Kapitel 3

Ein Statistisches Intermezzo

"Strange events permit themselves the luxury of occurring." (Charlie Chan)

"Conditioning is the soul of statistics." (Joe Blitzstein)

Unsere Umwelt produziert am laufenden Band zufällige Ergebnisse wie Wolken, Aktienkurse, Herzinfarkte oder Schmetterlinge. Wir interessieren uns zum Beispiel dafür, ob wir aus einer Wolkenstimmung auf baldigen Regen schließen können, ob es einen Zusammenhang zwischen Börsencrashs und Herzinfarkten gibt, oder ob sich die berühmten "Schmetterlinge im Bauch" auf den Prüfungserfolg auswirken.

Um solch breit gefassten Fragestellungen empirisch untersuchen zu können benötigen wir ein abstraktes Modell, d.h. auch ein mathematisches Instrumentarium, welches es uns gestattet, Phänomene wie oben geschildert mathematisch zu beschreiben.

Grob vereinfacht können wir uns die Welt, und damit auch die 'Wirtschaft', als einen riesigen datengenerierenden Prozess vorstellen, die laufend Ergebnisse wie Aktienkurse und Herzinfarkte produziert. Wir wollen hier Methoden entwickeln, die uns später helfen sollen einige Teilaspekte dieses äußerst komplexen Gebildes zu analysieren. Dabei geht es vor allem darum, wie wir aus den durch vielen Zufallsstörungen überlagerten Beobachtungen auf tiefer liegende Gesetzmäßigkeiten schließen können, die dem datengenerierenden Prozess zugrunde liegen, und wie wir diese Gesetzmäßigkeiten aus den beobachteten Daten schätzen können.

Dazu benötigen wir mehr als eine vage Intuition, wir benötigen ein paar theoretische Grundlagen. Deshalb werden wir diesem Kapitel ein grundlegendes statistische Denkmodell entwickeln, das allem Folgenden zugrunde liegt.

Das erste Problem besteht darin, dass uns die Natur ihre Ergebnisse nicht unmittelbar als fix und fertige Zahlen liefert, sondern z.B. in Form von Wolken oder Schmetterlingen. Um diese in ein mathematisches Gerüst zu bringen benötigen wir ein sehr allgemeines Konzept, nämlich Mengen. Mit Mengen kann man zwar fast beliebige Ergebnisse beschreiben, aber sie haben einen entscheidenden Nachteil, der Umgang mit ihnen ist umständlich, man kann nicht einfach mit ihnen 'rechnen'. Das Konzept der Zufallsvariablen wird es uns ermöglichen, ganz allgemeine Zufallsereignisse in die Zahlenmenge abzubilden. Der allgemein Beweis, dass dies generell möglich ist, wurde von Stochastikern wie z.B. Andrey Nikolaevich Kolmogorov (1903

– 1987) in den dreißiger Jahren des letzten Jahrhunderts erbracht. Tatsächlich sind Zufallsvariablen ziemlich komplexe mathematische Gebilde, was uns hier aber nicht weiter zu kümmern braucht, der Umgang mit ihnen ist denkbar einfach.

Wir werden im folgenden Abschnitt zuerst das Konzept der Zufallsvariablen ein bisschen ausführlicher erläutern, uns dann mit deren Verteilungen und Momenten (z.B. Erwartungswerte und Varianzen) beschäftigen, die wir später für Hypothesentests benötigen werden.

Zu Ihrer Beruhigung, wir werden auch in diesem Kapitel nicht wirklich in die Tiefe gehen, sondern der Intuition wieder den Vorrang gegenüber mathematischer Strenge einräumen. Manche Konzepte werden trotzdem zumindest anfänglich etwas abstrakt anmuten, aber diese Abstraktion hat einen hohen Ertrag, sie erlaubt es uns ein generelles Modell zu entwickeln, auf dessen Grundlage wir spätere Anwendungen aufbauen können.

3.1 Zufallsexperimente und deren Ergebnisse

Der logische Ausgangspunkt für die folgenden Überlegungen liefert das Gedankenmodell eines Zufallsexperiments. Ein Zufallsexperiment (*'random experiment'*) in unserem Sinne ist ein spezieller 'Datenerzeugender Prozess' (DGP), der die folgenden drei Bedingungen erfüllt:

Zufallsexperiment:

1. alle möglichen Versuchsausgänge, d.h. die Menge aller möglichen *Elementarereignisse* (Ergebnisse) des Experiments sind a priori bekannt;
2. das Ergebnis einer einzelnen Durchführung des Experiments kann nicht mit Sicherheit vorhergesagt werden, aber es gibt eine bestimmte Regelmäßigkeit bei wiederholten Durchführungen; und
3. das Experiment kann unter identischen Bedingungen beliebig oft wiederholt werden.

Klassische Zufallsexperimente sind zum Beispiel das Werfen einer Münze, das Ziehen einer Karte aus einem Stapel, Roulette oder Black Jack. Man beachte, dass es sich dabei nicht um ein Experiment im üblichen Sinne handeln muss, wir denken dabei bloss an ein Phänomen, dessen einzelne Ausgänge im Einzelfall nicht mit Sicherheit vorhergesagt werden können, obwohl *bei wiederholten Ausführungen* ein beschreibbares Muster erkennbar ist. Beim wiederholten Werfen einer fairen Münze erwarten wir z.B., dass Wappen und Zahl etwa gleich häufig auftreten. Da die Resultate von Zufallsexperimenten häufig keine Zahlen sind betrachten wir die einzelnen möglichen Ausgänge ganz allgemein als Elemente einer Menge.

Die Menge aller möglichen Ausgänge eines Zufallsexperiments wird *Ergebnismenge* oder Menge aller möglichen *Elementarereignisse* (*'outcomes set'*) genannt, und wird

häufig mit dem Symbol Ω bezeichnet. Beispiele für Elementarereignisse sind das Geschlecht der nächsten Person, die zur Tür hereinkommt, welche Partei die nächste Wahl gewinnt, die Inflationsrate im nächsten Monat, kurzum, alle Ereignisse, die als Ausgänge eines Zufallsexperimentes interpretiert werden können. Für das Werfen einer Münze besteht $\Omega = \{\text{Wappen}, \text{Zahl}\}$ aus den Elementarereignissen $\{\text{Wappen}\}$ und $\{\text{Zahl}\}$.

Wenn wir eine Karte aus einem gemischten Stapel ziehen und uns für die Farbe der Karte interessieren ist $\Omega = \{\heartsuit, \clubsuit, \diamondsuit, \spadesuit\}$, und $\heartsuit \in \Omega$ bedeutet $\heartsuit$ ist ein Element von Ω.

Die Anzahl der möglichen Ergebnisse eines Zufallsexperiments kann eine endlich große Zahl sein, wie in den oben aufgezählten Beispielen, aber die Anzahl der Elemente von Ω kann auch unendlich groß sein. In diesem Fall kann man weiter unterscheiden, ob Ω *abzählbar* oder *überabzählbar* viele Ergebnisse enthält.

Im Fall einer unendlich großen, aber abzählbaren Menge von Ergebnissen kann jedem Elementarereignis eine natürliche Zahl $\mathbb{N}$ zugeordnet werden; ein Beispiel wäre die Anzahl der Würfe die benötigt wird, bis die erste Sechs gewürfelt wird.

In den späteren Anwendungen werden wir uns hauptsächlich für Zufallsexperimente interessieren, deren Menge von Elementarereignissen Ω eine überabzählbare Anzahl von Elementen enthält, zum Beispiel das Einkommen einer zufällig ausgewählten Person, welches jeden beliebigen Wert innerhalb eines Intervalls annehmen kann. Für die Abbildung solcher Mengen wird in der Regel die Menge der reellen Zahlen $\mathbb{R}$ (bzw. ein Intervall daraus) benötigt.

Wir werden uns in diesem Kapitel hauptsächlich mit endlichen Ergebnismengen beschäftigen, ganz einfach weil dies einfacher ist. Für überabzählbar große Ergebnismengen wird ein mathematisches Instrumentarium benötigt, welches wir hier nicht voraussetzen wollen. Mathematiker haben aber gezeigt, dass die Intuition für Zufallsexperimente mit einer endlichen Anzahl von möglichen Versuchsausgängen zum größten Teil auch für Zufallsexperimente gilt, deren Ergebnismenge Ω eine überabzählbare Anzahl von Elementen enthält.

Häufig sind wir nicht an einem einzelnen Elementarereignis interessiert, sondern an *"interessierenden Ereignissen"*, zum Beispiel könnten wir uns beim Roulette für die Menge der geraden Zahlen größer 15 interessieren, oder beim Pokern dafür, ein 'Full House' zu ziehen.

Ereignisse (*'events'*) setzen sich aus einem oder mehreren Elementarereignissen zusammen. Formal wird ein Ereignis A als eine Teilmenge der Ergebnismenge Ω definiert, d.h. $A \subset \Omega$.[1] Beispielsweise setzt sich beim Würfeln das Ereignis "Werfen einer geraden Augenzahl" $A = \{2, 4, 6\}$ aus den Elementarereignissen $\{2\}$, $\{4\}$ und $\{6\}$ zusammen.

Wir sagen ein Ereignis A tritt ein, wenn bei der Durchführung des Zufallsexperiments genau eines der in A enthaltenen Elementarereignisse eintritt. Zum Beispiel tritt das Ereignis "Werfen einer geraden Augenzahl" genau dann ein, wenn eines der Elementarereignisse $\{2\}$, $\{4\}$ oder $\{6\}$ gewürfelt wird.

[1] $A \subset \Omega$ wenn jedes Element von A auch ein Element von Ω ist, bzw. etwas abstrakter $A \subset \Omega$ wenn für jedes $a \in A$ impliziert $a \in \Omega$.

Kernbegriffe:

1. **Zufallsexperiment:** ein unter identischen Bedingungen beliebig oft wiederholbarer Vorgang, der nach einer genau definierten Vorschrift ausgeführt wird, und dessen Ergebnis nicht eindeutig im Voraus bestimmt werden kann (d.h. vom "Zufall" bestimmt wird), z.B. das Werfen einer Münze, die Ziehung einer Lottozahl oder einer Zufallsstichprobe, etc.

2. **Ergebnisraum eines Zufallsexperiments (bzw. Menge der Elementarereignisse):** Menge aller möglichen Ausgänge eines Zufallsexperiments. Die Menge aller Elementarereignisse wird in der Literatur häufig mit $\mathbf{\Omega}$ bezeichnet. So ist z.B. für das Werfen eines Würfels $\mathbf{\Omega} = \{1, 2, 3, 4, 5, 6\}$, und die einzelnen Elementarereignisse sind $\{1\}$, $\{2\}$, $\{3\}$, ..., $\{6\}$.

3. **Ereignis**: eine beliebige Teilmenge des Ergebnisraums. Ein Ereignis setzt sich aus einem oder mehreren Elementarereignissen zusammen. Beispielsweise setzt sich beim Würfeln das Ereignis "Werfen einer geraden Augenzahl" $A = \{2, 4, 6\}$ aus den Elementarereignissen $\{2\}$, $\{4\}$ und $\{6\}$ zusammen. Wir sagen Ereignis A ist eingetreten, wenn ein Element von A realisiert wurde.

Nun gehen wir einen Schritt weiter und betrachten zwei Ereignisse, z.B. Ereignis A das Würfeln einer geraden Augenzahl und Ereignis B, das Würfeln Augenzahl > 3.

Wenn wir zwei beliebige Ereignisse A und B betrachten können wir die Vereinigung A und B $(A \cup B)$ oder den Durchschnitt $A \cap B$ definieren.

Die *Vereinigung zweier Ereignisse* A und B $(A \cup B)$ ist die Menge aller Elementarereignisse, die zu A *oder* B gehören (vgl. Abbildung 3.1).

Der *Durchschnitt zweier Ereignisse* A und B $(A \cap B)$ ist die Menge aller Elementarereignisse, die zu A *und* B gehören (d.h. wenn A und B gemeinsam eintreten; vgl. Abbildung 3.2).

Ein unmögliches Ereignis wird durch die leere Menge $\emptyset$ dargestellt. Zwei Ereignisse schließen sich gegenseitig aus, wenn $A \cap B = \emptyset$.

Die *komplementäre Menge* zu A relativ zu einer Universalmenge Ω sind alle Elemente von Ω, die *nicht* in A enthalten sind (vgl. Abbildung 3.3).

Man beachte, dass A und die Komplementärmenge $\overline{A}$ eine *Partition* bilden, d.h. sie decken den Ergebnisraum Ω vollständig ab und sind disjunkt (die Schnittmenge ist die leere Menge): $A \cup \overline{A} = \Omega$, $A \cap \overline{A} = \emptyset$.

Beispiel: Wenn beim Würfeln $A = \{\text{Werfen einer geraden Augenzahl}\} = \{2, 4, 6\}$ und $B = \{\text{Werfen einer Augenzahl} \leq 3\} = \{1, 2, 3\}$, dann ist $A \cap B = \{2\}$, $A \cup B = \{1, 2, 3, 4, 6\}$, $\overline{A} = \{1, 3, 5\}$ und $\overline{B} = \{4, 5, 6\}$, $\overline{A} \cap \overline{B} = \{5\}$, $\overline{A} \cup \overline{B} = \{1, 3, 4, 5, 6\}$.

Mit Hilfe der Definition eines Ereignisses und der Mengenoperationen ist es möglich

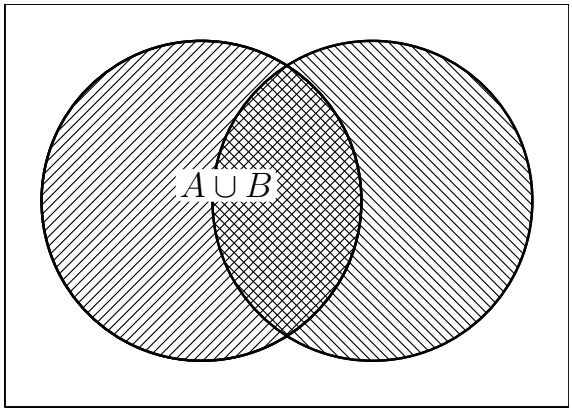


Abbildung 3.1: Vereinigung zweier Ereignisse (ODER - Ereignis)
$A \cup B := \{x : x \in A \text{ oder } x \in B\}$

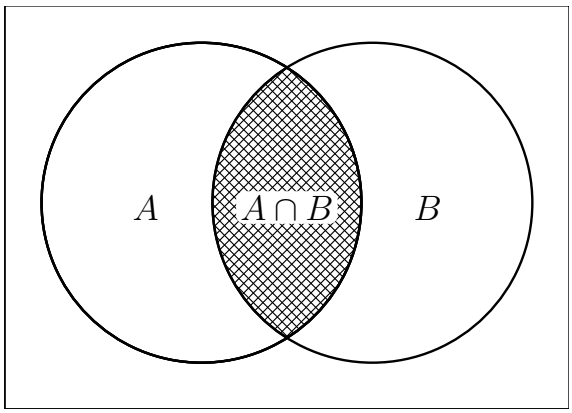


Abbildung 3.2: Durchschnitt zweier Ereignisse (UND - Ereignis)
$A \cap B := \{x : x \in A \text{ und } x \in B\}$

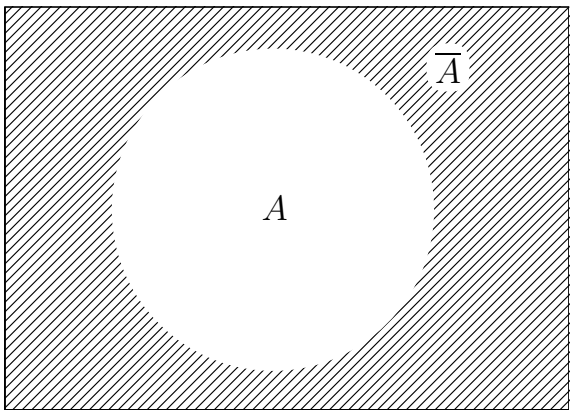


Abbildung 3.3: Komplementäre Menge, $\overline{A} := \{x : x \in \Omega \text{ und } x \notin A\}$

Zusammenfassung: zufällige Ereignisse können als Teilmengen der Menge aller Elementarereignisse Ω dargestellt werden.
Für $A \in \Omega$ und $B \in \Omega$ gilt

1. Schnittmenge (UND-Ereignis):

$$A \cap B := \{x \colon x \in A \text{ und } x \in B\}$$

2. Vereinigungsmenge (ODER-Ereignis):

$$A \cup B := \{x \colon x \in A \text{ oder } x \in B\}$$

3. Komplementäres Ereignis:

$$\overline{A} := \{x \colon x \in \Omega \text{ und } x \notin A\}$$

   entspricht der logischen Negation.

einen Ereignisraum[2] (*event space*, *sample space*) zu definieren. Ein **Ereignisraum** $\mathcal{A}$ enthält alle interessierenden Ereignisse und hat darüber hinaus eine mathematische Struktur. Wenn uns z.B. die Ereignisse A und B interessieren, enthält $\mathcal{A}$ zusätzlich zu den Ereignissen A und B die leere Menge $\emptyset$, die Ergebnismenge Ω sowie alle weiteren mit diesen Mengen über Mengenoperationen verknüpfte Mengen, wie z.B. $\overline{A}, \overline{B}, A \cup B, A \cap B$ etc. Dies ist aus mathematischen Gründen erforderlich, da dies später die Definition von Zufallsvariablen erlaubt, ist aber für das Folgende von geringer Bedeutung. In der Sprache der Mathematik bildet $\mathcal{A}$ eine sogenannte σ-Algebra, ein System von Mengen mit einer speziellen mathematischen Struktur. Den Elementen von $\mathcal{A}$ können Wahrscheinlichkeiten zugeordnet werden.

Wichtige Regeln für das Kombinieren von Ereignissen $A, B, C \in \Omega$ sind

1. *Kommutativgesetze:*

$$\begin{aligned} A \cap B &= B \cap A \\ A \cup B &= B \cup A \end{aligned}$$

2. *Assoziativgesetze:*

$$\begin{aligned} (A \cap B) \cap C &= A \cap (B \cap C) \\ (A \cup B) \cup C &= A \cup (B \cup C) \end{aligned}$$

[2]Unter einem Raum versteht man in der Mathematik ganz allgemein eine Menge mathematischer Objekte mit einer zusätzlichen mathematischen Struktur.

3. *Distributivgesetze:*[3]

$$\begin{aligned} A \cap (B \cup C) &= (A \cap B) \cup (A \cap C) \\ A \cup (B \cap C) &= (A \cup B) \cap (A \cup C) \end{aligned}$$

Im Weiteren wird es nun darum gehen, wie Ereignissen $A \in \Omega$ Wahrscheinlichkeiten $P(A)$ zugeordnet werden können.

3.2 Wahrscheinlichkeit

Was bedeutet eine Aussage wie z.B. *"die Wahrscheinlichkeit, dass es morgen regnet, beträgt 40 Prozent"*? Offensichtlich gibt es heute keine Möglichkeit mit Sicherheit vorauszusagen, ob es morgen regnen wird oder nicht, und übermorgen können wir mit Sicherheit sagen, ob es dann geregnet haben wird oder nicht.

Tatsächlich ist die Frage, was man unter *Wahrscheinlichkeit* verstehen soll, keineswegs so trivial wie man meinen möchte; einige der größten Philosophen und Mathematiker haben sich vergebens bemüht, eine einfache Beschreibung des 'Wesens' von Wahrscheinlichkeit zu finden.

3.2.1 Laplacesche Wahrscheinlichkeitsdefinition

Eine der ältesten Definitionen von Wahrscheinlichkeit geht auf den Mathematiker Pierre-Simon Marquis de Laplace (1749-1827) zurück und wird manchmal auch Lotterie-Definition oder 'naive Wahrscheinlichkeitsdefinition' genannt

$$P(A) = \frac{\text{Anzahl der günstigen Fälle}}{\text{Anzahl aller gleichmöglichen Fälle}}$$

wobei zwei Ereignisse als gleichmöglich bezeichnet werden, wenn man das Eintreten aller Ereignisse für 'gleich wahrscheinlich' hält. So ist z.B. beim Würfeln die Wahrscheinlichkeit für das Ereignis A "Werfen einer geraden Augenzahl"

$$P(A) = \frac{3}{6} = 0.5$$

Diese Wahrscheinlichkeitsdefinition ist allerdings nur für Zufallsexperimente mit *gleichwahrscheinlichen* Elementarereignissen anwendbar. Wenn Sie sich z.B. fragen, mit welcher Wahrscheinlichkeit Sie die nächste Prüfung bestehen, so gibt es einen günstigen Fall, Sie bestehen die Prüfung, und zwei mögliche Fälle, Sie bestehen die Prüfung oder Sie bestehen sie nicht. Daraus den Schluss zu ziehen, dass Sie die nächste Prüfung mit 50 Prozent Wahrscheinlichkeit bestehen werden, könnte sich als gefährlich erweisen. Außerdem wäre nach dieser Logik die Wahrscheinlichkeit, die nächste Prüfung mit einem 'sehr gut' zu bestehen, ebenfalls 50 Prozent, was offensichtlich unsinnig ist. Trotzdem leistet diese naive Wahrscheinlichkeitsdefinition für einfache Beispiele mit gleichwahrscheinlichen Ereignissen manchmal nützliche

[3] In der Schulmathematik entspricht dies dem Ausklammern (bzw. Herausheben) und Ausmultiplizieren.

Dienste, z.B. wenn es um einfache Stichprobenziehungen geht. Für allgemeinere Anwendungen ist sie allerdings ungeeignet, dafür benötigen wir die weiter unten diskutierte axiomatische Definition von Wahrscheinlichkeit.

Heute dominieren zwei Auffassungen von Wahrscheinlichkeit die Szene, die frequentistische Auffassung und subjektive Auffassung, deren bekannteste die Bayessche Sicht ist.

3.2.2 Frequentistische Wahrscheinlichkeitsdefinition

Wenn ein Zufallsexperiment unter identischen Bedingungen beliebig oft wiederholt werden kann und wir die relative Häufigkeit eines Ereignisses A nach n Durchführungen des Experiments mit n_A/n bezeichnen, dann versteht man unter der frequentistischen Definition den Grenzwert dieser relativen Häufigkeit, wenn die Anzahl der Experimente gegen Unendlich geht

$$P(A) = \lim_{n\to\infty} \left(\frac{n_A}{n}\right)$$

Dieser Wahrscheinlichkeitsbegriff ist in der Statistik immer noch am gebräuchlichsten und liegt auch diesem Skript zugrunde.

Die Aussage, dass es morgen mit 40% Wahrscheinlichkeit regnet, würde ein Vertreter der frequentistischen Sichtweise folgendermaßen interpretieren: wenn wir den morgigen Tag unter identischen Bedingungen unendlich oft wiederholen würden, dann könnten wir damit rechnen, dass es an 40% dieser Tage regnet (*"und ewig grüßt das Murmeltier ..."*). Natürlich geht niemand davon aus, dass dies wirklich möglich ist, häufig steht implizit die Vorstellung dahinter, dass es in der Vergangenheit an 40% der Tage mit ähnlichen Bedingungen geregnet hat. Hier wird die konzeptionelle Ähnlichkeit des frequentistischen Wahrscheinlichkeitsbegriffs mit relativen Häufigkeiten deutlich. Dieser Zugang versucht eine möglichst objektive Beschreibung von stochastischen Phänomenen zu finden.

Allerdings ist auch diese Auffassung von Wahrscheinlichkeit mit Problemen behaftet, z.B. ist es praktisch unmöglich unendlich viele Wiederholungen eines Zufallsexperiments durchzuführen, und gerade in den Wirtschafts- und Sozialwissenschaften ist häufig nicht einmal eine einzige exakte Wiederholung möglich (Geschichte wiederholt sich nicht). Außerdem gibt es keine Garantie, dass die Wahrscheinlichkeiten tatsächlich konvergieren. Relative Häufigkeiten sind deshalb bestenfalls als eine Annäherung an die gesuchte Wahrscheinlichkeit.

Bekannte Vertreter dieser Interpretation von Wahrscheinlichkeit sind Richard von Mises (der Bruder des Ökonomen), R.A. Fisher und Jerzy Neyman.

3.2.3 Subjektive Wahrscheinlichkeitsdefinitionen

Bereits im Beispiel mit der Regenwahrscheinlichkeit bereitet die frequentistische Wahrscheinlichkeitsdefinition Schwierigkeiten, und in vielen angewandten Fällen

Tabelle 3.1: Klassische versus Bayessche Statistik

	Classical	*Bayesian*
Goal:	To be objective	To express also subjective biases and intuition
For:	Making statements in a society	Making the best decision for oneself
Analogous to:	Rules of evidence	Self-help tool
To be used when you try:	To make a point	To make a decision

Quelle: Gilboa, Itzhak, *Making Better Decisions*

(z.B. Geschichte, Politik, Wirtschaft) ist die Vorstellung einer beliebigen Wiederholbarkeit realitätsfremd.

Vertreter von subjektiven Wahrscheinlichkeitsdefinitionen (z.B. Leonard J. Savage) bezweifeln die Existenz einer *objektiven* Wahrscheinlichkeit, ihrer Auffassung nach handelt es sich letztendlich um 'vernünftige Glaubensaussagen' (*"probability is viewed as representing a degree of reasonable belief with the limiting values of zero being complete disbelief or disproof and of one being complete belief or proof."* Zellner 1984, 6).

Quantifizieren lassen sich subjektive Wahrscheinlichkeiten z.B., indem man beobachtet, welche *Wettchancen* jemand einem Ereignis einräumen würde.

Die bedeutendste Richtung innerhalb subjektiver Wahrscheinlichkeitsauffassungen ist die *Bayessche Wahrscheinlichkeitstheorie*. Der bayessche Wahrscheinlichkeitsbegriff setzt keine unendlich oft wiederholbaren Zufallsexperimente voraus, anstelle von 'wahren Parametern der Grundgesamtheit' tritt der 'Grad vernünftiger Glaubwürdigkeit', der jeweils mit den verfügbaren Informationen 'upgedated' wird.

Subjektive Wahrscheinlichkeiten werden häufig für Entscheidungsmodelle unter Unsicherheit verwendet. Der Entscheidungstheoretiker Itzhak Gilboa vergleicht den *frequentistischen Ansatz* mit einem Gerichtsverfahren, bei dem versucht wird zu einem möglichst objektiven und nachvollziehbaren Ergebnis zu kommen, während er die Relevanz des *bayesschen Ansatz* eher bei der Modellierung individuellen Entscheidungsverhaltens sieht.

3.2.4 Axiomatische Wahrscheinlichkeitsdefinition

Für unsere Zwecke benötigen wir allerdings keine inhaltliche Interpretation von Wahrscheinlichkeit, für eine rein mathematische Behandlung reicht die *Axiomatische Wahrscheinlichkeitsdefinition* aus, die wesentlich auf *A.N. Kolmogorov* (1903 – 1987) zurückgeht. Dabei wird nicht versucht das 'Wesen' von Wahrscheinlichkeit zu ergründen, sondern es werden lediglich die erforderlichen mathematische Eigenschaften definiert.

Im Folgenden verstehen wir unter Wahrscheinlichkeit ganz allgemein ein Maß zur Quantifizierung der Sicherheit bzw. Unsicherheit eines Zufallsexperiments. Konkret

geht es darum, den Elementen der Ereignismenge $\mathcal{A}$ die dazugehörigen Wahrscheinlichkeiten zuzuordnen.

Die **Axiomatische Wahrscheinlichkeitsdefinition** umfasst die folgenden **drei Axiome**:

1. $P(\Omega) = 1$

   Da die Ergebnismenge Ω alle Elementarereignisse eines Zufallsexperiments enthält ist Ω ein sicheres Ereignis (d.h. eine Ereignis aus Ω wird mit Wahrscheinlichkeit 1 eintreten);

2. $P(A) \geq 0$ für alle Ereignisse $A \in \mathcal{A}$

   Die Wahrscheinlichkeit $P(A)$ des Ereignisses A ist eine reelle, nichtnegative Zahl; gemeinsam mit 1. folgt $0 \leq P(A) \leq 1$.

3. Seien A und B sich *gegenseitig ausschließende* Ereignisse, dann gilt für die Vereinigungsmenge $A \cup B$

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

   Dies gilt allgemeiner auch für beliebig viele Ereignisse.

   Wenn $\{A\}_{j=1}^{\infty}$ eine Folge sich *gegenseitig ausschließender* Ereignisse in $\mathcal{A}$ ist, dann gilt für die Vereinigung $A = \bigcup_{j=1}^{\infty} A_j$

   $P(A) = \sum_{j=1}^{\infty} P(A_j)$.

   Für eine endliche Menge mit J sich wechselseitig ausschließenden Ereignissen $A_1, A_2, \ldots A_J$ bedeutet dies, dass die Wahrscheinlichkeit dafür, dass eines dieser Ereignisse eintritt (A_1 oder A_2 oder $\ldots A_J$) gleich der Summe der Einzelwahrscheinlichkeiten ist: $P(A_1) + P(A_2) + \cdots + P(A_J)$.

Kurzer Exkurs: "*Was sind Axiome?* Axiome bilden das Rückgrat jeder mathematischen Disziplin. Axiome beinhalten Aussagen, die nicht begründet oder bewiesen werden. Ausgehend von einem Axiomensystem werden dann aber alle weiteren Aussagen bewiesen. Sinnvolle und konsistente Axiomensysteme zu postulieren gehört zu den schwierigsten Aufgaben in der Mathematik. Von den Anfangen der (modernen) Wahrscheinlichkeitsrechnung im 17. Jahrhundert (Briefwechsel von Blaise Pascal und Pierre de Fermat im Jahr 1654) dauerte es fast 300 Jahre, bis Kolmogorov im Jahr 1933 die axiomatischen Grundlagen der Wahrscheinlichkeitstheorie begründete." (Achim Zeileis)

Aus diesen drei Axiomen kann der *Additionssatz* hergeleitet werden:

Additionssatz:
Wenn sich zwei Ereignisse A und B *nicht* ausschließen gilt

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

Die Logik dieses Satzes erschließt sich unmittelbar aus Abbildung 3.1, um eine 'Doppelzählung' zu vermeiden muss die Durchschnittsfläche einmal abgezogen werden. Das lässt sich natürlich auch formal zeigen.

*Beweis:** Dies folgt unmittelbar aus den Axiomen. Dazu beachten wir, dass sich A und die Komplementärmenge $\overline{A}$ gegenseitig ausschließen. Deshalb kann das Ereignis $A \cup B$ auch geschrieben werden als

$$A \cup B = A \cup (\overline{A} \cap B))$$

(vgl. Abbildung 3.1), und da sich die Ereignisse ausschließen sind die Wahrscheinlichkeiten

$$P(A \cup B) = P(A) + P(\overline{A} \cap B) \tag{3.1}$$

Ebenso kann B als Vereinigungsmenge zweier sich gegenseitig ausschließender Ereignisse angeschrieben werden

$$B = (A \cap B) \cup (\overline{A} \cap B)$$

Weil sich die Ereignisse ausschließen sind die Wahrscheinlichkeiten

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

Wenn wir dies umschreiben zu $P(\overline{A} \cap B) = P(B) - P(A \cap B)$ und in Gleichung (3.1) einsetzen erhalten wir das gewünschte Ergebnis $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. ■

Man beachte, dass $P(\cdot)$ sogenannte Mengenfunktionen sind, die Elementen der Ereignismenge $\mathcal{A}$ reelle Zahlen zwischen Null und Eins zuordnen; $P(\cdot) : \mathcal{A} \mapsto [0, 1]$. Dies sind keine üblichen Funktionen $f : \mathbb{R} \mapsto \mathbb{R}$ die wir aus der Schule kennen und mit denen man 'rechnen' kann! Wir werden gleich sehen, dass erst Zufallsvariablen dieses Problem lösen werden, erst diese gestatten die Definition von Wahrscheinlichkeits- und Dichtefunktionen $\Pr : \mathbb{R} \mapsto [0, 1]$, die uns die Anwendung des üblichen mathematischen Instrumentariums ermöglichen, d.h., das 'Rechnen mit dem Zufall'. Aber vorher müssen wir noch ein wichtiges Konzept vorstellen, welches für alles Folgende von zentraler Bedeutung ist.

3.2.5 Bedingte Wahrscheinlichkeiten

In der Regel beobachten wir mehrere Ereignisse, und häufig sind wir daran interessiert, aus einem bereits eingetretenen Ereignis Schlussfolgerungen auf ein erst in der

Zukunft eintretendes Ereignis zu ziehen. Wir wollen aus beobachteten Fakten etwas über unbeobachtbare Ereignisse lernen.

Dies ist nur möglich, wenn die Wahrscheinlichkeit des Eintretens eines Ereignisses A vom Eintritt eines anderen Ereignisses B abhängt. Die Wahrscheinlichkeit für das Eintreten von A unter der Bedingung, dass Ereignis B vorher eingetreten ist oder gleichzeitig eintritt, wird *bedingte Wahrscheinlichkeit* $P(A|B)$ genannt. Sie ist für $P(B) > 0$ definiert als

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

Die Logik wird unmittelbar aus Abbildung 3.2 ersichtlich: wenn $A \cap B \neq \emptyset$ (wobei $\emptyset$ die leere Menge bezeichnet) erlaubt uns das Wissen, dass Ereignis B bereits eingetreten ist, eine genauere Einschätzung der Eintrittswahrscheinlichkeit von A.

Beispiel: Betrachten wir einen fairen Würfel und die Ereignisse
$A = \{1, 2, 3\}$ (würfeln einer Zahl kleiner 4), und
$B = \{2, 4, 6\}$ (würfeln einer geraden Zahl).

Angenommen es wurde einmal gewürfelt und wir wissen nur, dass eine gerade Zahl gewürfelt wurde (also B eingetreten ist), wie groß ist dann die Wahrscheinlichkeit, dass diese Zahl kleiner als 4 ist?

Da $A \cap B = \{2\}$ ist $P(A \cap B) = 1/6$; $P(B) = 3/6$, deshalb ist $P(A|B) = (1/6)/(3/6) = 1/3$.

Aus der Definition der bedingten Wahrscheinlichkeit folgt der *Multiplikationssatz*.

Multiplikationssatz:
Für zwei beliebige Ereignisse A und B gilt

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

Mit Hilfe des Multiplikationssatzes kann die Wahrscheinlichkeit für das *gemeinsame* Eintreten von A *und* B (d.h. $P(A \cap B)$) berechnet werden.

Beispiel: Aus einer Urne mit insgesamt zwei weißen und zwei schwarzen Kugeln werden zwei Kugeln *ohne Zurücklegen* gezogen.

Die beiden interessierenden Ereignisse seien

$$\begin{aligned} A &= \{\text{die erste Kugel ist weiß}\} \\ B &= \{\text{die zweite Kugel ist weiß}\} \end{aligned}$$

Wie groß ist die Wahrscheinlichkeit, zwei weiße Kugeln zu ziehen?

Für die Ziehung der ersten Kugel ist $P(A) = \frac{1}{2}$.

Wenn aber die erste Kugel schon weiß war und ohne Zurücklegen gezogen wurde befinden sich in der Urne nur mehr eine weiße und zwei schwarze Kugeln, die Wahrscheinlichkeit in der zweiten Ziehung eine weiße Kugel zu ziehen beträgt also nur mehr 1/3, d.h. $P(B|A) = 1/3$.

Die *gemeinsame* Wahrscheinlichkeit ist also $P(A \cap B) = P(A) \cdot P(B|A) = \frac{1}{2} \times \frac{1}{3} = \frac{1}{6}$.

In diesem Beispiel hängt die Wahrscheinlichkeit bei der zweiten Ziehung davon ab, was bei der ersten Ziehung gezogen wurde (Frage: wie groß wäre die Wahrscheinlichkeit bei der zweiten Ziehung eine weiße Kugel zu ziehen, wenn bei der ersten Ziehung eine schwarze Kugel gezogen worden wäre?).

Aber nicht immer hängt die Wahrscheinlichkeit des Eintretens eines Ereignisses vom vorhergehenden Eintritt eines anderen Ereignisses ab. Wenn wir *mit Zurücklegen* gezogen hätten, würde die Wahrscheinlichkeit der zweiten Ziehung nicht vom Resultat der ersten Ziehung abhängen. Dies führt uns zur Definition *stochastischer Unabhängigkeit*.

3.2.6 Stochastische Unabhängigkeit

Ein wichtiger Spezialfall und ein zentrales Konzept der Statistik ist die *stochastische Unabhängigkeit*:

> Zwei Ereignisse A und B mit $P(A), P(B) > 0$ heißen *stochastisch unabhängig*, wenn die Wahrscheinlichkeit des Eintretens von Ereignis A nicht vom Eintreten oder Nichteintreten des Ereignisses B abhängt, d.h. wenn $P(A|B) = P(A)$.

Stochastische Unabhängigkeit:
Nur wenn Ereignisse A und B *stochastisch unabhängig* sind gilt

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

bzw.

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

da bei stochastischer Unabhängigkeit

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

In Worten, falls zwei Ereignisse stochastisch unabhängig sind ist die bedingte Wahrscheinlichkeit gleich der unbedingten Wahrscheinlichkeit. Die Kenntnis, dass B bereits eingetreten ist, erlaubt bei stochastischer Unabhängigkeit keine genaueren Aussagen über die Eintrittswahrscheinlichkeit von A.

Beispiel: Wir kehren zum vorhergehenden Beispiel mit der Urne zurück, die zwei weiße und zwei schwarzen Kugeln enthält, aber diesmal ziehen wir *mit Zurücklegen*.

Die beiden interessierenden Ereignisse seien wieder A: 'die erste Kugel ist weiß', und B: 'die zweite Kugel ist weiß'.

Wenn die Kugeln aber *mit Zurücklegen* gezogen werden ist $P(A) = \frac{1}{2}$ und $P(B) = \frac{1}{2}$, da sich das Verhältnis der weißen und schwarzen Kugeln in der Urne nicht ändert.

Die Wahrscheinlichkeit für das gemeinsame Ereignis $A \cap B$ ist also

$$P(A \cap B) = P(A) \cdot P(B) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

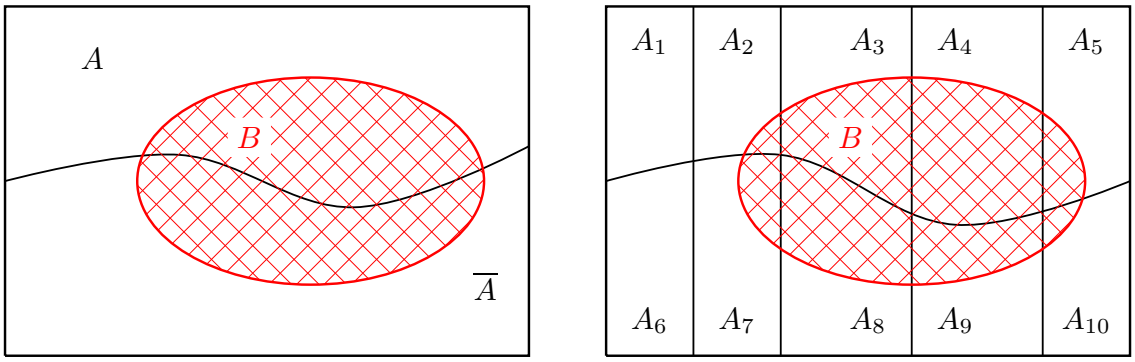


Abbildung 3.4: Theorem der Totalen Wahrscheinlichkeit; sowohl A und die Komplementärmenge $\overline{A}$ (links) als auch $A_1, \ldots, A_{10}$ im rechten Panel bilden eine Partition.

die Ereignisse sind also *unabhängig*.

Beispiel: Wie groß ist die Wahrscheinlichkeit mit einem fairen Würfel zwei Mal hintereinander eine Sechs zu würfeln?

Die beiden Ereignisse sind unabhängig, da das Ergebnis des ersten Wurfs keine Auswirkung auf den zweiten Wurf hat. Wenn $A = \{6 \text{ beim ersten Wurf}\}$ und $B = \{6 \text{ beim zweiten Wurf}\}$ ist

$$P(A \cup B) = P(A) \times P(B) = \frac{1}{6} \times \frac{1}{6} = 0.0278$$

Bei sehr häufiger Wiederholung de Versuchs rechnen wir also in weniger als 3% der Fälle zwei Mal hintereinander eine Sechs zu würfeln.

Dies kann auch für mehr als zwei Würfe verallgemeinert werden, die Wahrscheinlichkeit drei Mal hintereinander eine Sechs zu würfeln ist $P(A \cup B \cup C) = P(A) \times P(B) \times P(C) = \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} = 0.00463$.

3.2.7 Theorem der Totalen Wahrscheinlichkeit

Wir betrachten eine *Partition* des Ergebnisraums Ω, d.h. $A_1, A_2, \ldots, A_n$ sich gegenseitig ausschließende Ereignisse die den Ergebnisraum Ω zur Gänze ausfüllen, das heißt

$A_i \cap A_j = \emptyset$ für $i, j = 1, \ldots, n$ und $i \neq j$; sowie
$A_1 \cup A_2 \cup \cdots \cup A_n = \Omega$

Abbildung 3.4 zeigt im linken Panel, dass A und die Komplementärmenge $\overline{A}$ eine Partition bilden (die geschwungene horizontale Linie), und im rechten Panel eine Partition für 10 Ereignisse $A_1, \ldots, A_{10}$.

Damit kann jedes beliebige Ereignis B als Vereinigung sich gegenseitig ausschließender Ereignisse dargestellt werden, d.h.

$$B = (B \cap A) \cup (B \cap \overline{A})$$

oder allgemeiner (siehe rechtes Panel in Abbildung 3.4)

$$B = (B \cap A_1) \cup (B \cap A_2) \cup \cdots \cup (B \cap A_n)$$

Aufgrund des Additionssatzes für sich gegenseitig ausschließende Ereignisse gilt

$$P(B) = P(B \cap A_1) + P(B \cap A_2) + \cdots + P(B \cap A_n)$$

Aufgrund des *Multiplikatiossatzes* können wir dies mit bedingten Wahrscheinlichkeiten schreiben, und dies liefert uns den

Satz der Totalen Wahrscheinlichkeit:

$$\boxed{P(B) = P(A)P(B|A) + P(\overline{A})P(B|\overline{A})}$$

oder wieder allgemeiner

$$P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + \cdots + P(A_J)P(B|A_J)$$

Dies kann mit Hilfe des Summenzeichens etwas kürzer geschrieben werden

$$P(B) = \sum_{j=1}^{J} P(A_j)P(B|A_j)$$

Wenn nur die bedingten Wahrscheinlichkeiten $P(B|A_j)$ sowie die Wahrscheinlichkeiten des bedingenden Ereignisses $P(A_j)$ bekannt sind kann mit Hilfe des Theorems der Totalen Wahrscheinlichkeit die Gesamtwahrscheinlichkeit von B berechnet werden.

Beispiel: Angenommen Sie lassen sich auf eine bestimmte Krankheit testen. Von diesem Test ist bekannt, dass er bei tatsächlich Erkrankten in 90% der Fälle die Krankheit erkennt. Aber der Test ist nicht perfekt, in 5% der Fälle zeigt er auch bei Gesunden die Krankheit an (*false positive*). Weiters sei bekannt, dass ein Prozent der Bevölkerung an dieser Krankheit leidet.

Wie groß ist die Wahrscheinlichkeit ein positives Testergebnis zu erhalten, bevor Sie den Test gemacht haben?

Lösung: Wir bezeichnen das Ereignis 'eine Person ist krank' mit A, dann ist $P(A) = 0.01$ (ein Prozent der Bevölkerung leidet an dieser Krankheit). Die Gegenwahrscheinlichkeit ist $P(\overline{A}) = 0.99$, und da $A \cup \overline{A} = \emptyset$ und $A \cup \overline{A} = \Omega$ bildet dies eine Partition.

Sei B das Ereignis, dass der Test positiv anzeigt. Wir wissen, dass der Test 90% der tatsächlichen Erkrankungen erkennt, dies ist eine bedingte Wahrscheinlichkeit $P(B|A) = 0.9$ (d.h. bei 10% der tatsächlich Erkrankten zeigt der Test ein negatives Ergebnis).

In 5% der Fälle zeigt der Test auch bei Gesunden ($\overline{A}$) fälschlich die Krankheit an, also $P(B|\overline{A}) = 0.05$.

Damit können wir nun die unbedingte Wahrscheinlichkeit berechnen, dass der Test ein positives Ergebnis anzeigt

$$P(B) = P(A)\,P(B|A) + P(\overline{A})\,P(B|\overline{A}) = 0.01 \times 0.9 + 0.99 \times 0.05 = 0.0585$$

Wir erwarten also, dass dieser Test in 5.85% aller Fälle ein positives Testergebnis anzeigt, obwohl nur 1% der Bevölkerung tatsächlich an der Krankheit leidet.

3.2.8 Theorem von Bayes

Die einfachste Form des Theorems von Bayes folgt unmittelbar aus der Definition der bedingten Wahrscheinlichkeit:

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

für $P(A) > 0$ und $P(B) > 0$.

Nach dem Theorem der Totalen Wahrscheinlichkeit (siehe oben) ist

$$P(B) = P(A)\, P(B|A) + P(\overline{A})\, P(B|\overline{A})$$

Damit erhalten wir das **Theorem von Bayes**

$$\boxed{P(A|B) = \frac{P(B|A)\, P(A)}{P(A)\, P(B|A) + P(\overline{A})\, P(B|\overline{A})}}$$

Beachten Sie, dass links die bedingte Wahrscheinlichkeit $P(A|B)$ steht, und rechts nur die bedingten Wahrscheinlichkeit $P(B|A)$ bzw. $P(B|\overline{A})$ sowie unbedingte Wahrscheinlichkeiten vorkommen.

Dieses Ergebnis wird u.a. bei der Interpretation von Hypothesentests noch eine wichtige Rolle spielen.

*Hinweis:** Um das *Theorem von Bayes* etwas allgemeiner zu zeigen starten wir wieder mit einer *Partition* des Ergebnisraums Ω wie im rechten Panel in Abbildung 3.4, d.h. $A_1, A_2, \ldots, A_J$ sich gegenseitig ausschließende Ereignisse, die den Ergebnisraum Ω zur Gänze ausfüllen.

Aufgrund des Multiplikationssatzes gilt (für $j = 1, \ldots, J$)

$$\begin{aligned} P(A_j \cap B) &= P(B)P(A_j|B) \qquad \text{und symmetrisch} \\ P(A_j \cap B) &= P(A_j)P(B|A_j) \end{aligned}$$

Wenn wir die beiden rechten Seiten gleich setzen erhalten wir

$$\begin{aligned} P(B)P(A_j|B) &= P(B|A_j)P(A_j) \qquad \text{oder} \\ P(A_j|B) &= \frac{P(A_j)P(B|A_j)}{P(B)} \end{aligned}$$

Unter Verwendung des Theorems der Totalen Wahrscheinlichkeit $P(B) = \sum_{j=1}^{J} P(A_j)P(B|A_j)$ erhalten wir eine allgemeinere Form des *Theorems von Bayes*

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{P(B)} = \frac{P(B|A_j)P(A_j)}{\sum_{j=1}^{J} P(A_j)P(B|A_j)}$$

■

Das Theorem von Bayes hat viele Anwendungen, unter anderem im Bereich des *'machine learning'*; so verwenden z.B. viele *spam-filter* dieses Theorem um unerwünschte Emails auszusortieren.

Dieses Theorem zeigt auch, dass unser frequentistisch geprägtes Denken mit bedingten Wahrscheinlichkeiten häufig überfordert ist. Die folgenden beiden Beispiele demonstrieren dies.

Beispiel: Kehren wir nochmals zum vorhergehenden Test zurück. Dieser Test zeigt bei 90% der tatsächlich Erkrankten ein positives Ergebnis, aber bei 5% der Gesunden zeigt er fälschlich die Krankheit an (*false positive*). Es sei bekannt, dass ein Prozent der Bevölkerung an dieser Krankheit leidet.

Angenommen, Sie erhalten ein positives Resultat, wie groß ist dann die Wahrscheinlichkeit, dass Sie tatsächlich an dieser Krankheit erkrankt sind?

Viele glauben, dass diese Wahrscheinlichkeit 90% oder mehr betrage. Dies ist glücklicherweise falsch!

Sei B wieder das Ereignis 'Test zeigt positiv' und A das Ereignis 'Person ist krank'. Aus den Angaben wissen wir

$$\begin{aligned} P(B|A) &= 0.9 \\ P(B|\overline{A}) &= 0.05 \\ P(A) &= 0.01 \end{aligned}$$

Außerdem haben wir vorhin mit Hilfe des Theorems der Totalen Wahrscheinlichkeit berechnet, dass $P(B) = 0.0585$.

Wir wollen aber wissen, wie groß $P(A|B)$ ist, d.h. wie groß die Wahrscheinlichkeit ist tatsächlich erkrankt zu sein, gegeben der Test zeigt ein positives Ergebnis an.

Diese bedingte Wahrscheinlichkeit können wir einfach mit dem Theorem von Bayes berechnen

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \\ &= \frac{0.9 \times 0.01}{0.0585} = 0.15384 \end{aligned}$$

Gegeben Sie erhalten ein positives Testergebnis, dann beträgt die Wahrscheinlichkeit, dass Sie tatsächlich krank sind 15.38%.

Dies ist deutlich weniger als die ursprünglich befürchtete 90% Wahrscheinlichkeit, aber natürlich deutlich höher als die 1% Wahrscheinlichkeit die sie hatten, bevor Sie den Test durchführen ließen.

Wie wir sehen erhält ein Großteil der tatsächlich Erkrankten ein positives Ergebnis, aber ein Großteil der Personen mit einem positiven Ergebnis ist tatsächlich *nicht* krank!

Hinweis: Sie können sich dies auch einfacher vorstellen. Gehen wir von 10 000 Personen aus, davon ist 1%, d.h. 100 Personen, krank. Bei 90% dieser Erkrankten erkennt der Test die Krankheit, also bei 90 Personen. Die restlichen 9 900 Personen sind gesund, aber bei 5% der Gesunden wird dir Krankheit irrtümlich angezeigt (*false positive*), das sind 495 Personen.

Insgesamt erhalten also $495 + 90 = 585$ Personen ein positives Ergebnis, aber nur 90 davon sind tatsächlich krank. Also ist die bedingte Häufigkeit kranker Personen an allen Personen mit einem positiven Testergebnis gleich $90/585 = 0.1538$. Dies ist natürlich das gleiche Ergebnis, das wir mit Hilfe des Theorems von Bayes erhalten haben.

Frage: Was stimmt nicht an der folgenden Aussage: "Die meisten guten Tischtennisspieler sind Chinesen, deshalb ist ein Chinese wahrscheinlich ein guter Tischtennisspieler".

Beispiel: Das Ziegenproblem (engl. *'Monty-Hall problem'*) Angenommen Sie nehmen an einer Quizshow teil und der Showmaster Monty Hall zeigt Ihnen drei geschlossene Türen. Er erklärt Ihnen, dass hinter einer der Türen der Hauptpreis wartet, ein Auto, und hinter den beiden anderen Türen Ziegen. Falls Sie die Tür mit dem Auto erraten gehört Ihnen der Hauptpreis, anderenfalls eine Ziege.

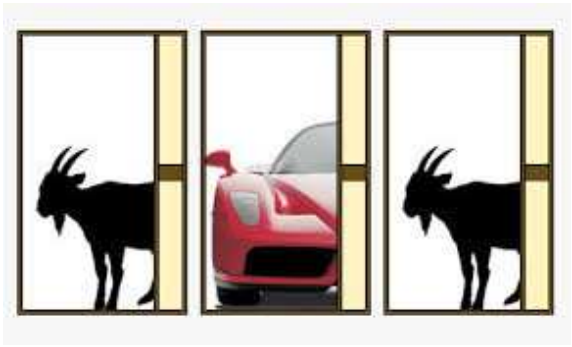


Abbildung 3.5: Das Ziegenproblem

Ihre Chance das Auto zu gewinnen beträgt also 1/3. Nachdem Sie Ihre Entscheidung getroffen haben bittet sie Monty um etwas Geduld. Er öffnet eine Tür, von der er weiß, dass sich dahinter eine Ziege verbirgt, und bietet Ihnen an, Ihre Wahl nochmals zu ändern.

Es stehen noch zwei Türen zur Auswahl, hinter einer verbirgt sich die zweite Ziege, hinter der anderen das Auto. Sollen Sie wechseln?

Auf den ersten Blick scheint es eine 50:50 Chance zu sein, also kein Grund zum wechseln.

Lösung: Tatsächlich können Sie in diesem Quiz durch das Wechseln der Tür Ihre Gewinnchance von 1/3 auf 2/3 erhöhen. Dies scheint auf den ersten Blick kontraintuitiv, und zahlreiche intelligente Köpfe haben sich mit diesem Problem abgequält.

Eine Möglichkeit dieses Problem zu verstehen bietet das Theorem von Bayes. Ohne Einschränkung der Allgemeingültigkeit konzentrieren wir uns auf Tür 1. Sei

$P(A)$ ... die a priori Wahrscheinlichkeit, dass sich hinter Tür 1 das Auto befindet. $P(A) = 1/3$.

$P(\overline{A})$ ... die Wahrscheinlichkeit, dass sich hinter Tür 1 kein Auto

befindet. $P(\overline{A}) = 1 - P(A) = 2/3$.

$P(Z|A)$ ... die Wahrscheinlichkeit, dass Monty eine Tür mit Ziege öffnet, wenn sich das Auto hinter Tür 1 befindet. Da Monty dies immer tut ist $P(Z|A) = 1$.

$P(Z|\overline{A})$ ... die Wahrscheinlichkeit, dass Monty eine Tür mit Ziege öffnet, wenn sich eine Ziege (also kein Auto) hinter der Tür 1 befindet. Natürlich ist auch $P(Z|\overline{A}) = 1$

Wir interessieren uns für $P(A|Z)$, also die Wahrscheinlichkeit das Auto zu erhalten, gegeben dass uns Monty eine Tür mit Ziege zeigt.

Das Theorem von Bayes sagt uns

$$\begin{aligned} P(A|Z) &= \frac{P(Z|A)\,P(A)}{P(Z)} = \frac{P(Z|A)\,P(A)}{P(Z|A)\,P(A) + P(Z|\overline{A})\,P(\overline{A})} \\ &= \frac{1 \times \frac{1}{3}}{1 \times \frac{1}{3} + 1 \times \frac{2}{3}} = \frac{\frac{1}{3}}{1} \\ &= \frac{1}{3} \end{aligned}$$

Wenn wir Türe 1 wählen, also *nicht* wechseln, bleibt die Wahrscheinlichkeit das Auto zu gewinnen 1/3. Die zweite Möglichkeit ist die Türe zu wechseln, und da sich das Auto nur hinter Türe 1 oder der von Monty nicht geöffneten Türe befinden kann, muss die Wahrscheinlichkeit bei Wechseln der Türe 2/3 betragen!

Der Schlüssel zum Verständnis liegt darin, dass Monty *immer* eine Türe mit Ziege öffnet, dadurch erhalten wir zusätzliche Information! □

3.2.9 Wahrscheinlichkeitsraum

Die axiomatische Definition von Wahrscheinlichkeit ermöglicht eine mathematische Beschreibung eines Zufallsexperiments, das Triple

$$[\Omega, \mathcal{A}, P(\cdot)]$$

bildet einen sogenannten *Wahrscheinlichkeitsraum* (*'probability space'*).

Unter einem Wahrscheinlichkeitsraum kann man die mathematische Beschreibung des zugrundeliegenden Zufallsexperiments verstehen. Damit werden zwar die relevanten Aspekte des zugrunde liegenden Zufallsexperiments formal beschrieben, aber wir können immer noch nicht unmittelbar damit 'rechnen', da er nur auf Mengen definiert ist! Erst das Konzept der Zufallsvariablen erlaubt uns die Abbildung relevanter Aspekte von Zufallsexperimenten in die reellen Zahlen.

3.3 Zufallsvariablen

Bisher haben wir Ereignisse untersucht, die Resultate von Zufallsexperimenten waren, und die ziemlich beliebige Mengen sein konnten. Diesen Ereignissen konnten wir zwar mit Hilfe von Mengenfunktionen Wahrscheinlichkeiten zuordnen und damit einige überraschende Resultate zeigen, aber der Umgang mit Mengen ist umständlich und die weiteren Möglichkeiten sind begrenzt.

Hier zeigen sich die Vorteile der axiomatischen Wahrscheinlichkeitsdefinition, sie ermöglicht die Definition von Zufallsvariablen, die uns erlauben diese große Hürde zu überwinden.

Sehr vereinfacht gesprochen sind Zufallsvariablen (*'random variables'*) Funktionen, die allen möglichen Ergebnissen eines Zufallsexperimentes (d.h. den Elementarereignissen oder Ereignissen) reelle Zahlen zuordnet. Diese Zuordnung geschieht derart, dass den Zahlen wieder die korrekten Wahrscheinlichkeiten des zugrunde liegenden Zufallsexperimentes zugeordnet werden können. In einem gewissen Sinne kann man sich also vorstellen, dass Zufallsvariablen eine Abbildung der relevanten Aspekte des dahinter liegenden Zufallsexperiments in die reellen Zahlen sind. Deshalb ermöglichen uns Zufallsvariablen mit den Resultaten von Zufallsexperimenten zu 'rechnen'.

Zufallsvariablen leisten in der Statistik etwas ähnliches wie Nutzenfunktionen in der Mikroökonomik. Auch Nutzenfunktionen können als Abbildung von Mengenkonzepten in die reellen Zahlen verstanden werden, eine auf Güterbündel definierte Präferenzordnung wird in die reellen Zahlen abgebildet, womit das Rechnen mit ihnen ganz wesentlich erleichtert wird.

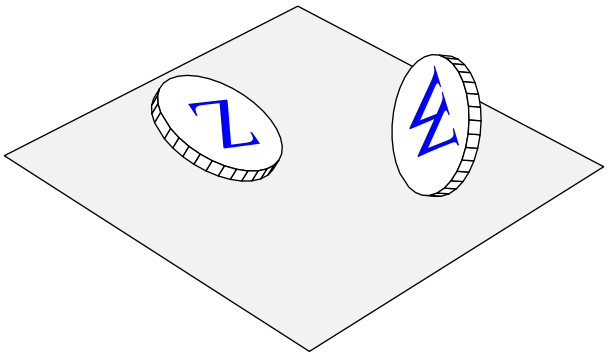
Im Grunde werden wir hier ähnlich vorgehen wie in der deskriptiven Statistik. Dort hatten wir *Realisationen* vorliegen, die z.B. auch das Ergebnis eines Zufallsexperiments sein konnten. Die Information dieser 'vielen Zahlen' haben wir in Form von relativen Häufigkeiten, Histogrammen und empirischen Verteilungsfunktionen komprimiert, und wir haben dafür Kennzahlen wie z.B. den Mittelwert und die Varianz berechnet.

Zufallsvariablen ordnen allen möglichen Ereignissen von Zufallsexperimenten Zahlen zu. Manche Zufallsvariablen können lediglich zwei unterschiedliche Werte annehmen (z.B. für das Geschlecht 1 = weiblich und 0 = nicht weiblich), andere hingegen überabzählbare viele Ausprägungen (z.B. Vermögen $\in \mathbb{R}$).

Auch die Information dieser 'vielen Zahlen' (d.h. mögliche Ausprägungen) können wir komprimieren, aber anstelle von relativen Häufigkeiten und Histogrammen in der deskriptiven Statistik verwenden wir nun Wahrscheinlichkeitsfunktionen (für diskrete Zufallsvariablen) und Dichtefunktionen (für stetige Zufallsvariablen), statt Mittelwerten berechnen wir als Lagemaße Erwartungswerte, und anstelle der empirischen Varianz können wir als Streuungsmaß die theoretische Varianz von Zufallsvariablen berechnen.

In der Statistik hat es sich eingebürgert Zufallsvariablen mit Großbuchstaben zu bezeichnen (z.B. X), während man für die Realisationen von Zufallsvariablen die entsprechenden Kleinbuchstaben verwendet (z.B. x). Die Wahrscheinlichkeit, dass eine Zufallsvariable X die Realisation x annimmt, wird geschrieben als $\Pr(X = x)$.

Zufallsexperiment:



Zufallsvariable: Abbildung in die reellen Zahlen

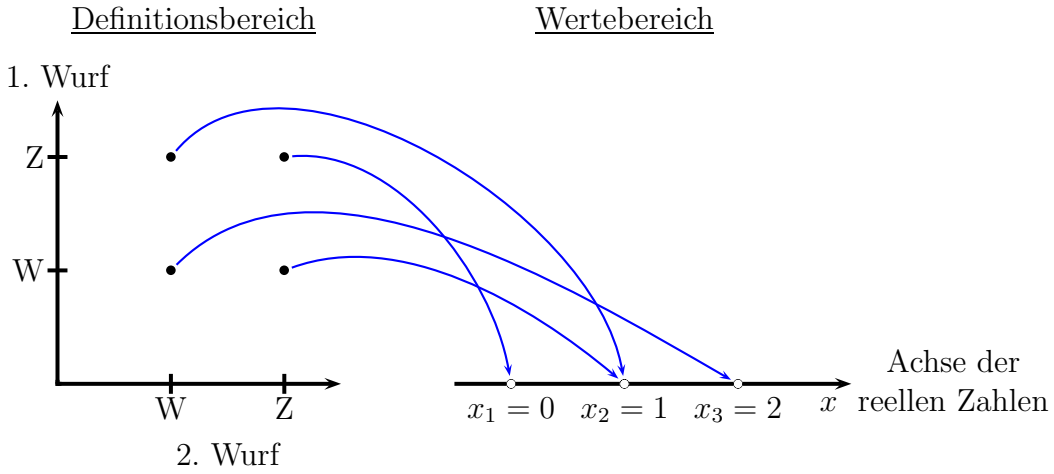


Abbildung 3.6: Definitions- und Wertebereich der Zufallsvariable X: "Anzahl Wappen" beim zweimaligen Werfen einer Münze (nach Bleymüller et al., 2002, 39f)

Aus Abbildung 3.6 ist ersichtlich, dass sich die Zufallsvariable X als Funktion auffassen lässt, die jedem Elementarereignis eine reelle Zahl zuordnet. Der *Definitionsbereich* ist der Ergebnisraum Ω des zugrundeliegenden Zufallsexperiments, und der *Wertebereich* ist die Menge der reellen Zahlen.

Achtung: Zufallsvariablen

1. beziehen sich immer auf die relevanten Ereignisse des zugrundeliegenden Zufallsexperiments (d.h. die Ereignismenge $\mathcal{A}$, und
2. sie beschreiben *alle möglichen* Ausgänge des zugrunde liegenden Zufallsexperiments.
3. die Abbildung der Ereignisse in die reellen Zahlen $\mathbb{R}$ erfolgt derart, dass diesen Zahlen wieder die korrekten Wahrscheinlichkeiten aus dem Zufallsexperiment zugeordnet werden können. Während die Zuordnung von (Teil-)Mengen zu Wahrscheinlichkeiten nur mit Mengenfunktionen $P(\cdot) : \mathcal{A} \mapsto [0, 1]$ möglich

ist, können Zufallsvariablen mit Hilfe reeller Funktionen[4] Wahrscheinlichkeiten zugeordnet werden. Um diesen Unterschied zu betonen verwenden für diese Wahrscheinlichkeiten das Symbol 'Pr', d.h. für diskrete Zufallsvariablen $\Pr(X) : \mathbb{R} \mapsto [0, 1]$, bzw. für stetige Zufallsvariablen $f(X) : \mathbb{R} \mapsto [0, 1]$.

Diskrete Zufallsvariablen: die Ereignismenge $\mathcal{A}$ enthält eine *abzählbare* Anzahl von Elementen, die in $\mathbb{N}$ abgebildet werden können.

Stetige Zufallsvariablen: die Ereignismenge Menge $\mathcal{A}$ enthält *überabzählbar* viele Elemente, die in $\mathbb{R}$ abgebildet werden.

Der Begriff *Zufallsvariable* ist eigentlich irreführend, denn im mathematischen Sinne handelt es sich dabei um keine Variable, sondern um eine Funktion. Darüber hinaus spielt der 'Zufall' für die mathematische Definition einer Zufallsvariable keine Rolle.

Aber den Ausprägungen einer Zufallsvariable können Wahrscheinlichkeiten zugeordnet werden; für eine Zufallsvariable X existiert für jede reelle Zahl c eine Wahrscheinlichkeit, dass X einen Wert annimmt, der kleiner oder gleich c ist, oder in anderen Worten, für $c \in \mathbb{R}$ existiert immer eine Wahrscheinlichkeit $\Pr(X \leq c)$ (diese Wahrscheinlichkeit kann aber auch Null oder Eins sein).

Dies führt uns zu den nächsten Konzepten, zu den Wahrscheinlichkeits-, Dichte- und Verteilungsfunktionen.

Aber vorher fassen wir nochmals zusammen: eine Zufallsvariable bildet alle möglichen Ausgänge des zugrunde liegenden Zufallsexperiments in die Menge der reellen Zahlen $\mathbb{R}$ derart ab, dass die Wahrscheinlichkeiten des zugrunde liegenden Zufallsexperiments korrekt 'übertragen' werden können. Deshalb müssen wir uns im Folgenden nicht mit den Ergebnissen des Zufallsexperiments abmühen, die beliebige Mengen sein können, sondern wir können mit deren Abbildung in den reellen Zahlen – d.h. den Zufallsvariablen – rechnen!

3.3.1 Wahrscheinlichkeitsraum und Zufallsvariablen*

Wir haben schon früher erwähnt, dass Zufallsvariablen ziemlich komplexe mathematische Gebilde sind. Eine wirkliche Einführung in das Konzept der Zufallsvariablen würde den Rahmen dieser Einführung bei weitem sprengen, aber da dieses Konzept für alles Folgende von derartiger Bedeutung ist wollen wir hier zumindest einige zentrale Begriffe kurz vorstellen. Die eilige Leserin kann diesen Abschnitt getrost überspringen ...

Ausgangspunkt der folgenden Überlegungen ist ein Zufallsexperiment, welches in einen Wahrscheinlichkeitsraum $[\Omega, \mathcal{A}, P(\cdot)]$ abgebildet werden kann. Ω die wieder die Ergebnismenge, $\mathcal{A}$ eine Ereignismenge und $P(\cdot)$ eine Mengenfunktion.

Die Ereignismenge $\mathcal{A}$ ist abgeschlossen bezüglich der Komplementbildung, der Vereinigungs- und Durchschnittsbildung. Das bedeutet, wenn eine dieser Mengenoperationen auf irgendein Element von $\mathcal{A}$ angewandt wird, ist das Ergebnis wieder ein Element von $\mathcal{A}$.

[4] Reelle Funktionen sind Abbildungen, in denen sowohl die Definitionsmenge als auch die Wertemenge Teilmengen von $\mathbb{R}$ sind.

Eine mögliche Ereignismenge ist immer die Potenzmenge, d.h. die Menge aller Teilmengen von Ω. Für einen einfachen Münzwurf mit den Elementarereignissen 'Wappen' (W) und 'Zahl' (Z) ist die Ereignismenge $\mathcal{A}_1 = \{\emptyset, \{K\}, \{W\}, \Omega\}$.

Für einen zweifachen Münzwurf mit $\Omega = \{(ZZ), (ZW), (WZ), (WW)\}$ ist die Ereignismenge schon deutlich komplexer, da sie neben den Elementarereignissen, $\emptyset$ und $\Omega = \{(ZZ), (ZW), (WZ), (WW)\}$ auch alle Durchschnitte, Vereinigungen und Komplemente davon enthält

$$\begin{aligned}\mathcal{A} \;=\; & \{\{\emptyset\}, \{(ZZ)\}, \{(ZW)\}, \{(WZ)\}, \{(WW)\}, \\ & \{(ZZ), (ZW)\}, \{(ZZ), (WZ)\}, \{(ZZ), (WW)\}, \{(ZW), (WZ)\}, \\ & \{(ZW), (WW)\}, \{(WZ), (WW)\}, \\ & \{(ZZ), (ZW), (WZ)\}, \{(ZZ), (ZW), (WW)\}, \\ & \{(ZZ), (WZ), (WW)\}, \{(ZW), (WZ), (WW)\}, \{\Omega\}\}\end{aligned}$$

Diese Potenzmenge enthält insgesamt bereits 16 Elemente, für praktische Anwendungen ist der Weg über die Potenzmengen häufig nicht gangbar. Glücklicherweise benötigt man selten die wirklichen Potenzmengen, meist reichen deutlich einfachere Ereignismengen.

Wenn wir uns z.B. beim zweimaligen Münzwurf für das Ereignis A "mindestens ein Wappen" interessieren ist $A = \{(WW), (WZ), (ZW)\}$ und der Ereignisraum $\mathcal{A}_W = \{\emptyset, A, \overline{A}, \Omega\} = \{\emptyset, \{(WW), (WZ), (ZW)\}, \{(ZZ)\}, \{(WW), (WZ), (ZW), (ZZ)\}\}$.

Die Ereignismenge $\mathcal{A}$ umfasst also alle interessierenden Ereignisse, und darüber hinaus neben $\emptyset$ und Ω auch die über Mengenoperationen damit verknüpften Mengen.

Im mathematischen Sinne bildet die Ereignismenge $\mathcal{A}$ eine σ-Algebra, sie besitzt eine bestimmte mathematische Struktur und erfüllt folgende Bedingungen: (1) $\Omega \in \mathcal{A}$, (2) wenn $A \in \mathcal{A}$ muss $\overline{A} \in \mathcal{A}$, und (3) wenn $A_j \in \mathcal{A}$ für $j = 1, 2, \ldots, J$ dann $\bigcup_{j=1}^{\infty} A_j \in \mathcal{A}$.

$P(\cdot)$ ist schließlich eine Mengen-Funktion vom Ereignisraum $\mathcal{A}$ in die reellen Zahlen zwischen Null und Eins, $P(\cdot) : \mathcal{A} \to [0, 1]$, die bestimmte Axiome erfüllt.

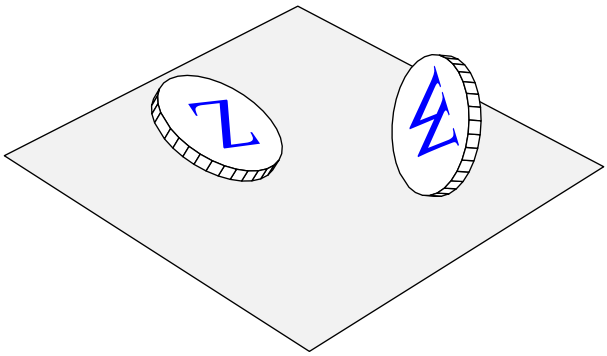
Abbildung 3.7 zeigt diesen Wahrscheinlichkeitsraum für ein sehr einfaches Zufallsexperiment mit nur vier diskreten Elementarereignissen.

Für solche einfachen Zufallsexperimente scheint dies ein bisschen viel Aufwand, aber der Vorteil dieser Herangehensweise liegt darin, dass dies auch für Mengen mit überabzählbar vielen Elementen verallgemeinert werden kann, und somit die Definition stetiger Zufallsvariablen ermöglicht.

Eine der großen Einsichten von A.N. Kolmogorov bestand darin, dass für dieses Problem eine damals noch relativ neues Teilgebiet der Mathematik anwendbar ist, die Maßtheorie, welche ursprünglich für ganz andere Zwecke entwickelt wurde (es ging v.a. um die Verallgemeinerung von elementargeometrischen Begriffen wie Streckenlänge, Flächeninhalt und Volumen, die es ermöglichte auch komplizierteren Mengen ein Maß zuzuordnen).

Im mathematischen Sinne ist eine Zufallsvariable eine *messbare* Funktion von einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \Pr(\cdot))$ in einen Messraum. Messbarkeit bedeutet dabei, dass das Urbild einer Menge wieder in einem bestimmten Mengensystem liegt, in unserem Fall eine Teilmenge der Ereignisalgebra $\mathcal{A}$ ist.

Zufallsexperiment:



Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P(\cdot))$**:**

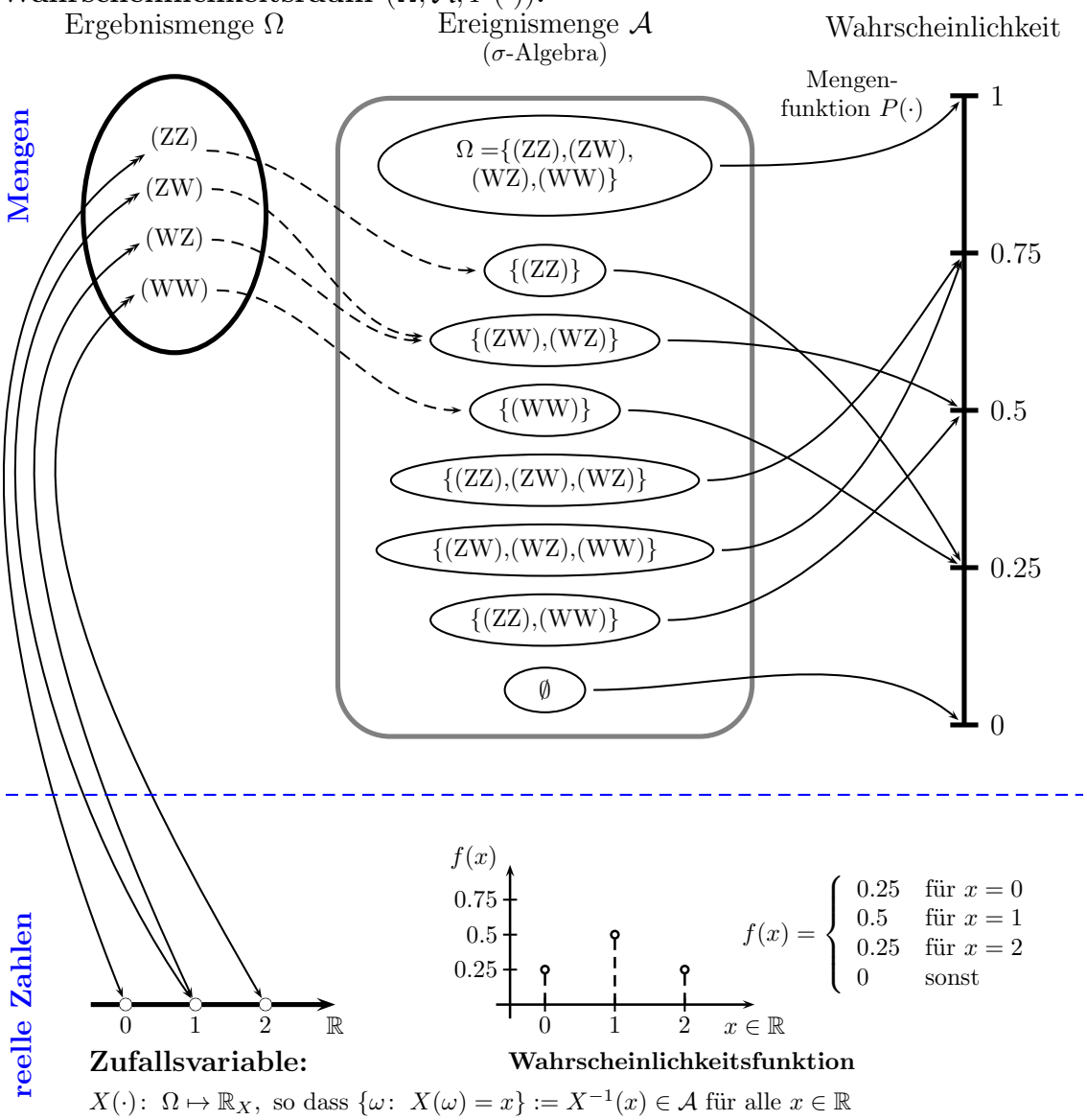


Abbildung 3.7: Zufallsexperiment Wurf mit zwei Münzen; Wahrscheinlichkeitsraum und Zufallsvariable für das interessierende Ereignis $X =$ Anzahl der Wappen (W). Die Abbildung in die reellen Zahlen erfolgt derart, dass das Urbild ein Element von $\mathcal{A}$ ist.

Damit kann eine stetige Zufallsvariable als eine Funktion $X(.) \to \mathbb{R}$ definiert werden, die (für stetige Ereignisse) folgende Bedingung erfüllt

$$\{\omega : X(\omega) \leq x\} := X^{-1}((-\infty, x]) \in \mathcal{A}) \quad \text{für alle } x \in \mathbb{R}$$

Zu Ihrer Beruhigung, für das Verständnis des Folgenden benötigen Sie dies nicht wirklich. Die mathematische Theorie hinter den Zufallsvariablen garantiert uns aber, dass wir den folgenden Ausführungen vertrauen können.

3.4 Wahrscheinlichkeits- und Verteilungsfunktionen einer einzelnen Zufallsvariable

Eine Zufallsvariable kann mindestens zwei, aber auch unendlich viele Ausprägungen annehmen. Ähnlich wie wir in der deskriptiven Statistik gegebene Beobachtungen in Form relativer Häufigkeitsverteilungen dargestellt haben, können wir in der Stochastik für diskrete Zufallsvariablen Wahrscheinlichkeitsfunktionen (bzw. für stetige Zufallsvariablen Dichtefunktionen) verwenden. Das Analogon für die Wahrscheinlichkeits- bzw. Dichtefunktionen der Stochastik sind die Histogramme in der deskriptiven Statistik.

Jedem Wert einer diskreten Zufallsvariable sind ein oder mehrere Elemente aus dem Ereignisraum des Zufallsexperiments zugeordnet. Da jedem möglichen Ereignis eines Zufallsexperiments eine Wahrscheinlichkeit zugeordnet ist, kann auch jedem diskreten Wert einer Zufallsvariable eine Wahrscheinlichkeit zugeordnet werden.

Stetige Zufallsvariablen werden in die reellen Zahlen abgebildet, deshalb werden nicht einzelnen Ausprägungen Wahrscheinlichkeiten zugeordnet, sondern Intervallen in $\mathbb{R}$.[5]

3.4.1 Wahrscheinlichkeits- und Verteilungsfunktion einer *diskreten* Zufallsvariable

Eine **Wahrscheinlichkeitsfunktion** (*'probability mass function' pmf*, auch Zähldichte genannt) ordnet jeder der abzählbar vielen Ausprägungen einer diskreten Zufallsvariable die dazugehörige Wahrscheinlichkeit zu.

Wenn wir die unterschiedlichen Ausprägungen einer diskreten Zufallsvariablen X mit $x_1, x_2, \ldots$ bezeichnen gibt die Wahrscheinlichkeitsfunktion $f(x_j)$ also die Wahrscheinlichkeiten ihres Auftretens an

$$f(x_j) = \Pr(X = x_j) \qquad \text{für } j = 1, 2, \ldots$$

Im Unterschied zur Mengenfunktion $P(\cdot)$ des Wahrscheinlichkeitsraums ist $f(x_j) = \Pr(X = x_j)$ eine reelle Funktion mit der man wie üblich 'rechnen' kann.

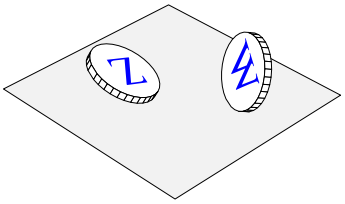
Jede Wahrscheinlichkeitsfunktion muss die folgenden beiden Eigenschaften erfüllen (J ist die Zahl der möglichen Ausprägungen, und j der Laufindex):

$$1) \qquad f(x_j) \geq 0 \qquad \text{für } j = 1, 2, \ldots, J$$

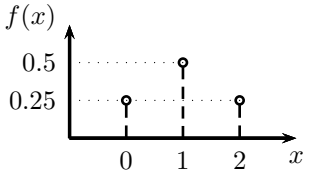
$$2) \qquad \sum_{j=1}^{J} f(x_j) = 1$$

Wie kommen wir nun zu den Wahrscheinlichkeiten? Diese Frage kann kaum allgemein beantwortet werden, aber einzelne Möglichkeiten sind z.B.

[5] Genau genommen ist die Wahrscheinlichkeit für $X = x$ immer gleich Null, da die reellen Zahlen 'unendlich dicht gepackt' sind, aber $a < X < b$ kann eine positive Wahrscheinlichkeit annehmen.

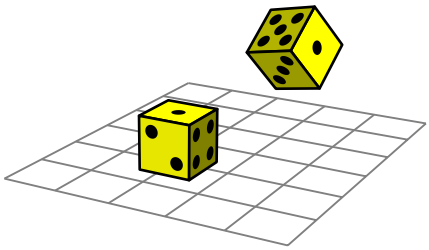


x	Elemente im Ergebnisraum	$f(x)$
0	(ZZ)	0.25
1	(ZW), (WZ)	0.5
2	(WW)	0.25

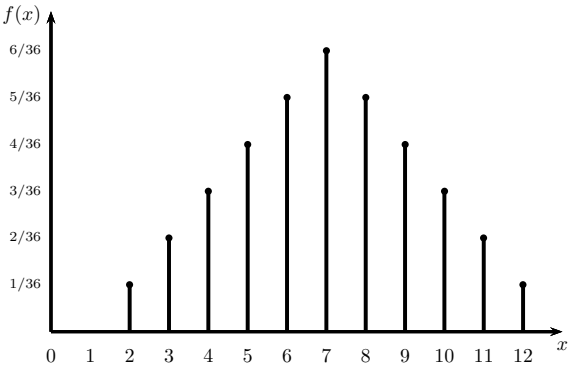


$$f(x) = \begin{cases} 0.25 & \text{für } X = 0 \\ 0.5 & \text{für } X = 1 \\ 0.25 & \text{für } X = 2 \\ 0 & \text{sonst} \end{cases}$$

Abbildung 3.8: Beispiel 1: Wahrscheinlichkeitsfunktion der Zufallsvariable X: "Anzahl der Wappen bei zweifachen Münzwurf".



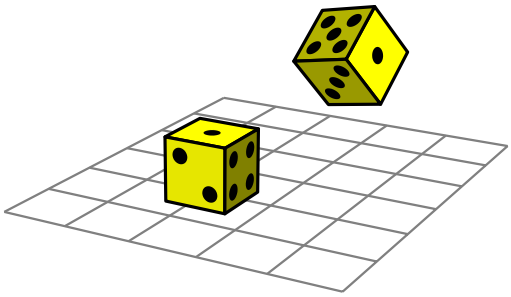
x	Elemente im Ereignisraum	$f(x)$
2	1 1	1/36
3	1 2, 2 1	2/36
4	1 3, 3 1, 2 2	3/36
5	1 4, 4 1, 2 3, 3 2	4/36
6	1 5, 5 1, 2 4, 4 2, 3 3	5/36
7	1 6, 6 1, 2 5, 5 2, 3 4, 4 3	6/36
8	2 6, 6 2, 3 5, 5 3, 4 4	5/36
9	3 6, 6 3, 4 5, 5 4	4/36
10	4 6, 6 4, 5 5	3/36
11	5 6, 6 5	2/36
12	6 6	1/36



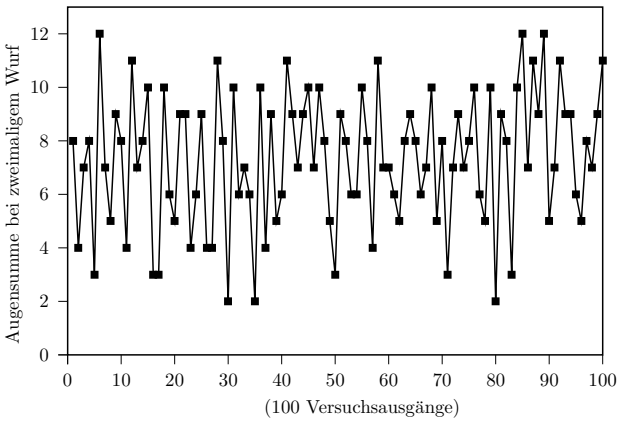
$$f(x) = \Pr(x) = \begin{cases} 1/36 & \text{für } X = 2 \\ 2/36 & \text{für } X = 3 \\ \vdots & \vdots \\ 1/36 & \text{für } X = 12 \\ 0 & \text{sonst.} \end{cases}$$

Abbildung 3.9: Beispiel 2: Wahrscheinlichkeitsfunktion der Zufallsvariablen X: "Augensumme bei einem Wurf mit zwei Würfeln".

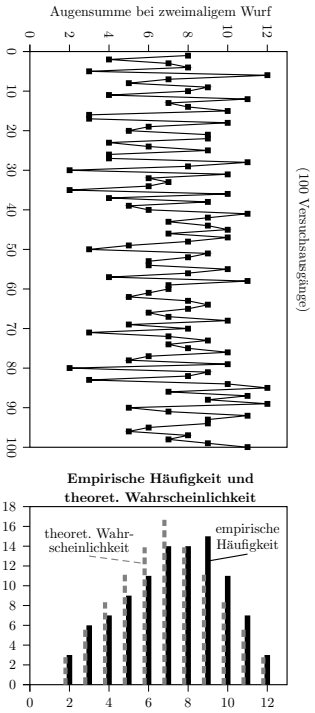
Zufallsexperiment: **“Anzahl der Augen beim zweimaligen Würfeln”**



Realisationen: Das zufällige Ergebnis von 100 Würfen



Deskriptiv:



Zufallsvariable:

x	Elemente im Ereignisraum	$f(x)$
2	1 1	1/36
3	1 2, 2 1	2/36
4	1 3, 3 1, 2 2	3/36
5	1 4, 4 1, 2 3, 3 2	4/36
6	1 5, 5 1, 2 4, 4 2, 3 3	5/36
7	1 6, 6 1, 2 5, 5 2, 3 4, 4 3	6/36
8	2 6, 6 2, 3 5, 5 3, 4 4	5/36
9	3 6, 6 3, 4 5, 5 4	4/36
10	4 6, 6 4, 5 5	3/36
11	5 6, 6 5	2/36
12	6 6	1/36

$$f(x) = \Pr(x) = \begin{cases} 1/36 & \text{für } X = 2 \\ 2/36 & \text{für } X = 3 \\ \vdots & \vdots \\ 1/36 & \text{für } X = 12 \\ 0 & \text{sonst.} \end{cases}$$

Abbildung 3.10: Empirische Häufigkeit (deskriptiv) versus Wahrscheinlichkeitsfunktion

1. in sehr einfachen Fällen können wir die Wahrscheinlichkeiten unmittelbar angeben, wenn wir das zugrunde liegende Zufallsexperiment kennen. Abbildungen 3.8 und 3.9 geben zwei Beispiele dafür. Dieser Fall ist selten, die Prozesse, die uns interessieren, sind meist deutlich komplexer.

2. In manchen Fällen können wir zwar nicht unmittelbar die Wahrscheinlichkeiten angeben, aber aus theoretischen Überlegungen und praktischen Erfahrungen können wir vermuten, welche theoretische Verteilung sich zur Beschreibung eignet. Interessiert uns für das Zufallsexperiment 'zweifacher Münzwurf' eine andere Zufallsvariable Y *"mindestens ein Wappen wird geworfen"* sind nur zwei Ausgänge möglich, nämlich $X = 0$ oder $X = 1$. Die Wahrscheinlichkeitsfunktion wird deshalb durch eine Bernoulli-Verteilung $f(x;\theta) = \theta^x(1-\theta)^{1-x}$ beschrieben, wobei $0 \leq \theta \leq 1$ ein Parameter der Verteilung ist.

   Für stetige Zufallsvariablen werden wir häufig eine Normalverteilung annehmen, da viele in der Natur beobachtbare Merkmalsausprägungen (z.B. Körpergröße) näherungsweise normalverteilt sind und sie darüber hinaus zahlreiche angenehme Eigenschaften aufweist.

3. In vielen Fällen ist es gar nicht erforderlich eine spezifische Verteilung anzunehmen. Für bestimmte Schätzfunktionen wie z.B. Regressionskoeffizienten ist bekannt, dass diese als bestimmte Funktionen von Momenten der Verteilung geschrieben werden können, für die zentrale Grenzwertsätze gelten. Deshalb konvergiert die Verteilung dieser Schätzfunktionen mit zunehmender Stichprobengröße gegen die Normalverteilung, unabhängig davon, wie die ursprünglichen Zufallsvariablen verteilt sind, sofern bestimmte Annahmen (z.B. Unabhängigkeit) erfüllt sind.

Verteilungsfunktion Eine (kumulative) Verteilungsfunktion $F(x)$ (*cumulative distribution function*) gibt die Wahrscheinlichkeit dafür an, dass eine Zufallsvariable X *höchstens* den Wert x annimmt. Wenn die Ausprägungen x_j (mit $j = 1, 2, \ldots, J$) aufsteigend nach ihrem Wert geordnet sind erhalten wir durch kumulieren (d.h. durch jeweiliges addieren aller vorhergehenden Werte)

$$F(x_j) = \Pr(X \leq x_j) \ = f(x_1) + f(x_2) + \cdots + f(x_j) = \sum_{l=1}^{j} f(x_l)$$

Abbildung 3.11 zeigt die Verteilungsfunktion für die Zufallsvariable X: "Augensumme bei einem Wurf mit zwei Würfeln" von unserem obrigen Beispiel.

Hinweis: Die Verteilungsfunktion ist das stochastische Analogon zur empirischen Verteilungsfunktion, die wir in der deskriptiven Statistik auf Grundlage relativer Häufigkeiten *für Realisationen* berechnet haben.

Übung: Wie lautet die Wahrscheinlichkeits- und Verteilungsfunktion für das *Produkt* der Augenzahlen bei zwei Würfen mit einem Würfel?

x	$f(x)$	$F(x)$
2	1/36	1/36
3	2/36	3/36
4	3/36	6/36
5	4/36	10/36
6	5/36	15/36
7	6/36	21/36
8	5/36	26/36
9	4/36	30/36
10	3/36	33/36
11	2/36	35/36
12	1/36	36/36

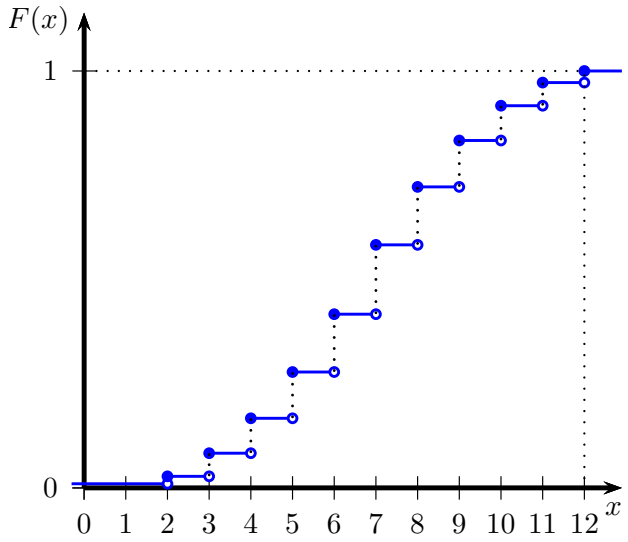


Abbildung 3.11: Verteilungsfunktion der Zufallsvariablen X: "Augensumme bei einem Wurf mit zwei Würfeln".

3.4.2 Dichte- und Verteilungsfunktion einer *stetigen* Zufallsvariable

Eine (Wahrscheinlichkeits-)Dichtefunktion (*'density functions for continuous random variables'*) ist das Analogon zur Wahrscheinlichkeitsfunktion für stetige Zufallsvariablen. Ein wesentlicher Unterschied besteht darin, dass Wahrscheinlichkeiten nur als Fläche unter der Dichtefunktion berechnet werden können.

Wenn $f(x)$ eine Dichtefunktion ist, dann ist die Wahrscheinlichkeit dafür, dass X einen Wert in einem beliebigen Intervall $[a, b]$ (mit $a < b$ und $a, b \in \mathbb{R}$) annimmt, gleich

$$\Pr(a < X < b) = \int_a^b f(x)dx$$

Da stetige Zufallsvariablen innerhalb jedes Intervalls überabzählbar viele Ausprägungen annehmen können ist die Wahrscheinlichkeit dafür, dass die Zufallsvariable X einen exakten Wert x annimmt, immer gleich Null – $\Pr(X = x) = 0$ – denn die Fläche über einem Punkt ist Null! Man beachte, dass $f(x)$ zwar berechnet werden kann und $f(x) \geq 0$, dass dieser Wert aber nicht als Wahrscheinlichkeit interpretiert werden darf!

Die Fläche unter einem Intervall der Dichtefunktion gibt also an, mit welcher Wahrscheinlichkeit Ereignisse, die diesem Intervall der Zufallsvariable zugeordnet sind, eintreten (siehe Abbildung 3.12).

Eine Dichtefunktion muss folgende beiden Bedingungen erfüllen

$$1) \qquad f(x) \geq 0$$

$$2) \qquad \int_{-\infty}^{\infty} f(x)\, dx = 1$$

$$3) \qquad \Pr(a \leq X \leq b) = \int_a^b f(x)\, dx$$

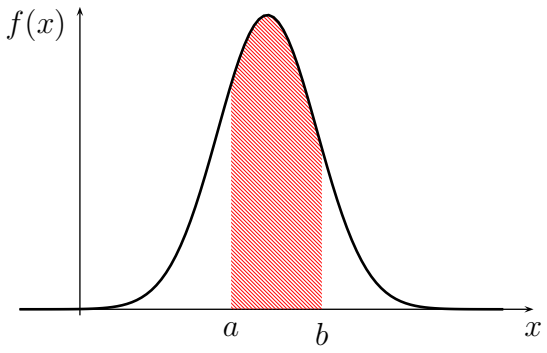


Abbildung 3.12: Dichtefunktion einer stetigen Zufallsvariablen.

Ein Beispiel mit der Verteilung der log(Stundenlöhne) für Österreich finden Sie in Abbildung 3.13.

Beispiel: Ist die Funktion

$$f(x) = \begin{cases} \frac{1}{9}\,x^2 & \text{für } 0 \le x \le 3 \\ 0 & \text{sonst} \end{cases}$$

eine Dichtefunktion?

1. Offensichtlich ist $f(x) \geq 0$ für alle x im Bereich 0 bis 3.
2. Das Integral von 0 bis 3 ist [6]

$$\int_0^3 \frac{1}{9}\,x^2 dx = \frac{1}{27}x^3\Big|_0^3 = \frac{27}{27} - 0 = 1$$

3. Die Wahrscheinlichkeit, dass X zwischen 0 und 1 liegt, ist z.B.

$$\int_0^1 \frac{1}{9}\,x^2 dx = \frac{1}{27}x^3\Big|_0^1 = \frac{1}{27} - 0 = \frac{1}{27}$$

¶

Verteilungsfunktion Analog zum diskreten Fall existiert auf für stetige Zufallsvariablen eine *Verteilungsfunktion* $F(x) = \Pr(X \leq x)$.
Die Verteilungsfunktion herhalten wir, indem wir die Dichtefunktion von $-\infty$ bis x kumulieren (d.h. x ist die obere Integralgrenze). Deshalb benötigen wir für das Integral eine andere Laufvariable v

$$F(x) = \int_{-\infty}^{x} f(v)dv$$

Verteilungsfunktionen haben folgende Eigenschaften:

[6] $\int X^n dX = \frac{1}{1+n}\,X^{n+1} + c, \quad (n \neq -1)$

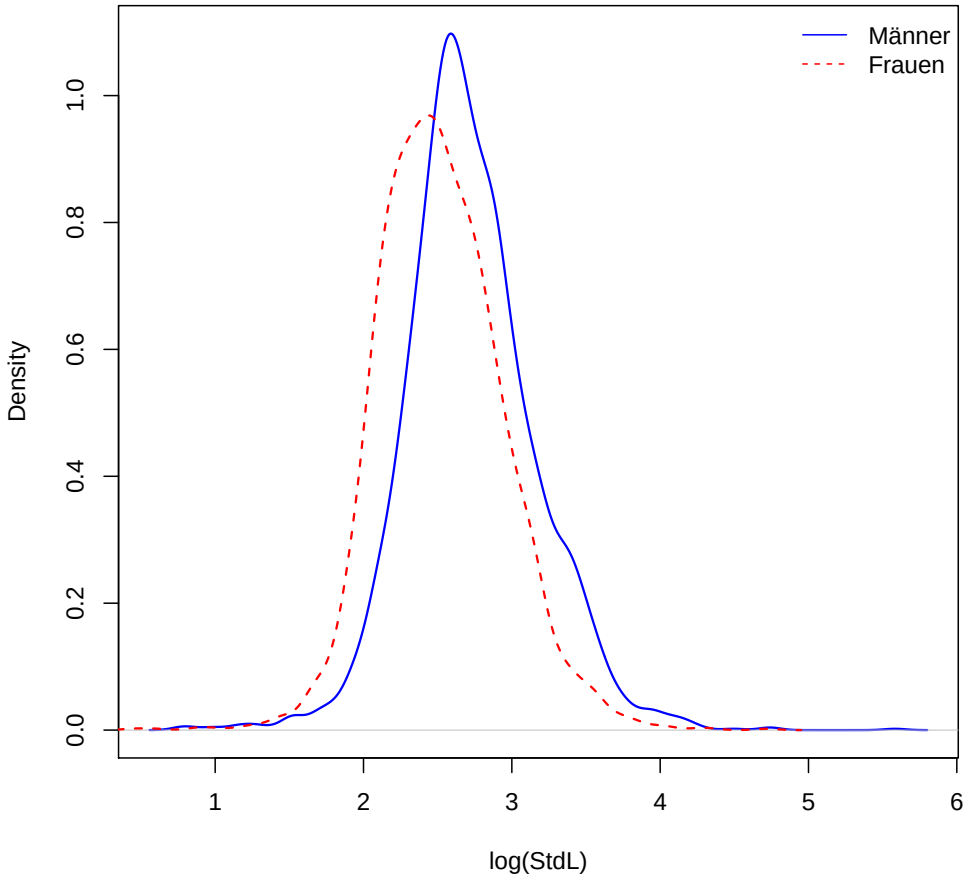


Abbildung 3.13: Beispiel: Verteilung der log(Stundenlöhne) unselbständig Beschäftigter in Österreich 2009 (Datenquelle: EU-SILC Daten)

1. $0 \leq F(x) \leq 1$;
2. $F(x)$ ist monoton wachsend, d.h. für $x_1 < x_2$ gilt $F(x_1) \leq F(x_2)$;
3. $\lim_{x \to -\infty} F(x) = 0$;
4. $\lim_{x \to +\infty} F(x) = 1$;
5. $F(x)$ ist stetig.

Beispiel: (nach Bleymüller et al., 2002, 42f) Ist die Funktion

$$f(x) = \begin{cases} 0.5 - 0.125x & \text{für } 0 \leq x \leq 4 \\ 0 & \text{sonst} \end{cases}$$

eine Dichtefunktion? Wie lautet die Verteilungsfunktion?

Offensichtlich ist $f(x)$ eine Dichtefunktion, denn

$$f(x) \geq 0 \quad \text{für alle } x$$

und

$$\begin{aligned} \int_{-\infty}^{+\infty} f(x)dx &= \int_{-\infty}^{0} f(x)dx + \int_{0}^{4} f(x)dx + \int_{4}^{+\infty} f(x)dx \\ &= 0 + \int_{0}^{4} f(x)dx + 0 \\ &= \int_{0}^{4} (0.5 - 0.125x)dx \\ &= \left[0.5x - \frac{0.125}{2}\, x^2\right]_0^4 \\ &= 2 - 1 = 1 \end{aligned}$$

Die Wahrscheinlichkeit, dass X z.B. einen Wert zwischen 1 und 2 annimmt, ist

$$\begin{aligned} \Pr(1 \leq X \leq 2) &= \int_{1}^{2} f(x)dx \\ &= \int_{1}^{2} (0.5 - 0.125x)dx \\ &= \left[0.5x - \frac{0.125}{2}\, x^2\right]_1^2 \\ &= 0.75 - 0.4375 = 0.3125 \end{aligned}$$

Die *Verteilungsfunktion* $F(x)$ erhält man

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f(v)dv \; = \int_{0}^{x} (0.5 - 0.125v)dv \\ &= \left[0.5v - \frac{0.125}{2}\, v^2\right]_0^x \\ &= 0.5x - 0.0625x^2 \end{aligned}$$

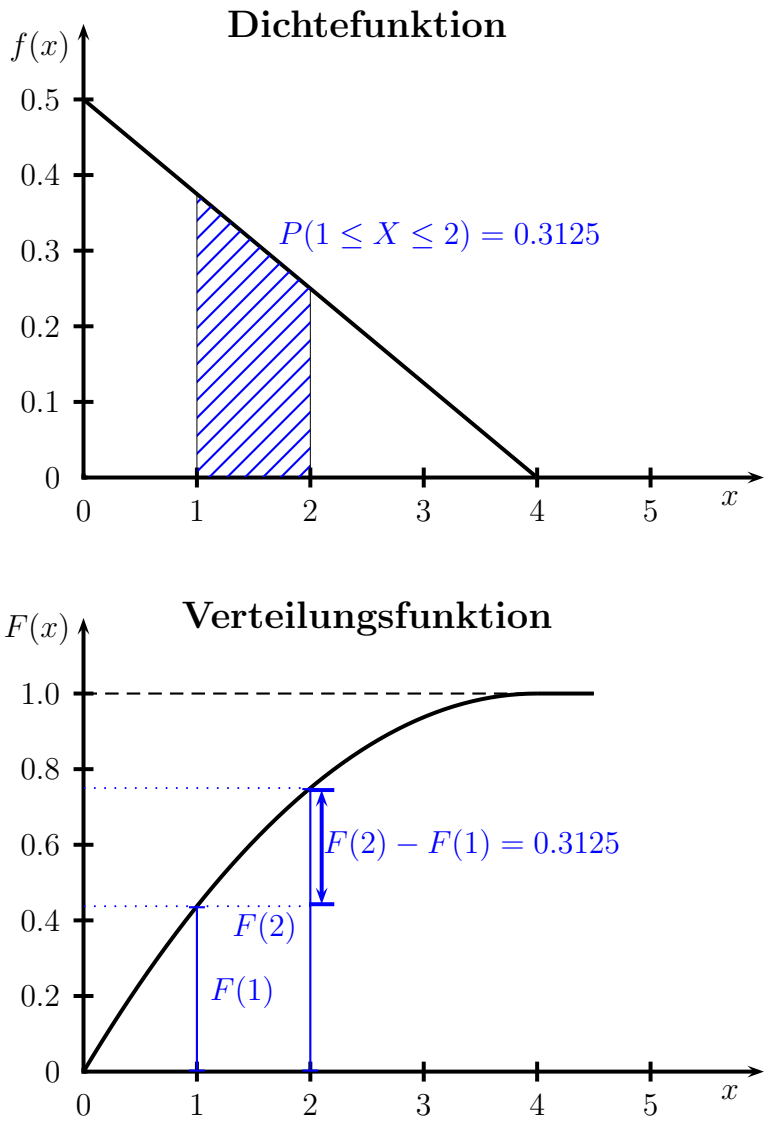


Abbildung 3.14: Dichte- und Verteilungsfunktion einer stetigen Zufallsvariablen.

also

$$F(x) = \begin{cases} 0 & \text{für } x < 0 \\ 0.5x - 0.0625x^2 & \text{für } 0 \leq x \leq 4 \\ 1 & \text{für } x > 4 \end{cases}$$

Die Wahrscheinlichkeit dafür, dass X zwischen 1 und 2 liegt, kann auch mit Hilfe der Verteilungsfunktion berechnet werden:

$$\begin{aligned} \Pr(1 \leq X \leq 2) &= F(2) - F(1) \\ &= 0.75 - 0.4375 \\ &= 0.3125 \end{aligned}$$

Dieses Beispiel ist in Abbildung 3.14 dargestellt.

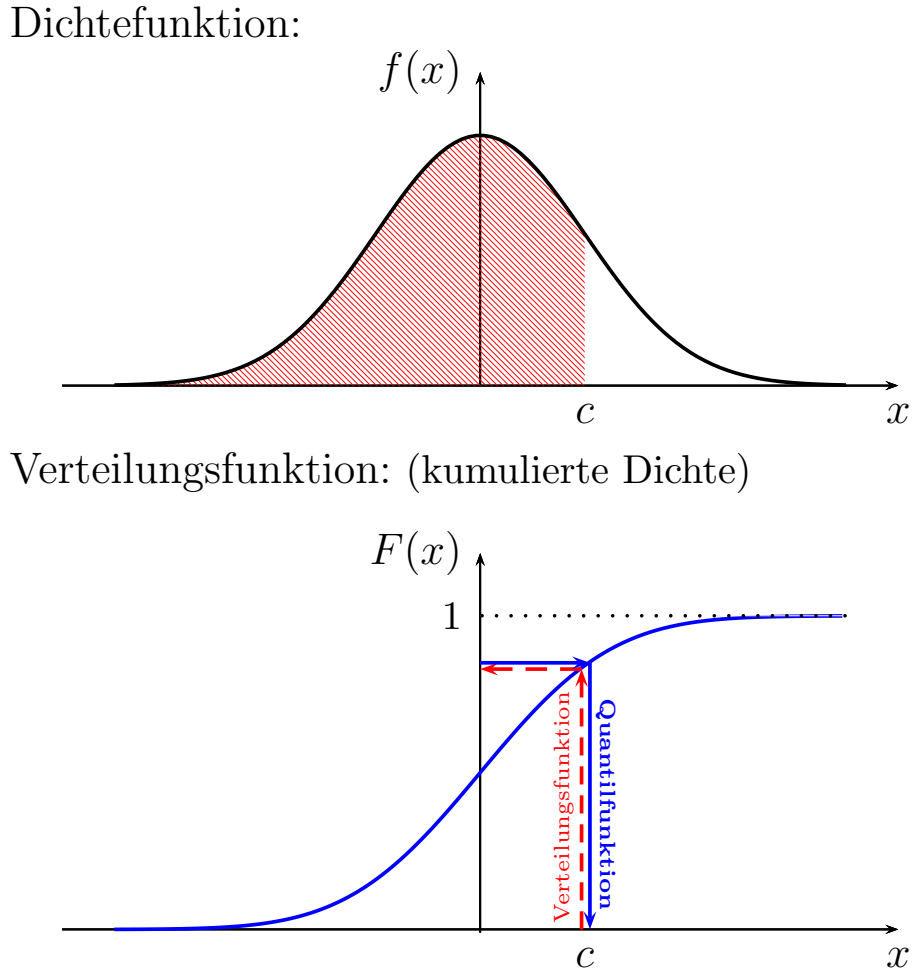


Abbildung 3.15: Eine Quantilfunktion ist die Umkehrfunktion einer Verteilungsfunktion.

Quantilfunktion

Wenn F eine Verteilungsfunktion[7] ist, dann heißt die Umkehrfunktion (Inverse) $F^{-1}(w)$ *Quantilfunktion.*

Beispiel: Eine gegebene Verteilungsfunktion für die Körpergröße einer bestimmten Ethnie ordne einer Körpergröße von 180 cm ein Wert von 0.75 zu. Dies erlaubt uns die Aussage, dass 75% dieser Ethnie eine Körpergröße von 180 cm *oder kleiner* aufweisen.

Die zugehörige *Quantilfunktion* (inverse Verteilungsfunktion) an der Stelle 0.75 ordnet die Körpergröße zu, so dass mehr als 75% der Ethnie eine Körpergröße kleiner als w aufweist.

Formal ist die Quantilfunktion $F^{-1} : (0, 1) \to \mathbb{R}$ einer Verteilungsfunktion F definiert durch

$$F^{-1}(w) := \inf \{x \in \mathbb{R} \mid F(x) \geq w\}$$

[7] Zur Erinnerung: Verteilungsfunktionen $F : \mathbb{R} \to \mathbb{R}$ haben folgende Eigenschaften: a) monotion wachsend, b) rechtsseitig stetig und c) $\lim_{x \to -\infty} F(x) = 0$ und $\lim_{x \to +\infty} F(x) = 1$

Mit Hilfe der Quantilfunktion können Quantile wie z.B. der Median, Quartile oder Perzentile berechnet werden.

3.5 Gemeinsame Wahrscheinlichkeitsfunktionen (*'Joint Probability Density Functions'*)

Die meisten Zufallsexperimente erzeugen mehr als eine Zufallsvariable, und wir interessieren uns hier vor allem für Zusammenhänge zwischen solchen Zufallsvariablen, z.B. für den Zusammenhang zwischen Bildung und Einkommen, oder dem Preis und dem Alter von Gebrauchtautos.

Beispiel: Wir gehen wieder von einem einfachen Zufallsexperiment 'zweifacher Münzwurf' aus. Die Ergebnismenge ist

$$\Omega = \{(ZZ), (WZ), (ZW), (WW)\}$$

Als nächstes definieren wir zwei Zufallsvariablen X und Y

$$X = \begin{cases} 1 & \text{wenn 'Wappen beim ersten Wurf'} \\ 0 & \text{sonst} \end{cases}$$

und

$$Y = \begin{cases} 1 & \text{wenn 'mindestens ein Wappen bei zwei Würfen'} \\ 0 & \text{sonst} \end{cases}$$

Zwei von vier Elementarereignissen erfüllen die Bedingung 'Wappen beim ersten Wurf' $\{(WZ), (WW)\}$, und drei Elementarereignisse erfüllen die Bedingung 'mindestens ein Wappen bei zwei Würfen' $\{(WZ), (ZW), (WW)\}$.

Damit können wir die (unbedingten) Wahrscheinlichkeitsfunktionen für diese beiden diskreten Zufallsvariablen hinschreiben

$$f_x(x) = \begin{cases} 0.5, & \text{für } X = 0 \\ 0.5, & \text{für } X = 1 \end{cases} \quad \text{und} \quad f_y(y) = \begin{cases} 0.25, & \text{für } Y = 0 \\ 0.75, & \text{für } Y = 1 \end{cases}$$

Aber wir können auch die gemeinsamen Wahrscheinlichkeiten angeben, z.B. ist die Wahrscheinlichkeit für $X = 0$ (d.h. beim ersten Wurf *kein Wappen*, also eine Zahl) und $Y = 0$ (d.h. bei beiden Würfen *kein Wappen* zu erhalten) gleich 0.25, denn nur das Element $\{(ZZ)\}$ aus Ω erfüllt diese Bedingung, also ist $f(0,0) = \Pr(X = 0, Y = 0) = 0.25$. Ähnlich können wir die anderen Wahrscheinlichkeiten ermitteln und als gemeinsame Wahrscheinlichkeitsfunktion $f(x, y)$ in Tabellenform anschreiben, wobei die erste Spalte die möglichen Ausprägungen von X und die erste Zeile die möglichen Ausprägungen von Y bezeichnet.

X\Y	0	1	$f_x(x)$
0	0.25	0.25	0.5
1	0	0.5	0.5
$f_y(y)$	0.25	0.75	1

Die Wahrscheinlichkeit dafür, dass wir beim ersten Wurf *ein Wappen* ($X = 1$), und bei beiden Würfen *kein Wappen* ($Y = 0$) erhalten, ist Null, d.h. $f(1, 0) = \Pr(X = 1, Y = 0) = 0$.

Die Bedingung $X = 1$ *und* $Y = 1$ erfüllen nur die Elementarereignisse $\{(WZ), (WW)\}$, also ist $f(X = 1 \text{ und } Y = 1) = 0.5$.

Wie man einfach erkennen kann erhält man durch Aufsummieren der Wahrscheinlichkeiten die Randwahrscheinlichkeiten (*'marginal probability'*), die gemeinsam die Randverteilungen bilden

Randverteilungen (*Marginal Probability Function*)

$$\begin{aligned} f_x(x) &= \sum_y f(x, y) \qquad \text{Randverteilung von } X \\ f_y(y) &= \sum_x f(x, y) \qquad \text{Randverteilung von } Y \end{aligned}$$

bzw. analog für stetige Zufallsvariablen

$$\begin{aligned} f_x(x) &= \int_{-\infty}^{+\infty} f(x, y)dy \qquad \text{Randverteilung von } X \\ f_y(y) &= \int_{-\infty}^{+\infty} f(x, y)dx \qquad \text{Randverteilung von } Y \end{aligned}$$

Man beachte, dass diese Randverteilungen wieder die univariaten Verteilungen sind die wir erhalten haben, als wir die beiden Zufallsvariablen X und Y unabhängig voneinander untersucht haben.

In Tabelle 3.2 finden Sie eine etwas allgemeinere Schreibweise der gemeinsamen Wahrscheinlichkeitsfunktion zweier *diskreter* Zufallsvariablen, wobei Zufallsvariable X insgesamt J verschiedene Ausprägungen annehmen kann (mit $j = 1, \ldots, J$), und Zufallsvariable Y insgesamt L Ausprägungen hat (mit $l = 1, \ldots, L$).

Die gemeinsame Dichte ist $f(x_j, y_l) = \Pr(X = x_j, Y = y_l)$, das heißt, $f(x_j, y_l)$ gibt im Fall diskreter Zufallsvariablen die Wahrscheinlichkeit dafür an, dass die Zufallsvariable X den Wert x_j und die Zufallsvariable Y gleichzeitig den Wert y_l annimmt (mit $j = 1, \ldots, J$ und $l = 1, \ldots, L$).

Das Analogon zur gemeinsamen Wahrscheinlichkeitsfunktion zweier diskreten Zufallsvariablen in der deskriptiven Statistik ist natürlich die relative Häufigkeitsverteilung (Kontingenztabelle).

Selbstverständlich kann dies auch auf höhere Dimensionen erweitert werden, z.B. $f(x, y, z)$, aber diese Wahrscheinlichkeitsfunktionen können nicht mehr einfach grafisch dargestellt werden.

Natürlich muss auch für gemeinsame Wahrscheinlichkeitsfunktionen wieder gelten, dass

$$f(x_j, y_l) \geq 0 \qquad \text{für } j, l = 1, 2, \ldots$$

und

$$\sum_{j=1}^{J} \sum_{l=1}^{L} f(x_j, y_l) = 1$$

Tabelle 3.2: Gemeinsame Wahrscheinlichkeitsfunktion zweier diskreten Zufallsvariablen X und Y mit Randverteilungen, wobei J und L die Anzahl der Ausprägungen der Zufallsvariablen X und Y angeben.

$X \backslash Y$	y_1	y_2	$\dots$	y_L	$f_x(x)$
x_1	$f(x_1, y_1)$	$f(x_1, y_2)$	$\dots$	$f(x_1, y_L)$	$\sum_l f(x_1, y_l)$
x_2	$f(x_2, y_1)$	$f(x_2, y_2)$	$\dots$	$f(x_2, y_L)$	$\sum_l f(x_2, y_l)$
x_3	$f(x_3, y_1)$	$f(x_3, y_2)$	$\dots$	$f(x_3, y_L)$	$\sum_l f(x_3, y_l)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_J	$f(x_J, y_1)$	$f(x_J, y_2)$	$\dots$	$f(x_J, y_L)$	$\sum_l f(x_J, y_l)$
$f_y(y)$	$\sum_j f(x_j, y_1)$	$\sum_j f(x_j, y_2)$	$\dots$	$\sum_j f(x_j, y_L)$	1

bzw. für stetige Zufallsvariablen

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y)\, dx\, dy = 1$$

Gemeinsame Verteilungsfunktion: In analoger Weise ist auch die *gemeinsame Verteilungsfunktion* zweier diskreter Zufallsvariablen definiert,

$$F(x, y) = \Pr(X \leq x, Y \leq y)$$

definiert.

Sie gibt an, mit welcher Wahrscheinlichkeit die Zufallsvariable X *höchstens* den Wert x *und* die Zufallsvariable Y *höchstens* den Wert y annimmt.

Analog für stetige Zufallsvariablen

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(v, w) dv\, dw$$

3.5.1 Bedingte Wahrscheinlichkeitsfunktion (*Conditional Probability Density Function*)

Angenommen ein Zufallsexperiment erzeugt zwei Zufallsvariablen X und Y, und wir kennen bereits die Realisation von $X = x$, wissen aber noch nichts über Y. Erlaubt uns dies eine bessere Einschätzung der Wahrscheinlichkeiten für Y?

Kehren wir noch einmal zurück zu unserem früheren Beispiel mit dem zweifachen Münzwurf, wobei $X = 1$ wenn beim ersten Wurf ein Wappen geworfen wurde und Null sonst, und $Y = 1$ wenn bei beiden Würfen mindestens ein Wappen geworfen wurde. Die gemeinsame Wahrscheinlichkeitsfunktion mit den Randverteilungen haben wir bereits früher ermittelt und wird hier zur Bequemlichkeit noch einmal wiedergegeben

X\Y	0	1	$f_x(x)$
0	0.25	0.25	0.5
1	0	0.5	0.5
$f_y(y)$	0.25	0.75	1

Angenommen wir wissen, dass beim ersten Wurf eine Zahl geworfen wurde ($X = 0$), ändert dies unsere Einschätzung für die Wahrscheinlichkeit bei zwei Würfen mindestens ein Wappen zu werfen? Offensichtlich ja, denn wenn wir bereits mit dem ersten Wurf eine Zahl erhalten haben sind die beiden Ereignisse $\{(WZ)\}$ und $\{(WW)\}$ aus $\Omega = \{(ZZ), (ZW), (WZ), (WW)\}$ unmöglich! Mit diesem Vorwissen $X = 0$ ist die Wahrscheinlichkeit überhaupt kein Wappen zu werfen $Y = 0$ gleich 0.5, wir schreiben dies

$$\Pr(Y = 0|X = 0) = 0.5 \quad \text{bzw.} \quad \Pr(Y = 1|X = 0) = 0.5$$

und sagen, die Wahrscheinlichkeit für $Y = 0$ *gegeben* $X = 0$ ist 0.5, oder besser, die *bedingte Wahrscheinlichkeit* für $Y = 0$ *gegeben* $X = 0$ ist 0.5, und analog für $Y = 1$.

Wenn wir auf X konditionieren halten wir gewissermaßen die X bei den jeweiligen Ausprägungen 'fest', damit wir wieder Wahrscheinlichkeiten für Y bei $X = x$ erhalten müssen wir (zeilenweise) durch die Randwahrscheinlichkeiten $f_x(x)$ dividieren

Bedingte Wahrscheinlichkeit von Y für gegebene X:

X\Y	0	1	
0	0.5	0.5	1
1	0	1	1

Wenn wir bereits wissen, dass $X = 0$, dann ist die bedingte Wahrscheinlichkeit für $Y = 0$ gleich 0.5, d.h. $f(Y = 0|X = 0) = 0.5$.

Ebenso gut können wir die bedingte Wahrscheinlichkeit von X für gegebene Y berechnen, indem wir (spaltenweise) durch die Randwahrscheinlichkeiten $f_y(y)$ dividieren.

Bedingte Wahrscheinlichkeit von X für gegebene Y:

X\Y	0	1
0	1	1/3
1	0	2/3
$f_y(y)$	1	1

Gegeben $Y = 1$ ist die Wahrscheinlichkeit für $X = 0$ gleich 1/3, d.h. $f(X = 0|Y = 1) = 1/3$.

Etwas allgemeiner können wir für diskrete Zufallsvariablen schreiben

Bedingte Wahrscheinlichkeitsfunktion von Y:

$$f(y|X = x) = \Pr(Y = y|X = x) = \frac{f(x, y)}{f_x(x)} \quad \text{für } f_x(x) > 0$$

Bedingte Wahrscheinlichkeitsfunktion von X:

$$f(x|Y=y) = \Pr(X=x|Y=y) = \frac{f(x,y)}{f_y(y)} \quad \text{für } f_y(y) > 0$$

d.h. wir erhalten die bedingte Wahrscheinlichkeit indem wir die gemeinsame Wahrscheinlichkeit durch die entsprechende Randwahrscheinlichkeit dividieren (sofern diese positiv ist).

Warum? Wir können die bedingte Wahrscheinlichkeit $f(y|x=0)$ als eine *gewichtete* gemeinsame Wahrscheinlichkeit für $f(y, x=0)$ mit der Randwahrscheinlichkeit $f_x(x)$ als Gewicht vorstellen. Die Gewichtung ist erforderlich, damit die Summe der bedingten Wahrscheinlichkeiten wieder Eins ergibt ($\sum_y f(y|x=0) = 1$ und $\sum_y f(y|x=1) = 1$).

3.5.2 Stochastische (bzw. statistische) Unabhängigkeit

Zwei Zufallsvariablen X und Y sind stochastisch unabhängig, wenn

$$f(x,y) = f_x(x) f_y(y)$$

bzw. unter Verwendung der Definition der bedingten Dichte $f(y|X=x) = \frac{f(y,X=x)}{f_x(x)}$

$$f(y|X=x) = f_y(y)$$

Man beachte, dass für stetige Zufallsvariablen $f(x,y)$ oder auch $f(y|X=x)$ keine Wahrscheinlichkeit angibt (oder genauer, jedem Punkt ist die Wahrscheinlichkeit Null zugeordnet), trotzdem implizieren diese Bedingungen stochastische Unabhängigkeit.

Für diskrete Zufallsvariablen können wir auch anschaulicher schreiben

$$\Pr(X=x_j, Y=y_l) = \Pr(X=x_j)\Pr(Y=y_l)$$

Übung: Gegeben sei folgende diskrete Wahrscheinlichkeitsverteilung, berechnen Sie die bedingten Wahrscheinlichkeiten. Sind die bedingten Wahrscheinlichkeiten gleich den unbedingten Wahrscheinlichkeiten?

		Werte von Y		
		0	1	$f_x(x)$
Werte von X	1	0	1/3	1/3
	2	1/3	0	1/3
	3	0	1/3	1/3
	$f_y(y)$	1/3	2/3	1

3.6 Erwartungswerte (*'expected values'*)

Wahrscheinlichkeitsfunktionen sind wie normale Häufigkeitsverteilungen durch bestimmte Parameter charakterisiert. Die ersten zwei Momente sind der Erwartungswert $\mathrm{E}(X)$ und die Varianz $\mathrm{var}(X)$, häufig abgekürzt als μ und σ^2.

Der **Erwartungswert** einer Zufallsvariable ist die *mit den Eintrittswahrscheinlichkeiten gewichtete Summe aller möglichen Ausprägungen einer Zufallsvariable.*

$$\begin{aligned} \mathrm{E}(X) &= \sum_{j=1}^{J} x_j f(x_j) \qquad \text{für diskrete ZV} \\ \mathrm{E}(X) &= \int_{-\infty}^{\infty} x f(x)\, dx \quad \text{für stetige ZV} \end{aligned}$$

Achtung:

1. Beim Erwartungswert wird über *alle möglichen* Ausprägungen der Zufallsvariable aufsummiert, gewichtet mit den Wahrscheinlichkeiten.
2. Erwartungswerte beziehen sich niemals auf Realisationen (z.B. Stichprobenbeobachtungen), sondern auf Zufallsvariablen!
   Das Analogon für den Erwartungswert von Zufallsvariablen für Realisationen ist der Mittelwert.

Beispiel für diskrete ZV: Erwartungswert der Augenzahl beim Würfeln

$$\mathrm{E}(X) = \mu_X = 1 * \frac{1}{6} + 2 * \frac{1}{6} + 3 * \frac{1}{6} + 4 * \frac{1}{6} + 5 * \frac{1}{6} + 6 * \frac{1}{6} = 3.5$$

Man beachte, dass der Erwartungswert einer Zufallsvariable X, also $\mathrm{E}(X)$, *keine* Zufallsvariable ist. Jeder, der den Erwartungswert der Augenzahl berechnet wird auf das gleiche Ergebnis kommen (wenn er sich nicht verrechnet), da ist kein Zufallselement enthalten!
Um zu betonen, dass der Erwartungswert eine feste Zahl ist, wird er häufig mit μ bezeichnet (d.h. $\mathrm{E}(X) := \mu_X$); der Mittelwert einer Stichprobe wird als $\bar{x}$ geschrieben.
Für die Erwartungswerte von Zufallsvariablen gilt ebenso wie für Mittelwerte, dass die Summe der Abweichungen davon immer Null ist

$$\sum_j [x_j - \mathrm{E}(X)] f(x_j) = \sum_j x_j f(x_j) - \mathrm{E}(X) \sum_j f(x_j) = \mathrm{E}(X) - \mathrm{E}(X) = 0$$

das $\mathrm{E}(X)$ eine Konstante ist und $\sum_j f(x_j) = 1$.

Beispiel für stetige ZV: Erwartungswert der Dichtefunktion $f(x) = x^2/9$ für $0 \leq x \leq 3$:

$$\begin{aligned} \mathrm{E}(X) &= \int_0^3 x \left(\frac{1}{9} x^2\right) dx = \int_0^3 \frac{x^3}{9} dx \\ &= \left.\frac{x^4}{36}\right|_0^3 \\ &= \frac{81}{36} = \frac{9}{4} = 2.25 \end{aligned}$$

3.6.1 Rechnen mit Erwartungswerten

Erwartungswerte sind gewichtete Summen, deshalb kann mit dem Erwartungswertoperator $\mathrm{E}(\cdot)$ 'sehr ähnlich' gerechnet werden wie mit dem Summenzeichen.

- **Für eine Zufallsvariable X und $c \in \mathbb{R}$ gilt:**

$$\begin{aligned}
\mathrm{E}(c) &= c \qquad \text{für c = konst., weil } \sum f(x) = 1 \\
\mathrm{E}(cX) &= c\,\mathrm{E}(X) \qquad \text{für c const.} \\
\mathrm{E}\left[\mathrm{E}(X)\right] &= \mathrm{E}(X) \qquad \text{weil } \mathrm{E}(X) := \mu = \text{konst.} \\
\mathrm{E}\left[g(X)\right] &= \int_{-\infty}^{\infty} g(x) f(x)\, dx \qquad \text{für eine Funktion } g(\cdot)
\end{aligned}$$

  Warum?

  Weil z.B. $\mathrm{E}(c) = \sum_j c f(x_j) = c \sum_j f(x_j) = c$; $\mathrm{E}(cX) = \sum_j c x_j f(x_j) = c \sum_j x_j f(x_j) = c\,\mathrm{E}(X)$;
  $\mathrm{E}(X)$ ist eine Konstante, also ist $\mathrm{E}\left[\mathrm{E}(X)\right] = \sum_j \mathrm{E}(X) f(x_j) = \mathrm{E}(X) \sum_j f(x_j) = \mathrm{E}(X)$.

  Beispiel: Wenn X die Augenzahl eines fairen Würfels ist, wie groß ist der Erwartungswert von $g(X) = X^2$?

$$\mathrm{E}(X^2) = 1^2 \frac{1}{6} + 2^2 \frac{1}{6} + 3^2 \frac{1}{6} + 4^2 \frac{1}{6} + 5^2 \frac{1}{6} + 6^2 \frac{1}{6} = 15.1\dot{6}$$

  Natürlich ist $\mathrm{E}(X^2) = 15.1\dot{6} \neq [\mathrm{E}(X)]^2 = 3.5^2 = 12.25$.

- **Für eine diskrete Zufallsvariable X und $a, b \in \mathbb{R}$ gilt:**

$$\mathrm{E}(a + bX) = a + b\,\mathrm{E}(X)$$

  Beweis:

$$\begin{aligned}
\mathrm{E}(a + bX) &= \sum_{j=1}^{J} (a + b x_j) f(x_j) \\
&= \sum_j a f(x_j) + \sum_j b x_j f(x_j) \\
&= a \sum_j f(x_j) + b \sum_j x_j f(x_j) \\
&= a + b\,\mathrm{E}(X)
\end{aligned}$$

  Dies gilt analog auch für stetige Zufallsvariablen.

Bislang haben wir ausschließlich univariate Wahrscheinlichkeitsverteilungen untersucht. Nun wollen wir das Konzept für multivariate Fälle erweitern. Angenommen wir haben zwei Zufallsvariablen X und Y mit einer gemeinsamen Verteilung $f(x, y)$.

$$\begin{aligned} \mathrm{E}\left[g(X,Y)\right] &= \sum_{j=1}^{\infty}\sum_{l=1}^{\infty} g(x_j, y_l) f(x_j, y_l) \qquad \text{für diskrete ZV} \\ \mathrm{E}\left[g(X,Y)\right] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f(x,y)\, dx\, dy \quad \text{für stetige ZV} \end{aligned}$$

- **Der Erwartungswert einer Summe von Zufallsvariablen ist gleich der Summe der Erwartungswerte**, d.h.

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

Warum?

$$\begin{aligned} \mathrm{E}(X+Y) &= \sum_j \sum_l (x_j + y_l) f(x_j, y_l) = \sum_j \sum_l x_j f(x_j, y_l) + \sum_j \sum_l y_j f(x_j, y_l) \\ &= \sum_j \left(x_j \sum_l f(x_j, y_l) \right) + \sum_l \left(y_l \sum_j f(x_j, y_l) \right) \\ &= \sum_j x_j f_x(x_j) + \sum_l y_l f_y(y_l) \\ &= \mathrm{E}(X) + \mathrm{E}(Y) \end{aligned}$$

für $j = 1, \ldots, J$ und $l = 1, \ldots, L$.

Dies kann einfach verallgemeinert werden. Der Erwartungswert einer Linearkombination von Zufallsvariablen ist gleich der Linearkombination der Erwartungswerte, d.h. für $a_1, \ldots, a_J \in \mathbb{R}$ und $X_1, \ldots, X_J$ Zufallsvariablen: $\mathrm{E}(a_1 X_1 + a_2 X_2 + \cdots + a_J X_J) = a_1 \mathrm{E}(X_1) + a_2 \mathrm{E}(X_2) + \cdots + a_n \mathrm{E}(X_J)$, bzw.

$$\mathrm{E}\left(\sum_{j=1}^{J} a_j X_j \right) = \sum_{j=1}^{J} a_j \mathrm{E}(X_j)$$

- **Wenn X und Y zwei *stochastisch unabhängige* Zufallsvariablen sind, dann ist der Erwartungswert ihres Produktes gleich dem Produkt der Erwartungswerte**, d.h.

$$\boxed{\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)}$$

Actung, dies gilt *nur* für stochastisch unabhängige Zufallsvariablen!

Beweis:

$$\mathrm{E}(XY) = \sum_j \sum_l (x_j y_l) f(x_j, y_l)$$

wenn X und Y unabhängig sind gilt: $f(x_j, y_l) = f_x(x_j) f_y(y_l)$. Deshalb:

$$\mathrm{E}(XY) = \sum_j \sum_l x_j y_l f_x(x_j) f_y(y_l) = \sum_j x_j f_x(x_j) \sum_l y_l f_y(y_l) = \mathrm{E}(X)\,\mathrm{E}(Y)$$

Achtung: Der Erwartungswert einer Summe ist *immer* gleich der Summe der Erwartungswerte, hingegen ist der Erwartungswert eines Produktes im allgemeinen nur dann gleich dem Produkt der Erwartungswerte, wenn die Variablen stochastisch unabhängig sind!

Beispiel:

X\Y	0	1	$f_x(x)$
0	0.25	0.25	0.5
1	0	0.5	0.5
$f_y(y)$	0.25	0.75	1

$$\begin{aligned}
\mathrm{E}(X) &= 0 \times 0.5 + 1 \times 0.5 = 0.5 \\
\mathrm{E}(Y) &= 0 \times 0.25 + 1 \times 0.75 = 0.75 \\
\mathrm{E}(XY) &= 0 \times 0 \times 0.25 + 0 \times 1 \times 0.25 + \\
&\quad 1 \times 0 \times 0 + 1 \times 1 \times 0.5 = 0.5 \\
\Rightarrow \quad \mathrm{E}(XY) &\neq \mathrm{E}(X)\,\mathrm{E}(Y)
\end{aligned}$$

Wie wir schon früher gesehen haben sind diese beiden Zufallsvariablen nicht stochastisch unabhängig.

3.6.2 Varianz

Die **Varianz** σ_X^2 einer Zufallsvariablen X ist definiert als

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

d.h. für diskrete Zufallsvariablen $\sigma_X^2 = \sum_{j=1}^{J} (x_j - \mu)^2 f(x_j)$

Die Varianz kann auch folgendermaßen berechnet werden:

$$\begin{aligned}
\sigma_X^2 &= \mathrm{E}\left[X - \mathrm{E}(X)\right]^2 = \mathrm{E}(X-\mu)^2 = \mathrm{E}(X^2 - 2\mu X + \mu^2) \\
&= \mathrm{E}(X^2) - 2\mu\,\mathrm{E}(X) + \mu^2 = \mathrm{E}(X^2) - 2\mu\mu + \mu^2 \\
&= \mathrm{E}(X^2) - \mu^2 \\
&= \mathrm{E}(X^2) - \left[\mathrm{E}(X)\right]^2
\end{aligned}$$

also

$$\boxed{\mathrm{var}(X) := \sigma_X^2 = \mathrm{E}\left[X - \mathrm{E}(X)\right]^2 = \mathrm{E}(X^2) - \left[\mathrm{E}(X)\right]^2}$$

- **Für eine Zufallsvariable X und $a, b \in \mathbb{R}$ gilt:**

$$\boxed{\operatorname{var}(a + bX) = b^2 \operatorname{var}(X)}$$

Beweis:

$$\begin{aligned}
\operatorname{var}(a + bX) &= \mathrm{E}\left[(a + bX) - \mathrm{E}(a + bX)\right]^2 \\
&= \mathrm{E}\left[a + bX - a - b\,\mathrm{E}(X)\right]^2 \\
&= \mathrm{E}\left[b\,(X - \mathrm{E}(X))\right]^2 \\
&= b^2\,\mathrm{E}\left[X - \mathrm{E}(X)\right]^2 \\
&= b^2 \operatorname{var}(X)
\end{aligned}$$

Die Varianz einer Konstanten a ist natürlich immer Null, $\operatorname{var}(a) = 0$. Dies gilt auch für stetige Zufallsvariablen.

Beispiel: Gesucht ist die Varianz einer Zufallsvariablen, deren Dichtefunktion durch $f(x) = x^2/9$ für $0 \leq x \leq 3$ gegeben ist.

Wir verwenden $\operatorname{var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$ und berechnen zuerst $\mathrm{E}(X^2)$:

$$\begin{aligned}
\mathrm{E}(X^2) &= \int_0^3 x^2 \left(\frac{1}{9}\,x^2\right) dx = \int_0^3 \frac{x^4}{9} dx \\
&= \left.\frac{x^5}{45}\right|_0^3 \\
&= \frac{243}{45} = 5.4
\end{aligned}$$

Da $\mathrm{E}(X) = 9/4$ (siehe voriges Beispiel) ist $[\mathrm{E}(X)]^2 = (9/4)^2$. Wir erhalten also

$$\operatorname{var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 = \frac{243}{45} - \left(\frac{9}{4}\right)^2 = 0.34$$

3.6.3 Kovarianz

Die **Kovarianz** ist definiert als

$$\boxed{\operatorname{cov}(X, Y) := \mathrm{E}\left[[X - \mathrm{E}(X)]\,[Y - \mathrm{E}(Y)]\right] = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y)}$$

das zweite Gleichheitszeichen gilt weil

$$\begin{aligned}
\operatorname{cov}(X, Y) &= \mathrm{E}\left[[X - \mathrm{E}(X)]\,[Y - \mathrm{E}(Y)]\right] \\
&= \mathrm{E}\left[XY - Y\,\mathrm{E}(X) - X\,\mathrm{E}(Y) + \mathrm{E}(X)\,\mathrm{E}(Y)\right] \\
&= \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y)
\end{aligned}$$

- *Wenn X und Y zwei stochastisch unabhängige Zufallsvariablen sind, dann ist die Kovarianz zwischen X und Y immer gleich Null* ($\operatorname{cov}(X, Y) = 0$)).

Da bei stochastischer Unabhängigkeit gilt $\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)$ folgt dies unmittelbar aus obiger Definition.

Achtung: Eine Kovarianz von Null impliziert aber umgekehrt nicht stochastische Unabhängigkeit, wie man sich anhand des folgenden Beispiels für eine diskrete Wahrscheinlichkeitsverteilung verdeutlichen kann:

Beispiel: Gegeben sei folgende diskrete Wahrscheinlichkeitsverteilung:

		Werte von Y		
		0	1	$f(x)$
Werte von X	1	0	1/3	1/3
	2	1/3	0	1/3
	3	0	1/3	1/3
	$f(y)$	1/3	2/3	1

Die Kovarianz ist $\mathrm{cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y)$

$$\begin{aligned}
\mathrm{E}(XY) &= \sum_j \sum_l (x_j y_l) f(x_j, y_l) \\
&= 1*0*0 + 1*1*\frac{1}{3} + 2*0*\frac{1}{3} + 2*1*0 + 3*0*0 + 3*1*\frac{1}{3} \\
&= \frac{4}{3} \\
\mathrm{E}(X) &= 1*\frac{1}{3} + 2*\frac{1}{3} + 3*\frac{1}{3} = 2 \\
\mathrm{E}(Y) &= 0*\frac{1}{3} + 1*\frac{2}{3} = \frac{2}{3} \\
\mathrm{cov}(X,Y) &= \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y) = \frac{4}{3} - 2\,\frac{2}{3} = 0
\end{aligned}$$

Die Variablen X und Y sind offenbar nicht stochastisch unabhängig, da $f(x_j)f(y_l) \neq f(x_j, y_l)$. Trotzdem ist die Kovarianz Null! Kovarianzen messen nur die lineare Abhängigkeit!

- Wenn X und Y zwei Zufallsvariablen sind gilt:

  Die Varianz der Summe (bzw. Differenz) zweier Zufallsvariablen ist

$$\begin{aligned}
\mathrm{var}(X+Y) &= \mathrm{var}(X) + \mathrm{var}(Y) + 2\,\mathrm{cov}(X,Y) \\
\mathrm{var}(X-Y) &= \mathrm{var}(X) + \mathrm{var}(Y) - 2\,\mathrm{cov}(X,Y)
\end{aligned}$$

  Warum? Erinnern Sie sich, $(a \pm b)^2 = a^2 + b^2 \pm 2ab$

$$\begin{aligned}
\mathrm{var}(X-Y) &= \mathrm{E}\left[(X-Y) - \mathrm{E}(X-Y)\right]^2 \\
&= \mathrm{E}\left[(X - \mathrm{E}(X)) - (Y - \mathrm{E}(Y))\right]^2 \\
&= \mathrm{E}\left[X - \mathrm{E}(X)\right]^2 + \mathrm{E}\left[Y - \mathrm{E}(Y)\right]^2 - \\
&\quad 2\,\mathrm{E}\left[(X - \mathrm{E}(X)\right]\left[(Y - \mathrm{E}(Y)\right] \\
&= \mathrm{var}(X) + \mathrm{var}(Y) - 2\,\mathrm{cov}(X,Y)
\end{aligned}$$

Wenn X und Y stochastisch unabhängig sind ist $\text{cov}(X,Y) = 0$. deshalb gilt $\text{var}(X+Y) = \text{var}(X) + \text{var}(Y)$

- Die Varianz einer Summe ist die Summe der Varianzen plus zwei Mal die Summe aller Kovarianzen zwischen den Zufallsvariablen der ursprünglichen Summe (analog zu $(a+b+c)^2 = a^2+b^2+c^2+2ab+2ac+2bc$). Für insgesamt K Zufallsvariablen (mit $k = 1, \ldots, K$)

$$\text{var}\left(\sum_{k=1}^{K} X_k\right) = \sum_{k} \text{var}(X_k) + \sum_{k=1}^{K} \sum_{\substack{j=1 \\ j \neq k}}^{K} \text{cov}(X_k, X_j)$$

weil $\text{cov}(X_k, X_j) = \text{cov}(X_j, X_k)$.

Die positive Quadratwurzel der Varianz einer Zufallsvariablen X heißt **Standardabweichung**: $\sigma_X = +\sqrt{\text{var}(X)}$.

Übungsbeispiele:

1. Zeigen Sie, dass $\text{cov}[X, 2X] = 2\,\text{var}(X)$.
2. Zeigen Sie, dass $\text{cov}[X, (Y+Z)] = \text{cov}(X,Y) + \text{cov}(X,Z)$.
3. Zeigen Sie, dass für konstante a und b gilt
$$\text{cov}[X, (a+bX)] = b\,\text{var}(X)$$
4. Zeigen Sie, dass für konstante a_1, b_1, a_2 und b_2 gilt
$$\text{cov}[a_1 + b_1 X, a_2 + b_2 Y] = b_1 b_2\,\text{cov}(X,Y)$$

3.6.4 Korrelationskoeffizient

Die Kovarianz hängt von den Maßeinheiten der Variablen ab und ist deshalb manchmal schwierig zu interpretieren.

Der **Korrelationskoeffizient** (corr) hat diesen Problem nicht, er ist ist unabhängig von den zugrunde liegenden Maßeinheiten, in denen X und Y gemessen wurde:

$$\text{corr}(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\,\text{var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Der Korrelationskoeffizient ist eine dimensionslose Zahl, die immer zwischen Null und Eins liegt

$$-1 \leq \text{corr}(X,Y) \leq 1$$

Da die Varianzen immer positiv sind hat der Korrelationskoeffizient immer das gleiche Vorzeichen wie die Kovarianz.

Bei einem perfekten negativen linearen Zusammenhang nimmt er den Wert -1 an, bei einem perfekten positiven linearen Zusammenhang den Wert $+1$. Bei stochastischer Unabhängigkeit ist $\text{cov}(X,Y)$ Null, deshalb ist in diesem Fall auch der Korrelationskoeffizient Null.

Übung: Zeigen Sie, dass $\text{corr}[X,(a+bX)] = 1$, und $\text{corr}[X,(a-bX)] = -1$.

Beweis* für $-1 \leq \text{corr}(X,Y) \leq 1$.

Beginnen wir mit dem Vorzeichen der Varianz

$$\text{var}(X+bY) = \text{var}(X) + 2b\,\text{cov}(X,Y) + b^2\,\text{var}(Y) \geq 0$$

weil Varianzen nie negativ werden können.

Da dies für alle b gilt, muss es auch für ein spezielles b gelten. Der Trick besteht nun in der Wahl eines speziellen b, welches uns neue Einsichten liefert. Ein solches b ist

$$b = -\frac{\text{cov}(X,Y)}{\text{var}(Y)}$$

(mit $\text{var}(Y) > 0$), denn wenn wir dieses in die obige Varianz $\text{var}(X+bY)$ einsetzen folgt

$$\begin{aligned}
\text{var}(X+bY) &= \text{var}(X) - \frac{2\,\text{cov}(X,Y)}{\text{var}(Y)}\,\text{cov}(X,Y) + \frac{\text{cov}(X,Y)^2}{\text{var}(Y)^2}\,\text{var}(Y) \\
&= \text{var}(X) - \frac{2\,\text{cov}(X,Y)^2}{\text{var}(Y)} + \frac{\text{cov}(X,Y)^2}{\text{var}(Y)} \\
&= \text{var}(X) - \frac{\text{cov}(X,Y)^2}{\text{var}(Y)} \geq 0
\end{aligned}$$

Dies muss wieder größer gleich Null sein, da eine Varianz nicht negativ werden kann.

Daraus folgt aber

$$\text{var}(X)\,\text{var}(Y) \geq \text{cov}(X,Y)^2$$

oder

$$\frac{\text{cov}(X,Y)^2}{\text{var}(X)\,\text{var}(Y)} \leq 1$$

Die Wurzel des linken Ausdrucks ist der Korrelationskoeffizient $r = \text{corr}(X,Y) = \text{cov}(X,Y)/\sqrt{\text{var}(X)\,\text{var}(Y)}$, deshalb muss gelten

$$r = \text{corr}(X,Y) \leq |1|$$

bzw. $-1 \leq \text{corr}(X,Y) \leq 1$. Dies ist ein Spezialfall der Cauchy-Schwarz Ungleichung (siehe Appendix). ■

Weiters gilt für konstante a_1, b_1, a_2 und b_2, wenn $b_1 b_2 > 0$

$$\text{corr}(a_1 + b_1 X, a_2 + b_2 Y) = \text{corr}(X,Y)$$

und wenn $b_1 b_2 < 0$

$$\text{corr}(a_1 + b_1 X, a_2 + b_2 Y) = -\,\text{corr}(X,Y)$$

3.6.5 Bedingte Erwartungswerte

Die bedingten Erwartungswerte spielen in der Ökonometrie eine herausragende Rolle, da die gefitteten Werte $\hat{y}$ des Regressionsmodells als bedingte Erwartungswerte interpretiert werden können.

Der bedingte Erwartungswert einer Zufallsvariablen Y wird unter der Voraussetzung berechnet, dass noch zusätzliche Informationen über den Ausgang des zugrunde liegenden Zufallsexperiments verfügbar ist, z.B. dass $X = x$.

Im wesentlichen werden sie gleich berechnet wir die unbedingten Erwartungswerte, als gewichtete Summe über alle möglichen Ausprägungen von Y, aber als Gewichte dienen nun die *bedingten Wahrscheinlichkeiten* (bzw. Dichten für stetige ZV).

$$\begin{aligned} \mathrm{E}(Y|X=x) &= \sum_{j=1}^{J} y_j f(y_j|X=x) \qquad \text{für diskrete ZV} \\ \mathrm{E}(Y|X=x) &= \int_{-\infty}^{\infty} y f(y|X=x)dy \qquad \text{für stetige ZV} \end{aligned}$$

Sie können intuitiv als ein Analogon zu den bedingten Mittelwerten der deskriptiven Statistik angesehen werden.

Beispiel: Angenommen wir interessieren uns für die Einkommenssituation in Abhängigkeit vom Geschlecht. Dem Geschlecht X ordnen wir die Ausprägungen $X = 1$ für weiblich und $X = 0$ für männlich zu, und den Einkommen Y ordnen wir $Y = 0$ für 'niedrig', $Y = 1$ für 'mittel' und $Y = 2$ für 'hoch' zu.

Die bivariate Wahrscheinlichkeitsverteilung $f(x, y)$ sei

			Werte von X männl. 0	weibl. 1	$f(y)$
Werte von Y	niedrig	0	0.1	0.2	0.3
	mittel	1	0.4	0.1	0.5
	hoch	2	0.0	0.2	0.2
		$f(x)$	0.5	0.5	1

Die Wahrscheinlichkeit, dass eine Frau ein mittleres Einkommen hat, beträgt z.B. 0.1, d.h. $\Pr(Y = 1, X = 1) = 0.1$. $f(x)$ und $f(y)$ bezeichnet die Randwahrscheinlichkeiten.

Die **Randwahrscheinlichkeit** von Y ist z.B. (für m mögliche Ausprägungen)

$$f_y(y) = \Pr(Y = y) = \sum_{j=1}^{J} \Pr(X = x_j, Y = y)$$

Die unbedingten Erwartungswerte sind

$$\begin{aligned} \mathrm{E}(X) &= \sum_j \sum_l x_j f(x_j, y_l) = \sum_j x_j f_x(x_j) = 0.5 \\ \mathrm{E}(Y) &= \sum_l \sum_j y_l f(x_j, y_l) = \sum_l y_l f_y(y_l) = 0.9 \end{aligned}$$

Die **bedingte Wahrscheinlichkeit** von Y, gegeben $X = x$ ist

$$\Pr(Y = y|X = x) = \frac{\Pr(X = x, Y = y)}{\Pr(X = x)} \quad \text{bzw. einfacher} \quad f(y|x) = \frac{f(x, y)}{f(x)}$$

Die Wahrscheinlichkeit, dass eine Person ein niedriges Einkommen hat, gegeben diese Person ist ein Mann, beträgt z.B. $\Pr(Y = 0|X = 0) = 0.1/0.5 = 0.2$ oder 20%.

Die bedingte Wahrscheinlichkeitsverteilung für Y gegeben X ist für dieses Beispiel also

			$f(y\|X=0)$	$f(y\|X=1)$
	niedrig	0	0.2	0.4
Werte von Y	mittel	1	0.8	0.2
	hoch	2	0.0	0.4
			1	1

Den **bedingten Erwartungswert** von Y, gegeben X, erhalten wir, indem wir die Ausprägungen von Y mit den auf X *bedingten* Wahrscheinlichkeiten gewichten und aufsummieren

$$\mathrm{E}(Y|X = x) = \sum_{j=1}^{J} y_j \Pr(Y = y_j|X = x) = \sum_j y_j f(y_j|x)$$

Für obiges Beispiel:

$$\begin{aligned} \mathrm{E}(Y|X = 0) &= 0 \times 0.2 + 1 \times 0.8 + 2 \times 0.0 = 0.8 \\ \mathrm{E}(Y|X = 1) &= 0 \times 0.4 + 1 \times 0.2 + 2 \times 0.4 = 1 \end{aligned}$$

Die **bedingte Erwartungswertfunktion** (*Conditional Expectation Function*, CEF) von Y ordnet schließlich jeder Ausprägung von X den bedingten Erwartungswert von Y zu

$$\mathrm{E}(Y|X = x) = \begin{cases} 0.8 & \text{für } X = 0 \\ 1 & \text{für } X = 1 \end{cases}$$

Übung: Der bedingte Erwartungswert von X, gegeben Y, macht in diesem Beispiel inhaltlich nicht sehr viel Sinn, aber Sie können aber trotzdem versuchen ihn zu berechnen.

Zur Kontrolle

$$\mathrm{E}(X|Y = y) = \begin{cases} 2/3 & \text{für } Y = 0 \\ 0.2 & \text{für } Y = 1 \\ 1 & \text{für } Y = 2 \end{cases}$$

Die stochastische bedingte Erwartungswertfunktion

Wir haben bisher die bedingten Erwartungswerte für *gegebene* Ausprägungen von X berechnet, d.h. $\mathrm{E}(Y|X = x_j)$. Diese bedingten Erwartungswerte sind wie die unbedingten Erwartungswerte Konstante, d.h. deterministische Größen.

Können wir auch für ein 'beliebiges' X die bedingte Erwartung berechnen? Können wir auf eine Zufallsvariable konditionieren?

Die Theorie dahinter ist ziemlich komplex, u.a. weil den einzelnen Ausprägungen x_j unterschiedliche Wahrscheinlichkeiten zugeordnet sind (die Randwahrscheilichkeiten von X). Deshalb muss man auf die σ-Algebra des zugrunde liegenden Zufallsexperiments Bezug nehmen, doch man kann (mit Hilfe der Maßtheorie und des Satzes von Radon-Nikodým) zeigen, dass eine solche stochastische bedingte Erwartungswertfunktion existiert und einige für das Folgende wichtige Eigenschaften hat.

Um deutlich zu machen, dass hier auf eine Zufallsvariable konditioniert wird schreibt man für den stochastischen bedingten Erwartungswert häufig

$$\mathrm{E}(Y|\sigma(X))$$

Da wir in diesem Fall auf eine Zufallsvariable konditionieren ist auch der stochastische bedingte Erwartungswert eine Zufallsvariable!

Eigenschaften der bedingten Erwartungswertfunktion

Die folgenden Eigenschaften gelten auch für die stochastische bedingte Erwartungswertfunktion.

1. **Linearität:** Seien X, Y und Z Zufallsvariablen und $a, b \in \mathbb{R}$ Konstante

   $$\mathrm{E}(aX + bY|Z = z) = a\,\mathrm{E}(X|Z = z) + b\,\mathrm{E}(Y|Z = z)$$

2. **Das einfache Gesetz der iterierten Erwartungen:** Erinnern wir uns, in der deskriptiven Statistik haben wir gezeigt, dass die mit den Anteilen gewichtete Summe der bedingten Mittelwerte der unbedingte Mittelwert ist.

   Wenn z.B. wie in Tabelle 2.7 (Dummy Variablen) der Anteil von Männern und Frauen je 0.5 ist, und der durchschnittliche Stundenlohn von Männern 15 Euro und von Frauen 12.5 Euro ist, dann ist durchschnittliche Stundenlohn über alle Personen $0.5 \times 15 + 0.5 \times 12.5 = 13.75$.

   Ein analoges Gesetz gilt auch für Zufallsvariablen. Für zwei Zufallsvariablen Y und X gilt

   $$\mathrm{E}(Y) = \mathrm{E}_x[\mathrm{E}(Y|X = x)]$$

   d.h. *der (unbedingte) Erwartungswert der bedingten Erwartungswerte ist der unbedingte Erwartungwert* (E_x soll bedeuten, dass der äußere Erwartungswert über die X gebildet wird).

   Für obiges Beispiel haben wir bereits den unbedingten Erwartungswert $\mathrm{E}(Y) = 0.9$ und die bedingten Erwartungswerte $\mathrm{E}(Y|X = 0) = 0.8$ sowie

$\mathrm{E}(Y|X=1)=1$ berechnet. Außerdem haben wir auch die Randverteilungen $f_x(X=0)=0.5$ und $f_x(X=1)=0.5$.

Daraus folgt

$$\mathrm{E}(Y)=\mathrm{E}[\mathrm{E}(Y|X)]=\sum_j \mathrm{E}(Y|X=x_j)\,f(x_j)=0.8\times 0.5+1\times 0.5=0.9$$

bzw.

$$\begin{aligned}\mathrm{E}(X)=E_y[\mathrm{E}(X|Y)] &= \sum_j \mathrm{E}(X|Y=y_j)\,f(y_j)\\ &= 2/3\times 0.3+0.2\times 0.5+1\times 0.2=0.5\end{aligned}$$

Beispiel: Kehren wir nochmals zu dem früheren Beispiel zurück:

		Werte von X männl. 0	weibl. 1	$f_y(y)$
Werte von Y	0	0.1	0.2	0.3
	1	0.4	0.1	0.5
	2	0.0	0.2	0.2
	$f_x(x)$	0.5	0.5	1

Der unbedingte Erwartungswert von Y ist $\mathrm{E}(Y)=0\times 0.3+1\times 0.5+2\times 0.2=0.9$.

Die auf X bedingten Erwartungswerte von Y sind $\mathrm{E}(Y|X=0)=0.8$ und $\mathrm{E}(Y|X=1)=1$ (siehe oben).

Nach dem Gesetz der iterierten Erwartungen

$$\mathrm{E}(Y)=\mathrm{E}_x[\mathrm{E}(Y|X=x)]=0.5\times 0.8+0.5\times 1=0.9$$

Auch der $\mathrm{E}(X)=0.5$ kann so berechnet werden

$$\mathrm{E}(X)=\mathrm{E}_y[\mathrm{E}(X|Y=y)]=2/3\times 0.3+0.2\times 0.5+1\times 0.2=0.5$$

Exkurs:* Das Gesetz der iterierten Erwartungen für stetige Zufallsvariablen (Beweisskizze)

$$\begin{aligned}f(y|x) &= \frac{f(x,y)}{fx}\\ f(x,y) &= f(y|x)\,f_x(x)\\ \int f(x,y)\,dx=f_y(y) &= \int f(y|x)\,f_x(x)\,dx\end{aligned}$$

Zusammenhang zwischen Randdichte aus bedingter Dichte:

$$f_y(y)=\int f(y|x)f_x(x)\,dx$$

$$
\begin{aligned}
\mathrm{E}(Y) &= \int y f_y(y) dy \\
&= \int y \left[\int f(y|x) f_x(x)\, dx \right] dy \\
&= \int \int y f(y|x) f_x(x)\, dx\, dy \\
&= \int \underbrace{\int y f(y|x) dy}_{\mathrm{E}(Y|x)}\ f_x(x) dx \qquad \text{(Vertauschen von } dy \text{ und } dx\text{)} \\
&= \int \mathrm{E}(Y|x) f_x(x) dx \\
&= \mathrm{E}_x[\mathrm{E}(Y|X)]
\end{aligned}
$$

■

3. ***'Taking out what is known property':*** Seien $g(X)$ und $h(Y)$ Funktionen der Zufallsvariablen X, Y, dann gilt

$$\mathrm{E}[(g(X)h(Y)|X = x] = g(X)\,\mathrm{E}[h(Y)|X = x]$$

Intuitiv können wir uns vorstellen, dass durch die Konditionierung auf $X = x$ gewissermaßen X 'festgehalten' wird wird, und damit auch $g(X)$, weshalb $g(X)$ als Konstante vor den Erwartungswertoperator gezogen werden kann.

Als Spezialfall sehen wir uns $\mathrm{E}(XY|X = x)$ für diskrete X und Y an, wobei $j = 1, \ldots, J$ die Ausprägungen von X und $l = 1, \ldots, L$ die Ausprägungen von Y indexiert.

$$
\begin{aligned}
\mathrm{E}(XY|X = x_j) &= \sum_{l=1}^{L} x_j y_l \Pr(y_l, x_j|X = x_j) \\
&= x_j \sum_{l=1}^{L} y_l \Pr(y_l|X = x_j) \\
&= x_j\,\mathrm{E}(Y|X = x_j)
\end{aligned}
$$

Da wir die Zufallsvariable X bei der Ausprägung x_j 'festhalten' wird ist $\Pr(y_l, x_j|X = x_j) = \Pr(y_l|X = x_j)$, und weil die Summation über l läuft können wir x_j als Konstante vor das Summenzeichen ziehen (wir untersuchen XY für einen spezifischen Wert von X).

Beispiel: wir setzen mit dem vorhergehenden Beispiel fort

$\mathrm{E}(XY|X = 1) = 0 \times 1 \times 0.4 + 1 \times 1 \times 0.2 + 2 \times 1 \times 0.4 = 1.$

Mit der 'Taking out what is known property' für $x_j = 1$:

$x_j\,\mathrm{E}(Y|X = x_j) = 1 \times \mathrm{E}(Y|X = 1) = 1 \times 1 = 1$

Oder $\mathrm{E}(XY|Y = 1) = 1 \times 0 \times 0.8 + 1 \times 1 \times 0.2 = 0.2$

bzw. mit der 'Taking out what is known property' für $y_j = 1$:

$y_j\,\mathrm{E}(X|Y = y_j) = 1 \times 0.2 = 0.2$

Beispiel: diese Eigenschaft der bedingten Erwartungswertfunktion ist besonders für Regressionsfunktionen mit stochastischen X von Bedeutung (siehe Spanos 1999, 364f).

Für bivariat normalverteilte Zufallsvariablen X und Y kann man zeigen, die bedingte Erwartungswertfunktion immer linear ist

$$\mathrm{E}(Y|\sigma(X)) = \beta_1 + \beta_2 X$$

d.h. die bedingten Erwartungswerte liegen exakt auf einer Geraden. Für nicht normalverteilte Zufallsvariablen gilt dies manchmal zumindest approximativ.

Wir zeigen nun, dass in diesem Fall eine einfache Beziehung zwischen den Parametern β_1 und β_2 und den Momenten der gemeinsamen Verteilung von X und Y existiert.

Aufgrund des einfachen Gesetzes der iterierten Erwartungen gilt $\mathrm{E}_x(\mathrm{E}(Y|X)) = \mathrm{E}(Y) = \beta_1 + \beta_2 \mathrm{E}(X)$, oder

$$\beta_1 = \mathrm{E}(Y) - \beta_2 \mathrm{E}(X)$$

Außerdem folgt aus dem Gesetzes der iterierten Erwartungen und der *'taking out what is known property'*, dass

$$\mathrm{E}(XY) = \mathrm{E}[\mathrm{E}(XY|\sigma(X))] = \mathrm{E}[X\,\mathrm{E}(Y|\sigma(X))]$$

Einsetzen von $\beta_1 = \mathrm{E}(Y) - \beta_2 \mathrm{E}(X)$ gibt

$$\begin{aligned}
\mathrm{E}(XY) &= \mathrm{E}\left[X(\beta_1 + \beta_2 X)\right] \\
&= \mathrm{E}[X(\underbrace{\mathrm{E}(Y) - \beta_2 \mathrm{E}(X)}_{\beta_1} + \beta_2 X)] \\
&= \mathrm{E}\left\{X\,\mathrm{E}(Y) + \beta_2\left[X^2 - X\,\mathrm{E}(X)\right]\right\} \\
&= \mathrm{E}(X)\,\mathrm{E}(Y) + \beta_2\left[\mathrm{E}(X^2) - \mathrm{E}(X)\,\mathrm{E}(X)\right] \\
\underbrace{\mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y)}_{\mathrm{cov}(X,Y)} &= \beta_2[\underbrace{\mathrm{E}(X^2) - [\mathrm{E}(X)]^2}_{\mathrm{var}(X)}]
\end{aligned}$$

daraus folgt

$$\beta_2 = \frac{\mathrm{cov}(X,Y)}{\mathrm{var}(X)}$$

Man beachte, dass sich dies auf die PRF und die Momente der Grundgesamtheit bezieht, nicht auf die SRF!

Für die bivariate Wahrscheinlichkeitsverteilung

		Werte von X		
		männl.	weibl.	
		0	1	$f_y(y)$
Werte von Y	0	0.1	0.2	0.3
	1	0.4	0.1	0.5
	2	0.0	0.2	0.2
	$f_x(x)$	0.5	0.5	1

haben wir bereits die unbedingten Erwartungswerte $\mathrm{E}(Y) = 0.9$ und $\mathrm{E}(X) = 0.5$ sowie die bedingte Erwartungswertfunktion

$$\mathrm{E}(Y|X = x) = \begin{cases} 0.8 & \text{für } X = 0 \\ 1 & \text{für } X = 1 \end{cases}$$

berechnet.

Wir können nun auch die *lineare Approximation* an diese bedingten Erwartungswerte berechnen; dazu benötigen wir $\mathrm{E}(X^2) = 0^2 \times 0.5 + 1^2 \times 0.5 = 0.5$ und $\mathrm{E}(XY) = \sum_{j=1}^{2} \sum_{l=1}^{3} x_j y_l f(x_j, y_l) = 0 \times 0 \times 0.1 + 0 \times 1 \times 0.2 + 1 \times 0 \times 0.4 + 1 \times 1 \times 0.1 + 2 \times 0 \times 0 + 2 \times 1 \times 0.2 = 0.5$.

Also

$$\beta_2 = \frac{\mathrm{cov}(X,Y)}{\mathrm{var}(X)} = \frac{\mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y)}{\mathrm{E}(X^2) - [\mathrm{E}(X)]^2} = \frac{0.5 - 0.5 \times 0.9}{0.5 - 0.5^2} = \frac{0.05}{0.25} = 0.2$$

und

$$\beta_1 = \mathrm{E}(Y) - \beta_2\,\mathrm{E}(X) = 0.9 - 0.2 \times 0.5 = 0.8$$

Als lineare Approximation an die bedingte Erwartungswertfunktion erhalten wir also

$$\mathrm{E}(Y|\sigma(X)) \overset{\text{lin}}{\approx} 0.8 + 0.2X$$

Offensichtlich ist in diesem Fall die bedingte Erwartungswertfunktion tatsächlich linear, denn die lineare Approximation $\mathrm{E}(Y|X = 0) = 0.8 + 0.2 \times 0 = 0.8$ und $\mathrm{E}(Y|X = 1) = 0.8 + 0.2 \times 1 = 1$ liefert exakt die gleichen Werte wie früher, aber das ist natürlich ein Zufall, oder um den eingangs erwähnten Charlie Chan zu zitieren, *"Strange events permit themselves the luxury of occurring"*. □

4. Die bedingte Erwartungswertfunktion ist der beste *'mean squared errors'* Prediktor

$$\mathrm{E}\left[Y - \mathrm{E}(Y|\sigma(X))\right]^2 \leq \mathrm{E}\left[Y - g(X)\right]^2 \qquad \text{für alle } g(\cdot)$$

Die Distanz $\mathrm{E}\left[Y - g(X)\right]^2 < \infty$ heißt *'mean squared error'* (MSE). Von allen möglichen Funktionen $g(X)$ liefert der bedingte Erwartungswert $\mathrm{E}(Y|\sigma(X))$ den kleinsten MSE.

3.6.6 Bedingte Varianz

Neben den bedingten Erwartungswerten ist in der Ökonometrie v.a. die bedingte Varianz von Bedeutung. Sie ist definiert als der bedingte Erwartungswert der quadratischen Abweichung der Zufallsvariablen von ihrem bedingten Erwartungswert.

Die Berechnung erfolgt wieder analog zu normalen Varianzen, nur dass nun jeweils die bedingten Wahrscheinlichkeiten (bzw. bedingten Dichten) für die Gewichtung verwendet werden.

$$\begin{aligned}
\operatorname{var}(Y|X=x) &= \mathrm{E}\left\{[Y-\mathrm{E}(Y|X=x)]^2\,|X=x\right\} \\
&= \sum_y [Y-\mathrm{E}(Y|X=x)]^2 f(Y|X=x) \qquad \text{für diskrete ZV} \\
&= \int_{-\infty}^{+\infty} [Y-\mathrm{E}(Y|X=x)]^2 f(Y|X=x) \qquad \text{für stetige ZV}
\end{aligned}$$

Auch für bedingte Varianzen gilt:

- Nichtnegativität: $\operatorname{var}(Y|X=x) \geq 0$
- Lineare Transformationen: $\operatorname{var}(a+bY|X=x) = b^2 \operatorname{var}(Y|X=x)$ für $a, b \in \mathbb{R}$
- Verschiebungssatz: $\operatorname{var}(Y|X=x) = \mathrm{E}(Y^2|X=x) - [\mathrm{E}(Y|X=x)]^2$

Intuitiv können wir uns wieder vorstellen, dass es sich um die Varianz einer Untergruppe handelt, die durch die Eigenschaft $X=x$ definiert ist.

Beispiel Die bedingte Varianz von Y für $X=0$ ist für obiges Beispiel

$$\begin{aligned}
\operatorname{var}(Y|X=0) &= \sum_j (y_j - \mathrm{E}(Y|X=0)^2 \Pr(y_j|X=0) \\
&= (0-0.8)^2 \times 0.2 + (1-0.8)^2 \times 0.8 + (2-0.8)^2 \times 0.0 \\
&= 0.16
\end{aligned}$$

Die **bedingte Varianzfunktion** (*'scedastic function'*) ordnet jeder möglichen Ausprägungen von X die entsprechende bedingte Varianz zu.

Hinweis: Wenn X, Y bivariat normalverteilt sind ist die bedingte Varianzfunktion $\operatorname{var}(Y|X=x) = \sigma^2$, d.h. konstant und damit unabhängig von x, d.h. *homoskedastisch*.

Dies gilt fast nur für die Normalverteilung, für die allermeisten anderen theoretischen Verteilungen ändert sich die bedingte Varianz mit den x, d.h. die bedingte Varianzfunktion ist eine Funktion der x (vgl. **?**, 342fff). Dieser Fall wird *Heteroskedastizität* genannt und wird uns später noch ausführlich beschäftigen.

Beispiel Kehren wir nochmals zurück zum früheren Zahlenbeispiel

Für die bivariate Wahrscheinlichkeitsverteilung

		Werte von X männl. 0	weibl. 1	$f_y(y)$
Werte von Y	0	0.1	0.2	0.3
	1	0.4	0.1	0.5
	2	0.0	0.2	0.2
	$f_x(x)$	0.5	0.5	1

haben wir bereits die bedingte Erwartungswertfunktion berechnet

$$\mathrm{E}(Y|X=x) = \begin{cases} 4/5 & \text{für } X=0 \\ 1 & \text{für } X=1 \end{cases}$$

Die bedingten Varianzen sind

$$\begin{aligned} \mathrm{var}(Y|X=0) &= \mathrm{E}(Y^2|X=0) - [\mathrm{E}(Y|X=0)]^2 \\ &= \left[0^2 \times \frac{1}{5} + 1^2 \times \frac{4}{5} - 2^2 \times \frac{0}{5}\right] - \left[\frac{4}{5}\right]^2 = \frac{4}{25} = 0.16 \\ \mathrm{var}(Y|X=1) &= \mathrm{E}(Y^2|X=1) - [\mathrm{E}(Y|X=1)]^2 \\ &= \left[0^2 \times \frac{2}{5} + 1^2 \times \frac{1}{5} - 2^2 \times \frac{2}{5}\right] - [1]^2 = \frac{4}{5} = 0.8 \end{aligned}$$

Die bedingte Varianzfunktion ist also

$$\mathrm{var}(Y|X=x) = \begin{cases} 0.16 & \text{für } X=0 \\ 0.8 & \text{für } X=1 \end{cases}$$

Da in diesem Beispiel die bedingte Varianz davon abhängt, welche Ausprägung X annimmt, liegt Heteroskedastizität vor.

Übungsbeispiele:

1. Angenommen, ein eigenartiger Würfel mit 3 Seiten werde zweimal geworfen. Die Augenzahl sei 1,2 oder 3. Z_1 sei die Augenzahl des ersten Wurfes, und Z_2 die Augenzahl des zweiten Wurfes. Weiters sei $X = Z_1 + Z_2$, und $Y = Z_1 - Z_2$. (Lösungen ohne Gewähr!)

   (a) Berechnen Sie die gemeinsame Wahrscheinlichkeitsfunktion von X und Y sowie die Randverteilungen.
   Lösung:

X/Y	-2	-1	0	1	2	$f_x(x)$
2	0	0	1/9	0	0	1/9
3	0	1/9	0	1/9	0	2/9
4	1/9	0	1/9	0	1/9	3/9
5	0	1/9	0	1/9	0	2/9
6	0	0	1/9	0	0	1/9
$f_y(y)$	1/9	2/9	3/9	2/9	1/9	1

   (b) Berechnen Sie den Erwartungswert und die Varianz von Y.

   (c) Berechnen Sie die Kovarianz zwischen X und Y.

   (d) Sind X und Y statistisch unabhängig? (Lösg.: nein)

   (e) Berechnen Sie den bedingten Erwartungswert von Y für $X = 3$. (Lösg.: 0)

   (f) Berechnen Sie den bedingten Erwartungswert von X für $Y = 0$. (Lösg.: 4)

2. Gegeben sei folgende diskrete Wahrscheinlichkeitsverteilung:

		Werte von Y		
		1	3	9
Werte von X	2	1/8	1/24	1/12
	4	1/4	1/4	0
	6	1/8	1/24	1/12

   (a) Berechnen Sie die Randverteilungen von X und Y (d.h. die *marginal probability density functions*).

   (b) Berechnen Sie die bedingte Wahrscheinlichkeitsfunktion von Y für $X = 2$ (d.h. die *conditional probability density function*).

   (c) Berechnen Sie die Kovarianz zwischen X und Y.

   (d) Sind X und Y statistisch unabhängig?

   (aus GHJ, S. 59) *Einige Lösungen:* $\Pr(X = 2 \text{ und } Y = 9) = 2/24$, $\Pr(X = 2|Y = 9) = \frac{2/24}{4/24} = \frac{1}{2}$, $E(X) = 4$, $\mathrm{E}(Y) = 3$, $\mathrm{E}(Y|X = 2) = 1\times\frac{1}{2}+3\times\frac{1}{6}+9\times\frac{1}{3} = 4$, $\mathrm{var}(X) = 2$, $\mathrm{var}(Y) = 8$, $E(XY) = 12$, $\mathrm{cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y) = 12 - 4\times 3 = 0$ (!), $\mathrm{var}(Y|X = 2) = \mathrm{E}(Y^2|X = 2) - (\mathrm{E}(Y|X = 2))^2 = 29 - 16 = 13$.

$$\mathrm{var}(Y|X) = \begin{cases} 13 & \text{für } X = 2 \\ 1 & \text{für } X = 4 \\ \dots & \text{für } X = 6 \end{cases}$$

3. Eine Zufallsvariable sei gleichverteilt im Intervall $[0, 1] : X \sim G(0, 1)$, d.h. die Dichtefunktion lautet $f(X) = 1$.

   (a) Wie lautet die zugehörige Verteilungsfunktion?

   (b) Berechnen Sie $\Pr(0.1 \leq X \leq 0.9)$

   (c) Berechnen Sie $\mathrm{E}[X]$

   (d) Berechnen Sie $\mathrm{var}[X]$

   (e) Berechnen Sie $\mathrm{E}[a + bX]$

   (f) Berechnen Sie $\mathrm{var}[a + bX]$.

3.A Appendix

3.A.1 Cauchy-Schwarz Ungleichung

Seien X und Y zwei Zufallsvariablen, dann ist die Cauchy-Schwarz Ungleichung

$$|\,\mathrm{E}(XY)| \leq \sqrt{\mathrm{E}(X^2)\,\mathrm{E}(Y^2)}$$

Für den Beweis definieren wir $W = Y + bX$, wobei b eine Konstante ist. Dann ist

$$\mathrm{E}(W^2) = \mathrm{E}(Y^2) + 2b\,\mathrm{E}(XY) + b^2\,\mathrm{E}(X^2) \geq 0$$

da $W^2 \geq 0$ und deshalb auch $\mathrm{E}(W^2) \geq 0$. Dies muss für jedes b gelten, also z.B. auch für

$$b = -\frac{\mathrm{E}(XY)}{\mathrm{E}(X^2)}$$

(Erwartungswerte sind deterministische Größen). Einsetzen gibt

$$\begin{aligned}
\mathrm{E}(W^2) &= \mathrm{E}(Y^2) - \left(2\frac{\mathrm{E}(XY)}{\mathrm{E}(X^2)}\right)\mathrm{E}(XY) + \left(\frac{\mathrm{E}(XY)}{\mathrm{E}(X^2)}\right)^2 \mathrm{E}(X^2) \\
&= \mathrm{E}(Y^2) - \frac{2[\mathrm{E}(XY)]^2}{\mathrm{E}(X^2)} + \frac{[\mathrm{E}(XY)]^2}{\mathrm{E}(X^2)} \\
&= \mathrm{E}(Y^2) - \frac{[\mathrm{E}(XY)]^2}{\mathrm{E}(X^2)} \geq 0
\end{aligned}$$

Achtung, wir benötigen nur $\mathrm{E}(W^2) \geq 0$, dies gilt für jedes b, also auch für dieses spezielle b!

Deshalb gilt

$$\mathrm{E}(X^2)\,\mathrm{E}(Y^2) \geq [\mathrm{E}(XY)]^2$$

bzw.

$$|\,\mathrm{E}(XY)| \leq \sqrt{\mathrm{E}(X^2)\,\mathrm{E}(Y^2)}$$