mathbb{N}_0 \rightarrow [0, 1]$, für die

$$f_Z(k) = \begin{cases} \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}} & \text{für } k = \max(0, M + n - N), \dots, \min(M, n) \\ 0 & \text{sonst} \end{cases}$$

gilt, heißt *hypergeometrische Verteilung* mit Parametern N , M und n . Für die hypergeometrische Verteilung mit diesen Parametern schreiben wir auch $\text{Hypergeo}(M, N, n)$.

Die charakterisierende Eigenschaft der Binomialverteilung aus Abschnitt 6.3 ist, dass sie zur Beschreibung der Ausgänge eines Zufallsexperiments verwendet werden kann, das n -mal unabhängig durchgeführt wird. In Analogie zu den Standardmodellen der Kombinatorik aus Kapitel 3 entspricht dies der Situation des Urnenziehens *mit Zurücklegen* (da ja dann vor jedem Ziehen die ursprüngliche Situation wiederhergestellt wird, wie es für unabhängiges Wiederholen notwendig ist).

Die hypergeometrische Verteilung entspricht in diesem Sinn dem Ziehen *ohne Zurücklegen*, wenn also die Wiederholungen des Zufallsexperiments nicht wie bei der Binomialverteilung unabhängig sind. Diese Überlegung wurde bereits in Beispiel 6.4 zur Herleitung der W.Funktion der hypergeometrischen Verteilung verwendet. Zwei konkrete Beispiele sollen die Verwendung dieser Verteilung illustrieren.

Beispiel 6.5 Beim österreichischen Lotto werden 6 von 45 Kugeln gezogen; diese 6 können als "besonders" betrachtet werden. Mit der hypergeometrischen Verteilung kann die W. für Sechser, Fünfer, Vierer etc. berechnet werden. So ist die W. einen Sechser zu tippen genau die W., 6 von 6 besonderen Kugeln aus einer Urne mit 45 Kugeln zu ziehen. Die für die hypergeometrische Verteilung benötigten Parameter sind somit $N = 45$, $M = 6$, und $n = 6$. Damit erhalten wir als W. für einen Sechser

$$P_Z(6) = \frac{\binom{6}{6} \binom{39}{0}}{\binom{45}{6}} = \frac{1}{\binom{45}{6}} = \frac{1}{8145060} = 1.22 \cdot 10^{-7}.$$

Auf dieses Ergebnis kann man auch ohne hypergeometrische Verteilung kommen, da ja nur eine von $\binom{45}{6}$ Ziehungen einen Sechser liefert. Für die W. der anderen Gewinnziehungen ist es aber hilfreich, die hypergeometrische Verteilung zu verwenden. Es ergeben sich die Zahlen

			k			
	3	4	5	6		
$P_Z(k)$	$\frac{\binom{6}{3}\binom{39}{3}}{\binom{45}{6}}$	$\frac{\binom{6}{4}\binom{39}{2}}{\binom{45}{6}}$	$\frac{\binom{6}{5}\binom{39}{1}}{\binom{45}{6}}$	$\frac{\binom{6}{6}\binom{39}{0}}{\binom{45}{6}}$		□
	$= 0.0224$	$= 0.00136$	$= 2.87 \cdot 10^{-5}$	$= 1.22 \cdot 10^{-7}$		

Beispiel 6.6 In einem Teich mit 100 Fischen werden zur Untersuchung des Revierverhaltens an einer bestimmten Stelle 10 Fische gefangen, markiert, und wieder freigelassen. Einige Tage später werden an derselben Stelle wiederum 10 Fische gefangen. Es wird festgestellt, dass 8 von diesen 10 markiert sind. Wie wahrscheinlich ist dieses Ergebnis, wenn angenommen wird, dass sich die Fische innerhalb der Tage zwischen Markierung und zweitem Fang gleichmässig über den Teich verteilen?

Lösung: Sei Z die ZV, die die Anzahl der markierten gezogenen Fische angibt. Über die Analogie mit dem Urnenmodell erkennt man, dass Z hypergeometrisch verteilt ist mit Parametern $N = 100$, $M = 10$ und $n = 10$. Die gesuchte W. ist

$$P_Z(8) = \frac{\binom{10}{8}\binom{90}{2}}{\binom{100}{10}} = 1.04 \cdot 10^{-8}.$$

Dieses Ereignis ist somit sehr unwahrscheinlich, wenn man eine hypergeometrische Verteilung verwendet (was der Fall ist, wenn man eine gründliche Durchmischung der Fischpopulation annimmt). Die Zahlen sprechen also dafür, dass sich die Fische nicht innerhalb weniger Tage gleichmäßig im See verteilen.

In den Naturwissenschaften werden oftmals verwandte Aufgabenstellungen untersucht: Meist ist N unbekannt, kann aber mit ähnlichen Experimenten unter Verwendung der hypergeometrischen Verteilung geschätzt werden. □

6.5 Poissonverteilung

Wie wir aus dem letzten Abschnitt wissen, ist eine ZV, die die Anzahl des Auftretens eines Ereignisses bei n -maliger Wiederholung des Experiments misst binomialverteilt. Ein Spezialfall der Binomialverteilung tritt ein, wenn das zugrundeliegende Experiment *sehr oft* wiederholt wird, wobei die W. des Eintretens des gewünschten Ereignisses *sehr klein* wird. Die ZV ist selbstverständlich auch in diesem Fall binomialverteilt, sie nähert sich aber der Poissonverteilung, die wie folgt definiert ist.

Definition 6.4 (Poissonverteilung)

Eine Verteilung P_Z mit W.Funktion $f_Z : \mathbb{N}_0 \rightarrow [0, 1]$, für die

$$f_Z(k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

gilt, heißt *Poissonverteilung* mit Parameter λ . Für die Poissonverteilung mit diesem Parameter schreiben wir auch $\text{Poisson}(\lambda)$.

Die Poissonverteilung wird auch als “Verteilung der seltenen Ereignisse” bezeichnet und kann überall dort eingesetzt werden, wo bei einem Zufallsexperiment ein gewünschtes Ereignis selten (aber doch) eintritt. Der Parameter λ gibt dabei an, wie oft im Durchschnitt über einen längeren Zeitraum, oder über öfteres Wiederholen, das gewünschte Ereignis dann doch eintritt. Beispiele dafür wären etwa die Anzahl der eingehenden Anrufe in einer Notrufzentrale: In jeder einzelnen Sekunde ist die W. eines Anrufs gering, in einem Zeitraum von 10 Minuten gehen aber im Durchschnitt 4 Anrufe ein—in diesem Fall ist $\lambda = 4$. Ebenso können Aussagen über die Anzahl von Tippfehlern in einem Skriptum gemacht werden: Die W. eines Fehlers ist bei jedem einzelnen Buchstaben sehr gering, kann aber in einem ganzen Skriptum im Durchschnitt einen bestimmten Wert ausmachen, der dann als λ -Wert der Poissonverteilung verwendet werden kann. Ebenso lassen sich viele Ereignisse in den Naturwissenschaften durch Poissonverteilungen beschreiben: so ist etwa die Anzahl der Chromosomenschäden bei Bestrahlung organischer Zellen mit Röntgenstrahlen, oder die Anzahl der emittierten α -Teilchen pro Zeiteinheit bei radioaktivem Zerfall Poisson-verteilt.

Der oben bereits angesprochene Zusammenhang zwischen Binomial- und Poissonverteilung wird im folgenden Beispiel dargelegt, und in Satz 6.1 näher spezifiziert.

Beispiel 6.7 Nehmen wir an, dass an der Sponsionsfeier eines Studiengangs der FH Hagenberg an einem speziellen Termin (etwa am 30. Juli) genau 600 Personen teilnehmen werden. Wie groß ist die W., dass an diesem Tag genau k der 600 Personen Geburtstag haben?

Lösung: Wir nehmen zur Vereinfachung an, dass keine Person am 29. Februar Geburtstag hat, und dass alle anderen Geburtstage gleich wahrscheinlich sind. Damit ergibt sich das einfache Modell

$$\Omega = \{1. \text{ Jänner}, \dots, 31. \text{ Dezember}\} \setminus \{29. \text{ Februar}\};$$

das W.Maß auf Ω ist die Gleichverteilung. Für das Ereignis $A = \{30. \text{ Juli}\}$ ist somit (wie für alle anderen Tage) $P(A) = \frac{1}{365}$.

Es sei $Z : \Omega \rightarrow \mathbb{N}_0$ die ZV, die die Anzahl der Personen mit Geburtstag am 30. Juli misst. Die Aufgabenstellung dieses Beispiels ist analog zu der in den Beispielen 6.2 und 6.3: Das Experiment “Test, ob Person am 30. Juli Geburtstag hat” wird 600 mal unabhängig durchgeführt. Für jedes Experiment ist die W. des gewünschten Resultats $\frac{1}{365}$. Wie wir aus Beispiel 6.1 wissen, ist die Anzahl der gewünschten Resultate in dieser Versuchsanordnung binomialverteilt; somit ist auch Z $\text{Bin}(600, \frac{1}{365})$ -verteilt.

Dieses Beispiel eignet sich speziell dazu, die Ähnlichkeit von Binomial- und Poissonverteilung für lange Versuchsreihen (hier $n = 600$) und kleine W. (hier $p = \frac{1}{365}$) aufzuzeigen. Der Parameter λ der Poissonverteilung ergibt sich dabei aus der Multiplikation der Parameter n und p der Binomialverteilung. Die folgende Tabelle zeigt die weitgehende Übereinstimmung, die mit diesen beiden Verteilungen erreicht wird. Zur Erinnerung: Die in den Tabellen angegebenen Verteilungen sind

$$\text{Bin}(600, \frac{1}{365})(k) = \binom{600}{k} \left(\frac{1}{365}\right)^k \left(\frac{364}{365}\right)^{600-k} \quad \text{und}$$

$$\text{Poisson}\left(\frac{600}{365}\right)(k) = e^{-\frac{600}{365}} \frac{\left(\frac{600}{365}\right)^k}{k!}$$

k	Bin(600, $\frac{1}{365}$)	Poisson($\frac{600}{365}$)	k	Bin(600, $\frac{1}{365}$)	Poisson($\frac{600}{365}$)
0	0.1928	0.1932	4	0.0587	0.0588
1	0.3178	0.3177	5	0.0192	0.0193
2	0.2615	0.2611	6	0.0052	0.0053
3	0.1432	0.1431	7	0.0012	0.0012

□

Diese Annäherung der Binomialverteilung an die Poissonverteilung für großes n und kleines p gilt auch allgemein, wie im folgenden Satz festgehalten.

Satz 6.1 (Poisson'scher Grenzwertsatz) Für $k \in \mathbb{N}_0$ gilt

$$\text{Bin}(n, p)(k) \xrightarrow[\substack{n \rightarrow \infty \\ p \rightarrow 0 \\ np = \lambda > 0}]{\hspace{1.5cm}} \text{Poisson}(\lambda)(k),$$

in Worten: Für große n und kleine p , deren Produkt λ nicht null ist, wird die Binomialverteilung durch die Poissonverteilung approximiert.

Beweis Nicht exakt (speziell in der Verwendung des Gleichheitszeichens—eigentlich sind Grenzwerte gemeint), aber dafür hoffentlich einfacher nachvollziehbar: Zuerst muss man wissen, dass die Exponentialfunktion e^x als

$$e^x = \lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n$$

definiert werden kann. Der Rest ergibt sich durch Umformen der W.Funktion der Binomialverteilung (speziell durch Herausheben).

$$\begin{aligned} \text{Bin}(n, p)(k) &= \binom{n}{k} p^k (1-p)^{n-k} = \frac{n(n-1) \cdots (n-k+1)}{k!} p^k (1-p)^{n-k} \\ &= \frac{(np)^k}{k!} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right) (1-p)^{n-k} \\ &= \frac{(np)^k}{k!} \underbrace{\left(1 - \frac{1}{n}\right)}_{\downarrow 1} \underbrace{\left(1 - \frac{2}{n}\right)}_{\downarrow 1} \cdots \underbrace{\left(1 - \frac{k-1}{n}\right)}_{\downarrow 1} \underbrace{\left(1 - \frac{np}{n}\right)^n}_{\downarrow e^{-np}} \underbrace{(1-p)^{-k}}_{\downarrow 1} \\ &\xrightarrow{n \rightarrow \infty} \frac{(np)^k}{k!} e^{-np} = \text{Poisson}(np)(k). \end{aligned}$$

Die Pfeile in der vorletzten Zeile geben an, gegen welchen Wert der Ausdruck für $n \rightarrow \infty$ konvergiert. □

Um zu zeigen, wie genau sich natürliche Vorgänge durch die Poissonverteilung beschreiben lassen, untersuchen wir den radioaktiven Zerfall im nächsten Beispiel genauer.

Beispiel 6.8 Die Physiker Rutherford und Geiger führten einen Versuch zur Messung der Emission von α -Teilchen durch. Dabei wurde in jedem von $m = 2608$ Zeitabschnitten (insgesamt etwa 7.5 Sekunden) die Anzahl der emittierten Teilchen registriert—meist zwischen ein und sechs Teilchen pro Zeitabschnitt. In allen m Zeitabschnitten zusammen wurden $N = 10094$ emittierte Teilchen gemessen. Die Messungen von Rutherford und Geiger haben bestätigt, dass sich die in der Praxis erhaltenen Werte sehr wenig von denen unterscheiden, die mit Hilfe theoretischer Überlegungen (Poissonverteilung) vorhergesagt werden können.

Zur theoretischen Berechnung benötigt man den Wert des Parameters λ ; dies ist der durchschnittliche Wert von emittierten Teilchen pro Zeiteinheit, also $\lambda = \frac{N}{m} = 3.8704$. Rutherford und Geiger stellten eine Tabelle auf, in der sie notierten, in wievielen der m Zeitabschnitte gerade k Teilchen emittiert wurden. Die untenstehende Tabelle liefert diese Zahlen m_k , und als Vergleich dazu die mit Hilfe einer $\text{Poisson}(3.8704)$ -verteilten ZV berechneten theoretischen Werte. Für den Vergleich muss die Verteilung, die ja nur W. zwischen 0 und 1 liefert, noch mit der Anzahl der Zeiteinheiten multipliziert werden.

k	m_k	$m \text{Poisson}(\lambda)(k)$	k	m_k	$m \text{Poisson}(\lambda)(k)$
0	57	54.4	5	408	393.5
1	203	210.5	6	273	253.8
2	383	407.4	7	139	140.3
3	525	525.5	8	45	67.9
4	532	508.4	≥ 9	43	46.3

□

6.6 Gemeinsame Verteilungen

Unter der gemeinsamen Verteilung von mehreren ZV versteht man die W. für das gemeinsame Auftreten mehrerer Ereignisse, die bei einem Zufallsexperiment beobachtet werden. Wenn etwa Z_1 und Z_2 zwei verschiedene Maßzahlen eines Experiments sind, so interessieren oftmals nicht nur $P_{Z_1}(A)$ und $P_{Z_2}(B)$, also die W. des Auftretens zweier Ereignisse A (für Z_1) und B (für Z_2), sondern auch die W. des *gemeinsamen* Auftretens von A und B . Die beiden ZV können dabei als Komponenten eines ZV-Vektors $Z = (Z_1, Z_2)$ aufgefasst werden, wie dies in Abschnitt 5.2 eingeführt wurde.

Die Frage nach der W. Verteilung mehrdimensionaler ZV motiviert die folgende Definition.

Definition 6.5 (Gemeinsame Verteilung von Zufallsvariablen)

Gegeben seien zwei ZV $Z_1 : \Omega \rightarrow \mathbb{N}_0$ und $Z_2 : \Omega \rightarrow \mathbb{N}_0$ sowie zwei beliebige Ereignisse $A, B \subseteq \mathbb{N}_0$. Dann nennt man

$$P_{Z_1, Z_2}(A, B) := P(\{Z_1 \in A\} \cap \{Z_2 \in B\})$$

die *gemeinsame Verteilung* von Z_1 und Z_2 , also die W., dass Z_1 in A und Z_2 in B liegt. Wir werden wie schon zuvor die Mengennotation für ein-elementige

Mengen A und B ignorieren; wir schreiben also für $k, j \in \mathbb{N}_0$

$$P_{Z_1, Z_2}(k, j) := P_{Z_1, Z_2}(\{k\}, \{j\})$$

Ein erstes Beispiel einer gemeinsamen Verteilung von ZV wurde bereits in Beispiel 5.6 präsentiert.

Beispiel 6.9 Aus einer Urne mit vier Kugeln, die mit den Zahlen 1 bis 4 durchnummeriert sind, wird zweimal ohne Zurücklegen gezogen. Dabei bezeichne Z_1 die Zahl der ersten und Z_2 die Zahl der zweiten gezogenen Kugel. Mit Hilfe der gemeinsamen Verteilung kann dieses Zufallsexperiment folgendermaßen beschrieben werden.

Wir interessieren uns für die gemeinsame Verteilung von Z_1 und Z_2 , die beides ZV auf dem Ereignisraum $\Omega = \{1, 2, 3, 4\}^2 \setminus \{(1, 1), (2, 2), (3, 3), (4, 4)\}$ sind. Dabei ist für ein Paar $(k, j) \in \Omega$ $Z_1(k, j) = k$ und $Z_2(k, j) = j$, die ZV sind also Projektionen auf die Komponenten von Ω . Die W., dass beim ersten Ziehen die Kugel mit der Zahl k und beim zweiten die mit der Zahl j gezogen wird, lässt sich dann als

$$P_{Z_1, Z_2}(k, j) = P(\{Z_1 = k\} \cap \{Z_2 = j\})$$

ausdrücken. Da jede Kombination (k, j) von gezogenen Kugeln gleich wahrscheinlich ist, ist Ω ein Laplace-Raum; damit ist auch schon die W. auf den Elementarereignissen $(k, j) \in \Omega$ festgelegt:

$$P_{Z_1, Z_2}(k, j) = P(k, j) = \frac{1}{|\Omega|} = \frac{1}{12}. \quad \square$$

Beispiel 6.10 (Fortsetzung von Beispiel 2.10) Zur Illustration des Konzepts der gemeinsamen Verteilung von ZV betrachten wir einen Text, den wir in überlappende Segmente der Länge 2 zerlegen. Von jedem Paar von Buchstaben gebe die Zufallsvariable X den ersten, und die Zufallsvariable Y den zweiten Buchstaben an, wobei wir nur die 26 Kleinbuchstaben und das Leerzeichen verwenden.

Die gemeinsame Verteilung von X und Y ist dann die relative Häufigkeit von Buchstabenkombinationen. Für den Text dieses Skriptums ist das Ergebnis einer solchen Analyse graphisch in Abbildung 6.3 zu sehen; dabei bilden die ersten Buchstaben die Zeilen, und die zweiten Buchstaben die Spalten der Abbildung. Man kann erkennen, dass über alle Buchstabenkombinationen der Wert von $P(X = \mathbf{e}, Y = \mathbf{n})$ am größten ist. Weitere häufige Kombinationen sind etwa “er” oder der Buchstabe n am Wortende. □

Beispiel 6.11 (Fortsetzung von Beispiel 6.10) An dieser Stelle ist es sinnvoll, sich nochmals den Unterschied zwischen *bedingten* und *gemeinsamen* W. vor Augen zu führen. Wir illustrieren diesen Unterschied anhand des Datenmaterials aus Abbildung 6.3. So ist etwa

$$P(Y = \mathbf{h} | X = \mathbf{c}) = 0.763$$

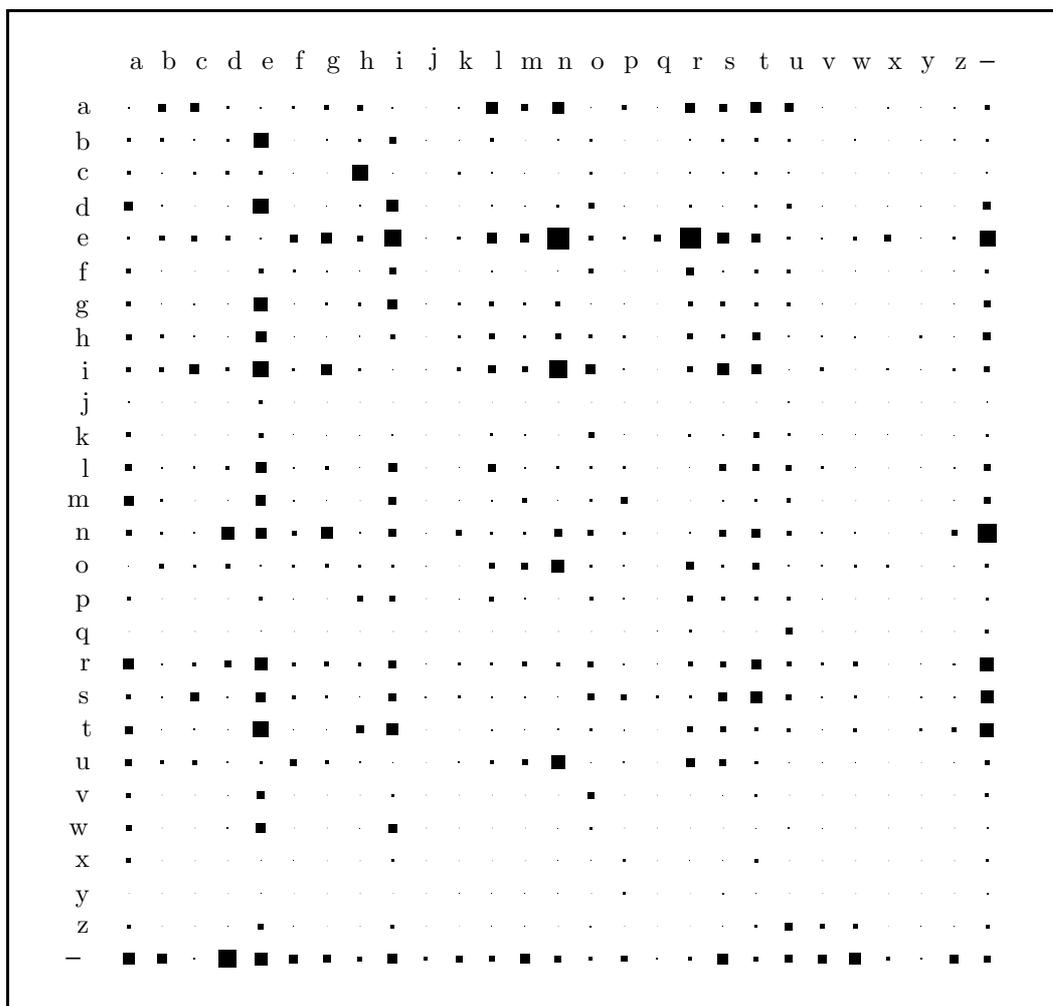


Abbildung 6.3: Gemeinsame Verteilung der beiden Zufallsvariablen, die den ersten bzw. zweiten Buchstaben einer zweielementigen Buchstabenkombination angeben.

die Wahrscheinlichkeit, dass der Buchstabe h nach dem Buchstaben c erscheint. Graphisch sind diese bedingten Wahrscheinlichkeiten in Abbildung 6.4 links zu sehen. In dieser Abbildung werden Verteilungen bedingt auf X -Werte dargestellt; daher repräsentiert jede *Zeile* dieser Graphik eine Wahrscheinlichkeitsverteilung.

In Abbildung 6.4 rechts sind die bedingten Wahrscheinlichkeiten $P(X | Y)$ graphisch aufgelistet. Es ist etwa

$$P(X = c | Y = h) = 0.513$$

die Wahrscheinlichkeit, dass ein c vor einem h auftritt. Wie man sieht, ist dies *nicht* derselbe Wert wie $P(Y = h | X = c)$, da nach c sehr oft h vorkommt, aber vor h auch öfter andere Buchstaben auftreten. In Abbildung 6.4 rechts sind auf Y bedingte Wahrscheinlichkeiten dargestellt; somit ist jede *Spalte* dieser Abbildung eine Wahrscheinlichkeitsverteilung. \square

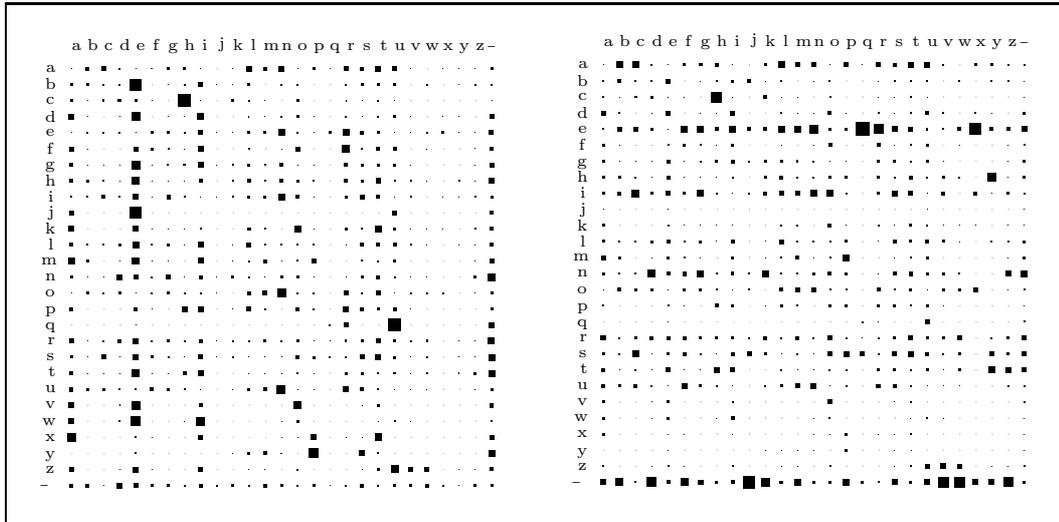


Abbildung 6.4: Die bedingten Wahrscheinlichkeiten $P(Y|X)$ (links) und $P(X|Y)$ (rechts) für die gemeinsame Verteilung aus Abbildung 6.3. Die Werte von X bilden die Zeilen, die von Y die Spalten der Matrizen. Die Einträge jeder Zeile der linken und jeder Spalte der rechten Abbildung summieren zu 1.

Definition 6.6 (W.Funktion der gemeinsamen Verteilung)

Gegeben seien zwei ZV Z_1 und Z_2 mit gemeinsamer Verteilung P_{Z_1, Z_2} . Dann nennt man

$$f_{Z_1, Z_2}(k, j) := P_{Z_1, Z_2}(\{k\}, \{j\})$$

die *gemeinsame W.Funktion* von Z_1 und Z_2 .

Wie im eindimensionalen Fall ist durch die W.Funktion die Verteilung eindeutig bestimmt, wie man durch Nachrechnen sehen kann:

$$\begin{aligned} P_{Z_1, Z_2}(A, B) &= P(\{Z_1 \in A\} \cap \{Z_2 \in B\}) = P\left(\bigcup_{k \in A} \{Z_1 = k\} \cap \bigcup_{j \in B} \{Z_2 = j\}\right) \\ &= \sum_{k \in A} \sum_{j \in B} P(\{Z_1 = k\} \cap \{Z_2 = j\}) = \sum_{k \in A} \sum_{j \in B} f_{Z_1, Z_2}(k, j). \end{aligned}$$

Ebenso gilt die vom eindimensionalen Fall bekannte Normiertheit der W.Funktion:

$$\sum_{k \geq 0} \sum_{j \geq 0} f_{Z_1, Z_2}(k, j) = 1.$$

Wenn die gemeinsame Verteilung zweier ZV bekannt ist, so lässt sich die Verteilung der beiden ZV (jede für sich selbst betrachtet) leicht berechnen. Diese Verteilungen werden als Randverteilungen bezeichnet.

Definition 6.7 (Randverteilung)

Die Verteilungen P_{Z_1} und P_{Z_2} zweier ZV Z_1 und Z_2 , deren gemeinsame Verteilung P_{Z_1, Z_2} bekannt ist, nennt man *Randverteilungen*.

Das Berechnen von Randverteilungen aus der gemeinsamen Verteilung erfolgt über das Aufsummieren aller möglichen Kombinationen der anderen Komponente der gemeinsamen Verteilung, wie der folgende Satz besagt.

Satz 6.2 Die Randverteilungen P_{Z_1} und P_{Z_2} einer gemeinsamen Verteilung P_{Z_1, Z_2} sind gegeben durch

$$P_{Z_1}(A) = P_{Z_1, Z_2}(A, \mathbb{N}_0) \quad \text{bzw.} \quad P_{Z_2}(B) = P_{Z_1, Z_2}(\mathbb{N}_0, B).$$

Für die W.Funktionen f_{Z_1} und f_{Z_2} von P_{Z_1} bzw. P_{Z_2} gilt

$$f_{Z_1}(k) = \sum_{j \geq 0} f_{Z_1, Z_2}(k, j) \quad \text{bzw.} \quad f_{Z_2}(j) = \sum_{k \geq 0} f_{Z_1, Z_2}(k, j).$$

Hierbei ist zu beachten, dass “ \mathbb{N}_0 ” bzw. die Summationsindizes “ $j \geq 0$ ” und “ $k \geq 0$ ” in der Weise zu interpretieren sind, dass über den gesamten Wertebereich von Z_1 bzw. Z_2 summiert wird. Wenn dieser Wertebereich eingeschränkt ist, wird selbstverständlich nur über diesen beschränkten Bereich summiert.

Die gemeinsame Verteilung bestimmt somit eindeutig die Randverteilungen; umgekehrt können wir allerdings *nicht* von den Randverteilungen auf die gemeinsame Verteilung schließen. Darauf gehen wir aber nicht näher ein; in Beispiel 6.13 ist diese Situation kurz dargestellt.

Beispiel 6.12 (Fortsetzung von Beispiel 6.9) Die W., beim ersten Ziehen die Zahl k zu erhalten, ist im Modell mit den gemeinsamen W.

$$\begin{aligned} P_{Z_1}(\{k\}) &= f_{Z_1}(k) = \sum_{j \neq k} f_{Z_1, Z_2}(k, j) \\ &= \frac{1}{12} + \frac{1}{12} + \frac{1}{12} = \frac{1}{4}, \end{aligned}$$

ebenso ist $P_{Z_2}(\{j\}) = \sum_{k \neq j} f_{Z_1, Z_2}(k, j) = \frac{1}{4}$. Diese Werte stimmen mit denjenigen überein, die man bei getrenntem Betrachten von Z_1 und Z_2 erhält. Dabei ist $f_{Z_1}(k) = \frac{1}{4}$ für $k = 1, \dots, 4$ unmittelbar klar. Wenn das Ergebnis dieses ersten Ziehens nicht bekannt ist, so gilt ebenfalls $f_{Z_2}(j) = \frac{1}{4}$ für $j = 1, \dots, 4$, da etwa $f_{Z_2}(1)$ die drei gleich wahrscheinlichen Elementarereignisse $(2, 1), (3, 1), (4, 1)$ aus dem 12-elementigen Ereignisraum $\Omega = \{(k, j) \mid k \neq j\}$ umfasst. \square

6.7 Unabhängigkeit von Zufallsvariablen

Ein wichtiger Begriff für ZV ergibt sich in Analogie zur Definition der Unabhängigkeit zweier Ereignisse in Definition 4.2.

Definition 6.8 (Unabhängigkeit von Zufallsvariablen)

Gegeben seien zwei ZV $Z_1 : \Omega \rightarrow \mathbb{N}_0$ und $Z_2 : \Omega \rightarrow \mathbb{N}_0$. Wenn für alle $A, B \subseteq \mathbb{N}_0$

$$P_{Z_1, Z_2}(A, B) = P_{Z_1}(A)P_{Z_2}(B)$$

gilt, nennt man Z_1 und Z_2 *unabhängig*.

Die Unabhängigkeit zweier ZV lässt sich einfach aus der W.Funktion ihrer gemeinsamen Verteilung ablesen.

Satz 6.3 Zwei ZV $Z_1 : \Omega \rightarrow \mathbb{N}_0$ und $Z_2 : \Omega \rightarrow \mathbb{N}_0$ mit gemeinsamer W.Funktion f_{Z_1, Z_2} sind genau dann unabhängig, wenn für alle Paare $(k, j) \in \mathbb{N}_0^2$

$$f_{Z_1, Z_2}(k, j) = f_{Z_1}(k)f_{Z_2}(j)$$

gilt.

Beispiel 6.13 Gegeben seien zwei ZV X_1 und X_2 mit gemeinsamer Verteilungsdichte $f_{X_1, X_2}(k, j)$, sowie zwei weitere ZV Y_1 und Y_2 mit gemeinsamer W.Funktion $f_{Y_1, Y_2}(k, j)$. Diese Funktionen sind in folgenden Tabellen aufgelistet:

		k	
	f_{X_1, X_2}	0	1
j	0	$\frac{1}{6}$	$\frac{1}{6}$
	1	$\frac{1}{3}$	$\frac{1}{3}$

		k	
	f_{Y_1, Y_2}	0	1
j	0	$\frac{1}{3}$	0
	1	$\frac{1}{6}$	$\frac{1}{2}$

Man kann unschwer nachprüfen, dass die Randverteilung von X_1 identisch ist mit der Randverteilung von Y_1 :

$$\begin{aligned} f_{X_1}(0) &= f_{X_1, X_2}(0, 0) + f_{X_1, X_2}(0, 1) = \frac{1}{6} + \frac{1}{3} = \frac{1}{2} \\ f_{X_1}(1) &= f_{X_1, X_2}(1, 0) + f_{X_1, X_2}(1, 1) = \frac{1}{6} + \frac{1}{3} = \frac{1}{2} \\ f_{Y_1}(0) &= f_{Y_1, Y_2}(0, 0) + f_{Y_1, Y_2}(0, 1) = \frac{1}{3} + \frac{1}{6} = \frac{1}{2} \\ f_{Y_1}(1) &= f_{Y_1, Y_2}(1, 0) + f_{Y_1, Y_2}(1, 1) = 0 + \frac{1}{2} = \frac{1}{2}, \end{aligned}$$

ebenso ist die Randverteilung von X_2 identisch mit der von Y_2 :

$$\begin{aligned} f_{X_2}(0) &= f_{X_1, X_2}(0, 0) + f_{X_1, X_2}(1, 0) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3} \\ f_{X_2}(1) &= f_{X_1, X_2}(0, 1) + f_{X_1, X_2}(1, 1) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3} \\ f_{Y_2}(0) &= f_{Y_1, Y_2}(0, 0) + f_{Y_1, Y_2}(1, 0) = \frac{1}{3} + 0 = \frac{1}{3} \end{aligned}$$

$$f_{Y_2}(1) = f_{Y_1, Y_2}(0, 1) + f_{Y_1, Y_2}(1, 1) = \frac{1}{6} + \frac{1}{2} = \frac{2}{3}.$$

Hiermit ist also ein Fall gegeben, in dem ZV mit gleichen Randverteilungen nicht die gleichen gemeinsamen Verteilungen haben. In diesem Beispiel soll aber auf eine anderen Sachverhalt hingewiesen werden: Es soll die Unabhängigkeit von X_1 und X_2 bzw. die Unabhängigkeit von Y_1 und Y_2 untersucht werden.

Wenn X_1 und X_2 unabhängig sind, muss mit Satz 6.3 gelten, dass

$$f_{X_1, X_2}(k, j) = f_{X_1}(k)f_{X_2}(j)$$

für alle $k, j = 0, 1$ erfüllt ist. Dies lässt sich leicht nachrechnen. Für Y_1 und Y_2 ist diese Bedingung allerdings *nicht* erfüllt: So ist etwa

$$\frac{1}{2} = f_{Y_1, Y_2}(1, 1) \neq f_{Y_1}(1)f_{Y_2}(1) = \frac{1}{2} \cdot \frac{2}{3} = \frac{1}{3}.$$

Aus dieser Rechnung ergibt sich also, dass X_1 und X_2 unabhängig sind, Y_1 und Y_2 hingegen nicht (obwohl die Randverteilungen beider Paare von ZV identisch sind). Man sieht also, dass in der gemeinsamen Verteilung viel mehr Information steckt als in den Randverteilungen—nämlich die gesamte Information. \square

Kapitel 7

Erwartungswert und Varianz

Wie wir im letzten Kapitel gesehen haben, wird durch eine ZV $Z : \Omega \rightarrow \mathbb{N}_0$ ein Merkmal eines Experiments E mit Ereignisraum Ω beobachtet. Durch die zufällige Natur von E hängt das Merkmal $Z(\omega)$ natürlich vom Ausgang ω von E ab und ist damit ebenfalls zufällig. Wir betrachten dazu das folgende Beispiel.

Beispiel 7.1 Ein Experiment E sei gegeben durch $\Omega = \{\omega_1, \dots, \omega_n\}$ und $P(\omega_i) = p_i$ für $i = 1, \dots, n$. Ein Spieler wettet auf den Ausgang des Experiments und erhält bei Ausgang ω_i genau a_i Geldeinheiten. Wir betrachten als ZV zum Experiment E den Gewinn des Spielers, also

$$\begin{aligned} Z : \Omega &\rightarrow \mathbb{N}_0 \\ \omega_i &\mapsto Z(\omega_i) := a_i. \end{aligned}$$

Welchen Gewinn kann sich der Spieler erwarten? Um das Spiel fair zu gestalten, hat der Spieler diesen zu erwartenden Gewinn vor Beginn des Spiels als Einsatz zu tätigen. Dann ist zu erwarten, dass der Spieler auf lange Sicht gegen die Bank weder gewinnt noch verliert.

Lösung: Diese Art von Problemstellung ist charakteristisch für die Entwicklung der W.Rechnung im 18. Jahrhundert. Da es sicher nicht genügt, sich auf ein einzelnes Spiel zu konzentrieren (dessen Ausgang ja zufällig ist), betrachtet man N Spiele. Aufgrund der W. p_i wird für großes N etwa $N \cdot p_1$ mal der Ausgang ω_1 eintreten, $N \cdot p_2$ mal der Ausgang ω_2 usw. bis $N \cdot p_n$ mal der Ausgang ω_n . Über alle N Experimente aufsummiert erhält der Spieler also $N \cdot p_1 \cdot a_1 + \dots + N \cdot p_n \cdot a_n$ Geldeinheiten, bei einem einzelnen Spiel also im Durchschnitt $p_1 \cdot a_1 + \dots + p_n \cdot a_n$ Geldeinheiten. \square

So plausibel diese Erklärung auch klingt, ist sie doch nicht imstande, für folgendes Spiel den korrekten Einsatz zu berechnen.

Beispiel 7.2 (St. Petersburg Paradoxon) Ein Spieler wirft eine ideale Münze so lange, bis zum ersten Mal KOPF erscheint. Wenn dies beim k -ten Wurf passiert, erhält er 2^k Geldeinheiten. Wieviel soll der Einsatz für dieses Spiel sein?

Lösung: Die W., dass KOPF das erste Mal beim k -ten Werfen eintritt, ist $(\frac{1}{2})^k$. Mit der Argumentation von Beispiel 7.1 wäre der zu tätigende Einsatz

$$2 \frac{1}{2} + 4 \frac{1}{4} + 8 \frac{1}{8} + \dots = 1 + 1 + \dots = \infty,$$

ein sicherlich zu hoher Einsatz. \square

7.1 Erwartungswert

Die Überlegung aus Beispiel 7.1 und der Einwand aus Beispiel 7.2 führten in der W.Rechnung zur Einführung des Erwartungswerts, dessen Definition beide Punkte berücksichtigt.

Definition 7.1 (Erwartungswert)

Sei $Z : \Omega \rightarrow \mathbb{N}_0$ eine ZV. Der *Erwartungswert* E von Z ist definiert als

$$E(Z) = \sum_{k \geq 0} k P_Z(k) = \sum_{k \geq 0} k f_Z(k),$$

also als die durch die W.Funktion $f_Z(k)$ gewichtete Summe aller möglichen Werte k der ZV. Der Erwartungswert ist nur definiert, wenn obige Summe existiert, also echt kleiner als unendlich ist.

Die Existenz eines Erwartungswerts ist nicht für alle ZV der Fall, wie etwa Beispiel 7.2 zeigt. Wie man aus der Definition sieht, hängt der Erwartungswert einer ZV Z einzig und allein von der Verteilung P_Z von Z ab, nicht aber von der Funktion $Z : \Omega \rightarrow \mathbb{N}_0$ selbst. ZV, die die gleiche Verteilung aufweisen, haben damit auch den gleichen Erwartungswert. Wir betrachten anfangs einige einfache Beispiele, bevor wir auf Erwartungswerte von speziellen Verteilungen näher eingehen.

Beispiel 7.3 Die ZV Z gebe die Augenzahl beim einmaligen Werfen eines Würfels an. Der Erwartungswert von Z ist

$$E(Z) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6} = \frac{7}{2}. \quad \square$$

Beispiel 7.4 Ein Spieler trägt folgendes Spiel gegen die Bank aus: Er setzt einen Betrag a ein und wirft zwei Würfel. Wenn die Augensumme kleiner oder gleich acht ist, gewinnt die Bank und kassiert den Einsatz, bei Augensummen zwischen neun und elf (inklusive) zahlt die Bank den dreifachen Einsatz, und bei einer Doppelsechs erhält der Spieler den achtfachen Einsatz zurück. Ist dieses Spiel auf lange Sicht für die Bank rentabel?

Lösung: Der Ereignisraum des Zufallsexperiments ist $\Omega = \{1, \dots, 6\}^2$ mit der Gleichverteilung als W.Maß. Als ZV nehmen wir den Gewinn der Bank, da wir zur Beantwortung der Aufgabenstellung berechnen müssen, ob der Erwartungswert dieser ZV positiv oder negativ ist, ob also ein Gewinn oder ein Verlust zu erwarten ist. Es interessieren uns also die drei Ereignisse $A =$ "Augensumme kleiner gleich acht", $B =$ "Augensumme zwischen neun und elf (inklusive)" und $C =$ "Doppelsechs". Die ZV, die den Gewinn der Bank angibt, ist

$$Z : \Omega \rightarrow \mathbb{Z}$$

$$(i, j) \mapsto Z(i, j) := \begin{cases} a & \text{für } (i, j) \in A \\ -2a & \text{für } (i, j) \in B \\ -7a & \text{für } (i, j) \in C \end{cases}$$

Zu berechnen ist nun $E(Z)$. Dabei ist zu beachten, dass wir in diesem Beispiel den Wertebereich der ZV von \mathbb{N}_0 auf \mathbb{Z} geändert haben, da wir auch negative Werte betrachten. Diese Änderung beeinflusst den weiteren Rechenvorgang nicht.

Da $Z(i, j)$ nur die Werte $a, -2a$ oder $-7a$ annimmt, kann man leicht in Definition 7.1 einsetzen. Dazu benötigt man allerdings noch $P_Z(k)$ für $k = a, -2a, -7a$. Da ja $P_Z(k) = P(\{Z = k\})$ ist, ergeben sich die Werte $P_Z(a) = P(A) = \frac{26}{36}$, $P_Z(-2a) = P(B) = \frac{9}{36}$ und $P_Z(-7a) = P(C) = \frac{1}{36}$. Die einzelnen W. erhält man durch Abzählen aller möglichen Kombinationen, die die gewünschten Augensummen ergeben. Mit Einsetzen in die Definition ergibt das dann

$$\begin{aligned} E(Z) &= aP_Z(a) - 2aP_Z(-2a) - 7aP_Z(-7a) \\ &= a\frac{26}{36} - 2a\frac{9}{36} - 7a\frac{1}{36} = \frac{a}{36} > 0 \end{aligned}$$

Somit macht die Bank im Durchschnitt pro Spiel $\frac{1}{36}$ des Einsatzes Gewinn. \square

Beispiel 7.5 Eine Münze werde 100 mal geworfen. Die ZV Z gebe an, wie oft dabei das Ereignis ZAHL eintritt. Was ist der Erwartungswert von Z ?

Lösung: Wie wir aus Abschnitt 6.3 wissen, ist Z binomialverteilt mit Parametern $n = 100$ und $p = \frac{1}{2}$. Einsetzen der W.Funktion f_Z einer $\text{Bin}(100, \frac{1}{2})$ -verteilten ZV Z in die Definition des Erwartungswerts liefert

$$E(Z) = \sum_{k=0}^{100} k f_Z(k) = \sum_{k=0}^{100} k \binom{100}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{100-k} = \left(\frac{1}{2}\right)^{100} \sum_{k=0}^{100} k \binom{100}{k}$$

An dieser Stelle muss man sich überlegen, dass $\sum_{k=0}^n k \binom{n}{k} = n \cdot 2^{n-1}$ ist (etwa durch einen Induktionsbeweis), und kommt so zum wenig überraschenden Ergebnis

$$E(Z) = \left(\frac{1}{2}\right)^{100} \sum_{k=0}^{100} k \binom{100}{k} = 2^{-100} \cdot 100 \cdot 2^{99} = 50. \quad \square$$

Das Ergebnis des letzten Beispiels deckt sich mit dem, was man auch ohne zu rechnen "erraten" hätte. Der Erwartungswert binomialverteilter ZV wird im nächsten Satz allgemein ausgedrückt.

Satz 7.1 Für eine $\text{Bin}(n, p)$ -verteilte ZV Z gilt

$$E(Z) = np.$$

Beweis als Übung. \square

Beispiel 7.6 (Fortsetzung von Beispiel 6.2) Die ZV Z aus Beispiel 6.2 ist $\text{Bin}(10, \frac{4}{11})$ -verteilt. Damit ist $E(Z) = \frac{40}{11}$ und es ist zu erwarten, dass auf lange Sicht (bei oftmaligem Wiederholen des Experiments) am häufigsten etwas weniger als 4 rote Kugeln gezogen werden. \square

Zu fast jeder Verteilung existiert der Erwartungswert und lässt sich in einfacher Form (wie oben) schreiben. Dies ist auch für die hypergeometrische Verteilung der Fall. Wir geben zuerst das Resultat an, das wir im Anschluss beweisen.

Satz 7.2 Für eine Hypergeom(N, M, n)-verteilte ZV Z gilt

$$E(Z) = n \frac{M}{N}.$$

Beweis Es ist mit der Definition des Erwartungswert somit nachzurechnen, dass gilt:

$$\sum_{k=0}^n k \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}} = n \frac{M}{N}. \quad (\text{ErwHG})$$

Zum Beweis dieser Gleichung benötigt man mehrere Fakten: Erstens gelten

$$k \binom{M}{k} = M \binom{M-1}{k-1} \quad \text{und} \quad \binom{N}{n} = \frac{N}{n} \binom{N-1}{n-1}$$

für $N, M, n, k \geq 1$, wie man durch Einsetzen in die Definition des Binomialkoeffizienten nachrechnen kann. Die zweite Tatsache, die man für den Beweis benötigt ist, dass die hypergeometrische Verteilung eben eine *Verteilung* ist, alle W. also zu 1 aufsummieren. Speziell ist dies auch für Hypergeom($N-1, M-1, n-1$) der Fall, es gilt also

$$\sum_{k=0}^{n-1} \frac{\binom{M-1}{k} \binom{N-M}{n-1-k}}{\binom{N-1}{n-1}} = 1.$$

Zum Beweis von Gleichung (ErwHG) formen wir folgendermaßen um:

$$\begin{aligned} \sum_{k=0}^n k \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}} &= \frac{1}{\binom{N}{n}} \sum_{k=1}^n k \binom{M}{k} \binom{N-M}{n-k} \\ &= \frac{1}{\binom{N}{n}} \sum_{k=1}^n M \binom{M-1}{k-1} \binom{N-M}{n-k} \\ &= \frac{n}{N} M \underbrace{\frac{1}{\binom{N-1}{n-1}} \sum_{k=0}^{n-1} \binom{M-1}{k} \binom{N-M}{n-k-1}}_{=1} \\ &= n \frac{M}{N}. \quad \square \end{aligned}$$

Beispiel 7.7 (Fortsetzung von Beispiel 6.6) Wir betrachten nochmals das Beispiel vom Fischteich mit 100 Fischen, von denen 10 markiert wurden. Wenn sich die Fische im Teich nach mehreren Wochen gut durchmischt haben, welche Anzahl von markierten Fischen unter 10 neuerlich gefangenen ist zu erwarten?

Lösung: Mit Satz 7.2 ergibt sich für den Erwartungswert der ZV Z , die die Anzahl markierter Fische unter den 10 gefangenen angibt der Wert

$$E(Z) = 10 \frac{10}{100} = 1. \quad \square$$

Ebenso wie für die Binomialverteilung und die hypergeometrische Verteilung kann man auch den Erwartungswert einer Poisson-verteilten ZV leicht angeben. Wir betrachten zunächst ein Beispiel.

Beispiel 7.8 (Fortsetzung von Beispiel 6.8) Gegeben ist die ZV Z , die die Anzahl der emittierten α -Teilchen in einem kleinen Zeitabschnitt angibt. Mit Hilfe von Definition 7.1 und den Daten der Tabelle in Beispiel 6.8 ergibt sich das folgende Resultat, für das man die Werte aus der Tabelle natürlich noch durch $m = 2608$ dividieren muss, um die W.Funktion zu erhalten. Man erhält

$$E(Z) = \sum_{k \geq 0} k \cdot f_Z(k) \approx 1 \cdot \frac{210.5}{m} + \dots + 9 \frac{46.3}{m} = 3.8603.$$

Der errechnete Wert ist also praktisch ident mit dem Parameter $\lambda = 3.8704$ der Poissonverteilung; die Abweichung ergibt sich daraus, dass wir ja nicht über alle k summiert haben. \square

Die im obigen Beispiel zu beobachtende Übereinstimmung zwischen dem Erwartungswert und dem Parameter einer poissonverteilten ZV gilt auch allgemein.

Satz 7.3 Für eine Poisson(λ)-verteilte ZV Z gilt

$$E(Z) = \lambda.$$

Beweis als Übung. \square

Das Ergebnis des letzten Satzes war zu erwarten, da wir in Satz 6.1 den Zusammenhang zwischen Binomial- und Poissonverteilung gesehen haben, der durch $\lambda = np$ gegeben ist. Mit Satz 7.1 ist der Erwartungswert einer $\text{Bin}(n, p)$ -verteilten ZV gleich np , sodass ein Erwartungswert von λ für eine Poisson(λ)-verteilte ZV plausibel ist.

Beispiel 7.9 Die in Beispiel 6.7 angegebene ZV Z ist $\text{Bin}(600, \frac{1}{365})$ - und damit auch annähernd Poisson($\frac{600}{365}$)-verteilt. In beiden Fällen ist $E(Z) = \frac{600}{365} = 1.6438$; somit ist zu erwarten, dass unter 600 Personen zwischen ein und zwei Personen an einem bestimmten Tag Geburtstag haben. \square

Als Vorbereitung auf den nächsten Satz betrachten wir folgendes Beispiel.

Beispiel 7.10 Ein Glücksspiel wird folgendermaßen veranstaltet: In einer Urne befinden sich vier mit den Ziffern 1 bis 4 nummerierte Kugeln, die hintereinander aus der Urne gezogen werden. Für jede Kugel, deren Ziffer mit der Stelle, an der sie gezogen wurde übereinstimmt, erhält der Spieler eine Geldeinheit. Was ist der zu erwartende Gewinn?

Lösung: Die brute-force Methode zur Lösung dieser Aufgabe ist, alle $4! = 24$ möglichen Ziehungen und die damit verbundenen Gewinne aufzulisten und daraus dann den Erwartungswert zu berechnen. Eine solche Tabelle ist

ω	$Z(\omega)$	ω	$Z(\omega)$	ω	$Z(\omega)$	ω	$Z(\omega)$
1 2 3 4	4	2 1 3 4	2	3 1 2 4	1	4 1 2 3	0
1 2 4 3	2	2 1 4 3	0	3 1 4 2	0	4 1 3 2	1
1 3 2 4	2	2 3 1 4	1	3 2 1 4	2	4 2 1 3	1
1 3 4 2	1	2 3 4 1	0	3 2 4 1	1	4 2 3 1	2
1 4 2 3	1	2 4 1 3	0	3 4 1 2	0	4 3 1 2	0
1 4 3 2	2	2 4 3 1	1	3 4 2 1	0	4 3 2 1	0

Jede der möglichen Sequenzen ω ist gleich wahrscheinlich mit $P(\omega) = \frac{1}{24}$. Der Erwartungswert ist

$$E(Z) = 1 \cdot \frac{8}{24} + 2 \cdot \frac{6}{24} + 4 \cdot \frac{1}{24} = 1.$$

Man kann diesen Erwartungswert noch auf elegantere Weise berechnen, wie wir in Beispiel 7.18 sehen werden. \square

Für die Berechnung des Erwartungswerts von ZV, die keine spezielle Verteilung aufweisen, gelten einfache algebraische Regeln. Die erste von ihnen ist gewissermaßen der Grundstock, auf dem die meisten anderen aufbauen.

Satz 7.4 (Erwartungswert einer Funktion einer ZV)

Für den Erwartungswert einer Funktion $g : \mathbb{N}_0 \rightarrow \mathbb{N}_0$ einer ZV $Z : \Omega \rightarrow \mathbb{N}_0$ gilt

$$E(g(Z)) = \sum_{k \geq 0} g(k) f_Z(k).$$

Beweis Nicht ganz einfach, aber der Wichtigkeit des Resultats entsprechend trotzdem erwähnt.

$$\begin{aligned} E(g(Z)) &= \sum_{k \geq 0} k f_{g(Z)}(k) = \sum_{k \geq 0} k P(\{g(Z) = k\}) \\ &= \sum_{k \geq 0} k \sum_{\substack{j \geq 0 \\ g(j)=k}} P(\{Z = j\}) = \sum_{k \geq 0} \sum_{\substack{j \geq 0 \\ g(j)=k}} g(j) P(\{Z = j\}) \\ &= \sum_{j \geq 0} g(j) f_Z(j). \end{aligned}$$

Die entscheidende Einsicht ist hier, dass die W. $P(\{g(Z) = k\})$ als die Summe von W. $\sum_{g(j)=k} P(\{Z = j\})$ geschrieben werden kann. \square

Aus dem letzten Satz folgt unmittelbar der folgende, der ebenfalls vielen weiteren Resultaten zugrundeliegt.

Satz 7.5 (Linearität des Erwartungswerts) Gegeben seien zwei beliebige Zahlen $a, b \in \mathbb{R}$ und eine ZV $Z : \Omega \rightarrow \mathbb{N}_0$, deren Erwartungswert existiert. Dann gilt

$$E(aZ + b) = aE(Z) + b.$$

Beweis Mit Satz 7.4 einfach nachzurechnen; als Übung. \square

7.2 Varianz

In vielen Fällen ist man beim zufälligen Verhalten eines Experiments nicht nur daran interessiert, bestimmte Merkmale zu beobachten und den zu erwartenden Wert dieser Merkmale zu berechnen; oft ist auch von Bedeutung, wie stark ein Merkmal "streut", d.h., wie stark es vom zu erwartenden Wert abweicht. Für diese zu erwartende Abweichung wird der Begriff der *Varianz* wie folgt eingeführt.

Definition 7.2 (Varianz, Standardabweichung)

Für eine ZV $Z : \Omega \rightarrow \mathbb{N}_0$ ist die *Varianz* von Z die zu erwartende quadratische Abweichung von Z und $E(Z)$, also

$$\text{Var}(Z) = E((Z - E(Z))^2).$$

Dabei ist vorauszusetzen, dass die Erwartungswerte existieren; andernfalls ist die Varianz nicht definiert. Die positive Wurzel der Varianz bezeichnet man als *Standardabweichung*.

Auch bei der Varianz soll nochmals darauf hingewiesen werden, dass dieser Wert nur von der Verteilung P_Z und nicht von Z selbst abhängt. Wir werden später sehen, dass sich für die bisher bekannten Verteilungen einfache Formeln für die Varianz ergeben. In allgemeinen Fällen kann man die Varianz auch wie im folgenden Satz ausrechnen.

Satz 7.6 Für die Varianz einer ZV Z gilt

$$\text{Var}(Z) = E(Z^2) - (E(Z))^2,$$

wenn die Erwartungswerte existieren.

Beweis Ausnutzen der Linearität des Erwartungswerts:

$$\begin{aligned} \text{Var}(Z) &= E((Z - E(Z))^2) = E(Z^2 - 2ZE(Z) + (E(Z))^2) \\ &= E(Z^2) - 2E(Z)E(Z) + (E(Z))^2 \\ &= E(Z^2) - (E(Z))^2. \end{aligned}$$

Erwähnenswert ist hier nur, dass der Erwartungswert $E(Z)$ eine Konstante ist und somit $E(E(Z)) = E(Z)$ ist. Diese und alle anderen angewandten Rechenregeln folgen direkt aus Satz 7.5. \square

Beispiel 7.11 Sei Z die ZV, die die Augenzahl beim einmaligen Werfen eines Würfels angibt. Aus Beispiel 7.3 wissen wir, dass $E(Z) = \frac{7}{2}$ ist. Für die Varianz von Z gilt mit Satz 7.6

$$\begin{aligned} \text{Var}(Z) &= E(Z^2) - (E(Z))^2 \\ &= \sum_{k=1}^6 k^2 \frac{1}{6} - \left(\frac{7}{2}\right)^2 = \frac{91}{6} - \frac{49}{4} = \frac{35}{12}. \end{aligned} \quad \square$$

Aufgrund des Quadrats in der Definition der Varianz lässt sich die Linearität nicht übernehmen. Statt dessen gilt folgender Satz.

Satz 7.7 Gegeben seien eine ZV $Z : \Omega \rightarrow \mathbb{N}_0$, deren Varianz existiert, und zwei beliebige Zahlen $a, b \in \mathbb{R}$. Dann gilt

$$\text{Var}(aZ + b) = a^2 \text{Var}(Z).$$

Beweis Entweder mit Satz 7.6 oder direkt nachrechnen; als Übung. \square

Als Beispiel für die letzten Sätze über das Rechnen mit Erwartungswert und Varianz berechnen wir Erwartungswert und Varianz der diskreten Gleichverteilung aus Abschnitt 6.1 und der Bernoulliverteilung aus Abschnitt 6.2.

Beispiel 7.12 Gegeben sei eine ZV Z , die auf der Menge $\{0, \dots, 100\}$ gleichverteilt ist. Somit ist $f_Z(k) = \frac{1}{101}$ für $k = 0, \dots, 100$ und der Erwartungswert ist

$$E(Z) = \sum_{k=0}^{100} k f_Z(k) = \frac{1}{101} \sum_{k=1}^{100} k = \frac{100 \cdot 101}{2 \cdot 101} = 50,$$

liegt also in der Mitte der Menge. Für die Varianz errechnet man mit Kenntnis der Formel $\sum_{k=1}^n k^2 = \frac{n(n+1)(2n+1)}{6}$ und Sätzen 7.5 und 7.6

$$\begin{aligned} \text{Var}(Z) &= E(Z^2) - (E(Z))^2 = \sum_{k=1}^{100} k^2 f_Z(k) - 50^2 \\ &= \frac{1}{101} \cdot \frac{100 \cdot 101 \cdot 201}{6} - 50^2 \\ &= 850. \end{aligned}$$

Die Standardabweichung ist somit $\sqrt{850} = 29.1548$. Die allgemeinen Formeln lassen sich aus diesem konkreten Beispiel ableiten. \square

Beispiel 7.13 Eine Bernoulli-verteilte ZV Z nimmt mit W. $1 - p$ den Wert 0 und mit W. p den Wert 1 an. Somit ist ihr Erwartungswert

$$E(Z) = 0 \cdot (1 - p) + 1 \cdot p = p,$$

der Erwartungswert des Quadrats von Z ist mit Satz 7.5

$$E(Z^2) = 0^2 \cdot (1 - p) + 1^2 \cdot p = p,$$

woraus sich für die Varianz

$$\text{Var}(Z) = E(Z^2) - (E(Z))^2 = p - p^2 = p(1 - p)$$

ergibt. \square

Um die Varianz in einigen konkreten Beispielen berechnen zu können, geben wir sie für die Binomial-, hypergeometrische und Poisson-Verteilung an.

Satz 7.8 Für die Varianz einer $\text{Bin}(n, p)$ -verteilten ZV Z gilt

$$\text{Var}(Z) = np(1 - p).$$

Beweis Als Übung. □

Beispiel 7.14 (Fortsetzung von Beispiel 6.2) Gegeben ist eine $\text{Bin}(10, \frac{4}{11})$ -verteilte ZV. Mit Satz 7.8 ist ihre Varianz $10 \frac{4}{11} \frac{7}{11} = \frac{280}{121} = 2.3140$ und ihre Standardabweichung $\sqrt{2.3140} = 1.5212$. Im Mittel weicht die Anzahl der gezogenen roten Kugeln also nur um etwa 1.5 vom Erwartungswert 3.6364 ab. □

Satz 7.9 Für die Varianz einer $\text{Hypergeom}(N, M, n)$ -verteilten ZV Z gilt

$$\text{Var}(Z) = n \frac{M}{N} \left(1 - \frac{M}{N}\right) \left(\frac{N - n}{N - 1}\right).$$

Beweis Ähnlich dem Beweis von Satz 7.2, aber mit komplizierteren Umformungen; wird deswegen weggelassen. □

Beispiel 7.15 (Fortsetzung von Beispiel 7.7) Wir haben gesehen, dass beim Fangen von zehn Fischen in einem Teich mit zehn von 100 markierten Fischen zu erwarten ist, dass einer von zehn gefangenen Fischen markiert sein wird. Mit oben angegebener Formel erhält man für $N = 100$, $M = 10$ und $n = 10$ folgende Varianz der ZV Z , die die Anzahl der markierten gefangenen Fische angibt:

$$\text{Var}(Z) = 10 \frac{10}{100} \left(1 - \frac{10}{100}\right) \left(\frac{100 - 10}{100 - 1}\right) = \frac{9}{11}. \quad \square$$

Zuletzt fehlt uns noch die Varianz einer $\text{Poisson}(\lambda)$ -verteilten ZV. Diese ist im nächsten Satz angegeben.

Satz 7.10 Für die Varianz einer $\text{Poisson}(\lambda)$ -verteilten ZV Z gilt

$$\text{Var}(Z) = \lambda.$$

Beweis Als Übung. □

Die Symmetrie zwischen Binomial- und Poissonverteilung, die bei den Erwartungswerten gegeben ist, setzt sich also bei den Varianzen nicht fort. Allerdings gilt die Annäherung von Binomial- durch Poissonverteilung nur für kleine p (also $1 - p \approx 1$), womit wieder annähernd sowohl $\lambda = np$ und $\lambda = np(1 - p)$ gelten.

Beispiel 7.16 (Fortsetzung von Beispiel 6.7.) Die Anzahl der Personen, die unter 600 an einem bestimmten Tag Geburtstag haben, ist $\text{Bin}(600, \frac{1}{365})$ - und damit auch annähernd $\text{Poisson}(\frac{600}{365})$ -verteilt.

Wenn wir diese ZV Z als binomialverteilt annehmen, so ist ihre Varianz als $\text{Var}(Z) = 600 \cdot \frac{1}{365} \cdot \frac{364}{365} = 1.6393$ gegeben; dies unterscheidet sich nur geringfügig von Erwartungswert $E(Z) = 1.6438$ (siehe Beispiel 7.9).

Da für eine poissonverteilte ZV $\text{Var}(Z) = \lambda$ gilt (wobei ja $\lambda = np$ ist), ergibt sich 1.6438 als Varianz. Die beiden Varianzen sind somit annähernd gleich. \square

Der Zusammenhang zwischen Erwartungswert und Varianz lässt sich am einfachsten durch den folgenden Satz beschreiben. Dieser Satz wird vor allem dann verwendet, wenn die Verteilung einer Zufallsvariablen nicht bekannt ist, ihr Erwartungswert und Varianz aber schon (wie das möglich ist, werden wir in Kapitel 9 sehen). Bei bekannter Verteilung der ZV könnte man die gesuchte W. direkt ausrechnen und müsste sie nicht durch eine Ungleichung abschätzen.

Satz 7.11 (Ungleichung von Tschebyscheff) Für eine ZV Z und beliebiges $\varepsilon > 0$ gilt

$$P(|Z - E(Z)| \geq \varepsilon) \leq \frac{\text{Var}(Z)}{\varepsilon^2}.$$

Das bedeutet, dass die W. eines Abweichens von Z und $E(Z)$ durch $\text{Var}(Z)$ beschränkt wird; für ZV mit kleiner Varianz ist es demnach sehr wahrscheinlich, dass sie nicht weit von ihrem Erwartungswert abweichen.

Beispiel 7.17 In einer Fabrik werden pro Tag eine bestimmte Anzahl Waren hergestellt; die ZV Z messe diese Anzahl. Wenn bekannt ist, dass $E(Z) = 50$ und $\text{Var}(Z) = 25$ ist, was kann dann über die W. gesagt werden, dass die Produktion an einem Tag zwischen 41 und 59 Stück liegt?

Lösung: Mit der Ungleichung von Tschebyscheff ist

$$P(\{|Z - 50| \geq 10\}) \leq \frac{25}{10^2} = \frac{1}{4}$$

und somit

$$P(\{|Z - 50| < 10\}) \geq 1 - \frac{1}{4} = \frac{3}{4};$$

die W., dass zwischen 41 und 59 Stück der Ware produziert werden, ist also mindestens 75%. \square

7.3 Kovarianz und Korrelationskoeffizient

Wie in Abschnitt 6.6 erläutert wurde, steckt in der gemeinsamen Verteilung zweier ZV X und Y die gesamte Information über X und Y . Die W.Funktion der gemeinsamen Verteilung kann graphisch veranschaulicht werden, um Aussagen über den Zusammenhang zwischen X und Y zu treffen. Eine kompaktere Maßzahl liefert die Kovarianz, die wie folgt definiert ist.

Definition 7.3 (Kovarianz)

Gegeben seien zwei ZV $X : \Omega \rightarrow \mathbb{N}_0$ und $Y : \Omega \rightarrow \mathbb{N}_0$, deren Erwartungswerte existieren. Dann nennt man

$$\text{Cov}(X, Y) = E((X - E(X))(Y - E(Y)))$$

die *Kovarianz* von X und Y .

Meist berechnet man die Kovarianz über die folgende Formel, die sich aus einer Umformung der Definition ergibt.

Satz 7.12 Für die Kovarianz zweier ZV $X : \Omega \rightarrow \mathbb{N}_0$ und $Y : \Omega \rightarrow \mathbb{N}_0$ gilt

$$\text{Cov}(X, Y) = E(X \cdot Y) - E(X)E(Y).$$

Beweis Einfaches Nachrechnen:

$$\begin{aligned} \text{Cov}(X, Y) &= E((X - E(X))(Y - E(Y))) \\ &= E(X \cdot Y - YE(X) - XE(Y) + E(X)E(Y)) \\ &= E(X \cdot Y) - E(Y)E(X) - E(X)E(Y) + E(X)E(Y) \\ &= E(X \cdot Y) - E(X)E(Y). \end{aligned} \quad \square$$

An dieser Stelle muss betont werden, dass wir bisher Erwartungswert und Varianz nur für eindimensionale ZV betrachtet haben. Zum Berechnen des Erwartungswerts einer Funktion zweier ZV—wie im Term $E(X \cdot Y)$ in Satz 7.12—benötigt man ein Resultat, das den grundlegenden Satz 7.4 für zweidimensionale ZV verallgemeinert.

Satz 7.13 Für den Erwartungswert einer Funktion $g : \mathbb{N}_0 \rightarrow \mathbb{N}_0$ zweier ZV $X : \Omega \rightarrow \mathbb{N}_0$ und $Y : \Omega \rightarrow \mathbb{N}_0$ mit gemeinsamer W.Funktion $f_{X,Y}$ gilt

$$E(g(X, Y)) = \sum_{k \geq 0} \sum_{j \geq 0} g(k, j) f_{X,Y}(k, j).$$

Aus diesem Satz folgen mehr oder weniger unmittelbar einige Resultate, die in diesem und späteren Kapitel noch mehrmals benötigt werden. So ergibt sich für den Erwartungswert der Summe von ZV folgender Satz.

Satz 7.14 Seien $X_1, \dots, X_n : \Omega \rightarrow \mathbb{N}_0$ ZV auf Ω . Dann gilt

$$E(X_1 + \dots + X_n) = E(X_1) + \dots + E(X_n).$$

Beweis Für diesen n -dimensionalen Fall braucht man die Verallgemeinerung von Satz 7.13; da dieser aber völlig analog zum zweidimensionalen Fall ist, beschränken wir uns der Einfachheit halber hier auf die Summe zweier ZV. Dieser Beweis lässt sich ohne großen Aufwand auf mehrere ZV erweitern.

$$\begin{aligned} E(X_1 + X_2) &= \sum_{k \geq 0} \sum_{j \geq 0} (k + j) f_{X_1, X_2}(k, j) \\ &= \sum_{k \geq 0} \sum_{j \geq 0} (k f_{X_1, X_2}(k, j) + j f_{X_1, X_2}(k, j)) \\ &= \sum_{k \geq 0} k \underbrace{\sum_{j \geq 0} f_{X_1, X_2}(k, j)}_{f_{X_1}(k)} + \sum_{j \geq 0} j \underbrace{\sum_{k \geq 0} f_{X_1, X_2}(k, j)}_{f_{X_2}(j)} \\ &= E(X_1) + E(X_2). \quad \square \end{aligned}$$

Beispiel 7.18 (Fortsetzung von Beispiel 7.10) Zum Ergebnis $E(Z) = 1$ kann man noch auf eine weitere Art kommen. Dazu betrachten wir die ZV Y_1, \dots, Y_4 , die als

$$Y_i : \Omega \rightarrow \mathbb{N}_0$$

$$\omega \mapsto \begin{cases} 1 & \text{falls Ziffer } i \text{ an } i\text{-ter Stelle in } \omega \\ 0 & \text{sonst} \end{cases}$$

definiert sind; sie ergeben in Summe gerade Z . Wir sind also an $E(\sum_{i=1}^4 Y_i)$ interessiert. Dazu betrachten wir zuerst einmal die Erwartungswerte der einzelnen Y_i . Weil Y_i nur die Werte 0 oder 1 annehmen kann, ist auch der Erwartungswert einfach zu berechnen:

$$E(Y_i) = \frac{6}{24} = \frac{1}{4}$$

für $i = 1, \dots, 4$, weil jede Ziffer in sechs der 24 möglichen Sequenzen an der richtigen Stelle steht. Mit Satz 7.5 ist

$$E(Z) = E\left(\sum_{i=1}^4 Y_i\right) = \sum_{i=1}^4 E(Y_i) = \frac{1}{4} + \frac{1}{4} + \frac{1}{4} + \frac{1}{4} = 1,$$

und wir kommen zum gleichen Ergebnis wie in Beispiel 7.10. \square

Zurück zu unserer Diskussion der Kovarianz, die ein Maß für das zu erwartende gemeinsame Abweichen zweier ZV von ihren Erwartungswerten ist. Eine einfache Interpretation der Kovarianz ist im nächsten Beispiel gegeben.

Definition 7.4 (Unkorreliertheit)

Gegeben seien zwei ZV $X, Y : \Omega \rightarrow \mathbb{N}_0$. Wenn für X und Y

$$\text{Cov}(X, Y) = 0$$

gilt, so nennt man die beiden ZV *unkorreliert*. Ist $\text{Cov}(X, Y) > 0$, dann sind X und Y *positiv*, bei $\text{Cov}(X, Y) < 0$ *negativ korreliert*.

Man beachte, dass laut Satz 7.12 zwei ZV X und Y genau dann unkorreliert sind, wenn

$$E(X \cdot Y) = E(X)E(Y)$$

gilt. Wir betrachten zuerst ein allgemein gehaltenes Beispiel, um das Konzept der Unkorreliertheit zu illustrieren.

Beispiel 7.19 Bei einem Zufallsexperiment werden zwei (nicht disjunkte) Ereignisse A und B beobachtet; so könnte etwa beim Experiment Würfeln $A =$ "gerade Augenzahl" und $B =$ "Augenzahl sechs" sein. Seien dazu X und Y ZV, deren Werte angeben, ob A bzw. B eingetreten ist:

$$X = \begin{cases} 1 & \text{wenn } A \text{ eintritt} \\ 0 & \text{sonst} \end{cases} \quad Y = \begin{cases} 1 & \text{wenn } B \text{ eintritt} \\ 0 & \text{sonst} \end{cases}$$

Dann ist die ZV $X \cdot Y$ genau dann eins, wenn sowohl A als auch B eintreten. Die Kovarianz dieser beiden ZV ist

$$\begin{aligned} \text{Cov}(X, Y) &= E(X \cdot Y) - E(X)E(Y) \\ &= P(X = 1 \wedge Y = 1) - P(X = 1)P(Y = 1) \end{aligned}$$

Jetzt können wir nachrechnen, wann X und Y positiv korreliert sind:

$$\begin{aligned} \text{Cov}(X, Y) > 0 &\Leftrightarrow \\ P(X = 1 \wedge Y = 1) &> P(X = 1)P(Y = 1) \Leftrightarrow \\ \frac{P(X = 1 \wedge Y = 1)}{P(X = 1)} &> P(Y = 1) \Leftrightarrow \\ P(Y = 1 | X = 1) &> P(Y = 1). \end{aligned}$$

Dieses Ergebnis liefert folgende einfache Interpretation: X und Y sind genau dann positiv korreliert, wenn das Eintreten von A (also $X = 1$) das Eintreten von B (also $Y = 1$) wahrscheinlicher macht. Aufgrund der Symmetrie der Kovarianz gilt das Argument natürlich auch umgekehrt: Das Eintreten von B macht das Eintreten von A wahrscheinlicher. \square

Manche Gleichungen in obigem Beispiel erinnern an die Definition der Unabhängigkeit von Ereignissen. In der Tat ist es so, dass unabhängige ZV unkorreliert sind, wie der folgende Satz besagt.

Satz 7.15 Zwei unabhängige ZV sind unkorreliert.

Beweis Wir müssen nachprüfen, dass aus der Unabhängigkeit von X und Y folgt, dass $E(X \cdot Y) = E(X)E(Y)$ ist. Aufbauend auf den bisher erarbeiteten Resultaten erhält man

$$\begin{aligned} E(X \cdot Y) &\stackrel{\text{Satz 7.13}}{=} \sum_{k \geq 0} \sum_{j \geq 0} k j f_{X,Y}(k, j) \stackrel{\text{Satz 6.3}}{=} \sum_{k \geq 0} \sum_{j \geq 0} k j f_X(k) f_Y(j) \\ &= \sum_{k \geq 0} k \left(\sum_{j \geq 0} j f_Y(j) \right) f_X(k) = \left(\sum_{k \geq 0} k f_X(k) \right) \left(\sum_{j \geq 0} j f_Y(j) \right) \\ &= E(X)E(Y). \quad \square \end{aligned}$$

Die Umkehrung obigen Satzes gilt im allgemeinen *nicht*: Die Unkorreliertheit zweier ZV impliziert keinesfalls ihre Unabhängigkeit, wie wir im nächsten Beispiel sehen werden.

Beispiel 7.20 Eine zweidimensionale ZV (X, Y) nimmt die vier Wertepaare $(0,1)$, $(1,0)$, $(0,-1)$ und $(-1,0)$ mit gleicher W. $\frac{1}{4}$ an. Wir rechnen nach, dass X und Y zwar unkorreliert, aber nicht unabhängig sind.

Die hier auftretenden ZV haben den Wertebereich $\{-1, 0, 1\}$, welcher ausserhalb von \mathbb{N}_0 liegt. Dennoch können alle Sätze mit geeigneter Modifikation der Summationindizes verwendet werden.

Für die Unkorreliertheit müssen wir nachprüfen, dass

$$E(X \cdot Y) = E(X)E(Y)$$

ist. Für die linke Seite dieser Gleichung ergibt sich mit

$$f_{X,Y}(k, j) = \begin{cases} \frac{1}{4} & \text{für } (k, j) \in \{(0, 1), (1, 0), (0, -1), (-1, 0)\} \\ 0 & \text{sonst} \end{cases}$$

für den Erwartungswert des Produkts

$$\begin{aligned} E(X \cdot Y) &= \sum_{k \in \{-1, 0, 1\}} \sum_{j \in \{-1, 0, 1\}} k j f_{X,Y}(k, j) \\ &= \frac{1}{4}((-1) \cdot 0 + 0 \cdot 1 + 0 \cdot (-1) + 1 \cdot 0) \\ &= 0. \end{aligned}$$

Zur Berechnung des Produkts der Erwartungswerte benötigt man

$$f_X(-1) = \sum_{j \in \{-1, 0, 1\}} f_{X,Y}(-1, j) = \frac{1}{4},$$

analog ergeben sich $f_X(0) = \frac{1}{2}$, $f_X(1) = \frac{1}{4}$, $f_Y(-1) = \frac{1}{4}$, $f_Y(0) = \frac{1}{2}$ und $f_Y(1) = \frac{1}{4}$. Diese W.Funktionen kann man auch mühelos an einer Skizze ablesen.

Mit Satz 6.2 ist dann

$$E(X) = E(Y) = (-1) \cdot \frac{1}{4} + 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{4} = 0.$$

Somit sind X und Y unkorreliert.

Für die Unabhängigkeit von X und Y müsste für alle $k, j \in \{-1, 0, 1\}$

$$f_{X,Y}(k, j) = f_X(k)f_Y(j)$$

gelten; dies ist aber etwa für

$$\frac{1}{4} = f_{X,Y}(1, 0) \neq f_X(1)f_Y(0) = \frac{1}{4} \cdot \frac{1}{2}$$

nicht der Fall. □

In Analogie zu Satz 7.14 über den Erwartungswert der Summe von ZV gilt folgender Satz über die Varianz.

Satz 7.16 Seien $X, Y : \Omega \rightarrow \mathbb{N}_0$ ZV auf Ω . Dann gilt

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y),$$

speziell ist also für n unkorrelierte ZV $X_1, \dots, X_n : \Omega \rightarrow \mathbb{N}_0$ ZV

$$\text{Var}(X_1 + \dots + X_n) = \text{Var}(X_1) + \dots + \text{Var}(X_n).$$

Beispiel 7.21 Die ZV Z gebe die Augensumme nach zehnmaligem unabhängigen Werfen eines Würfels an. Dann lässt sich Z als Summe $Z = X_1 + \dots + X_{10}$ schreiben, wobei die X_i die Augenzahlen des i -ten Würfels angeben. Da die X_i unabhängig sind, kann Satz 7.16 angewandt werden und es gilt mit dem Resultat von Beispiel 7.11

$$\begin{aligned} \text{Var}(Z) &= \text{Var}(X_1 + \dots + X_{10}) = \text{Var}(X_1) + \dots + \text{Var}(X_{10}) = 10 \cdot \frac{35}{12} \\ &= \frac{175}{6}. \end{aligned} \quad \square$$

Für manche Überlegungen spielt der Absolutbetrag der Kovarianz zweier ZV keine Rolle; man möchte nur wissen, ob die beiden ZV korreliert sind oder nicht. Für diesen Fall definiert man den Korrelationskoeffizienten wie folgt.

Definition 7.5 (Korrelationskoeffizient)

Seien $X, Y : \Omega \rightarrow \mathbb{N}_0$ zwei ZV mit positiver Varianz. Dann nennt man

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$$

den *Korrelationskoeffizienten* von X und Y .

Durch die Division durch die Wurzeln der Varianzen erreicht man eine Normierung der Kovarianz; man kann zeigen, dass der Korrelationskoeffizient immer im Intervall

$[-1, 1]$ liegt. Ausserdem gilt, dass $\text{Corr}(X, Y)$ genau dann -1 oder 1 wird, wenn X und Y linear abhängig sind, sich also Y als $Y = kX + d$ schreiben lässt. In diesem Fall ist die Steigung positiv, wenn $\text{Corr}(X, Y) = 1$ ist, und sie ist negativ für $\text{Corr}(X, Y) = -1$.

Beispiel 7.22 Ein Würfel wird zweimal geworfen. Dabei geben die ZV X und Y die Zahl der geworfenen Einsen bzw. Zweien an. Man berechne den Korrelationskoeffizienten von X und Y .

Lösung: Zur Berechnung von $\text{Corr}(X, Y)$ benötigt man $\text{Cov}(X, Y)$ sowie $\text{Var}(X)$ und $\text{Var}(Y)$. Für $\text{Cov}(X, Y) = E(X \cdot Y) - E(X)E(Y)$ braucht man auch diese Erwartungswerte; $E(X \cdot Y)$ lässt sich wie in Satz 7.13 über die gemeinsame W.Funktion $f_{X,Y}$ berechnen.

Wir berechnen also zuerst $f_{X,Y}$. So ist etwa

$$f_{X,Y}(0, 0) = P(X = 0 \wedge Y = 0) = \left(1 - \frac{2}{6}\right)^2 = \frac{16}{36}$$

und

$$f_{X,Y}(1, 0) = P(X = 1 \wedge Y = 0) = \frac{1}{6} \cdot \frac{4}{6} + \frac{4}{6} \cdot \frac{1}{6} = \frac{8}{36}.$$

Die anderen Werte berechnet man analog; sie sind in folgender Tabelle zusammengefasst:

		k		
f_{X_1, X_2}		0	1	2
	0	$\frac{16}{36}$	$\frac{8}{36}$	$\frac{1}{36}$
j	1	$\frac{8}{36}$	$\frac{2}{36}$	0
	2	$\frac{1}{36}$	0	0

Somit ergeben sich für die W.Funktion der Randverteilungen $f_X(0) = f_Y(0) = \frac{25}{36}$, $f_X(1) = f_Y(1) = \frac{10}{36}$ und $f_X(2) = f_Y(2) = \frac{1}{36}$ und daraus sofort die Erwartungswerte $E(X) = E(Y) = \frac{12}{36}$. Damit ist die Varianz

$$\text{Var}(X) = E(X^2) - (E(X))^2 = 1 \cdot \frac{10}{36} + 2^2 \cdot \frac{1}{36} - \left(\frac{12}{36}\right)^2 = \frac{10}{36},$$

ebenso ist $\text{Var}(Y) = \frac{10}{36}$. Für den Erwartungswert des Produkts von X und Y ergibt sich aus der Tabelle

$$E(X \cdot Y) = 0 \cdot 0 \cdot \frac{16}{36} + 0 \cdot 1 \cdot \frac{8}{36} + \dots + 2 \cdot 2 \cdot 0 = \frac{2}{36},$$

wobei nur einer der Summanden ungleich null ist. Somit ist

$$\text{Cov}(X, Y) = E(X \cdot Y) - E(X)E(Y) = \frac{2}{36} - \frac{12}{36} \cdot \frac{12}{36} = -\frac{1}{18}$$

und der Korrelationskoeffizient ist

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} = \frac{-\frac{1}{18}}{\frac{10}{36}} = -\frac{1}{5}.$$

Die beiden ZV X und Y sind somit sehr schwach negativ korreliert. □

Zum Abschluss noch eine Bemerkung über Korrelation und Korrelationskoeffizienten: Im alltäglichen Gebrauch von Korrelationen wird manchmal der Fehler gemacht, aus der Korrelation zweier ZV einen kausalen Zusammenhang zwischen Ereignissen abzuleiten, die von diesen ZV beschrieben werden. So sei etwa X die ZV, die die Schuhgröße eines Kindes misst; die ZV Y sei ein Maß für die Lesefähigkeit dieses Kindes. Wenn eine positive Korrelation zwischen X und Y besteht, so kann dies leicht fehlinterpretiert werden: Größere Füße bewirken höhere Lesefähigkeit des Kindes. Dies ist ein Trugschluss, denn die Korrelation besagt eben *nichts* über den kausalen Zusammenhang. In unserem Beispiel ist es wahrscheinlich, dass ein dritter Faktor die Verbindung zwischen X und Y herstellt: Wenn ein Kind älter wird, wachsen natürlich auch seine Füße—und im selben Maß nimmt auch seine Lesefähigkeit zu. Somit kann man die Korrelation zwischen X und Y erklären, ohne dass eine der beiden ZV die andere direkt beeinflusst.

Stetige Verteilungen

Wir betrachten in diesem Kapitel die Normalverteilung, das Standardbeispiel einer stetigen Verteilung. Da es sich dabei eben um eine *stetige* Verteilung handelt, bei der also die ZV nicht nur diskrete Werte annehmen kann, muss man zuerst den formellen Apparat von Kapitel 2 in einigen Einzelheiten verändern. Dabei geht es (informell ausgedrückt) um nichts anderes als die Summenzeichen, die beim Berechnen von W. und Erwartungswerten auftreten, durch Integrale zu ersetzen.

Wir betrachten die dabei auftretenden Änderungen am einfachen Beispiel der stetigen Gleichverteilung.

8.1 Unterschiede zu diskreten Verteilungen

Bei der Einführung von Begriffen wie *mathematische Wahrscheinlichkeit* und *Wahrscheinlichkeitsraum* in Definition 2.2 haben wir eine diskrete Menge Ω zugrundegelegt, auf deren Teilmengen eine Funktion P definiert ist, die den Teilmengen von Ω (den Ereignissen) W.Werte zwischen 0 und 1 zuweist.

In diesem Kapitel werden wir die Einschränkung, dass Ω diskret sei aufgeben und die sich daraus ergebenden stetigen W.Räume behandeln. Dabei ist zu beachten, dass an die Stelle der Potenzmenge von Ω eine σ -Algebra über Ω tritt. Eine σ -Algebra über Ω ist eine Menge von Teilmengen von Ω , die gegenüber allen gängigen Mengenoperationen abgeschlossen ist. Sie ist damit fast soviel wie die Potenzmenge von Ω , aber eben nicht ganz. Diese Einschränkung ist dadurch motiviert, dass es keine geeigneten W.Maße P geben würde, die auf der Potenzmenge von Ω definiert sind. Für unsere Untersuchungen ist dieser Unterschied aber nur technischer Natur, und wir können die von uns verwendete σ -Algebra wie eine Potenzmenge behandeln.

Definition 8.1 (Stetiger Wahrscheinlichkeitsraum)

Sei Ω eine beliebige Menge und \mathcal{A} eine σ -Algebra über Ω . Zusammen mit einer Funktion $P : \mathcal{A} \rightarrow [0, 1]$, die die Bedingungen (normiert) und (σ -additiv) aus Definition 2.2 erfüllt, nennt man das Tripel (Ω, \mathcal{A}, P) einen *stetigen Wahrscheinlichkeitsraum*.

Wir beschränken uns hier auf reelle W.Räume: es ist also $\Omega \subseteq \mathbb{R}$ und $\mathcal{A} = \mathcal{B}(\mathbb{R})$. Dabei bezeichnet $\mathcal{B}(\mathbb{R})$ die *Borel'sche σ -Algebra* über \mathbb{R} : die kleinste σ -Algebra, die alle Intervalle enthält.

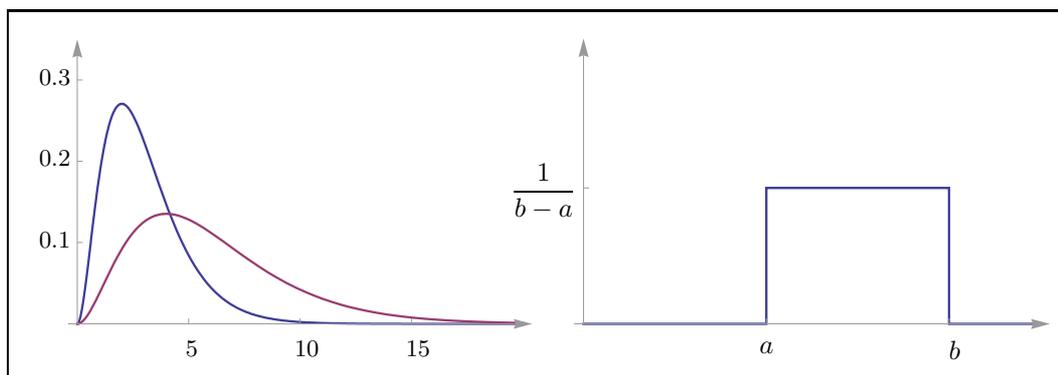


Abbildung 8.1: Graphen der Dichtefunktionen aus Beispiel 8.1.

Jede Menge, die im normalen mathematischen Alltag vorkommt, ist ein Element der Borel'schen σ -Algebra. Dies gilt auch für alle Mengen, die in dieser Vorlesung noch behandelt werden; wir werden also $\mathcal{B}(\mathbb{R})$ im formalen Apparat verwenden, ihr aber ebensowenig Beachtung schenken wie bisher $\text{Pot}(\mathbb{N}_0)$.

Der wichtigste Unterschied zwischen diesem Kapitel und Kapitel 2 ergibt sich bei der Verwendung von *Dichten* zur Berechnung von W.: War bis jetzt die W. eines Elementarereignisses ω durch die W.Funktion von ω gegeben, so werden wir gleich sehen, dass bei stetigen W.Maßen die Elementarereignisse W. null haben! Mit einiger Übung wird uns das aber nicht weiter stören, sondern sogar sehr einleuchtend vorkommen.

Definition 8.2 (Dichte)

Eine Funktion $f : \mathbb{R} \rightarrow \mathbb{R}_0^+$ heißt *Dichte*, wenn

$$\int_{\mathbb{R}} f(x) dx = 1$$

gilt.

Beispiel 8.1 Die Graphen einiger Dichten sind in Abbildung 8.1 zu sehen. Die linke Seite zeigt zwei Dichten der *Erlang-Verteilung*, definiert über

$$f(x) = \begin{cases} \frac{(\lambda x)^{n-1}}{(n-1)!} \lambda e^{-\lambda x} & \text{für } x > 0 \\ 0 & \text{für } x \leq 0 \end{cases}$$

für Parameterwerte $n = 3$ und $\lambda = 1$ bzw. $\lambda = 0.5$.

Die rechte Seite zeigt die Dichte der Gleichverteilung auf $[a, b]$:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{für } a \leq x \leq b \\ 0 & \text{sonst} \end{cases} \quad \square$$

Man beachte den Unterschied zwischen den Definitionen 2.3 und 8.2: Bei letzterer wird die Bedingung, dass die Dichte zu 1 aufsummieren muss explizit gefordert. In

Kapitel 2 war das noch eine Folgerung aus der Tatsache, dass $P(\{x\}) = f(x)$ ist. Der Zusammenhang zwischen Dichte und W.Maß ist für stetige W.Räume aber anders als der zwischen W.Funktionen und W.Maßen in diskreten W.Räumen.

Definition 8.3 (Wahrscheinlichkeit eines Intervalls)

Sei Ω ein reeller Ereignisraum, auf dem eine Dichte f definiert ist. Dann ist die W. jedes Intervalls $[a, b] \subseteq \Omega$ definiert als

$$P([a, b]) = \int_a^b f(x) dx.$$

Jede uns interessierende Menge kann als Vereinigung oder Durchschnitt von Intervallen geschrieben werden, womit obige Definition ausreicht, um im Folgenden W. zu berechnen. Mit obiger Definition gilt speziell

$$P(\{a\}) = \int_a^a f(x) dx = 0.$$

Bei genauer Überlegung wird man feststellen, dass dieses auf den ersten Blick unverständliche Resultat sehr wohl einen Sinn ergibt. Da die W. einzelner Punkte null ist macht es auch keinen Unterschied, ob wir in obiger Definition offene oder geschlossene Intervalle verwenden. Wir betrachten dazu folgendes Beispiel.

Beispiel 8.2 Auf einer Kreisscheibe wird vom Mittelpunkt zu einem beliebigen Punkt am Rand eine Linie gezogen. Anschließend wird die Kreisscheibe an die Wand gehängt und wie ein Glücksrad gedreht. Beobachtet wird dabei der Winkel φ (in Radian), den die aufgezeichnete Linie nach Stillstand mit der Horizontalen einnimmt.

Für jedes Intervall ist die W., dass φ in diesem Intervall liegt, proportional zur Größe des Intervalls. Der Ereignisraum Ω ist das Intervall $[0, 2\pi]$ und das W.Maß darauf die Gleichverteilung. Aus der Bedingung für die Dichte aus Definition 8.2 (oder aus Beispiel 8.1) folgt sofort, dass

$$f(x) = \begin{cases} \frac{1}{2\pi} & \text{für } x \in [0, 2\pi] \\ 0 & \text{sonst} \end{cases}$$

eine geeignete Dichte für dieses Beispiel ist. Die Dichte ist also auf ein Rechteck von Seitenlängen 2π und $\frac{1}{2\pi}$ konzentriert; man vergleiche dazu die rechte Seite von Abbildung 8.1.

Die W., dass φ im Intervall $[a, b]$ liegt, ist somit

$$P([a, b]) = \int_a^b f(x) dx = \frac{b-a}{2\pi}.$$

Auf dieses Ergebnis wäre man vermutlich auch ohne Dichte gekommen. Es ist etwa $P([\pi, \frac{3}{2}\pi]) = \frac{1}{4}$, wie erwartet.

Anhand dieses Beispiels kann man sich auch relativ einfach überlegen, dass $P(\{\omega\}) = 0$ die einzig sinnvolle W. auf Elementarereignissen ist. Das Argument ist ähnlich zu dem in Abschnitt 6.1. Angenommen, wir weisen jedem Elementarereignis ω eine W. $P(\{\omega\}) = c$ zu. Dann enthält jedes Intervall $[a, b]$ überabzählbar viele

Elementarereignisse, deren W. aufsummiert sicher größer als 1 ist; ein Widerspruch zu $P(\Omega) = 1$. \square

Ein weiterer, wenn auch nur kleiner Unterschied zwischen diskreten und stetigen W.Räumen liegt in der Definition von ZV. Definition 5.1 baut auf diskreten W.Räumen und den natürlichen Zahlen als Bildbereich auf; dies kann durch eine einfache Änderung der Definition erweitert werden.

Definition 8.4 (Stetige Zufallsvariable)

Eine *stetige Zufallsvariable* Z auf einem W.Raum (Ω, \mathcal{A}, P) ist eine Funktion

$$Z : \Omega \rightarrow \mathbb{R}.$$

Wie in Kapitel 5 wird durch eine ZV Z eindeutig ein W.Maß auf \mathbb{R} festgelegt: die Verteilung P_Z von Z . Wiederum ist diese Verteilung der ZV durch ihre Dichte schon eindeutig bestimmt; im Gegensatz zu Kapitel 5 ist jetzt allerdings die Gleichheit von W.Funktion und W. von Elementarereignissen nicht mehr gegeben. Wir verwenden dazu die Schreibweise

$$Z^{-1}(B) := \{Z \in B\} = \{\omega \in \Omega \mid Z(\omega) \in B\}$$

für das Urbild von einer Menge B unter Z .

Definition 8.5 (Verteilung einer stetigen Zufallsvariable)

Für beliebiges $B \subseteq \mathbb{R}$ ist die *Verteilung* einer stetigen ZV Z als

$$P_Z(B) = P(Z^{-1}(B)) = \int_{Z^{-1}(B)} f(x)dx = \int_B f_Z(x)dx.$$

Hierbei ist f die Dichte von P und f_Z die Dichte von P_Z .

In dieser Vorlesung ist ein näheres Eingehen auf das Finden von Dichten stetiger Verteilungen nicht zielführend. Das Hauptaugenmerk dieses Kapitels ist die Normalverteilung, von der man die Dichte als Funktion angeben kann. Diese Verteilung werden wir nach der folgenden Erweiterung der Definition des Erwartungswerts genauer behandeln.

Definition 8.6 (Erwartungswert einer stetigen Zufallsvariable)

Der *Erwartungswert* einer stetigen ZV Z ist definiert als

$$E(Z) = \int_{\mathbb{R}} x f_Z(x)dx.$$

Diese Definition ist also völlig analog zur Einführung des Erwartungswerts diskreter ZV in Definition 7.1. Der Rest von Kapitel 7 kann ebenfalls sofort übernommen werden, wenn die Summen durch Integrale ersetzt werden.

Beispiel 8.3 (siehe Beispiel 8.2) Wir berechnen den Erwartungswert und die Varianz des Winkels φ , dessen Dichte als

$$f(x) = \begin{cases} \frac{1}{2\pi} & \text{für } x \in [0, 2\pi] \\ 0 & \text{sonst} \end{cases}$$

gegeben ist. Wir haben den Erwartungswert nur für die Verteilung einer ZV definiert; um in diesem formalen Gerüst zu bleiben, kann man als ZV die Identitätsabbildung verwenden.

Es ergibt sich dann für den Erwartungswert nach Definition 8.6

$$\begin{aligned} E(Z) &= \int_0^{2\pi} x \frac{1}{2\pi} dx = \frac{1}{2\pi} \int_0^{2\pi} x dx \\ &= \frac{1}{2\pi} \frac{x^2}{2} \Big|_0^{2\pi} = \pi, \end{aligned}$$

also genau die Intervallmitte.

Für die Berechnung der Varianz nach Satz 7.6 benötigen wir $E(Z^2)$; dies ist mit Satz 7.5

$$E(Z^2) = \int_0^{2\pi} x^2 \frac{1}{2\pi} dx = \frac{1}{2\pi} \frac{x^3}{3} \Big|_0^{2\pi} = \frac{4\pi^2}{3},$$

woraus sich für die Varianz

$$\text{Var}(Z) = E(Z^2) - (E(Z))^2 = \frac{4\pi^2}{3} - \pi^2 = \frac{\pi^2}{3}. \quad \square$$

Die Ergebnisse des letzten Beispiels lassen sich einfach zu einer Formel für stetig gleichverteilte ZV verallgemeinern.

Satz 8.1 Für eine ZV Z , die auf dem Intervall $[a, b]$ gleichverteilt ist, gilt

$$\begin{aligned} E(Z) &= \frac{a+b}{2} \quad \text{und} \\ \text{Var}(Z) &= \frac{(b-a)^2}{12} \end{aligned}$$

Beweis Nachrechnen von Beispiel 8.3. □

Wir betrachten zum Abschluss dieser Überlegungen ein weiteres Beispiel einer einfachen stetigen Verteilung.

Beispiel 8.4 Die *Exponential-Verteilung* ist definiert über die Dichte

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{für } x \in [0, \infty) \\ 0 & \text{sonst.} \end{cases}$$

Einfaches Integrieren zeigt, dass diese Funktion eine Dichte definiert. Man kann über die bekannten Integrationsregel den Erwartungswert und die Varianz dieser

Verteilung nachrechnen. Für eine ZV X mit Exponentialverteilung gilt mit partieller Integration und der Regel von de l'Hôpital

$$\begin{aligned} E(X) &= \int_0^{\infty} x \lambda e^{-\lambda x} dx = -x e^{-\lambda x} \Big|_0^{\infty} + \int_0^{\infty} e^{-\lambda x} dx \\ &= 0 - \frac{1}{\lambda} e^{-\lambda x} \Big|_0^{\infty} = \frac{1}{\lambda}. \end{aligned}$$

Ebenso kann man für die Varianz von X nachrechnen:

$$\begin{aligned} E(X^2) &= \int_0^{\infty} x^2 \lambda e^{-\lambda x} dx = -x^2 e^{-\lambda x} \Big|_0^{\infty} + 2 \int_0^{\infty} x e^{-\lambda x} dx \\ &= 0 + 2 \left(-x \frac{1}{\lambda} e^{-\lambda x} \Big|_0^{\infty} + \frac{1}{\lambda} \int_0^{\infty} e^{-\lambda x} dx \right) = \frac{2}{\lambda^2}; \end{aligned}$$

somit gilt

$$\text{Var}(X) = E(X^2) - (E(X))^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

Die W. von beliebigen Intervallen lassen sich ähnlich berechnen:

$$P_X([a, b]) = \int_a^b \lambda e^{-\lambda x} dx = -e^{-\lambda x} \Big|_a^b = e^{-\lambda a} - e^{-\lambda b}. \quad \square$$

Nach diesen abschließenden Resultaten über stetige Verteilungen wird im nächsten Abschnitt die Normalverteilung und ihre herausragende Rolle in der W.Rechnung behandelt.

8.2 Die Normalverteilung

Die Normalverteilung lässt sich nicht auf einfache Weise als Verteilung der Ausgänge eines Zufallsexperiments einführen, wie wir das in Kapitel 6 für die Binomial-, hypergeometrische und Poisson-Verteilung getan haben. Wir definieren daher gleich zu Beginn die Normalverteilung über ihre Dichte und illustrieren dann erst die vielfältigen Verwendungsmöglichkeiten dieser Verteilung.

Definition 8.7 (Normalverteilung)

Eine stetige Verteilung P_Z mit Dichte $f_Z : \mathbb{R} \rightarrow \mathbb{R}_0^+$, für die

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

gilt, heißt *Normalverteilung* (oder *Gaußverteilung*) mit Erwartungswert μ und Varianz σ^2 . Abkürzend schreiben wir für P_Z meist $N(\mu, \sigma^2)$.

Die Dichte der Normalverteilung für einen Wert μ und zwei Werte für σ ist in Abbildung 8.2 zu sehen. Für die Normalverteilung gelten folgende Tatsachen, die nicht ganz einfach auszurechnen sind:

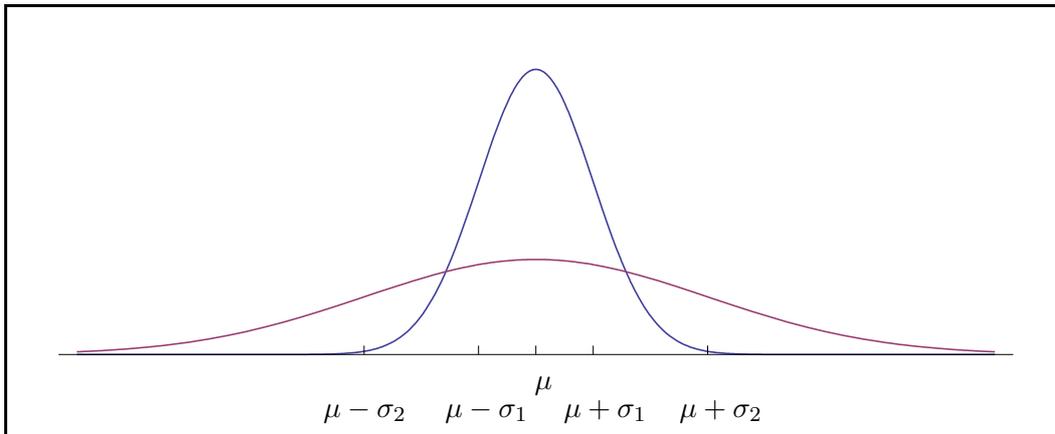


Abbildung 8.2: Zwei Dichten der Normalverteilung für verschiedene Werte der Standardabweichung σ .

- Die Dichte f_Z erfüllt die Bedingung $\int_{\mathbb{R}} f_Z(x) dx = 1$ (die ja an alle Dichten gestellt wird).
- Für eine $N(\mu, \sigma^2)$ -verteilte ZV Z gilt $E(Z) = \mu$ und $\text{Var}(Z) = \sigma^2$.
- Die Punkte $\mu - \sigma$ und $\mu + \sigma$ sind die Wendepunkte des Graphen der Dichte.

Aus dem quadratischen Term $(x - \mu)^2$ in Definition 8.7 ist hingegen einfach zu sehen, dass der Graph der Dichte um den Punkt μ symmetrisch ist.

Das Berechnen von W. einer $N(\mu, \sigma^2)$ -verteilten ZV Z ist auf analytischem Weg nicht machbar, d.h. zur Berechnung von

$$P_Z([a, b]) = \int_a^b f_Z(x) dx$$

muss das Integral auf der rechten Seite numerisch evaluiert werden. Natürlich ist es nicht sehr praktisch, für jede neue Kombination von μ und σ^2 diese Berechnung durchführen zu müssen. Der folgende Satz besagt, dass man mit *einer* Tabelle berechneter Werte für alle Kombinationen von μ und σ^2 auskommt:

Satz 8.2 Sei Z eine $N(\mu, \sigma^2)$ -verteilte ZV. Dann ist die ZV $Z' = \frac{Z - \mu}{\sigma} \sim N(0, 1)$ -verteilt. Die so erhaltene ZV nennt man *standardisierte Zufallsvariable*.

Beweis Einfaches Nachrechnen: Mit Satz 7.5 und der Bedingung $E(Z) = \mu$ ist

$$\begin{aligned} E(Z') &= E\left(\frac{Z - \mu}{\sigma}\right) \\ &= \frac{1}{\sigma} E(Z) - \frac{\mu}{\sigma} = 0; \end{aligned}$$

ebenso ist mit Satz 7.7 und der Bedingung $\text{Var}(Z) = \sigma^2$

$$\text{Var}(Z') = \text{Var}\left(\frac{Z - \mu}{\sigma}\right)$$

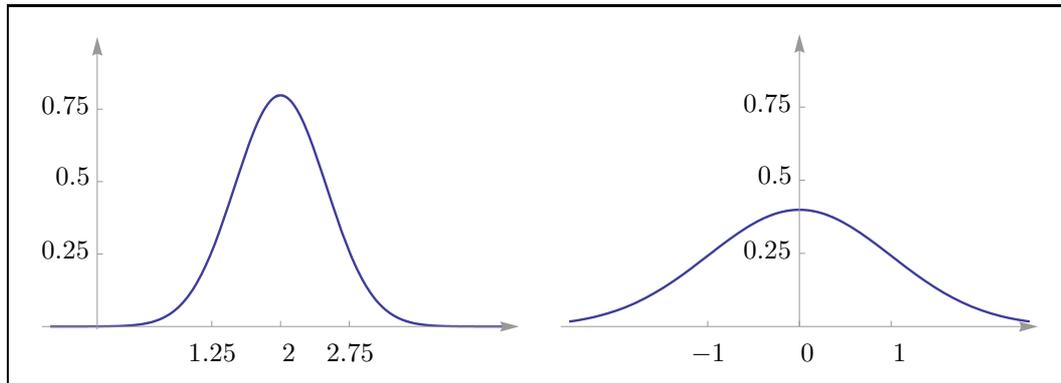


Abbildung 8.3: Veränderung der Dichte einer $N(2, 0.75^2)$ -verteilten ZV bei Standardisierung.

$$= \frac{1}{\sigma^2} \text{Var}(Z) = 1. \quad \square$$

In Abbildung 8.3 ist die Veränderung einer Dichte einer ZV durch Standardisierung zu erkennen.

Wie man im Folgenden sehen kann genügt es, Wertungswerte für $N(0, 1)$ -verteilte ZV zu tabellieren. Dazu benötigt man die *Verteilungsfunktion* der standardisierten Normalverteilung. Verteilungsfunktionen werden dabei wie folgt definiert.

Definition 8.8 (Verteilungsfunktion)

Sei Z eine ZV mit mit Verteilung P_Z und Dichte f_Z . Dann nennt man die Funktion

$$F_Z(x) = \int_{-\infty}^x f_Z(t) dt$$

die *Verteilungsfunktion* von P_Z .

Für die Dichte $\varphi(t)$ einer $N(0, 1)$ -verteilten ZV ergibt sich als Verteilungsfunktion

$$\Phi(x) = \int_{-\infty}^x \varphi(t) dt$$

Der Zusammenhang zwischen $\Phi(x)$ und $\varphi(x)$ ist graphisch in Abbildung 8.4 dargestellt: $\Phi(x)$ gibt die Fläche unter dem Graphen von $\varphi(t)$ zwischen $-\infty$ und x an.

Mit Hilfe dieser Verteilungsfunktion lassen sich W. von Intervallen normalverteilter ZV berechnen. Für $N(0, 1)$ -verteilte ZV Z' gilt

$$P_{Z'}([a, b]) = \int_a^b \varphi(t) dt = \int_{-\infty}^b \varphi(t) dt - \int_{-\infty}^a \varphi(t) dt = \Phi(b) - \Phi(a).$$

Andere normalverteilte ZV müssen zuerst standardisiert werden, bevor W. berechnet werden können. Sei dazu Z eine $N(\mu, \sigma^2)$ -verteilte ZV. Dann gilt

$$P_Z([a, b]) = P(Z \in [a, b]) = P\left(\frac{Z - \mu}{\sigma} \in \left[\frac{a - \mu}{\sigma}, \frac{b - \mu}{\sigma}\right]\right)$$

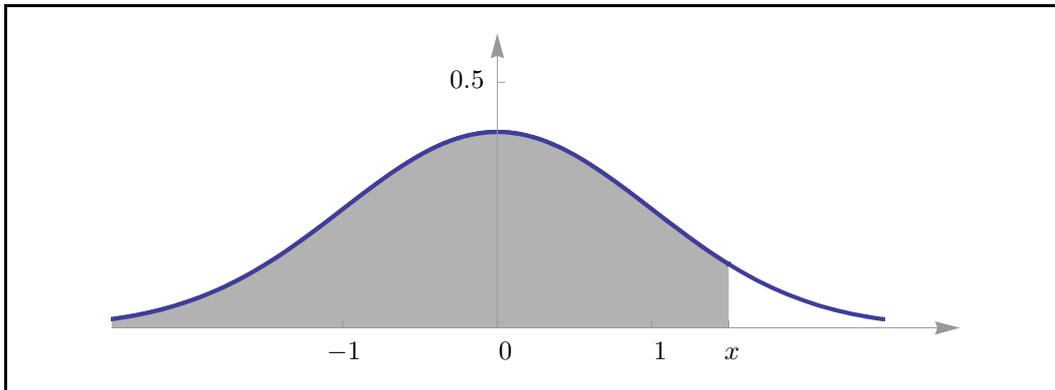


Abbildung 8.4: Zusammenhang zwischen Dichte $\varphi(t)$ (Graph) und Verteilungsfunktion $\Phi(x)$ (schraffierte Fläche) der Standard-Normalverteilung $N(0, 1)$.

$$= P_{Z'} \left(\left[\frac{a - \mu}{\sigma}, \frac{b - \mu}{\sigma} \right] \right) = \Phi \left(\frac{b - \mu}{\sigma} \right) - \Phi \left(\frac{a - \mu}{\sigma} \right).$$

Somit genügt es für die W.Berechnung *beliebiger* normalverteilter ZV, die Verteilungsfunktion $\Phi(x)$ tabelliert zu haben. Solche Tabellen (wie hier Tabelle 8.1) findet man in praktisch allen Büchern über W.Rechnung, wobei wegen der Symmetrie von $\varphi(t)$ und der sich daraus ergebenden Beziehung

$$\Phi(-x) = 1 - \Phi(x)$$

meist nur die positiven Werte tabelliert sind. Die Werte $\Phi(x)$ für Argumente x , die nicht in obiger Tabelle aufscheinen, berechnet man am besten mit linearer Interpolation: Sei etwa $a \leq x \leq b$, wobei die Werte von a und b in der Tabelle aufscheinen. Dann muss die relative Entfernung, in der sich x zu a und b befindet, gleich der relativen Entfernung des unbekanntes $\Phi(x)$ von $\Phi(a)$ und $\Phi(b)$ sein:

$$\frac{x - a}{b - a} = \frac{\Phi(x) - \Phi(a)}{\Phi(b) - \Phi(a)};$$

daraus kann man sich den Wert für $\Phi(x)$ ausrechnen.

Beispiel 8.5 Aus langjähriger Erfahrung weiß eine Fluglinie, dass die Flugzeit der Maschine von Frankfurt nach Boston normalverteilt ist mit $\mu = 7.5 \text{ h}$ und $\sigma = 15 \text{ min}$. Was ist die W., dass ein Flug länger als 8 h dauert?

Lösung: Laut Angabe ist die Flugdauer $X \sim N(7.5, 0.25^2)$ verteilt. Gesucht ist die W. $P(X > 8)$. Unter Verwendung einer standardisierten ZV X' ergibt sich

$$\begin{aligned} P(X > 8) &= P\left(\frac{X - 7.5}{0.25} > \frac{8 - 7.5}{0.25}\right) = P(X' > 2) \\ &= 1 - P(X' < 2) = 1 - \Phi(2) = 0.023. \end{aligned}$$

Die W., dass ein Flug länger als 8 h dauert ist somit etwa 2.3%. \square

Neben dieser direkten Berechnung von W. mit Hilfe der Normalverteilung können auch "inverse" Probleme gelöst werden. Bei diesen Problemstellungen ist eine W.

x	00	25	50	75	x	00	25	50	75
0.0	0.500	0.510	0.520	0.530	1.5	0.933	0.936	0.939	0.942
0.1	0.540	0.550	0.560	0.569	1.6	0.945	0.948	0.951	0.953
0.2	0.579	0.589	0.599	0.608	1.7	0.955	0.958	0.960	0.962
0.3	0.618	0.627	0.637	0.646	1.8	0.964	0.966	0.968	0.970
0.4	0.655	0.665	0.674	0.683	1.9	0.971	0.973	0.974	0.976
0.5	0.691	0.700	0.709	0.717	2.0	0.977	0.979	0.980	0.981
0.6	0.726	0.734	0.742	0.750	2.1	0.982	0.983	0.984	0.985
0.7	0.758	0.766	0.773	0.781	2.2	0.986	0.987	0.988	0.989
0.8	0.788	0.795	0.802	0.809	2.3	0.989	0.990	0.990	0.991
0.9	0.816	0.826	0.829	0.835	2.4	0.992	0.992	0.993	0.993
1.0	0.841	0.847	0.853	0.859	2.5	0.994	0.994	0.995	0.995
1.1	0.864	0.870	0.875	0.880	2.6	0.995	0.996	0.996	0.996
1.2	0.885	0.890	0.894	0.899	2.7	0.997	0.997	0.997	0.997
1.3	0.903	0.907	0.911	0.915	2.8	0.997	0.998	0.998	0.998
1.4	0.919	0.923	0.926	0.930	2.9	0.998	0.998	0.998	0.999

Tabelle 8.1: Tabellierte Werte der Verteilungsfunktion $\Phi(x)$ der Standard-Normalverteilung $N(0, 1)$ für $x \in [0, 2.975]$. Die ersten zwei Stellen von x sind neben den Zeilen, die nächsten zwei über den Spalten aufgetragen.

gegeben; gesucht ist ein Schwellenwert, ab dem bzw. bis zu diesem diese W. erreicht wird.

Beispiel 8.6 Der *Club der Hagenberger und Mühlviertler Programmier-Superstars* (CHUMPS) veranstaltet einen Programmierwettbewerb, um die Aufnahme in seinen elitären Klub zu regeln (nur jemand, der genügend Punkte erreicht, kann aufgenommen werden). Aus Erfahrung wissen die Mitglieder, dass die Scores dieses Wettbewerbs normalverteilt sind mit $\mu = 120$ und $\sigma = 20$.

Wie hoch muss der Schwellenwert für Aufnahme in diesen Klub angesetzt werden, damit nur 10% derer, die beim Wettbewerb mitmachen, mindestens diese Punktezahl erreichen?

Lösung: Sei X die $N(120, 20^2)$ -verteilte ZV, die die Scores der Programmierer angibt. Gesucht ist somit derjenige Wert c mit der Eigenschaft $P(X > c) \leq 0.1$. Standardisieren liefert

$$P(X > c) = P\left(\frac{X - 120}{20} > \frac{c - 120}{20}\right) = P\left(X' > \frac{c - 120}{20}\right).$$

Die Ungleichung

$$\begin{aligned} P\left(X' > \frac{c - 120}{20}\right) \leq 0.1 &\Leftrightarrow P\left(X' < \frac{c - 120}{20}\right) \geq 0.9 \\ &\Leftrightarrow \Phi\left(\frac{c - 120}{20}\right) \geq 0.9 \end{aligned}$$

lässt sich durch Zuhilfenahme von Tabelle 8.1 ungefähr lösen, indem man diese Tabelle “umgekehrt” liest, also denjenigen Wert x bestimmt, für den $\Phi(x) = 0.9$ gilt. Dies ist bei $x \approx 1.275$ der Fall; einen besseren Wert erhält man durch die oben angesprochene Proportionalitäts-Berechnung über das Verhältnis

$$\frac{x - 1.275}{1.3 - 1.275} = \frac{\Phi(x) - \Phi(1.275)}{\Phi(1.3) - \Phi(1.275)}.$$

Mit $\Phi(x) = 0.9$ erhält man daraus durch Umformen $x = 1.28125$. Einsetzen dieses Werts in die noch offene Ungleichung ergibt

$$\frac{c - 120}{20} = 1.28125 \quad \Leftrightarrow \quad c = 145.625.$$

Ein Programmierer muss somit mindestens 146 Punkte erreichen, um bei den CHUMPS aufgenommen zu werden. \square

Zum Abschluss dieser einführenden Überlegungen zur Normalverteilung sei noch das Resultat erwähnt, das die herausragende Stellung dieser Verteilung erläutert.

Satz 8.3 (Zentraler Grenzwertungssatz) Sei X_1, X_2, \dots eine Folge von unabhängigen, identisch verteilten ZV, d.h., für alle i gilt $E(X_i) = \mu$ und $\text{Var}(X_i) = \sigma^2$. Weiters sei $S_n = X_1 + \dots + X_n$ die Summe der ersten n ZV und S'_n die standardisierte Summe $S'_n = (S_n - E(S_n))/\sqrt{\text{Var}(S_n)}$. Dann gilt

$$\lim_{n \rightarrow \infty} P(S'_n \in [a, b]) = \Phi(b) - \Phi(a).$$

Diese Formel besagt, dass die standardisierte Summe von ZV (die eine beliebige Verteilung haben können!) im Limit standard-normalverteilt ist. Als Faustregel gilt: Obige Formel liefert ab etwa $n \geq 30$ brauchbare Resultate.

Erläuterungen Die Operation der *Standardisierung*, die in Satz 8.2 für eine $N(\mu, \sigma^2)$ -verteilte ZV eingeführt wurde, lässt sich auf beliebig verteilte ZV anwenden. Dabei wird der Erwartungswert der ZV abgezogen und durch die Standardabweichung dividiert. Man erhält dann eine ZV, die die gleiche Verteilung wie die ursprüngliche ZV aufweist, allerdings mit Erwartungswert 0 und Varianz (und Standardabweichung) 1.

In der Definition von S'_n im obigen Satz werden $E(S_n)$ und $\text{Var}(S_n)$ verwendet: Wegen Satz 7.14 über den Erwartungswert einer Summe von ZV ist

$$E(S_n) = E(X_1 + \dots + X_n) = E(X_1) + \dots + E(X_n) = n \cdot \mu;$$

ebenso gilt wegen der Unabhängigkeit der X_i Satz 7.16 und somit ist

$$\text{Var}(S_n) = \text{Var}(X_1 + \dots + X_n) = \text{Var}(X_1) + \dots + \text{Var}(X_n) = n \cdot \sigma^2.$$

Mit obigen Vereinfachungen für $E(S_n)$ und $\text{Var}(S_n)$ kann man die Aussage von Satz 8.3 auch als “ $X_1 + \dots + X_n$ ist $N(n \cdot \mu, n \cdot \sigma^2)$ -verteilt” bzw. als

$$\lim_{n \rightarrow \infty} P\left(\frac{X_1 + \dots + X_n - n\mu}{\sqrt{n\sigma^2}} \in [a, b]\right) = \Phi(b) - \Phi(a).$$

schreiben.

Nach dieser Fülle von theoretischen Resultaten folgen nun einige Beispiele zur Normalverteilung und zum zentralen Grenzwertungssatz.

Beispiel 8.7 Ein Zufallsexperiment, bei dem ein Ereignis A mit W. p eintritt und mit W. $1 - p$ nicht eintritt, werde n mal unabhängig wiederholt. Die ZV S_n gebe an, wie oft das Ereignis A in der Versuchsreihe eingetreten ist.

Wie wir schon aus Abschnitt 6.3 wissen, ist S_n $\text{Bin}(n, p)$ -verteilt. Für großes n und viele Werte von k kann das Berechnen von $\text{Bin}(n, p)(k)$ arbeitsaufwändig werden. So sei etwa die folgende W. zu berechnen: Ein Würfel wird 1000 mal geworfen. Was ist die W., dass dabei höchstens 160mal eine Sechs geworfen wird?

Mit Hilfe der Binomialverteilung ist diese W. als

$$\text{Bin}\left(1000, \frac{1}{6}\right)(\{0, \dots, 160\}) = \sum_{k=0}^{160} \binom{1000}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{n-k}$$

leicht anzuschreiben, aber nicht unbedingt leicht auszurechnen. Abhilfe schafft hier der zentrale Grenzwertungssatz. Wir gehen dabei auf die ZV X_i zurück, die Bernoulli-verteilt sind (siehe Abschnitt 6.2) und angeben, ob A bei der i -ten Wiederholung des Experiments eingetreten ist ($X_i = 1$) oder nicht ($X_i = 0$).

Für Erwartungswert und Varianz einer Bernoulli-verteliten ZV X_i gilt $E(X_i) = p$ und $\text{Var}(X_i) = p(1 - p)$ (siehe Beispiel 7.13). Wegen des zentralen Grenzwertungssatzes ist $S_n = X_1 + \dots + X_n$ normalverteilt mit Erwartungswert np und Varianz $np(1 - p)$; somit gilt

$$\begin{aligned} P(S_n \in [a, b]) &= P\left(S'_n \in \left[\frac{a - np}{\sqrt{np(1 - p)}}, \frac{b - np}{\sqrt{np(1 - p)}}\right]\right) \\ &= \Phi\left(\frac{b - np}{\sqrt{np(1 - p)}}\right) - \Phi\left(\frac{a - np}{\sqrt{np(1 - p)}}\right), \end{aligned}$$

wobei $S'_n = (S_n - E(S_n))/\sqrt{\text{Var}(S_n)}$ wiederum die standardisierte Summe ist.

Für unsere konkreten Zahlen $n = 1000$ und $p = \frac{1}{6}$ ergibt sich dann

$$\begin{aligned} P(S_n \in \{0, \dots, 160\}) &\approx \Phi\left(\frac{160 - 1000\frac{1}{6}}{\sqrt{1000\frac{1}{6}\frac{5}{6}}}\right) - \Phi\left(\frac{0 - 1000\frac{1}{6}}{\sqrt{1000\frac{1}{6}\frac{5}{6}}}\right) \\ &= \Phi(-0.565) - \Phi(-14.142) \\ &= 1 - \Phi(0.565) - (1 - \Phi(14.142)) = 0.286. \end{aligned}$$

Das echte Ergebnis ist 0.303, was einen relativen Fehler von $(0.303 - 0.286)/0.303 = 5.6\%$ ausmacht. Man kann das Ergebnis noch verbessern, indem man bei der Approximation durch die Normalverteilung noch einen Korrekturterm einfließen lässt, der den Unterschied zwischen diskreter (Binomial-) und stetiger (Normal-)Verteilung kompensieren soll. Dazu ersetzt man die diskreten Intervallgrenzen a und b durch $a - 0.5$ bzw. $b + 0.5$, macht also das stetige Intervall etwas größer.

Für unser Beispiel erhält man das wesentlich genauere Ergebnis

$$\begin{aligned} P(S_n \in \{0, \dots, 160\}) &\approx \Phi\left(\frac{160 + \frac{1}{2} - 1000\frac{1}{6}}{\sqrt{1000\frac{1}{6}\frac{5}{6}}}\right) - \Phi\left(\frac{0 - \frac{1}{2} - 1000\frac{1}{6}}{\sqrt{1000\frac{1}{6}\frac{5}{6}}}\right) \\ &= \Phi(-0.523) - \Phi(-14.185) \\ &= 1 - \Phi(0.523) - (1 - \Phi(14.185)) = 0.300. \quad \square \end{aligned}$$

Der in diesem Beispiel behandelte Zusammenhang zwischen Binomial- und Normalverteilung gilt auch allgemein, wie der nächste Satz besagt.

Satz 8.4 (Approximation durch Normalverteilung)

Für $np(1-p) \geq 10$ gilt näherungsweise

$$\text{Bin}(n, p) \approx N(np, np(1-p)),$$

für $n \frac{M}{N} (1 - \frac{M}{N}) \geq 10$ und $\frac{n}{N} < 0.05$ ist annähernd

$$\text{Hypergeom}(N, M, n) \approx N\left(n \frac{M}{N}, n \frac{M}{N} \left(1 - \frac{M}{N}\right) \frac{N-n}{N-1}\right),$$

und für $\lambda \geq 10$ gilt näherungsweise

$$\text{Poisson}(\lambda) \approx N(\lambda, \lambda).$$

Beweis Für Binomialverteilung: Zentralen Grenzwertungssatz wie in Beispiel 8.7 anwenden. Für hypergeometrische und Poisson-Verteilung gilt Ähnliches unter Verwendung der entsprechenden Erwartungswerte und Varianzen. \square

Zwei Anmerkungen zu diesem Satz: Zum ersten sollte für genauere Ergebnisse ein Korrekturterm verwendet werden, wie in Beispiel 8.7 für die Binomialverteilung demonstriert. Dieser ist für die beiden anderen Verteilungen vollkommen analog (also $+\frac{1}{2}$ für die obere Bereichsgrenze, $-\frac{1}{2}$ für die untere). Zum zweiten illustriert obiger Satz die zentrale Rolle der Normalverteilung, da durch sie die bekanntesten diskreten Verteilungen approximiert werden können. Für das praktische Rechnen ist er durch die weite Verbreitung von mathematischen Softwaresystemen allerdings nicht mehr so wichtig, da mit solchen Systemen auch händisch aufwändige Berechnungen einfach durchgeführt werden können.

Beispiel 8.8 Wie hoch ist die W., bei 1000-maligem Setzen auf einfache Chance beim Roulette (also etwa “rouge” oder “pair”) mindestens 500 mal zu gewinnen?

Lösung: Die ZV, die die Anzahl der Gewinne bei 1000-maligem Spielen angibt, ist $\text{Bin}(1000, \frac{18}{37})$ -verteilt. Die gesuchte W. ist mit Satz 8.4 und Korrekturterm

$$\begin{aligned} \text{Bin}\left(1000, \frac{18}{37}\right)(\{500, \dots, 1000\}) &\approx \Phi\left(\frac{1000 + \frac{1}{2} - \frac{18000}{37}}{\sqrt{1000 \frac{18}{37} \frac{19}{37}}}\right) - \Phi\left(\frac{500 - \frac{1}{2} - \frac{18000}{37}}{\sqrt{1000 \frac{18}{37} \frac{19}{37}}}\right) \\ &= \Phi(32.5209) - \Phi(0.8233) = 1 - 0.7948 \\ &= 0.205. \end{aligned}$$

Dieses Ergebnis stimmt auf die angegebenen drei Nachkommastellen mit dem exakten Ergebnis überein. \square

Beispiel 8.9 (siehe Beispiel 5.4) Eine Münze wird n mal geworfen. Man berechne für große n die Anzahl der Runs in den Münzwürfen.

Lösung: Für jedes i gebe die ZV Y_i an, ob das Resultat des i -ten Werfens KOPF ($Y_i = 1$) oder ZAHL ($Y_i = 0$) ist. Wir betrachten jeweils Paare von aufeinanderfolgenden Resultaten (Y_i, Y_{i+1}) um zu zählen, wie oft ein Run unterbrochen wurde. Sei dazu X_i für $i = 1, \dots, n-1$ definiert als

$$X_i = \begin{cases} 1 & \text{für } Y_i \neq Y_{i+1} \\ 0 & \text{für } Y_i = Y_{i+1} \end{cases}$$

Es ist ersichtlich, dass die ZV $S_n = 1 + X_1 + \dots + X_{n-1}$ gleich der Anzahl der Runs ist. Da alle X_i unabhängig und identisch verteilt sind (nämlich Bernoulli-verteilt mit Parameter $p = \frac{1}{2}$), lässt sich der zentrale Grenzwertsatz anwenden. S_n ist als Summe von unabhängigen und identisch verteilten ZV annähernd normalverteilt; die Parameter dieser Normalverteilung errechnet man als

$$E(S_n) = 1 + E(X_1) + \dots + E(X_{n-1}) = \frac{n+1}{2}$$

und

$$\text{Var}(S_n) = \text{Var}(X_1) + \dots + \text{Var}(X_{n-1}) = \frac{n-1}{4}.$$

Somit ist S_n für große n annähernd $N(\frac{n+1}{2}, \frac{n-1}{4})$ -verteilt. □

Einführung in die induktive Statistik

Bei unserer Behandlung der W.Theorie standen bis jetzt genau definierte Zufallsexperimente im Vordergrund, wie etwa Würfeln, Werfen von Münzen, oder das Ziehen aus einer Urne. Jedes dieser Experimente konnte genau beschrieben werden, und zwar auf zwei Ebenen:

- Die Verteilungen der ZV, die Ausgänge von Experimenten beschrieben, waren bekannt.
- Die Parameter der Verteilungen waren ebenfalls bekannt.

In diesem Kapitel verwenden wir die Theorie der W.Rechnung, um uns der Beschreibung von Zufallsexperimenten von der anderen Seite zu nähern: Wenn Ergebnisse von Zufallsexperimente in Form von numerischen Daten vorliegen, wir aber keine a-priori Kenntnisse über einen der folgenden Punkte haben:

- über die Verteilung der ZV, deren numerische Werte als Daten vorliegen,
- wenn die Verteilung zwar bekannt ist, nicht aber der (die) Parameter dieser Verteilung.

Wir werden diese Fragen hier nur streifen können, da eine genaue Behandlung weit in den Bereich der Statistik reichen würde. Das Hauptaufgabengebiet der *induktiven Statistik* (auch *schließende Statistik* genannt) liegt darin, ausgehend von einer gegebenen Datenmenge Rückschlüsse zu ziehen auf die Gesamtpopulation, aus der die Daten gewonnen wurden. Es geht also darum, wie etwa bei der Meinungsforschung, auf Basis weniger Daten Aussagen über eine größere Datenmenge machen zu können.

Im Gegensatz zur induktiven Statistik ist das Aufgabengebiet der *deskriptiven* (*beschreibenden*) Statistik die Problemstellung, eine Gesamtpopulation von Daten durch einige wenige Kennzahlen zu beschreiben. Man will hier also *nicht* von einer vorliegenden Datenmenge auf eine andere (größere) schließen, sondern hat alle Daten bereits zur Verfügung, und möchte diese möglichst kurz und prägnant beschreiben.

In diesem Kapitel werden wir uns hauptsächlich mit einigen Thematiken der induktiven Statistik auseinandersetzen. In Abschnitt 9.1 geben wir dennoch einen kurzen Überblick über wichtige Begriffe der deskriptiven Statistik.

9.1 Deskriptive Statistik

Wie bereits oben erwähnt, wird in der deskriptiven Statistik eine Menge $\{x_1, \dots, x_n\}$ von Messwerten als gegeben betrachtet. Zusätzlich bilden diese Messwerte eine abgeschlossene Grundgesamtheit, sind also speziell nicht nur ein Teil eines größeren Menge.

In der deskriptiven Statistik werden eine oder mehrere *Merkmale* einer Grundgesamtheit betrachtet. Diese Merkmale können unterschiedlicher Natur sein, je nachdem, welche *Merkmalsausprägungen* (*Beobachtungswerte*) möglich sind:

Metrische (quantitative) Merkmale Die Merkmalsausprägungen sind Zahlen oder allgemein Werte, auf denen eine Metrik definierbar ist. Je nach Zahlentyp können metrische Merkmale in *stetige* oder *diskrete* Merkmale unterteilt werden. Als *quasi-stetige* Merkmale bezeichnet man diejenigen Merkmale, deren Ausprägungen zwar diskret sind, aber leichter als stetige Merkmale zu behandeln sind (etwa große Geldbeträge).

Ordinale Merkmale Ausprägungen der Merkmale dieser Art können zwar miteinander verglichen werden (es besteht eine lineare Ordnung), es gibt aber—im Gegensatz zu metrischen Merkmalen—keine Metrik auf diesen Ausprägungen.

Nominale (qualitative) Merkmale Die Ausprägungen dieser Merkmale sind Werte, auf denen weder Metrik noch Ordnung sinnvoll zu definieren ist.

Wir betrachten ein Beispiel zur Erläuterung dieser Merkmalsklassen.

Beispiel 9.1 Bei allen Studierenden der FH Hagenberg sollen folgende Merkmale gemessen werden: Augenfarbe, Alter (in Jahren), Note in der erstsemestrigen Mathematikvorlesung, Monatseinkommen (in Euro), Körpergröße. Es handelt bei diesen Merkmalen um

- ein nominales Merkmal (Augenfarbe),
- ein ordinales Merkmal (Mathematik-Note),
- ein quantitativ-diskretes Merkmal (Alter in Jahren),
- ein quantitativ-stetiges Merkmal (Körpergröße), und
- ein quantitativ-quasi-stetiges Merkmal (Monatseinkommen).

Zu diskutieren wäre, ob eine Note ein ordinales oder metrisches Merkmal darstellt, ob also der Leistungsunterschied zwischen SGT1 und BEF3 gleich dem zwischen GUT2 und GEN4 ist (wie es bei einem metrischen Merkmal der Fall wäre), oder ob diese Unterschiede nicht vergleichbar sind—dies würde für ein ordinales Merkmal sprechen. \square

Zur Beschreibung von Merkmalen bieten sich Lage- und Streuungsmaßzahlen an. Die bekanntesten Lagemaßzahlen einer Menge $\{x_1, \dots, x_n\}$ sind das *arithmetische Mittel*

$$\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i$$

und der *Median* m , der als mittlerer Wert einer geordneten Zahlenfolge definiert ist. Mit $(x_{s_1}, \dots, x_{s_n})$ als aufsteigend sortierte Folge der Werte in der Datenmenge gilt

$$m := \begin{cases} x_{s_{(n+1)/2}} & \text{für ungerades } n \\ (x_{s_{n/2}} + x_{s_{n/2+1}})/2 & \text{für gerades } n; \end{cases}$$

Der Median ist somit für den Fall, dass kein mittlerer Wert existiert, als das arithmetische Mittel der beiden mittleren Werte definiert.

Die Summenoperation in der Definition des arithmetischen Mittels bedeutet, dass nur für metrische Daten das arithmetische Mittel berechnet werden kann. Beim Median benötigt man nur die Ordnung auf den Daten, sodass diese Lagemaßzahl auch für ordinale Merkmale berechnet werden kann. Da die Metrik bei der Medianberechnung keine Rolle spielt, ist der Median die robustere Maßzahl, wenn unterschiedlich große Werte zusammengefasst werden. Dies wird speziell in Beispiel 9.3 deutlich.

Beispiel 9.2 Die Grundmenge seien die $n = 8$ metrischen Merkmalsausprägungen $\{3, 2, 1, 5, 8, 7, 8, 8\}$. Das arithmetische Mittel ist $\bar{x} = \frac{21}{4} = 5.25$, der Median ist 6. \square

Beispiel 9.3 Im normalen Sprachgebrauch ist mit ‘‘Durchschnitt’’ meist das arithmetische Mittel gemeint, manchmal aber auch der Median. So kann etwa der Chef einer 10-köpfigen Firma damit angeben, dass das Nettogehalt in seiner Firma durchschnittlich € 5 000 beträgt. Meist verschweigt er dann, dass er seinen neun Mitarbeitern je € 2 000 bezahlt, und sich selbst € 32 000. Der Median dieser Datenmenge ist € 2 000 und damit eine bessere Wiedergabe des Gehaltsgefüges. \square

Als Streuungsmaßzahl einer Grundgesamtheit werden die *Populationsvarianz* s^2 und davon abgeleitet die *Populations-Standardabweichung* s verwendet, die als

$$s^2 := \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (9.1)$$

bzw.

$$s := \sqrt{s^2} \quad (9.2)$$

definiert sind. Man beachte, dass diese Kennzahlen zur Beschreibung der Streuung einer Grundgesamtheit verwendet werden. Für das *Schätzen* der Streuung einer Grundgesamtheit, wenn nur eine Datenauswahl vorliegt, muss etwas anders vorgegangen werden (siehe Abschnitt 9.3).

Beispiel 9.4 (Fortsetzung von Beispiel 9.2) Die Populationsvarianz der Datenmenge $\{3, 2, 1, 5, 8, 7, 8, 8\}$ mit $\bar{x} = \frac{21}{4}$ ist $s^2 = \frac{119}{16} = 7.4375$. \square

Beispiel 9.5 (Fortsetzung von Beispiel 9.3) Das empirische Streuen der Gehälter in der Firma mit Durchschnittsgehalt € 5 000 ist, gemessen als Standardabweichung, $s = \sqrt{81\,000\,000} = 9\,000$. Damit kann man erkennen, dass es große Unterschiede bei den Gehältern dieser Firma gibt. \square

9.2 Mathematische Grundbegriffe

Zum Beginn dieser genaueren Behandlung der induktiven Statistik legen wir Teile der Terminologie fest.

Definition 9.1 (Konkrete Stichprobe)

Unter einer *konkreten Stichprobe vom Umfang n* versteht man n Zahlen x_1, \dots, x_n , die ein bestimmtes Merkmal von n zufällig ausgewählten Objekten einer Grundgesamtheit angeben.

Von diesen konkreten Zahlen wird die Brücke zu theoretischen Überlegungen dadurch geschlagen, dass man x_1, \dots, x_n als *Realisierungen* (Funktionswerte) von n ZV X_1, \dots, X_n auffasst, die alle unabhängig und identisch verteilt sind—wobei aber von vornherein nichts über diese Verteilung gesagt werden kann.

Definition 9.2 (Mathematische Stichprobe)

Wenn man die Zahlenwerte einer konkreten Stichprobe vom Umfang n als Realisierungen von n ZV X_1, \dots, X_n betrachtet, die unabhängig und identisch P_X -verteilt sind, dann nennt man die ZV X_1, \dots, X_n eine *nach der Verteilung P_X ausgewählte mathematische Stichprobe vom Umfang n* .

Beispiel 9.6 Das Zählen der Anzahl von Streichhölzern in $n = 45$ Zündholzschachteln liefert folgendes Ergebnis:

k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_k	51	52	49	51	52	51	53	52	48	52	50	53	49	50	51
k	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
x_k	53	51	52	50	51	53	53	55	50	49	53	50	51	51	52
k	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
x_k	48	53	50	49	51	54	49	52	51	49	53	51	50	51	53

Jede Zahl x_i kann als Realisierung einer ZV aufgefasst werden, die einer zufällig ausgewählten Zündholzschachtel die Anzahl der darin enthaltenen Zündhölzer zuweist.

Um einen Überblick über das Datenmaterial zu erhalten, werten wir es hinsichtlich der absoluten und relativen Häufigkeiten der Füllmengen aus und kommen so zum Ergebnis in untenstehender Tabelle:

	Füllmenge							
	48	49	50	51	52	53	54	55
Anzahl des Auftretens	2	6	7	12	7	9	1	1
relative Häufigkeit	0.04	0.13	0.16	0.27	0.16	0.2	0.02	0.02

Ein erster Ansatz zur Beschreibung des Zahlenmaterials in Beispiel 9.6 könnte etwa folgendermaßen lauten: Da die Menge der Streichhölzer in den Zündholzschachteln nicht konstant ist, schwankt die Menge vermutlich aufgrund kleiner Unregelmäßigkeiten in der Verarbeitung. Die Summe solcher Störungen kann als annähernd

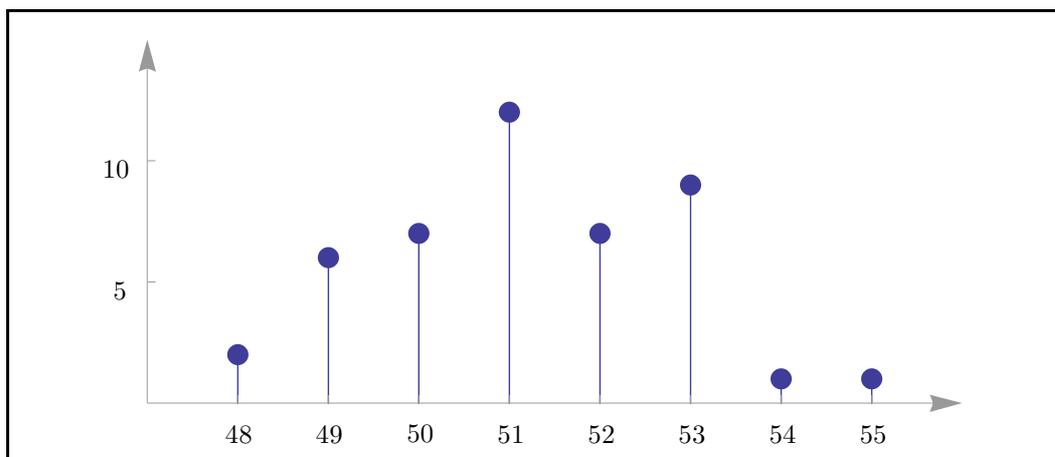


Abbildung 9.1: Häufigkeit von $k = 48, \dots, 55$ Streichhölzern in Zündholzschachteln.

normalverteilt betrachtet werden; wir fassen daher die x_i als (gerundete Versionen von) Realisierungen einer Normalverteilung mit unbekanntem Parametern μ und σ^2 auf. Die graphische Darstellung der Häufigkeit der Füllmengen zwischen 48 und 55 in Abbildung 9.1 unterstützt diese Vermutung.

Unabhängig davon, ob die Anzahl der Streichhölzer in den Schachteln nun normalverteilt ist oder nicht kann man sich fragen, wieviele Streichhölzer in einer Schachtel zu erwarten sind, und wie weit dieser Wert streut; mit anderen Worten: Wir können versuchen, Erwartungswert und Varianz der unbekanntem Verteilung zu bestimmen. Einige Ansätze dazu werden im nächsten Abschnitt beschrieben; zuvor führen wir noch die Begriffe *Statistik* und *Schätzer* ein. Die Bedeutung des Wortes Statistik ist etwas anders als im normalen Sprachgebrauch.

Definition 9.3 (Statistik)

Eine Statistik (im mathematischen Sinn) ist eine Abbildung

$$s : \mathbb{R}^n \rightarrow \mathbb{R}$$

$$(X_1, \dots, X_n) \mapsto s(X_1, \dots, X_n),$$

die einer Stichprobe vom Umfang n einen reellen Wert zuordnet.

Die obige Definition ist bewusst weit gefasst, um möglichst umfassend zu sein. Statistiken werden häufig verwendet, um das Datenmaterial in X_1, \dots, X_n ohne Informationsverlust (bezüglich spezieller Fragestellungen) zu komprimieren. Eine weitere Anwendung von Statistiken ist es, unbekannte Parameter von Verteilungen aufgrund einer Stichprobe zu berechnen. Da man dabei bei Vorliegen einer konkreten Stichprobe immer nur endlich viele zufällig ausgewählte Realisierungen der Verteilung vorliegen hat spricht man davon, die Parameter zu *schätzen*. So kann man, wie schon oben argumentiert, in Beispiel 9.6 die Anzahl der Streichhölzer in den Schachteln als normalverteilt mit unbekanntem Parametern μ und σ^2 ansehen; aufgrund einer Stichprobe können diese Parameter dann geschätzt werden.

Definition 9.4 (Schätzer)

Jede Statistik, die aufgrund einer nach der Verteilung P_X ausgewählten Stichprobe vom Umfang n einen unbekanntem Parameter θ von P_X berechnet, nennt man einen *Schätzer* für θ (meist geschrieben als $\hat{\theta}$).

9.3 Stichprobenmittel und Stichprobenvarianz

Der Übergang vom Datenmaterial in einer konkreten Stichprobe x_1, \dots, x_n zum theoretischen Modell mit der mathematischen Stichprobe X_1, \dots, X_n erlaubt es uns, theoretische Aussagen über das Datenmaterial und speziell über die Grundgesamtheit zu machen, aus der das Datenmaterial ausgewählt wurde. Die in diesem Abschnitt verwendeten Kennzahlen *Stichprobenmittel* und *Stichprobenvarianz* sind Beispiele von mathematischen Statistiken. Wenn man etwa Aussagen über Erwartungswert und Varianz der unbekanntem Verteilung der X_i machen will, so genügen dazu die folgenden zwei Statistiken.

Definition 9.5 (Stichprobenmittel)

Gegeben sei eine nach der Verteilung P_X ausgewählte mathematische Stichprobe X_1, \dots, X_n . Dann ist das *Stichprobenmittel* von X_1, \dots, X_n definiert als

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

Das Stichprobenmittel nimmt für Stichproben dieselbe Rolle ein, die der Erwartungswert für Verteilungen einnimmt. Dieser Zusammenhang kann noch präzisiert werden, wie der nächste Satz aussagt.

Satz 9.1 Für das Stichprobenmittel \bar{X} einer nach der Verteilung P_X ausgewählten Stichprobe um Umfang n gilt

$$E(\bar{X}) = E(X)$$

Beweis Mit Satz 7.5 und der Bedingung, dass die X_i P_X -verteilt sind (also $E(X_i) = E(X)$ ist), ergibt sich

$$E(\bar{X}) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = E(X). \quad \square$$

Dieser Satz besagt also, dass man den Erwartungswert der unbekanntem Verteilung P_X der ZV X_1, \dots, X_n durch das Stichprobenmittel gut schätzen kann: Es ist nämlich zu erwarten, dass das Stichprobenmittel gleich dem unbekanntem Erwartungswert ist. Diese Tatsache wird überall dort verwendet, wo das systematische

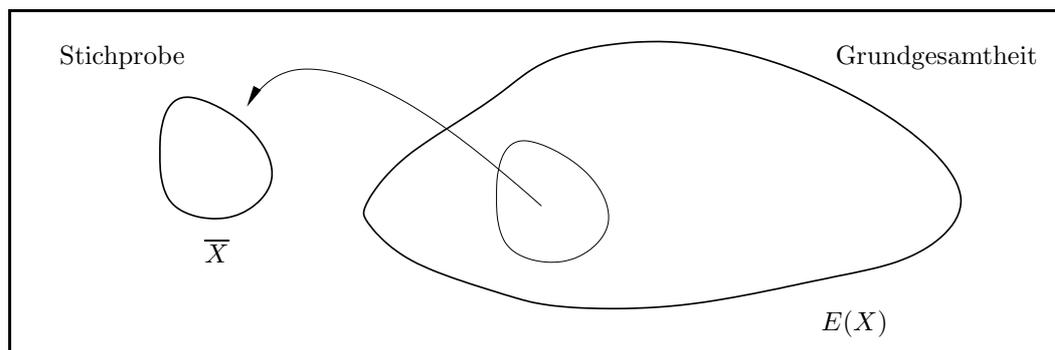


Abbildung 9.2: Das Stichprobenmittel ist ein erwartungstreuer Schätzer für den unbekanntem Erwartungswert in der Grundgesamtheit.

Erheben von Daten in einer Grundgesamtheit zu aufwändig wäre, und man sich mit dem Ziehen einer Stichprobe begnügen muss. Mit Satz 9.1 weiß man, dass der unbekanntem durchschnittliche Wert eines Merkmals der Grundgesamtheit (z.B. Körpergröße) dem durchschnittlichen Wert des Stichprobenmittels entspricht. Der aus *einer* konkreten Stichprobe gewonnene Wert ist somit ein Schätzwert, und zumindest nicht systematisch falsch ist. Diese Situation ist schematisch auch in Abbildung 9.2 zu sehen.

Die nächste Definition liefert eine kurze Übersicht über Anforderungen an die Qualität von Schätzern.

Definition 9.6 (Anforderungen an Schätzer)

Gegeben sei ein Schätzer $\hat{\theta}$ eines Parameters θ . $\hat{\theta}$ heißt *erwartungstreu*, wenn

$$E(\hat{\theta}) = \theta$$

ist. Eine Folge $\hat{\theta}_1, \hat{\theta}_2, \dots$ von Schätzern heißt *konsistent*, wenn

$$\lim_{n \rightarrow \infty} \hat{\theta}_n = \theta$$

gilt¹. $\hat{\theta}_1$ heißt *wirksamer* als $\hat{\theta}_2$, wenn

$$\text{Var}(\hat{\theta}_1) \leq \text{Var}(\hat{\theta}_2)$$

gilt.

Somit sind diejenigen Schätzer erwartungstreu, die im Durchschnitt den unbekanntem Parameter ergeben. Erwartungstreue ist die wichtigste Forderung an Schätzer unbekannter Parameter von Verteilungen, da damit ein systematischer Fehler vermieden wird. Zusätzlich ist man bestrebt, die Streuung eines Schätzers gering zu halten, da man aus *einer* Stichprobe ja nur *einen* Schätzwert erhält, der bei großer Streuung weit vom echten Wert abweichen kann. Aus dieser Forderung ergibt sich

¹eigentlich etwas komplizierter definiert, aber hier reicht diese vereinfachte Form.

der Begriff der Wirksamkeit eines Schätzers. Wenn kleine Varianz wichtiger ist als Erwartungstreue, werden in der Praxis auch nicht erwartungstreue Schätzer verwendet. Diese Schätzer sind aber dann meist konsistent.

Die Varianz des Stichprobenmittels lässt sich folgendermaßen berechnen.

Satz 9.2 Für die Varianz des Stichprobenmittels \bar{X} gilt

$$\text{Var}(\bar{X}) = \frac{\text{Var}(X)}{n},$$

das Mitteln mehrerer Werte liefert also eine Reduktion der Varianz.

Beweis Einfaches Nachrechnen liefert mit Satz 7.7 und Satz 7.16

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{\text{Var}(X)}{n}. \quad \square$$

Eine weitere Maßzahl einer Verteilung ist die Varianz, die aus einer Stichprobe wie folgt geschätzt werden kann.

Definition 9.7 (Stichprobenvarianz)

Gegeben sei eine nach der Verteilung P_X ausgewählte mathematische Stichprobe X_1, \dots, X_n . Die *Stichprobenvarianz* von X_1, \dots, X_n ist definiert als

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

Auf den ersten Blick mag der Faktor $\frac{1}{n-1}$ in dieser Definition verwundern (speziell verglichen mit der Definition der Populationsvarianz in Abschnitt 9.1). Man würde hier vielleicht eher $\frac{1}{n}$ erwarten. Der nächste Satz zeigt, dass sich dieser Faktor aus der Bedingung der Erwartungstreue ergibt.

Satz 9.3 Die Stichprobenvarianz S_X^2 einer nach der Verteilung P_X ausgewählten Stichprobe vom Umfang n ist ein erwartungstreuer Schätzer der Varianz von X , es gilt also

$$E(S_X^2) = \text{Var}(X)$$

Beweis Etwas umständlicher als der letzte Beweis; wir verwenden dieselben Zusammenhänge und berechnen zuerst

$$\begin{aligned} (n-1)S_X^2 &= \sum_{i=1}^n (X_i - \bar{X})^2 \\ &= \sum_{i=1}^n (X_i^2 - 2X_i\bar{X} + \bar{X}^2) \end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^n X_i^2 - 2\bar{X} \sum_{i=1}^n X_i + \sum_{i=1}^n \bar{X}^2 \\
&= \sum_{i=1}^n X_i^2 - 2\bar{X}(n\bar{X}) + n\bar{X}^2 \\
&= \sum_{i=1}^n X_i^2 - n\bar{X}^2,
\end{aligned}$$

dann ergibt sich aus der Linearität des Erwartungswerts und der Tatsache, dass $E(X_i^2) = E(X^2)$ ist

$$\begin{aligned}
E((n-1)S_X^2) &= E\left(\sum_{i=1}^n X_i^2\right) - nE(\bar{X}^2) \\
&= nE(X^2) - nE(\bar{X}^2).
\end{aligned}$$

Die beiden hier auftretenden Erwartungswerte kann man mit Satz 7.6 berechnen: so gilt

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

und

$$\text{Var}(\bar{X}) = E(\bar{X}^2) - (E(\bar{X}))^2.$$

wir wissen bereits, dass $\text{Var}(\bar{X}) = \frac{\text{Var}(X)}{n}$ und $E(\bar{X}) = E(X)$ ist (Satz 9.2 bzw. Satz 9.1), somit ist

$$E(X^2) = \text{Var}(X) + (E(X))^2$$

und

$$E(\bar{X}^2) = \frac{\text{Var}(X)}{n} + (E(X))^2;$$

einsetzen in die ursprüngliche Gleichung ergibt sich für den gesuchten Erwartungswert

$$\begin{aligned}
E((n-1)S_X^2) &= n(\text{Var}(X) + (E(X))^2) - n\left(\frac{\text{Var}(X)}{n} + (E(X))^2\right) \\
&= (n-1)\text{Var}(X).
\end{aligned}$$

Nach Division durch $n-1$ ergibt sich wie gefordert $E(S_X^2) = \text{Var}(X)$. □

Beispiel 9.7 (Fortsetzung von Beispiel 9.6) Unter der Annahme, dass die Anzahl der Streichhölzer in einer Schachtel normalverteilt mit unbekanntem Erwartungswert μ und unbekannter Varianz σ^2 ist, sind \bar{X} und S_X^2 erwartungstreue Schätzer für μ und σ^2 . Aus dem Datenmaterial ergeben sich die Werte $\bar{x} = 51.1556$ und $s_x^2 = 2.3655$. □

Die im obigen Beispiel verwendeten Symbole \bar{x} und s_x^2 sind dabei das *arithmetische Mittel* \bar{x} und die *empirische Varianz* s_x^2 , die als

$$s_x^2 := \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

definiert ist. Man beachte, dass s_x^2 sich um den Faktor $\frac{n-1}{n}$ von der Populationsvarianz s^2 unterscheidet, die in Abschnitt 9.1 behandelt wurde. Wie wir gesehen haben liegt der Unterschied darin, dass s_x^2 zum Schätzen der Varianz einer Grundgesamtheit herangezogen wird, der Berechnung von s^2 aber bereits eine Grundgesamtheit zugrundeliegt.

Beim Umgang mit konkreten Stichproben (also echtem Datenmaterial) rechnet man mit den Maßzahlen, die aus diesen Stichproben gewonnen wurden; für theoretische Aussagen wie Satz 9.1 und Satz 9.3 abstrahiert man jedoch von der konkreten zur mathematischen Stichprobe und kann dann allgemeinen Aussagen über Stichprobenmittel und Stichprobenvarianz machen.

In Analogie zu Abschnitt 7.3 kann man auch in der induktiven Statistik den Zusammenhang zwischen zwei ZV untersuchen.

9.4 Kovarianz und Korrelationskoeffizient einer Stichprobe

Im letzten Abschnitt haben wir nur eindimensionale Stichproben behandelt. Wenn zwei oder mehrere Merkmale von Ereignissen beobachtet werden, spricht man von mehrdimensionalen Stichproben.

Definition 9.8 (Zweidimensionale Stichproben)

Eine Folge von zweidimensionalen ZV $(X_1, Y_1), \dots, (X_n, Y_n)$, die alle unabhängig und identisch (nämlich $P_{X,Y}$ -) verteilt sind, nennt man eine *nach der Verteilung $P_{X,Y}$ ausgewählte mathematische Stichprobe vom Umfang n* .

Mit Kovarianz und Korrelationskoeffizienten können wir den linearen Zusammenhang zwischen den ZV X und Y bestimmen.

Definition 9.9 (Stichprobenkovarianz und -korr.koeffizient)

Für eine zweidimensionale Stichprobe $(X_1, Y_1), \dots, (X_n, Y_n)$ vom Umfang n nennt man

$$S_{X,Y} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

die *Kovarianz* von X und Y und

$$R_{X,Y} = \frac{S_{X,Y}}{\sqrt{S_X^2} \sqrt{S_Y^2}}$$

den *Korrelationskoeffizienten* von X und Y .

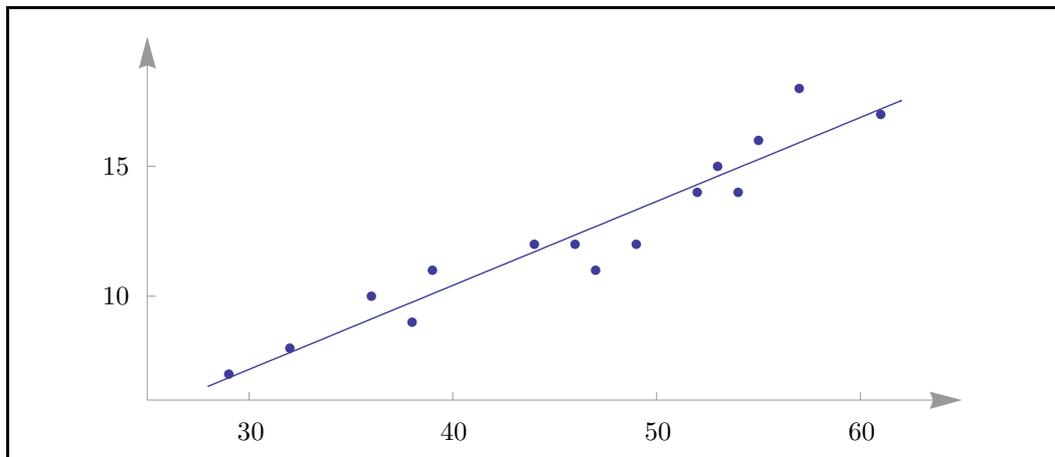


Abbildung 9.3: Datenpunkte und lineare Approximation der Abhängigkeit von X und Y aus Beispiel 9.8.

Diese beiden Maßzahlen werden für theoretische Überlegungen verwendet, wie etwa für den folgenden Satz 9.4; bei Vorliegen einer konkreten zweidimensionalen Stichprobe $(x_1, y_1), \dots, (x_n, y_n)$ rechnet man mit der *empirischen Kovarianz* $s_{x,y}$ und dem *empirischen Korrelationskoeffizienten* $r_{x,y}$:

$$s_{x,y} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}), \quad r_{x,y} = \frac{s_{x,y}}{\sqrt{s_x^2} \sqrt{s_y^2}}.$$

Wie schon in Abschnitt 7.2 erwähnt geben diese Werte Maße dafür an, in welchem Maß große Werte von X mit großen Werten von Y einhergehen. Wenn ein (annähernd) linearer Zusammenhang besteht, ist $R_{X,Y}$ nahe bei -1 oder 1 , abhängig davon, ob die Steigung der Gerade, die den Zusammenhang beschreibt, positiv oder negativ ist. Wenn sich der Zusammenhang nicht gut durch eine Gerade mit Steigung $k \neq 0$ beschreiben lässt, ist $R_{X,Y}$ etwa null. Wir werden im nächsten Abschnitt noch genauer auf den möglichen linearen Zusammenhang zwischen zwei ZV (und die Bedeutung des Korrelationskoeffizienten dabei) eingehen.

Satz 9.4 Für unabhängige ZV X und Y gilt

$$E(R_{X,Y}) = 0.$$

Zu beachten ist hierbei, dass man bei Vorliegen eines empirischen Korrelationskoeffizienten $r_{x,y}$ von null *nicht* auf die Unabhängigkeit von X und Y schließen kann, da die Aussage nur in die umgekehrte Richtung gilt. Wir werden obigen Satz allerdings dazu verwenden, um bei einem Ergebnis von $r_{x,y} \neq 0$ darauf schließen zu können, dass X und Y *nicht* unabhängig sind.

Beispiel 9.8 Ein Werkstoff wird vor der Verarbeitung einige Stunden in einer nicht klimatisierten Lagerhalle zwischengelagert. An 15 Tagen wurde die relative Feuchte

in der Lagerhalle (x_i) und der Feuchtigkeitsgehalt im Endprodukt (y_i) gemessen, um Aussagen über einen Zusammenhang dieser beiden Werte treffen zu können. Dabei wurden folgende Zahlen erhoben:

x_i	46	53	29	61	36	39	47	49	52	38	55	32	57	54	44
y_i	12	15	7	17	10	11	11	12	14	9	16	8	18	14	12

Wie schon zuvor fassen wir die konkreten Werte x_i und y_i als Realisierungen von ZV auf; in diesem Fall als Realisierungen einer zweidimensionalen ZV (X, Y) . Der Zusammenhang von X und Y kann durch Berechnung des Korrelationskoeffizienten bestimmt werden. Für obiges Datenmaterial erhält man $\bar{x} = 46.1333$, $\bar{y} = 12.4$, $s_X^2 = 91.981$, $s_Y^2 = 10.5429$ und damit ist $r_{X,Y} = 0.9547$. Mit Satz 9.4 ist damit die Abhängigkeit von X und Y zu erwarten. In Abbildung 9.3 ist zu sehen, dass der Zusammenhang zwischen X und Y annähernd linear ist. Wir werden in Abschnitt 10.1 sehen, wie der lineare Zusammenhang zwischen den x - und y -Werten zu berechnen ist. \square

Kapitel 10

Maximum Likelihood Schätzungen

Wie wir bereits in Kapitel 9 gesehen haben ist es das Ziel von Parameterschätzmethoden, auf Basis einer konkreten Stichprobe den (bestmöglichen) Wert eines oder mehrerer Parameter derjenigen Verteilung anzugeben, die die konkrete Stichprobe erzeugt hat. Um solche Berechnungen durchführen zu können benötigen wir noch etwas Terminologie.

Definition 10.1 (Likelihood-Funktion)

Seien X eine ZV mit W.Funktion bzw. Dichte $P(x; \theta)$, die durch θ parametrisiert ist, und x_1, \dots, x_n eine konkrete Stichprobe von n unabhängigen, wie X verteilten ZV. Dann nennt man

$$L(\theta) := P(\{x_1, \dots, x_n\}; \theta) = \prod_{k=1}^n P(x_k; \theta)$$

die *Likelihood-Funktion* von θ .

Es erscheint nun sinnvoll, als bestmöglichen Wert des Parameters θ denjenigen zu wählen, der die Daten maximal wahrscheinlich macht (bzw. im Fall stetiger Verteilungen die Dichten maximiert).

Definition 10.2 (Maximum-likelihood-Schätzer)

Seien $P(x; \theta)$ und x_1, \dots, x_n wie in Definition 10.1. Dann nennt man

$$\theta_{\text{ML}} := \arg \max_{\theta} L(\theta)$$

den *Maximum-Likelihood-Schätzer* von θ .

Je nach Verteilung kann die Bestimmung von θ_{ML} mathematisch kompliziert sein. Ein einfaches Beispiel soll helfen, diese Begriffe zu erläutern. Wir werden schnell sehen, dass selbst einfache Aufgabenstellungen einiges an technischem Aufwand benötigen.

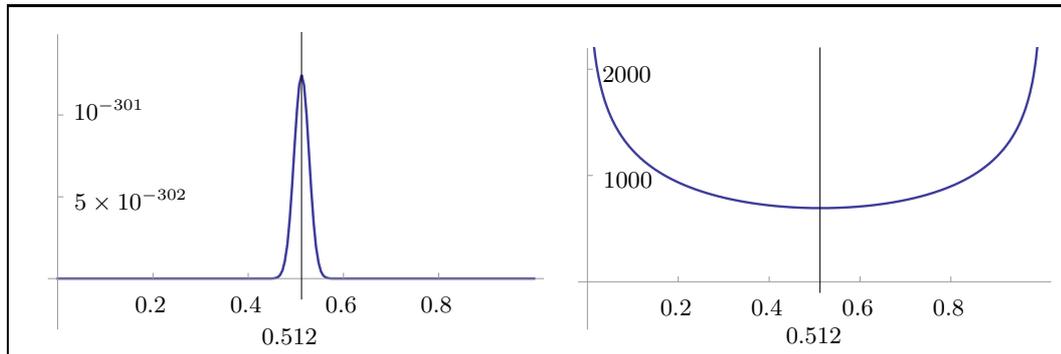


Abbildung 10.1: Links die Likelihood-Funktion $L(\theta)$ aus Beispiel 10.1, rechts $-\log L(\theta)$. Man kann erkennen, dass sich die Position des Optimums durch Logarithmieren nicht ändert.

Beispiel 10.1 Eine Münze werde 1000-mal geworfen. Bei jedem einzelnen Wurf sei θ die Wahrscheinlichkeit des Ereignisses KOPF. Auf Basis der Versuchsreihe x_1, \dots, x_{1000} soll der beste Wert von θ bestimmt werden.

Lösung: Wir benötigen die Verteilung von X_i , die beim i -ten Wurf das Auftreten von KOPF oder ZAHL angibt; es muss $P(X_i = \text{KOPF}) = \theta$ und $P(X_i = \text{ZAHL}) = 1 - \theta$ gelten. Wenn wir KOPF durch 1 und ZAHL durch 0 repräsentieren, kann man die Verteilung von X_i schreiben als

$$P(X_i = x; \theta) = \theta^x (1 - \theta)^{(1-x)}.$$

Mit dieser Definition erhält man wie gewünscht

$$P(\text{KOPF}) = P(X_i = 1; \theta) = \theta^1 (1 - \theta)^{(1-1)} = \theta$$

und

$$P(\text{ZAHL}) = P(X_i = 0; \theta) = \theta^0 (1 - \theta)^{(1-0)} = 1 - \theta.$$

Die Likelihood-Funktion von θ ist für observierte Daten x_1, \dots, x_n (die alle entweder 0 oder 1 sind) somit

$$\begin{aligned} L(\theta) &= P(\{x_1, \dots, x_n\}; \theta) \\ &= \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{(1-x_i)} = \theta^{x_1 + \dots + x_n} (1 - \theta)^{n - x_1 - \dots - x_n}. \end{aligned}$$

Dieser Ausdruck kann noch vereinfacht werden: So gibt $x_1 + \dots + x_n$ an, wie oft in der Versuchsreihe KOPF aufgetreten ist. Mit $K = x_1 + \dots + x_n$ ergibt sich für die Likelihood-Funktion von θ

$$L(\theta) = \theta^K (1 - \theta)^{n-K}.$$

Diese Funktion ist für eine konkrete Versuchsreihe, in der 512-mal KOPF und 488-mal ZAHL aufgetreten ist, in Abbildung 10.1 links zu sehen. Der senkrechte Strich bei 0.512 zeigt die Position des Maximums dieser Funktion.

Uns bleibt somit nur mehr, diesen plausiblen Wert auch numerisch zu bestimmen. Wir suchen den Wert θ , der den Ausdruck $\theta^K(1-\theta)^{n-K}$ maximiert. Rein rechnerisch ist es bei dieser Aufgabenstellung (wie auch bei vielen anderen Optimierungsproblemen der Statistik) leichter, das Minimum von $-\log(\theta^K(1-\theta)^{n-K})$ zu bestimmen: Durch das Logarithmieren wird nämlich das Produkt aus Definition 10.1 in eine Summe umgewandelt, und damit leichter zu behandeln. Die wichtige Erkenntnis dabei ist, dass sich die Position des Optimums auf der x -Achse durch das Logarithmieren nicht ändert. Dies ist graphisch in Abbildung 10.1 rechts zu sehen: Das Minimum von $-\log f(x)$ ist immer an der selben Stelle wie das Maximum von $f(x)$.

Für $L(\theta) = \theta^K(1-\theta)^{n-K}$ erhalten wir mit der Rechenregel $\log(ab) = \log(a) + \log(b)$ die Vereinfachung

$$-\log L(\theta) = -K \log \theta - (n - K) \log(1 - \theta).$$

Das Minimum dieses Ausdrucks befindet sich an der Stelle, an der die Ableitung null wird. Es gilt wegen $\log'(x) = 1/x$

$$-\frac{d \log L(\theta)}{d\theta} = -\frac{K}{\theta} + \frac{n - K}{1 - \theta}.$$

Nullsetzen dieses Ausdrucks liefert

$$-K(1 - \theta) + \theta(n - K) = 0$$

und mit kurzem Umformen schließlich das erwartete

$$\theta = \frac{K}{n}.$$

Wir erhalten somit für unser Zahlenmaterial den Maximum-Likelihood-Schätzwert $\theta_{\text{ML}} = 0.512$. \square

Das letzte Beispiel zeigt gleichzeitig die Mächtigkeit und den gravierenden Nachteil des Maximum-Likelihood-Ansatzes auf: Es ist rein aus den Daten möglich, einen optimalen (und noch dazu plausiblen) Wert des unbekanntes Parameters einer Verteilung zu bestimmen. Da dieser Wert aber *nur* von den Daten abhängt, unterliegt er auch der Schwankungsbreite der Daten: So kann bei der nächsten Versuchsreihe etwa $\theta_{\text{ML}} = 0.496$ und bei der nächsten $\theta_{\text{ML}} = 0.523$ bestimmt werden.

In den nächsten Beispielen werden wir uns mit Maximum-Likelihood-Schätzern für die Parameter der Normalverteilung beschäftigen.

Beispiel 10.2 Von einer konkreten Stichprobe x_1, \dots, x_n sei bekannt, dass sie von einer Normalverteilung generiert wurde. Von dieser Normalverteilung sei weiters σ^2 bekannt, nicht aber μ (es ist in vielen technischen Bereichen tatsächlich möglich, nur einen von mehreren Parametern einer Verteilung zu kennen). Man bestimme aus x_1, \dots, x_n den Maximum-Likelihood-Schätzer μ_{ML} für μ .

In diesem Beispiel ist die Likelihood-Funktion

$$L(\mu) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}},$$

und für $-\log L(\mu)$ gilt

$$\begin{aligned}
 -\log L(\mu) &= -\log\left(\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right) \\
 &= -\sum_{i=1}^n \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right) \\
 &= -\sum_{i=1}^n \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \sum_{i=1}^n \log\left(e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right) \\
 &= n \log(\sqrt{2\pi\sigma^2}) + \sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2} \\
 &= n \log(\sqrt{2\pi\sigma^2}) + \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2.
 \end{aligned}$$

Wir sind am Minimum dieses Ausdrucks interessiert; da der erste Teil nicht von den Daten x_1, \dots, x_n abhängt kann er weggelassen werden. Nullsetzen der Ableitung des verbleibenden Terms liefert wegen

$$-\frac{d \log L(\mu)}{d \mu} = -\frac{1}{2\sigma^2} \sum_{i=1}^n (-2(x_i - \mu))$$

die Gleichung

$$\frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0,$$

die umgeformt

$$\sum_{i=1}^n x_i = n\mu$$

und damit

$$\mu_{\text{ML}} = \frac{1}{n} \sum_{i=1}^n x_i$$

ergibt. Der im Maximum-Likelihood-Sinn beste Schätzwert des Parameters μ einer Normalverteilung ist damit der Mittelwert der Stichprobe x_1, \dots, x_n . Mit diesem Schätzwert kann durch eine ähnliche Rechnung der Maximum-Likelihood-Schätzwert

$$\sigma_{\text{ML}}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_{\text{ML}})^2$$

hergeleitet werden. □

10.1 Anwendung: Lineare Regression

Als Anwendung der Maximum-Likelihood-Methode, die wir in diesem Kapitel kennengelernt haben, betrachten wir nun zwei *Regressionsprobleme*. Von Regressionsproblemen spricht man normalerweise, wenn man auf Basis von Daten ein Modell

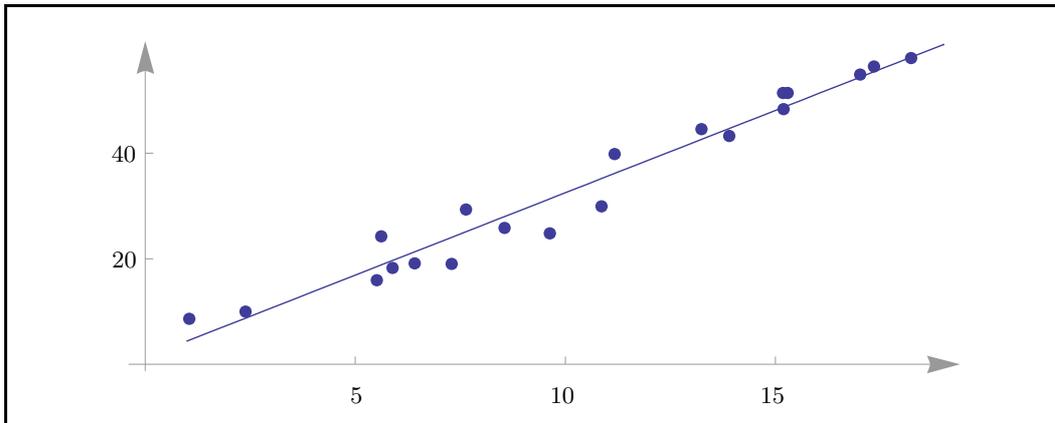


Abbildung 10.2: Illustration zur linearen Regression: Da die Werte y_i fehlerbehaftet sind, liegen die Datenpunkte (x_i, y_i) nicht auf der bestapproximierenden Gerade.

zur Vorhersage eines reellzahligen Wertes konstruiert (im Gegensatz zu *Klassifikationsproblemen*, bei denen das Modell die Klassenzugehörigkeit eines Datenpunktes entscheidet). Wir werden auch erkennen, dass die logistische Regression, die in Abschnitt 10.2 behandelt wird, trotz ihres Namens eigentlich ein Klassifikationsproblem löst.

Diese Unterscheidung wird klarer, wenn wir die Aufgabenstellungen genauer formalisieren. In beiden Fällen sind Datenpunkte x_1, \dots, x_n gegeben. Bei der *linearen* Regression sind zusätzlich noch reelle Werte y_1, \dots, y_n gegeben, die eine Funktion auf den Werten x_1, \dots, x_n annimmt; bei der *logistischen* Regression sind dies binäre Werte t_1, \dots, t_n , die die Zugehörigkeit der Datenpunkte zu einer von zwei Klassen anzeigen.

Wir werden zuerst den Fall der linearen Regression genauer betrachten. Ein Beispiel eines linearen Regressionsproblems (und seiner Lösung) ist in Abbildung 10.2 zu sehen. Es soll auf Basis der Daten x_1, \dots, x_n und y_1, \dots, y_n ein Modell der Abhängigkeit der y -Werte von den x -Werten gefunden werden, um für einen neuen Wert x_k einen “möglichst guten” Wert y_k vorhersagen zu können. Dabei stehen die Werte y_i mit den Werten x_i durch

$$y_i = f(x_i) + E_i$$

in Zusammenhang. Dabei ist f das Modell, das den Zusammenhang von x_i und y_i ausdrückt, und die E_i die ZV, die erklären sollen, warum kein direkter (deterministischer) Zusammenhang zwischen x_i und y_i besteht. In Abbildung 10.2 ist f eine lineare Funktion $y = kx + d$; aus den konkreten Daten x_1, \dots, x_n und y_1, \dots, y_n sollen die bestmöglichen Werte dieser beiden Parameter bestimmt werden. Ohne Fehler (wenn der Zusammenhang also nur $y_i = f(x_i)$ wäre) würden alle Datenpunkte in Abbildung 10.2 auf der Gerade liegen. Wir nehmen nun an, dass die E_i normalverteilt sind mit konstanten Parametern $\mu = 0$ und σ^2 . Das bedeutet, dass kein systematischer Fehler gemacht wird, und dass die Streubreite des Fehlers bei allen Datenpunkten gleich groß ist. Bei anderen Verteilungsannahmen ist die folgende Herleitung entsprechend zu modifizieren.

Wir betrachten die konkrete Stichprobe y_1, \dots, y_n als Realisierungen von ZV

Y_1, \dots, Y_n . Da die Fehler E_i $N(0, \sigma^2)$ -verteilt sind, folgt aus $y_i = f(x_i) + E_i$, dass die Y_i ebenfalls normalverteilt sind, und zwar mit Parametern $\mu_i = f(x_i)$ und konstantem σ^2 .

Wir drücken im Folgenden die Abhängigkeit des Modells $f(x)$ von den Parametern explizit aus, um den Maximum-Likelihood-Ansatz anwenden zu können. Für den einfachsten Fall wählen wir, wie schon in Abbildung 10.2 das Modell

$$f(x; \theta) = \theta_1 x + \theta_2.$$

Andere (komplizierte) Modell sind mit diesem Ansatz ebenso leicht behandelbar, solange sie *linear in den Parametern* sind. Das bedeutet, dass f (als Funktion von θ gesehen) linear ist.

Unter Verwendung des Modells $f(x; \theta)$ und unter der Annahme, dass die gemessenen y -Werte nur durch $N(0, \sigma^2)$ -verteilte Fehler vom vorhergesagten Wert $f(x; \theta)$ abweichen, ergibt sich als Dichtefunktion für die Verteilung der ZV Y_i in Abhängigkeit von der ZV X_i der Ausdruck

$$P(y | x; \theta, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-f(x;\theta))^2}{2\sigma^2}},$$

woraus man die Likelihood-Funktion

$$\begin{aligned} L(\theta) &= P(\{y_1, \dots, y_n\} | \{x_1, \dots, x_n\}; \theta, \sigma^2) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-f(x_i;\theta))^2}{2\sigma^2}} \end{aligned}$$

bestimmen kann. Minimieren von $-\log L(\theta)$ verläuft analog zur Herleitung des Resultats in Beispiel 10.2 und liefert

$$\begin{aligned} -\log L(\theta) &= -\log \left(\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-f(x_i;\theta))^2}{2\sigma^2}} \right) \\ &= [\dots] \\ &= n \log(\sqrt{2\pi\sigma^2}) + \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - f(x_i; \theta))^2. \end{aligned}$$

Zur Bestimmung des Minimums muss dieser Ausdruck nun nach θ abgeleitet und nullgesetzt werden. Obige Notation eignet sich nicht gut für weitere Berechnungen; wir wechseln nun zu Matrixnotation. Mit $\theta = (\theta_1, \theta_2)$ kann man das angenommene lineare Modell schreiben als

$$f(x; \theta) = \theta_1 x + \theta_2 = (\theta_1, \theta_2) \cdot \begin{pmatrix} x \\ 1 \end{pmatrix}.$$

Die gesamte Summe $\sum_{i=1}^n (y_i - f(x_i; \theta))^2$ kann man mit den Notationen

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \text{und} \quad X = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}$$

in Matrixschreibweise angeben als

$$\sum_{i=1}^n (y_i - f(x_i; \theta))^2 = (Y - X \cdot \theta)^T \cdot (Y - X \cdot \theta).$$

Der Vorteil der Matrixschreibweise liegt darin, dass man nach Vektoren wie nach "normalen" Variablen ableiten kann (auch wenn mathematisch eigentlich etwas mehr dahintersteckt). Man erhält

$$\begin{aligned} -\frac{d \log L(\theta)}{d\theta} &= \frac{d}{d\theta} \left(n \log(\sqrt{2\pi\sigma^2}) + \frac{1}{2\sigma^2} (Y - X \cdot \theta)^T \cdot (Y - X \cdot \theta) \right) \\ &= 0 + \frac{1}{2\sigma^2} (-2X^T \cdot (Y - X \cdot \theta)). \end{aligned}$$

Nullsetzen dieses Ausdrucks liefert

$$X^T \cdot (Y - X \cdot \theta) = 0,$$

welcher durch Ausmultiplizieren auf die Form

$$X^T \cdot Y = X^T \cdot X \cdot \theta$$

gebracht werden kann, woraus schließlich

$$\theta_{\text{ML}} = (X^T \cdot X)^{-1} \cdot X^T \cdot Y$$

folgt.

Beispiel 10.3 Durch die Datenpunkte

x_i	5	6	7	10	12	15	18	20
y_i	7.4	9.3	10.6	15.4	18.1	22.2	24.1	22.8

soll ein Modell der Form $f(x; \theta) = \theta_1 x + \theta_2$ gelegt werden. Die Maximum-Likelihood-Schätzwerte für den Parametervektor θ erhält man wie in obiger Herleitung. Für das Zahlenmaterial dieses Beispiels ist

$$X = \begin{pmatrix} 5 & 6 & 7 & 10 & 12 & 15 & 18 & 20 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}^T$$

und

$$Y = (7.4, 9.3, 10.6, 15.4, 18.1, 22.2, 24.1, 22.8)^T.$$

Damit ist

$$(X^T \cdot X)^{-1} = \begin{pmatrix} 0.0045 & -0.0524 \\ -0.0524 & 0.734 \end{pmatrix},$$

und insgesamt

$$\theta_{\text{ML}} = (X^T \cdot X)^{-1} \cdot X^T \cdot Y = \begin{pmatrix} 1.13 \\ 3.09 \end{pmatrix}.$$

Die Datenpunkte und die Gerade $y = 1.13x + 3.09$ sind graphisch in Abbildung 10.3 zu sehen; der empirische Korrelationskoeffizient $r_{x,y} = 0.968$ weist auch darauf hin, dass die Datenpunkte durch die lineare Funktion gut approximiert werden. \square

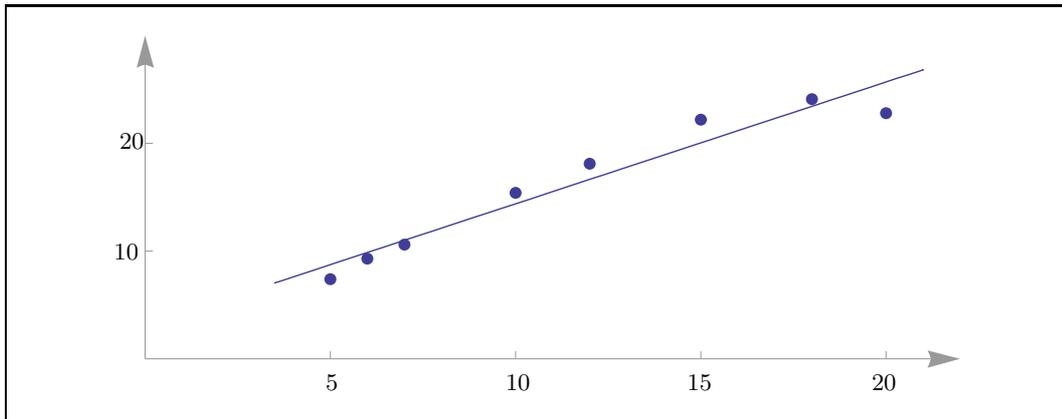


Abbildung 10.3: Die acht Datenpunkte aus Beispiel 10.3, zusammen mit dem durch den Maximum-Likelihood-Schätzwert θ_{ML} definierten linearen Modell.

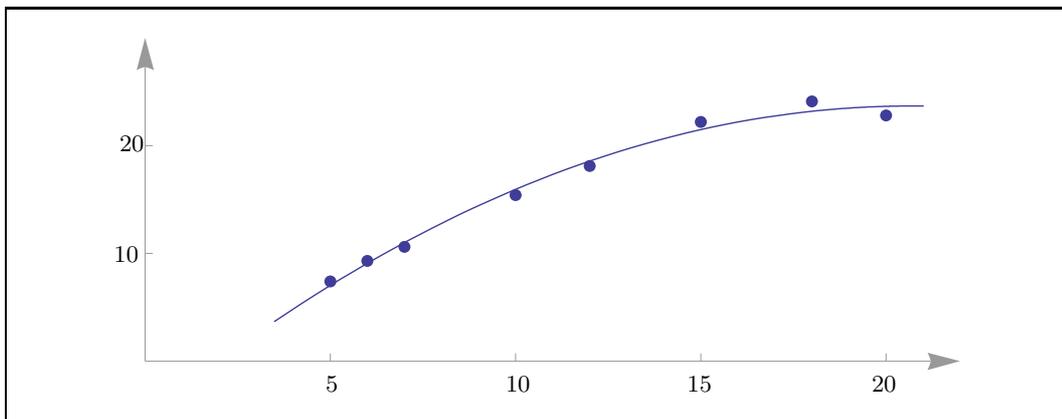


Abbildung 10.4: Quadratische Approximation der Datenpunkte aus Beispiel 10.3, ebenfalls durch den Maximum-Likelihood Ansatz bestimmt.

An dieser Stelle soll noch Folgendes vermerkt sein: Ein hoher Wert von $r_{x,y}$ bedeutet nicht notwendigerweise, dass sich der Zusammenhang zwischen X und Y optimal durch eine Gerade $Y = aX + b$ beschreiben lässt. Das Quadrat $r_{x,y}^2$ des Korrelationskoeffizienten gibt vielmehr (in einem sehr speziellen Sinn) an, um wieviel besser sich der Zusammenhang zwischen X und Y durch $Y = aX + b$ als nur durch eine Konstante $Y = b'$ ausdrücken lässt. Dadurch ist allerdings noch nicht ausgesagt, dass nicht etwa eine höhergradige Abhängigkeit besteht. Dies wird im nächsten Beispiel illustriert.

Beispiel 10.4 (Fortsetzung von Beispiel 10.3) Wir untersuchen nochmals die selben Daten wie im letzten Beispiel. Durch graphische Betrachtung der Datenpunkte kann man sich überzeugen, dass sich der Zusammenhang zwischen X und Y besser durch eine quadratische Funktion wiedergeben lässt. Dies entspricht dem quadratischen Modell

$$f(x; \theta) = \theta_2 x^2 + \theta_1 x + \theta_0 = (\theta_2, \theta_1, \theta_0) \cdot \begin{pmatrix} x^2 \\ x \\ 1 \end{pmatrix}$$

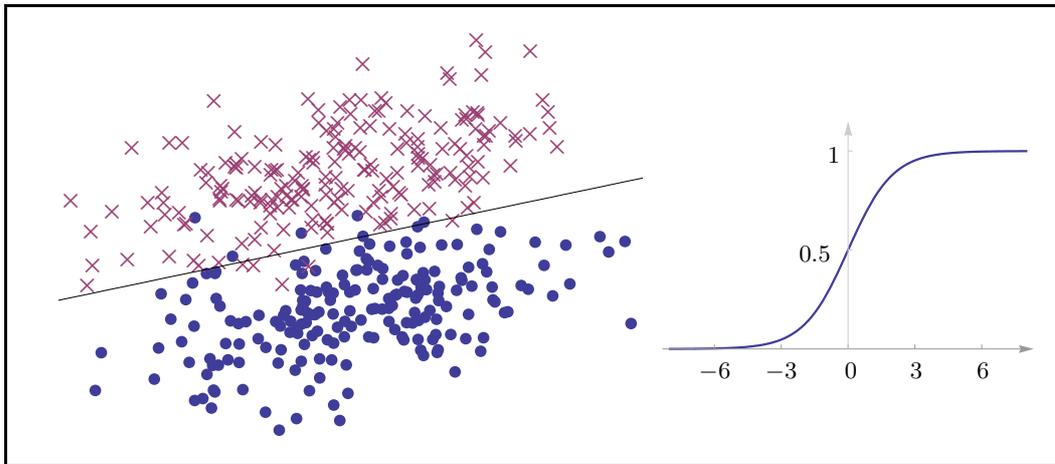


Abbildung 10.5: Illustration zur logistischen Regression: Links eine Menge von Datenpunkten, die zwei Klassen angehören, und die mittels logistischer Regression bestimmte optimale Trennlinie. Rechts ist der Graph der logistischen Funktion $f(x) = 1/(1 + e^{-x})$ zu sehen.

bei der Herleitung der linearen Regression; es muss nur die Matrix X entsprechend abgeändert werden. Die quadratische Approximation $y = -0.07x^2 + 2.80x - 5.27$ liefert eine bessere Annäherung an die Datenpunkte, wie graphisch in Abbildung 10.4 zu sehen ist. Dies lässt sich auch dadurch nachrechnen, dass man sowohl für das lineare als auch für das quadratische Modell die Summe der Fehlerquadrate berechnet (die ja durch die Modelle minimiert werden). Dies ergibt beim linearen Modell einen Wert von 18.86, beim quadratischen Modell aber nur 2.92. \square

10.2 Anwendung: Logistische Regression

Bei der *logistischen Regression* sind die vorherzusagenden Werte y_i nicht mehr reelle Zahlen, sondern binäre Klassenzugehörigkeitsindikatoren. Wir wollen nun auf Basis von x_1, \dots, x_n und y_1, \dots, y_n ein Modell finden, das für einen neuen Wert x_k die Klassenzugehörigkeit y_k bestimmt; dabei seien die beiden Klassen als 0 bzw. 1 kodiert. Mit dem hier präsentierten Ansatz ist es sogar möglich, eine *Wahrscheinlichkeit* der Klassenzugehörigkeit anzugeben; dies ist in den meisten Fällen einer rein dichotomen Entscheidung vorzuziehen. Das gesuchte Modell soll für gegebenen x -Wert

$$P(y = 1 | x)$$

repräsentieren; wegen $P(y = 0 | x) = 1 - P(y = 1 | x)$ können damit bei einem dichotomen Problem die Wahrscheinlichkeiten der Zugehörigkeit zu beiden Klassen bestimmt werden. Die Problemstellung der logistischen Regression ist in Abbildung 10.5 links dargestellt: Zwei Punktmengen sollen durch eine gerade Linie möglichst optimal getrennt werden; das Optimalitätskriterium wird später noch genauer angegeben. Die optimale Trennlinie repräsentiert die Menge aller Punkte x

mit

$$P(y = 1 | x) = P(y = 0 | x) = 0.5,$$

also die Menge aller Punkte, bei denen die Klassenzugehörigkeit nicht eindeutig festgestellt werden kann. Diese Linie trennt die Menge der Datenpunkte in zwei Teilmengen. Wir benötigen nun noch eine Funktion, um den beiden Regionen von Punkten mit

$$P(y = 1 | x) > 0.5 \quad \text{bzw.} \quad P(y = 1 | x) < 0.5$$

Wahrscheinlichkeiten zuordnen zu können. Dazu verwendet man die *logistische Funktion*

$$f(x) = \frac{1}{1 + e^{-x}}$$

deren Wertebereich (wie für eine Wahrscheinlichkeit gefordert) zwischen 0 und 1 liegt. Der Graph dieser Funktion ist in Abbildung 10.5 rechts zu sehen. Dabei wird im Modell x durch den Abstand von der Trennlinie ersetzt, sodass Punkte mit sehr großem Abstand Wahrscheinlichkeiten nahe bei 0 bzw. 1 haben, und Punkte nahe der Trennlinie Wahrscheinlichkeiten um die 0.5 ergeben. Da der Abstand eines zweidimensionalen Punktes $p = (p_1, p_2)$ von einer Trennlinie $\theta_1 x_1 + \theta_2 x_2 + \theta_0 = 0$ proportional zu $\theta_1 p_1 + \theta_2 p_2 + \theta_0$ ist, kann das logistische Modell mit den Abkürzungen $\bar{x} = (1, x_1, x_2)^T$ und $\theta = (\theta_0, \theta_1, \theta_2)^T$ geschrieben werden als

$$P(y = 1 | x; \theta) = \frac{1}{1 + e^{-\theta^T \cdot \bar{x}}}.$$

Man kann nachrechnen, dass unter bestimmten Verteilungsannahmen dieses Modell korrekt ist; wenn diese Annahmen nicht erfüllt sind, liefert das Modell immer noch eine gute Approximation.

Um im Folgenden die Notation zu vereinfachen, schreiben wir für die Modellausgabe kürzer

$$p(x; \theta) = P(y = 1 | x; \theta).$$

Für den Maximum-Likelihood-Ansatz benötigen wir einen Ausdruck für die $P(y | x; \theta)$, und nicht nur $P(y = 1 | x; \theta)$ (vergleiche dazu auch das Münzwurfen in Beispiel 10.1). Dies ist etwa durch

$$P(y | x; \theta) = p(x; \theta)^y \cdot (1 - p(x; \theta))^{1-y}$$

möglich, da daraus wie gewünscht

$$P(1 | x; \theta) = \frac{1}{1 + e^{-\theta^T \cdot \bar{x}}} \quad \text{und} \quad P(0 | x; \theta) = 1 - P(1 | x; \theta)$$

folgen.

Mit diesem Ausdruck erhält man die Likelihood-Funktion

$$\begin{aligned} L(\theta) &= P(\{y_1, \dots, y_n\} | \{x_1, \dots, x_n\}; \theta) \\ &= \prod_{i=1}^n \left(p(x_i; \theta)^{y_i} \cdot (1 - p(x_i; \theta))^{1-y_i} \right). \end{aligned}$$

An dieser Stelle benötigen wir zwei unterschiedliche Schreibweisen für die Größen, die in diesem Ausdruck vorkommen. Es gelten

$$p(x; \theta) = \frac{1}{1 + e^{-\theta^T \cdot x}} = \frac{e^{\theta^T \cdot x}}{1 + e^{\theta^T \cdot x}}$$

und

$$1 - p(x; \theta) = \frac{1}{1 + e^{\theta^T \cdot x}}.$$

Damit ist

$$\begin{aligned} -\log L(\theta) &= \sum_{i=1}^n \left(y_i \log p(x_i; \theta) + (1 - y_i) \log (1 - p(x_i; \theta)) \right) \\ &= \sum_{i=1}^n \left(y_i \log (e^{\theta^T \cdot x_i}) - y_i \log(1 + e^{\theta^T \cdot x_i}) - (1 - y_i) \log(1 + e^{\theta^T \cdot x_i}) \right) \\ &= \sum_{i=1}^n \left(y_i \theta^T \cdot x_i - \log(1 + e^{\theta^T \cdot x_i}) \right). \end{aligned}$$

Wie bei linearer Regression berechnen wir hier die Ableitung von $-\log L(\theta)$, um das Minimum dieser Funktion zu bestimmen. Dies liefert

$$\begin{aligned} -\frac{d \log L(\theta)}{d \theta} &= \sum_{i=1}^n \left(y_i x_i - \frac{1}{1 + e^{\theta^T \cdot x_i}} e^{\theta^T \cdot x_i} x_i \right) \\ &= \sum_{i=1}^n x_i (y_i - p(x_i; \theta)). \end{aligned}$$

Das Problem an dieser Stelle ist, dass nach Nullsetzen dieses Ausdrucks θ nicht freigestellt werden kann, es also keine exakte Lösung der Gleichung $-d \log L(\theta)/d \theta = 0$ gibt. Deshalb verwendet man zur Optimierung der Likelihood-Funktion mehrdimensionale Optimierungsverfahren, auf die wir hier aber nicht näher eingehen.

Beispiel 10.5 Gegeben seien 400 Datenpunkte in zwei Klassen, für die ein logistisches Regressionsmodell erstellt werden soll. Nach numerischer Optimierung des Logarithmus der Likelihood-Funktion $L(\theta)$ erhält man den Parametervektor $\theta = (-7.66, 2.36, 3.05)$. Damit ist die Menge aller Punkte

$$\{(x_1, x_2) \mid 2.36x_1 + 3.05x_2 - 7.66 = 0\}$$

die optimale Trennlinie zwischen den beiden Klassen, da an diesen Punkten x für die Klassenzugehörigkeit

$$P(y = 1 \mid x; \theta) = \frac{1}{1 + e^{-\theta^T \cdot \bar{x}}} = \frac{1}{1 + e^0} = \frac{1}{2}$$

gilt. In Abbildung 10.6 sind die beiden Punktmengen und diese Trennlinie dargestellt. Ebenfalls zu sehen sind die beiden Geraden, für die $P(y = 1 \mid x; \theta) = 0.75$ bzw. $P(y = 1 \mid x; \theta) = 0.25$ ist. \square

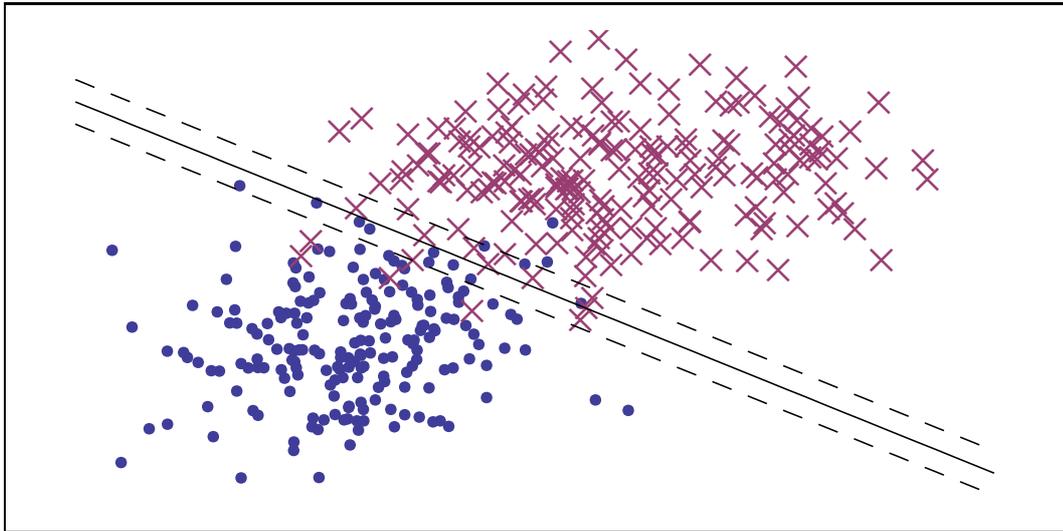


Abbildung 10.6: Die zwei Punktmengen aus Beispiel 10.5. Zu sehen sind außerdem die drei Linien mit $P(y = 1 | x; \theta) = 0.5$ (durchgezogen) und $P(y = 1 | x; \theta) = 0.25$ bzw. $P(y = 1 | \theta, x) = 0.75$ (gestrichelt).

Kapitel 11

Einführung in die Theorie der Konfidenzintervalle

Wir haben in Kapitel 9 und Kapitel 10 gesehen, wie man mit Statistiken auf Basis von Stichproben unbekannte Parameter von Verteilungen schätzen kann, die diesen Stichproben zugrunde liegen. Da diese Stichproben allerdings zufällig ausgewählt sind und damit auch zufälligen Schwankungen unterliegen, liefern diese sogenannten *Punktschätzer* Ergebnisse, die ebenfalls zufälligen Schwankungen unterliegen.

Beispiel 11.1 In einem Experiment möchte man auf Basis einer konkreten Stichprobe vom Umfang $n = 100$ den Parameter $p = P(\text{KOPF})$ einer möglicherweise unfairen Münze schätzen. Mit Beispiel 10.1 wissen wir, dass $\hat{p} = \frac{K}{n}$ der Maximum-Likelihood-Schätzer für p ist, wobei K die Anzahl von KOPF in n Versuchen bezeichnet. Gleichzeitig erfüllt \hat{p} auch die Definition des Stichprobenmittels, wenn $\text{KOPF} = 1$ und $\text{ZAHL} = 0$ gesetzt wird. Mit Satz 9.1 ist \hat{p} damit auch ein erwartungstreuer Schätzer. Bei einmaligem Durchführen des Experiments erhält man $\hat{p} = 0.47$, da beim 100-maligen Werfen der Münze 47 mal KOPF aufgetreten ist. Beim zweiten Durchführen des Experiments ergibt sich dann $\hat{p} = 0.51$. Wenn man dieses 100-malige Werfen selbst wieder 100 mal durchführt, erhält man eine Verteilung der \hat{p} -Werte, wie sie in Abbildung 11.1 zu sehen ist. \square

Beispiel 11.2 (Fortsetzung von Beispiel 11.1) Da die Varianz des Stichprobenmittels mit zunehmender Stichprobenanzahl n sinkt (siehe Satz 9.2), kann eine bessere Schätzung von $p = P(\text{KOPF})$ erreicht werden, indem man die Münze öfters wirft. In Abbildung 11.2 erkennt man, dass wie erwartet das 100-fache Wiederholen von $n = 1000$ -maligem Münzwerfen eine Menge von Schätzwerten liefert, die weniger weit streuen als die in Abbildung 11.1 gezeigten. Die Standardabweichung der beiden in Abbildungen 11.1 und 11.2 gezeigten Streuungen verhalten sich wie $\frac{\sqrt{1000}}{\sqrt{100}}$. \square

Man erkennt, dass das Ergebnis von Punktschätzungen von den konkreten Stichproben abhängt, ihrerseits zufällig ist, und innerhalb eines gewissen Bereichs streut. Die Idee von *Konfidenzintervallen* ist es nun, Punktschätzer durch *Intervallschätzer* zu ersetzen. Intervallschätzer liefern statt eines Wertes ein ganzes Intervall, in dem sich der unbekannte (zu schätzende) Wert mit gegebener Wahrscheinlichkeit befindet. Dazu verwenden wir folgende allgemein gehaltene Definition.

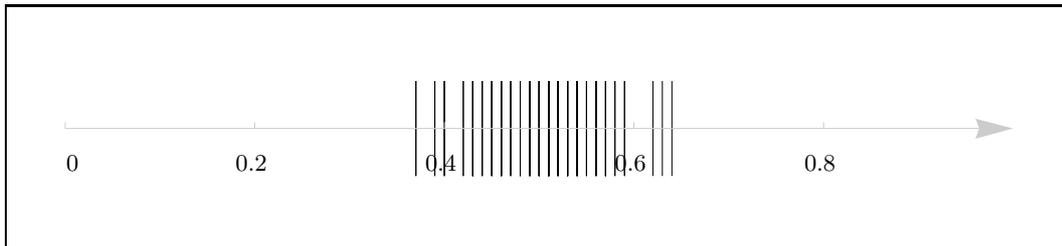


Abbildung 11.1: 100 Stichprobenmittel-Schätzer für $P(\text{KOPF})$ aus Beispiel 11.1. Da manche Werte öfters vorkommen, sind weniger als 100 Schätzwerte zu sehen.

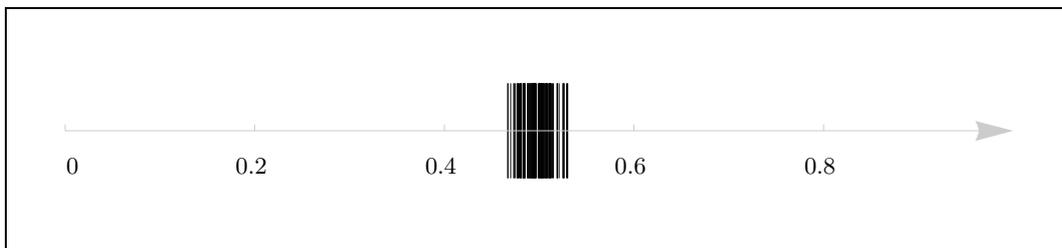


Abbildung 11.2: 100 Stichprobenmittel-Schätzer für $P(\text{KOPF})$ aus Beispiel 11.2; jeder einzelne Wert von \hat{p} ist Stichprobenmittel von 1000 Werten.

Definition 11.1 (Konfidenzintervall)

Unter einem *Konfidenzintervall* auf Niveau $1 - \alpha$ für einen unbekanntem Parameter θ einer Verteilung versteht man ein Intervall $[k_1, k_2]$ mit der Eigenschaft

$$P(\theta \in [k_1, k_2]) \geq 1 - \alpha.$$

Ein Konfidenzintervall auf Niveau $1 - \alpha$ wird oft auch als $(1 - \alpha)$ -Konfidenzintervall bezeichnet.

Die etwas umständlich erscheinende Formulierung des Niveaus über den Term $1 - \alpha$ hat seinen Ursprung in der Assoziation von Konfidenzintervallen und Testtheorie; dieser Zusammenhang wird im vorliegenden Skriptum kurz am Ende von Kapitel 12 angesprochen.

Durch die in obige Definition angegebene Bedingung an ein Konfidenzintervall ist es im Allgemeinen nicht möglich, die Intervallgrenzen k_1 und k_2 eindeutig zu bestimmen. Dies kann erst dadurch erreicht werden, dass man das Intervall symmetrisch um einen Punktschätzer $\hat{\theta}$ konstruiert. Die Grenzen können dann aus der Ungleichung

$$P(\theta \in [\hat{\theta} - k, \hat{\theta} + k]) \geq 1 - \alpha$$

berechnet. Meist wird als Punktschätzer der Maximum-Likelihood Schätzer θ_{ML} verwendet.

Der Vorteil von Konfidenzintervallen gegenüber Punktschätzern liegt also darin, dass ein ganzer Bereich angegeben werden kann *und* eine quantitative Aussage

darüber gemacht werden kann, wie sicher der unbekannte Parameter in diesem Intervall ist. Wenn man allerdings weitere Berechnungen auf Basis des zu schätzenden Parameters durchführen möchte, so sind Punktschätzer besser geeignet als Intervallschätzer.

Man beachte, dass der Berechnungsaufwand bei der Verwendung von Konfidenzintervallen höher ist als der von Punktschätzern, da auch noch die Streuung der Schätzer in Betracht gezogen werden muss. Dazu benötigt man Informationen über die Varianz der Schätzer. Da dies für den unbekannt Parameter p einer Binomialverteilung (wie in Beispielen 11.1 und 11.2) schwieriger ist als für eine Normalverteilung, werden wir zuerst Beispiele zu dieser Verteilung betrachten. Davor benötigen wir noch den Begriff des *Quantils* einer Verteilung.

Definition 11.2 (Quantil einer Verteilung)

Sei Z eine ZV mit Dichte f_Z . Dann nennt man den Wert c_α mit

$$\int_{-\infty}^{c_\alpha} f_Z(x) dx = \alpha$$

das α -Quantil dieser Verteilung.

Mit obiger Definition liegen $100 \cdot \alpha\%$ der Wahrscheinlichkeit einer Verteilung links von c_α . Für das α -Quantil c_α gilt somit die Gleichung

$$F_Z(c_\alpha) = \alpha,$$

wobei F_Z die Verteilungsfunktion zur Dichte f_Z bezeichnet.

Beispiel 11.3 Aus Tabellen kann man ablesen, dass $c_{0.975} = 1.96$ das 0.975-Quantil der Standard-Normalverteilung ist, und $c_{0.95} = 1.64$ das 0.95-Quantil. Links von diesen Werten befinden sich also 97.5% bzw. 95% der Wahrscheinlichkeitsmasse der Standard-Normalverteilung. \square

Mit dieser Notation können wir nun ein Konfidenzintervall für den Parameter μ einer Normalverteilung berechnen. Gegeben sei dazu eine Menge von Datenpunkten $\{x_1, \dots, x_n\}$, die normalverteilt sind mit bekannter Varianz σ^2 .

Wie wir bereits wissen, ist das Stichprobenmittel \bar{X} ein erwartungstreuer Schätzer für μ (siehe Beispiel 10.2). Wir berechnen ein um \bar{X} symmetrisches Konfidenzintervall, das die Bedingung

$$P(\bar{X} - k \leq \mu \leq \bar{X} + k) \geq 1 - \alpha \quad (11.1)$$

erfüllen soll. Für diese Berechnung benötigen wir noch die Verteilung von \bar{X} . Bekanntlich ist $\text{Var}(\bar{X}) = \text{Var}(X)/n$; da $\text{Var}(X) = \sigma^2$ bekannt ist, gilt also $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$. Somit ist $\bar{X}' = (\bar{X} - \mu)/\sqrt{\sigma^2/n}$ standard-normalverteilt.

Wir verwenden den Ansatz

$$P(-c \leq \bar{X}' \leq c) \geq 1 - \alpha, \quad (11.2)$$

um den unbekanntem Wert k in Ungleichung (11.1) zu berechnen. Wegen $\bar{X}' \sim N(0, 1)$ und $\Phi(-c) = 1 - \Phi(c)$ gilt

$$\begin{aligned} P(-c \leq \bar{X}' \leq c) &= 1 - \alpha \Leftrightarrow \Phi(c) - \Phi(-c) = 1 - \alpha \Leftrightarrow \Phi(c) = 1 - \frac{\alpha}{2} \\ &\Leftrightarrow c = c_{1-\alpha/2}, \end{aligned}$$

wobei $c_{1-\alpha/2}$ das $(1 - \frac{\alpha}{2})$ -Quantil der Standard-Normalverteilung bezeichne.

Wenn wir diesen Wert für c sowie $\bar{X}' = (\bar{X} - \mu)/\sqrt{\sigma^2/n}$ in Gleichung (11.2) einsetzen, so erhalten wir

$$P\left(-c_{1-\alpha/2} \leq \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \leq c_{1-\alpha/2}\right) \geq 1 - \alpha$$

und nach Umwandeln dieser Ungleichung in die gewünschte Form (11.1)

$$P\left(\bar{X} - c_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + c_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right) \geq 1 - \alpha.$$

Wir halten das Ergebnis dieser Berechnungen im folgenden Satz fest.

Satz 11.1 Sei $\{x_1, \dots, x_n\}$ eine normalverteilte konkrete Stichprobe vom Umfang n mit bekannter Varianz σ^2 und unbekanntem Erwartungswert μ . Das symmetrische Konfidenzintervall für μ auf Niveau $1 - \alpha$ ist

$$\left[\bar{x} - c_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + c_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right].$$

Beispiel 11.4 Für eine konkrete Stichprobe vom Umfang $n = 50$, Stichprobenmittel $\bar{x} = 6.74$ und als bekannt angenommene Varianz $\sigma^2 = 2.5$ erhält man für μ das 95%-Konfidenzintervall [6.30, 7.18].

Wenn wir annehmen, dass \bar{x} aus einer Stichprobe vom Umfang $n = 100$ berechnet wurde, so verändert sich das 95%-Konfidenzintervall zu [6.43, 7.05], wird also schmaler. Ebenso ändert sich das ursprüngliche Intervall, wenn wir als Niveau 90% verringern: In diesem Fall ist $c_{1-\alpha/2} = c_{0.95} = 1.64$, und das 90%-Konfidenzintervall ist [6.37, 7.11]. Das Konfidenzintervall wird also auch in diesem Fall schmaler. \square

Wie wir im letzten Beispiel gesehen haben, beruht auch die Lage des Konfidenzintervalls auf einem konkreten Punktschätzer, und ist damit auch wieder zufällig. Der Vorteil gegenüber reinen Punktschätzern liegt aber darin, dass der unbekannt zu schätzende Parameter in $1 - \alpha\%$ aller so konstruierten Konfidenzintervalle innerhalb des Intervalls liegt. Diese Tatsache ist graphisch in Abbildung 11.3 dargestellt.

Bevor wir exemplarisch zwei weitere Fälle der Verwendung von Konfidenzintervallen untersuchen, halten wir folgende wichtige Punkte fest: Die Breite des Konfidenzintervalls hängt ab von

- der Varianz der ZV X , aus deren Verteilung die konkrete Stichprobe gezogen wurde. Eine kleinere Varianz resultiert in einem schmälern Intervall;

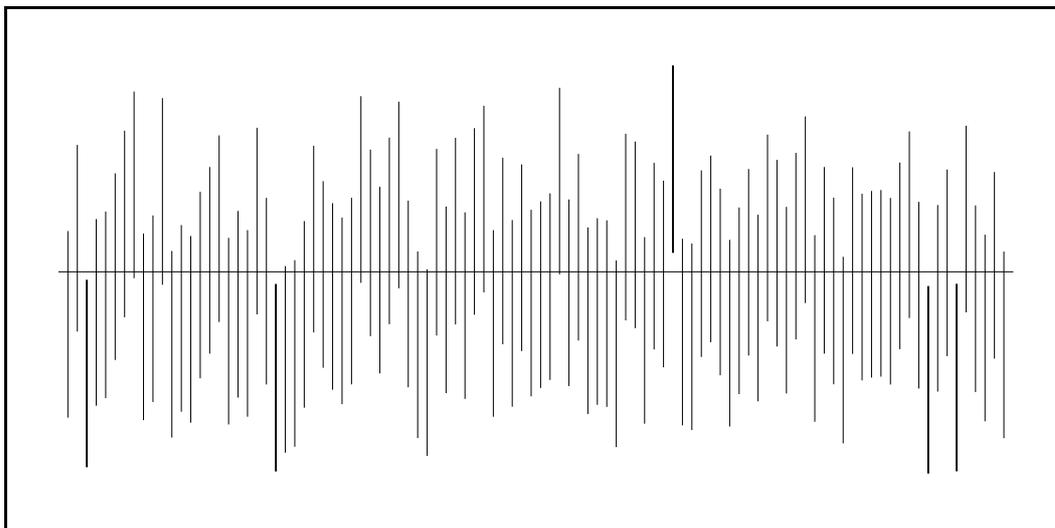


Abbildung 11.3: 100 95%-Konfidenzintervalle für einen unbekannt Parameter, dessen Lage durch den horizontalen Strich gekennzeichnet ist. Dieser Parameter liegt in nur 5 von 100 Fällen außerhalb des Konfidenzintervalls.

- dem Stichprobenumfang n . Das Intervall wird umso schmaler, je größer n ist;
- dem Niveau $1 - \alpha$. Je größer das Niveau ist, desto breiter wird das Konfidenzintervall.

Wir werden nun eine Variante der obigen Herleitung für das Konfidenzintervall des unbekannt Parameters μ einer Normalverteilung untersuchen. Dazu benötigen wir noch einige Definitionen.

Definition 11.3 (χ^2 -Verteilung)

Seien X_1, \dots, X_n unabhängige, standard-normalverteilte ZV. Dann bezeichnet man die Verteilung der ZV

$$Z = \sum_{i=1}^n X_i^2$$

als χ^2 -Verteilung mit n Freiheitsgraden.

Da wir die χ^2 Verteilung in diesem Kontext nur für die folgende Definition benötigen, werden wir nicht näher darauf eingehen.

Definition 11.4 (Student t -Verteilung)

Sei X eine standard-normalverteilte ZV, und Y eine von X unabhängige χ^2 -verteilte ZV mit n Freiheitsgraden. Dann bezeichnet man die Verteilung der ZV

$$Z = \frac{X}{\sqrt{\frac{Y}{n}}}$$

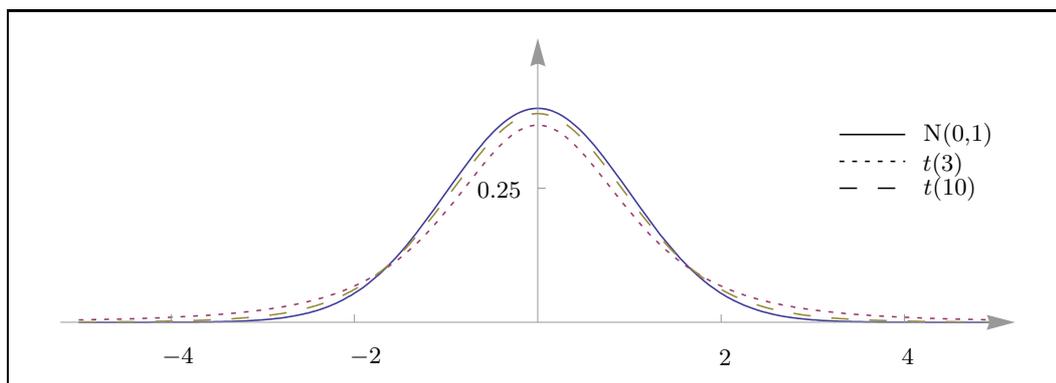


Abbildung 11.4: Dichten der Standard-Normalverteilung $N(0,1)$, sowie der Student t -Verteilung mit $n = 3$ und $n = 10$ Freiheitsgraden.

als *Student t -Verteilung* (oder kurz als *t -Verteilung*) mit n Freiheitsgraden. Für die Student t -Verteilung mit n Freiheitsgraden schreiben wir auch Student t_n .

Für große n (etwa $n > 30$) nähert sich die Student t -Verteilung der Standard-Normalverteilung an; für kleinere n streut die Student t -Verteilung (deutlich) mehr als die Standard-Normalverteilung. Dies ist auch in Abbildung 11.4 zu sehen.

Beim Schätzen des unbekanntes Parameters μ einer normalverteilten Stichprobe muss man etwas anders vorgehen, wenn die Varianz σ^2 nicht bekannt ist. Es sei wiederum eine $N(\mu, \sigma^2)$ -verteilte konkrete Stichprobe vom Umfang n gegeben, aus der ein Konfidenzintervall für μ berechnet werden soll. Im Gegensatz zu oben sei σ^2 diesmal nicht bekannt, sondern muss ebenfalls aus der Stichprobe geschätzt werden.

Mit Satz 9.3 ist die Stichprobenvarianz $S_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ ein erwartungstreuer Schätzer für σ^2 . Man könnte nun in Analogie zur Herleitung von Satz 11.1 annehmen, dass \bar{X} normalverteilt ist mit Parametern μ und S_X^2 . Dies trifft aber nicht zu! Vielmehr kann man nachrechnen, dass die standardisierte ZV

$$\bar{X}' = \frac{\bar{X} - \mu}{\sqrt{\frac{S_X^2}{n}}}$$

Student t -verteilt ist mit $n - 1$ Freiheitsgraden. Sei weiters $c_{1-\alpha/2; n-1}$ das $(1 - \alpha/2)$ -Quantil der Student t -Verteilung mit $(n - 1)$ Freiheitsgraden, und $S_X := \sqrt{S_X^2}$. Dann führt das Auflösen der Ungleichung

$$P\left(-c_{1-\alpha/2; n-1} \leq \frac{\bar{X} - \mu}{S_X/\sqrt{n}} \leq c_{1-\alpha/2; n-1}\right) \geq 1 - \alpha$$

zum Konfidenzintervall für μ , wenn σ^2 unbekannt ist.

Satz 11.2 Sei $\{x_1, \dots, x_n\}$ eine normalverteilte konkrete Stichprobe vom Umfang n , von der weder μ noch σ^2 bekannt sind. Das symmetrische Konfidenzin-

tervall für μ auf Niveau $1 - \alpha$ ist

$$\left[\bar{x} - c_{1-\alpha/2; n-1} \sqrt{\frac{s_x^2}{n}}, \bar{x} + c_{1-\alpha/2; n-1} \sqrt{\frac{s_x^2}{n}} \right],$$

wobei s_x^2 die empirische Varianz ist.

Beispiel 11.5 Wir betrachten nochmals das gleiche Zahlenmaterial wie in Beispiel 11.4, also $n = 50$, $\bar{x} = 6.74$, und ein Niveau von 95%. Der Unterschied sei nun, dass wir $\sigma^2 = 2.5$ nicht als gegeben annehmen, sondern als $s_x^2 = 2.5$ aus den Daten schätzen. Mit Tabellen oder Computer-Mathematik-Systemen kann man das Quantil $c_{0.975; 49} = 2.01$ ablesen bzw. berechnen. Damit erhalten wir das 95%-Konfidenzintervall $[6.29, 7.19]$, das im Gegensatz zum Konfidenzintervall für bekanntes σ^2 etwas breiter ist. Je kleiner n ist, desto grösser der Unterschied zwischen der Student t -Verteilung mit n Freiheitsgraden und der Standardnormalverteilung. Dementsprechend größer wird dann auch das Konfidenzintervall, wenn σ^2 durch s_x^2 geschätzt werden muss. Das ist auch dahingehend einleuchtend, dass ja durch die Schätzung von σ^2 ein weiterer nicht-deterministischer Faktor in die Rechnung einfließen muss. \square

Nach diesen Berechnungen von Konfidenzintervallen für μ wenden wir uns nun der Aufgabenstellung aus den einführenden Beispielen dieses Kapitels zu und berechnen ein symmetrisches $(1 - \alpha)\%$ -Konfidenzintervall für den unbekanntes Parameter $p = P(\text{KOPF})$ des Münzwürfens. Der dafür benötigte Punktschätzer $\hat{p} = \frac{K}{n}$ ist der erwartungstreue Maximum Likelihood-Schätzer für das unbekanntes p . Zu bestimmen ist somit, für gegebenes \hat{p} , ein k mit der Eigenschaft

$$P(\hat{p} - k \leq p \leq \hat{p} + k) \geq 1 - \alpha. \quad (11.3)$$

Wir wissen aus Satz 8.4, dass sich die Binomialverteilung unter gewissen Bedingungen ($np(1-p) \geq 10$) gut durch die Normalverteilung approximieren lässt; diese Bedingung ist für unsere Werte $n = 100$ bzw. $n = 1000$ erfüllt. Da $\hat{p} = \bar{X}$ ein erwartungstreuer Schätzer ist, gilt $E(\hat{p}) = p$. Die ZV X ist hier entweder 0 (bei ZAHL, mit W. $1-p$) oder 1 (bei KOPF, mit W. p); es gilt $\text{Var}(X) = p(1-p)$.

Mit Satz 9.2 über die Varianz von $\hat{p} = \bar{X}$ gilt $\text{Var}(\hat{p}) = \text{Var}(X)/n = p(1-p)/n$. Somit ist \hat{p} annähernd $N(p, \frac{p(1-p)}{n})$ -verteilt, und die standardisierte ZV $\hat{p}' = (\hat{p} - p)/\sqrt{p(1-p)/n}$ ist annähernd standard-normalverteilt.

Zur Bestimmung der um \hat{p} symmetrischen Intervallgrenzen kann man folgendermaßen umformen. Dabei gehen wir ähnlich wie bei der Herleitung von Satz (11.1) vor. Wegen $\hat{p}' \sim N(0, 1)$ gilt

$$P(-c \leq \hat{p}' \leq c) \geq 1 - \alpha \Leftrightarrow 2\Phi(c) \geq 2 - \alpha \Leftrightarrow c = c_{1-\alpha/2},$$

wobei $c_{1-\alpha/2}$ wiederum das $(1 - \frac{\alpha}{2})$ -Quantil der Standard-Normalverteilung ist. Einsetzen der Definition von \hat{p}' in die erste Ungleichung oben liefert nach Umformen

$$P\left(-c_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}} \leq \hat{p} - p \leq c_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}}\right) \geq 1 - \alpha.$$

Hier steht man vor dem Problem, dass diese Ungleichung erst durch Auflösen einer quadratischen Gleichung direkt auf die gewünschte Form (11.3) umgeformt werden kann, da p auf allen drei Seiten der Ungleichungen vorkommt. Hier kann man einen Trick anwenden: man ersetzt unter der Wurzel p durch den Schätzwert \hat{p} . Damit erhält man

$$P\left(\hat{p} - c_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + c_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \geq 1 - \alpha.$$

Wir halten dieses Ergebnis in verallgemeinerter Form für beliebige Binomialverteilungen fest.

Satz 11.3 Sei $\{x_1, \dots, x_n\}$ eine Bernoulli-verteilte konkrete Stichprobe vom Umfang n , für die $p = P(x_i = 1)$ unbekannt ist, sowie $\hat{p} = \bar{x}$. Für $np(1-p) \geq 10$ ist

$$\left[\hat{p} - c_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + c_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right].$$

ein approximatives symmetrisches Konfidenzintervall für p auf Niveau $1 - \alpha$.

Der Zusammenhang dieser Formulierung über Realisierungen einer Bernoulli-verteilten mathematischen Stichprobe X_1, \dots, X_n und der Binomialverteilung liegt darin, dass die ZV $Z = \sum_{i=1}^n X_i$ binomialverteilt ist mit Parametern p und n . Mit der Formel in Satz 11.3 lässt sich somit ein approximatives Konfidenzintervall für den Parameter p von beliebigen Binomialverteilung angeben.

Beispiel 11.6 (Fortsetzung von Beispielen 11.1 und 11.2) Für $n = 100$ und ein konkretes Resultat von $\hat{p} = 0.48$ erhält man somit das 95%-Konfidenzintervall $[0.382, 0.578]$. Wenn wir dasselbe \hat{p} bei $n = 1000$ Würfeln erhalten, so verändert sich das 95%-Konfidenzintervall auf $[0.449, 0.511]$, wird also schmaler.

Ebenso ändert sich die Größe des Intervalls, wenn wir das Niveau auf 99% anheben. Das zugehörige Quantil der Normalverteilung ist $c_{0.995} = 2.58$. Für $n = 100$ erhält man daraus das 99%-Konfidenzintervall $[0.351, 0.609]$, für $n = 1000$ das Intervall $[0.439, 0.521]$. Die Intervalle werden durch Hinaufsetzen des Niveaus also breiter. \square

Beispiel 11.7 Vor Einführen eines neuen Produkts soll durch eine Meinungsumfrage erhoben werden, welcher Anteil der Bevölkerung sich vorstellen könnte, das neue Produkt zu erwerben. Für diesen unbekanntem Anteil p der Bevölkerung soll ein 95% Konfidenzintervall konstruiert werden, das Breite 0.06 hat (sodass das Umfrageergebnis mit einer Schwankungsbreite von $\pm 3\%$ veröffentlicht werden kann). Wieviele Personen müssen befragt werden, um ein Intervall dieser Breite zu erhalten?

Lösung: Mit der Formel aus Satz (11.3) gilt für die Breite des Konfidenzintervalls

$$2 \cdot 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.06.$$

Umformen liefert

$$\hat{p}(1-\hat{p}) = 0.0002343 n.$$

Auch wenn \hat{p} unbekannt ist, kann man die linke Seite dieser Gleichung abschätzen: Der Term $\hat{p}(1 - \hat{p})$ wird durch $\hat{p} = 0.5$ maximiert mit einem Wert von 0.25. Somit müssen $n = 1067$ Personen befragt werden, um eine Schwankungsbreite von maximal 0.03 um den erhobenen Schätzwert zu erreichen.

Wenn etwa bekannt ist, dass sich zwischen 10% und 30% der Bevölkerung für das Produkt interessieren werden, kann obige Abschätzung verbessert werden: Der Term $\hat{p}(1 - \hat{p})$ wird in diesem Fall durch $\hat{p} = 0.3$ maximiert; Umformen liefert dann $n = 896$. \square

An dieser Stelle könnte man noch Konfidenzintervall für die Parameter σ^2 einer Normalverteilung, oder für Parameter von Poisson- und hypergeometrischer Verteilung angeben. Dies würde in dieser Einleitung zu weit führen. Die angegebenen Beispiele sollen vielmehr zeigen, wie man ausgehend von einem Punktschätzer und dem Wissen um die Verteilung dieses Punkteschätzers vorgehen muss, um für beliebige Verteilungen Konfidenzintervalle angeben zu können.

Kapitel 12

Einführung in die Testtheorie

Im Rahmen dieser Vorlesung haben wir uns hauptsächlich damit beschäftigt einen formalen Apparat aufzubauen, um W . für bestimmte Zufallsexperimente auszurechnen. Dabei waren die Zufallsexperimente klar definiert und das Lösen der Aufgaben bestand daraus, durch theoretische Überlegungen geeignete Verteilungen von Zufallsvariablen zu bestimmen und dann die gesuchten W . über diese Verteilungen zu berechnen.

Im Folgenden wird diese Situation ein wenig verallgemeinert. Wir betrachten Zufallsexperimente, die nicht vollständig bestimmt sind, und überprüfen dann, ob eine Annahme, die das Zufallsexperiment vollständig bestimmt macht, im Gegensatz zum betrachteten Zahlenmaterial steht. Was das konkret bedeuten soll, kann man am Einfachsten anhand eines Beispiels erläutern.

Beispiel 12.1 Spieler A bietet Spieler B folgende Wette an: Wenn beim 100-maligen Münzwerfen öfter KOPF als ZAHL erscheint, gewinnt er (Spieler A) die Wette, sonst Spieler B. Die Münze zum Durchführen der Wette wird von Spieler A gestellt; er behauptet, es wäre eine faire Münze. Bevor Spieler B sich auf die Wette einlassen will, möchte er noch überprüfen, ob die Münze auch wirklich fair ist. Dafür wirft er sie 100 mal; es erscheint 61 mal KOPF. Soll sich Spieler B auf diese Wette einlassen?

Lösung: Es geht bei der Aufgabenstellung also darum herauszufinden, wie wahrscheinlich bei einer fairen Münze ein Ergebnis von mindestens 61 mal KOPF ist. Wir lösen die Aufgabe folgendermaßen: Die Anzahl Z der auftretenden Köpfe in diesem Beispiel ist binomialverteilt mit $n = 100$. Wenn die Münze fair ist (also $p = \frac{1}{2}$ gilt), sind Ergebnisse um $E(Z) = 50$ wahrscheinlicher als Ergebnisse, die stark von diesem Erwartungswert abweichen. Wie wahrscheinlich ist es also, dass mit einer fairen Münze ein Ergebnis (KOPF oder ZAHL) mehr als 60 mal auftritt? Man erhält

$$\begin{aligned} P_Z(\{0, \dots, 39\} \cup \{61, \dots, 100\}) &= 1 - P_Z(\{40, \dots, 60\}) \\ &= 1 - 0.9648 = 0.0352 \end{aligned}$$

Ein solches Ergebnis ist mit einer fairen Münze also äußerst unwahrscheinlich, und Spieler B wird die Wette daher vermutlich ablehnen. \square

An dieser Aufgabe kann eine allgemeine Vorgehensweise abgelesen werden: Es soll überprüft werden, wie realistisch das Auftreten eines Ergebnisses unter einer bestimmten Annahme ist (z.B. $p = \frac{1}{2}$ bei obigem Beispiel). Wenn das Ergebnis

unter der Annahme nicht wahrscheinlich ist, ist die Annahme nicht realistisch und wird abgelehnt.

Etwas präziser ausgedrückt: Die Testtheorie beschäftigt sich mit dem Überprüfen von *Hypothesen*. Diese Hypothesen sind meist Annahmen über den Wert eines Parameters, so etwa die Hypothese

$$\theta = \theta_0,$$

dass ein unbekannter Parameter θ einen konkreten Wert θ_0 annehme. Bei dieser Hypothese (auch *Nullhypothese* genannt) ist die *Alternative* $\theta \neq \theta_0$ zweiseitig: entweder ist $\theta < \theta_0$ oder $\theta > \theta_0$. Eine andere Art von Hypothese ist

$$\theta \geq \theta_0,$$

dass eine unbekannter Parameter θ mindestens so groß wie θ_0 ist. Hier ist die Alternative $\theta < \theta_0$ nur einseitig.

Nach Festlegung der Hypothese (und damit auch der Alternative) wird die W. des eingetretenen Ergebnisses unter der Hypothese berechnet. Wenn diese W. sehr gering ist, ist die Hypothese keine gute Erklärung für das Ergebnis und kann verworfen werden. Ist die W. groß, so bleibt man bei der Hypothese. Bei diesem Vorgehen können zwei Arten von Fehlern gemacht werden:

- *Fehler 1. Art* Die Hypothese ist wahr, wird aber fälschlicherweise abgelehnt, weil das Ergebnis unter der Hypothese unwahrscheinlich ist.
- *Fehler 2. Art* Die Hypothese ist falsch, man bleibt aber fälschlicherweise bei ihr, da das Ergebnis nicht in genügend großem Widerspruch dazu steht.

Idealerweise wird ein Test so angelegt werden, dass die W. für Fehler 1. und 2. Art minimiert werden. Da es schwierig ist, die W. für beide Fehler gleichzeitig zu minimieren, konzentriert man sich darauf, Fehler 1. Art so gering wie möglich zu halten. Dementsprechend ist ein Test mit *Signifikanz* α einer, bei dem die W. eines Fehlers 1. Art höchstens α ist. Gebräuchliche Werte für α sind etwa 0.01, 0.025 oder 0.05.

Die Signifikanz des Tests aus Beispiel 12.1 ist 0.0352: Wenn wir die Hypothese $p = \frac{1}{2}$ verwerfen, machen wir nur in etwa 3.5% der Fälle einen Fehler 1. Art. Es ist zu beachten, dass über Fehler 2. Art nichts ausgesagt werden kann. Wenn bei obigem Beispiel etwa 55 mal KOPF erschienen wäre, so wäre man vermutlich bei der Hypothese geblieben, da $\text{Bin}(100, \frac{1}{2})(\{46, \dots, 54\}) = 0.6318$ und somit die W. für einen Fehler 1. Art 0.3682 gewesen wäre. Die W. für einen Fehler 2. Art, dass wir also der Hypothese $p = \frac{1}{2}$ bleiben, obwohl die Münze eigentlich nicht fair ist, können wir aber nicht berechnen. Damit werden wir uns weiter unten noch ausführlicher beschäftigen. Zuvor behandeln wir noch die Frage, wie man zu gegebener Signifikanz den Bereich finden kann, bei dem man die Hypothese ablehnt.

Beispiel 12.2 Wir wollen wie bei Beispiel 12.1 oben testen, ob eine Münze fair ist oder nicht. Es ist also wieder die Hypothese $p = \frac{1}{2}$. Wir wollen nun, bevor wir zur Überprüfung die Münze 100 mal werfen berechnen, ab welchem Ergebnis wir mit Signifikanz 0.05 die Hypothese verwerfen können. Sei also $Z \text{ Bin}(100, \frac{1}{2})$ -verteilt; wir möchten wissen, für welches k

$$1 - P_Z(\{50 - k, \dots, 50 + k\}) < 0.05$$

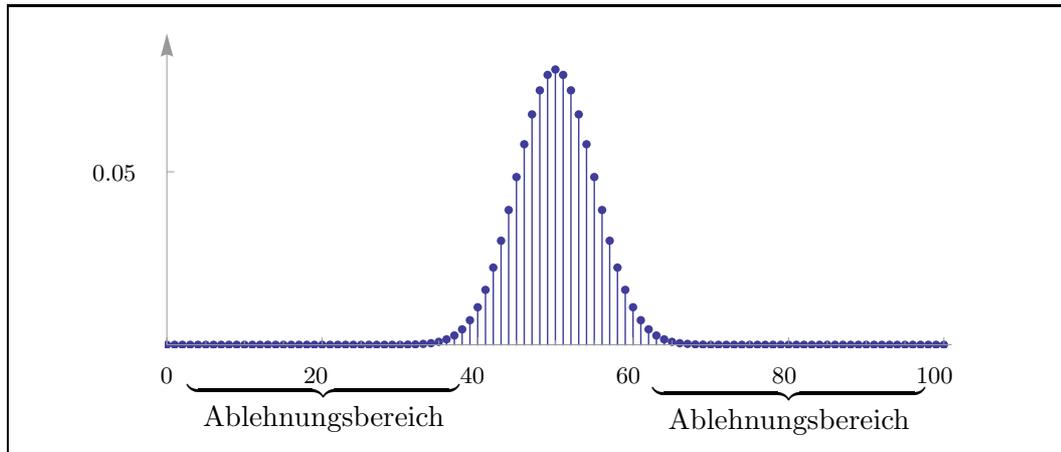


Abbildung 12.1: Ablehnungsbereich für $\alpha = 0.05$ bei der Hypothese $p = \frac{1}{2}$ in Beispiel 12.2. Die W. des Ablehnungsbereichs ist kleiner als α .

ist. Mit dem zentralen Grenzwertungssatz ist Z annähernd $N(50, 25)$ -verteilt, und es gilt (mit Korrekturterm für die Approximation der Binomialverteilung durch die Normalverteilung)

$$\begin{aligned} P_Z(\{50 - k, \dots, 50 + k\}) &= P(\{Z \in \{50 - k, \dots, 50 + k\}\}) \\ &= \Phi\left(\frac{50.5 + k - 50}{5}\right) - \Phi\left(\frac{49.5 - k - 50}{5}\right) \\ &= \Phi\left(\frac{k + 0.5}{5}\right) - \Phi\left(\frac{-k - 0.5}{5}\right) = 2\Phi\left(\frac{k + 0.5}{5}\right) - 1. \end{aligned}$$

Somit ergibt sich für die ursprüngliche Gleichung

$$\begin{aligned} 1 - \left(2\Phi\left(\frac{k - 0.5}{5}\right) - 1\right) &< 0.05 \\ \Phi\left(\frac{k - 0.5}{5}\right) &> 0.975. \end{aligned}$$

Mit Tabelle 8.1 ergibt sich also $\frac{k-0.5}{5} = 1.96$, somit ist $k > 9.3$.

Wenn also die Zahl des Auftretens von KOPF kleiner als 41 oder größer als 59 ist, kann man die Hypothese, dass die Münze fair ist, verwerfen, und nur in 5 von 100 gleichen Fällen damit einen Fehler machen. \square

Die Situation des letzten Beispiels ist graphisch in Abbildung 12.1 dargestellt: Wenn die Anzahl von KOPF in den Bereich über 59 (oder unter 41) fällt, verwirft man die Hypothese $p = \frac{1}{2}$ mit $\alpha = 0.05$. Für gegebenes α nennt man den Wertebereich der ZV, bei dem man die Hypothese ablehnt, *Ablehnungsbereich* oder *kritischer Bereich*. Dieser wird bei großem α groß, und bei kleinem α klein sein.

Es ist zu beachten, dass man mit statistischen Tests dieser Art nie eine Hypothese beweisen kann, weil der Fehler 2. Art nicht begrenzt ist. Will man also eine Aussage treffen, zu der man eine Fehlerabschätzung machen kann, muss man gegebenenfalls die Hypothese und Alternative vertauschen, und dann die Hypothese mit Fehlerw.

α verwerfen. Oft bleibt man bei der Hypothese, obwohl sie falsch ist, wie man am nächsten Beispiel sehen kann.

Beispiel 12.3 Eine Münze werde 100 mal geworfen. Um zu testen, ob die Ergebnisse zufällig sind, formuliert man als Hypothese $p = \frac{1}{2}$. Wie wir aus vorigem Beispiel wissen, kann man die Hypothese $p = \frac{1}{2}$ verwerfen, wenn eine der beiden Seiten mehr als 59 mal geworfen wird. Nehmen wir nun an, die Sequenz S der Würfe wäre

$$S = \text{KOPF, ZAHL, KOPF, ZAHL, KOPF, ZAHL, } \dots, \text{ ZAHL, KOPF, ZAHL,}$$

es treten also immer abwechseln KOPF und ZAHL auf. Somit ist sowohl die Anzahl von KOPF und ZAHL gleich 50, und man bleibt bei der Hypothese (was für *diese* Hypothese natürlich richtig ist). Die Schwierigkeit ergibt sich daraus, eine geeignete Hypothese zu finden, die man dann verwerfen kann, um zu zeigen, dass diese Sequenz nicht zufällig ist. Dazu bietet sich Beispiel 5.4 an: Dort haben wir berechnet, dass die Verteilung der ZV Z , die die Anzahl der Runs in einer Münzwurffolge der Länge n angibt, durch die Dichte

$$f_Z(k) = \frac{\binom{n-1}{k-1}}{2^{n-1}}$$

gegeben ist. Ein geeignete Hypothese, die man dann verwerfen kann, ist die folgende: Die Anzahl der Runs in S entspricht der Anzahl der Runs in einer Zufallsfolge. Wenn die Anzahl der Runs in S in etwa dem Erwartungswert der Anzahl der Runs in einer Zufallsfolge entspricht, wird man bei dieser Hypothese bleiben. Wenn diese Zahl aber zu stark vom Erwartungswert abweicht, wird man die Hypothese verwerfen.

Wir benötigen also den Erwartungswert für die Anzahl Z der Runs in einer Serie von n Münzwürfen. Dies ist mit dem Ergebnis von Beispiel 8.9

$$E(Z) = \frac{n+1}{2} = \frac{101}{2}.$$

Wir setzen nun die W. für einen Fehler 1. Art auf $\alpha = 0.05$. Wie wir uns ebenfalls in Beispiel 8.9 überlegt haben, ist Z aufgrund des zentralen Grenzwertungssatzes annähernd normalverteilt mit obigem Erwartungswert und Varianz $\frac{n-1}{4}$. Somit berechnet man den Ablehnungsbereich für $\alpha = 0.05$, ebenfalls wieder mit Korrekturterm, über

$$\begin{aligned} 1 - P\left(Z \in \left\{\frac{101}{2} - k, \dots, \frac{101}{2} + k\right\}\right) &< 0.05 \\ P\left(\left(\frac{101}{2} - \frac{1}{2} - k - \frac{101}{2}\right)\sqrt{\frac{4}{99}} \leq Z' \leq \left(\frac{101}{2} + \frac{1}{2} + k - \frac{101}{2}\right)\sqrt{\frac{4}{99}}\right) &> 0.95 \\ \Phi\left(\frac{2k+1}{\sqrt{99}}\right) - \Phi\left(-\frac{2k+1}{\sqrt{99}}\right) &> 0.95 \\ \Phi\left(\frac{2k+1}{\sqrt{99}}\right) &> 0.975. \end{aligned}$$

Aus Tabelle 8.1 erhält man dann $\frac{2k+1}{\sqrt{99}} = 1.96$, also $k = 9.25$. Somit wird man die Hypothese, dass S eine Zufallssequenz ist verwerfen, wenn Z größer als 59 oder kleiner als 41 ist. Für obiges S haben wir $Z = 100$, und wir verwerfen die Hypothese. \square

Beim statistischen Testen kann man jede Hypothese verwenden, unter der man die Verteilung einer ZV und damit den Ablehnungsbereich für gegebenes α berechnen kann. So kann man z.B. testen, ob zwei ZV die gleiche Verteilung haben, oder ob sie unabhängig sind. Darauf werden wir hier nicht näher eingehen, sondern behandeln noch ein Beispiel für stetige Verteilungen.

Beispiel 12.4 In einer Fabrik werden Kugellagerkugeln mit Durchmesser 5 mm produziert. Wir wissen, dass durch produktionsbedingte Schwankungen natürlich nicht jede Kugel genau Durchmesser 5 mm hat; die Durchmesser einzelner Kugeln können jedoch sicherlich als Realisierungen einer $N(\mu, \sigma^2)$ -verteilten ZV betrachtet werden. Aus langjähriger Erfahrung weiß man, dass in dieser Fabrik $\sigma = 0.01$ ist.

Zur Qualitätskontrolle beobachtet man nun bei 100 Kugeln die Durchmesser x_1, \dots, x_{100} und berechnet den Durchschnitt \bar{x} dieser Werte. Wie groß muss die Abweichung dieses Durchschnitts vom Sollwert von 5 mm sein, um mit $\alpha = 0.05$ die Hypothese $\mu = 5$ ablehnen zu können?

Lösung: Wir betrachten die x_i als Realisierungen von $N(\mu, \sigma^2)$ -verteilten unabhängigen ZV X_i . Wir wissen aus Satz 9.1 und Satz 9.2, dass das Stichprobenmittel \bar{X} dann $N(\mu, \frac{\sigma^2}{n})$ -verteilt ist. In unserem Beispiel ist $n = 100$. Wir werden die Hypothese $\mu = 5$ ablehnen, wenn \bar{X} mehr als einen bestimmten Wert von 5 abweicht; dieser Wert hängt natürlich von der gewählten Testsignifikanz $\alpha = 0.05$ ab. Wir verwerfen also die Hypothese, wenn $P(\{|\bar{X} - 5| > c\}) < 0.05$ ist. Dies ist wegen

$$\begin{aligned} P(\{|\bar{X} - 5| > c\}) &= 1 - P(\{|\bar{X} - 5| \leq c\}) \\ &= 1 - N\left(5, \frac{0.01^2}{100}\right) [5 - c, 5 + c] \\ &= 1 - N(0, 1) \left[\frac{5 - c - 5}{\frac{0.01}{\sqrt{100}}}, \frac{5 + c - 5}{\frac{0.01}{\sqrt{100}}} \right] \\ &= 1 - N(0, 1) \left[\frac{-10c}{0.01}, \frac{10c}{0.01} \right] = 1 - \left(2\Phi\left(\frac{10c}{0.01}\right) - 1 \right) \end{aligned}$$

bei $\Phi\left(\frac{10c}{0.01}\right) > 0.975$ der Fall. Aus Tabelle 8.1 erhält man $\frac{10c}{0.01} > 1.96$ und daraus $c > 0.00196$. Wenn also der Durchschnitt der gemessenen Kugeldurchmesser größer als 5.00196 oder kleiner als 4.99804 ist, wird man die Hypothese $\mu = 5$ verwerfen und damit nur in 5% der Fälle einen Fehler machen. \square

Dem interessierten Leser/der interessierten Leserin ist möglicherweise ein Zusammenhang zwischen Konfidenzintervallen und dem Bereich eines statistischen Tests aufgefallen, in dem die Hypothese nicht verworfen wird. Dies ist tatsächlich der Fall: Ein statistischer Test mit Signifikanz α für eine Nullhypothese $\theta = \theta_0$ wird genau dann abgelehnt werden, wenn das $(1 - \alpha)$ -Konfidenzintervall den Wert θ_0 nicht enthält. Wir betrachten dazu ein paar Beispiele.

Beispiel 12.5 (Fortsetzung von Beispiel 12.4) Mit der Formel aus Beispiel 11.4 ist das 95%-Konfidenzintervall für den unbekanntem (als 5 vermuteten) Wert μ gegeben durch

$$\left[\bar{x} - c_{0.975} \frac{\sigma}{\sqrt{n}}, \bar{x} + c_{0.975} \frac{\sigma}{\sqrt{n}} \right] = [\bar{x} - 0.00196, \bar{x} + 0.00196].$$

Ein Vergleich mit Beispiel 12.4 lässt erkennen: Wenn \bar{x} im Ablehnungsbereich $[-\infty, 4.99804] \cup [5.00196, \infty]$ dieses Tests liegt, dann ist der vermutete Wert von $\mu = 5$ nicht mehr im 95%-Konfidenzintervall um \bar{x} . \square

Beispiel 12.6 Die Nullhypothese $\mu = \mu_0$ soll auf Basis einer normalverteilten Stichprobe von gegebenem Umfang n mit gegebener Signifikanz α getestet werden. Man konstruiere einen geeigneten Test für diese Aufgabenstellung.

Lösung: Wenn die Nullhypothese zutrifft, ist die konkrete Stichprobe $N(\mu_0, \sigma^2)$ -verteilt, wobei das unbekannte σ^2 durch die Stichprobenvarianz S_X^2 geschätzt werden muss. Erst wenn der erwartungstreue Schätzer \bar{X} für μ zu weit vom Wert μ_0 abweicht, wird man sich gegen die Nullhypothese entscheiden. Wie wir bereits wissen (siehe Beispiel 11.5), ist unter der Nullhypothese $\mu = \mu_0$ das Stichprobenmittel \bar{X} Student t_{n-1} -verteilt mit $E(\bar{X}) = \mu_0$ und $\text{Var}(\bar{X}) = S_X^2/n$. Die Nullhypothese wird für all jene Werte von \bar{X} abgelehnt, für die

$$P(|\bar{X} - \mu_0| > c) < \alpha$$

gilt. Aus dieser Bedingung kann man sich die Grenze c ausrechnen, ab der die Nullhypothese verworfen wird. Umformen liefert

$$\begin{aligned} 1 - P(|\bar{X} - \mu_0| \leq c) &< \alpha \\ P(|\bar{X} - \mu_0| \leq c) &> 1 - \alpha \\ P_{\bar{X}}[\mu_0 - c, \mu_0 + c] &> 1 - \alpha \\ \text{Student } t_{n-1} \left[\frac{-c}{\sqrt{S_X^2/n}}, \frac{c}{\sqrt{S_X^2/n}} \right] &> 1 - \alpha \\ 2T_{n-1} \left(\frac{c\sqrt{n}}{S_X} \right) - 1 &> 1 - \alpha, \end{aligned}$$

wobei T_{n-1} die (um Null symmetrische) Verteilungsfunktion von Student t_{n-1} bezeichne. Daraus ergibt sich mit der Notation $c_{1-\alpha/2; n-1}$ für das $(1 - \frac{\alpha}{2})$ -Quantil der Student t_{n-1} -Verteilung folgender Wert für c :

$$c = c_{1-\alpha/2; n-1} \frac{S_X}{\sqrt{n}}.$$

Für konkrete Zahlen, etwa $\mu_0 = 10$, $n = 100$, $\alpha = 0.05$ und $S_X^2 = 9$ wird man bei einem konkreten Stichprobenmittel von $\bar{x} = 10.5$ die Nullhypothese noch nicht ablehnen, da man mit obiger Formel $c = 0.603$ erhält. Die Differenz $|\mu_0 - \bar{x}| = |10 - 10.5| = 0.5$ ist kleiner als $c = 0.603$, man wird somit bei der Nullhypothese $\mu_0 = 10$ bleiben.

Man beachte, dass obiger Wert für c identisch ist mit der in Beispiel 11.5 berechneten halben Länge des Konfidenzintervalls für den unbekannt Parameter μ , der aufgrund des Stichprobenmittels \bar{X} geschätzt werden soll. Auch hier gilt somit, dass die Nullhypothese $\mu = \mu_0$ genau dann abgelehnt wird, wenn dieser Wert nicht mehr im $(1 - \alpha)$ -Konfidenzintervall um \bar{X} enthalten ist. \square

Weiterführende Literatur

- [Beaumont, 1986] G.P. Beaumont. *Probability and Random Variables*. Ellis Horwood Limited, Chichester, 1986.
- [Bialas, 2005] W.F. Bialas. *Lecture Notes in Applied Probability*. University of Buffalo, 2005. Frei verfügbar unter <http://www.acsu.buffalo.edu/~bialas/>.
- [Blom, 1989] G. Blom. *Probability and Statistics: Theory and Applications*. Springer-Verlag, New York, 1989.
- [Dür, 1988] A. Dür. Skriptum zur Vorlesung Stochastische Methoden I. Institut für Mathematik, Universität Innsbruck, 1988.
- [Freudenthal, 1973] H Freudenthal. *Mathematics as an Educational Task*. D. Reidel Publishing Company, Dordrecht - Holland, 1973.
- [Grinstead and Snell, 1997] C.M. Grinstead and J.L. Snell. *Introduction to Probability*. American Mathematical Society, 1997. Frei verfügbar unter <http://www.dartmouth.edu/~chance/>.
- [Larson and Marx, 2006] R.J. Larson and M.L. Marx. *An Introduction to Mathematical Statistics and Its Applications*. Pearson Prentice Hall, Upper Saddle River, NJ, 4th edition, 2006.
- [MacKay, 2003] D.J.C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.
- [Ross, 1987] S.M. Ross. *Introduction to Probability and Statistics for Engineers and Scientists*. John Wiley & Sons, New York, 1987.
- [Székely, 1990] G.J. Székely. *Paradoxa. Klassische und neue Überraschungen aus Wahrscheinlichkeitsrechnung und mathematischer Statistik*. Verlag Harri Deutsch, Thun und Frankfurt am Main, 1990.
- [Weiss, 1992] P. Weiss. Stochastik I: Einführung in die Wahrscheinlichkeitsrechnung (Skriptum zur Vorlesung). Institut für Mathematik, Universität Linz, 1992.