Einführung

in die

elementare Stochastik

1 Einleitung

Die Bezeichnung Stochastik stammt vom griechischen Wort "$\sigma\tau\omega\chi\omega\sigma$" ab. Dieses bedeutet soviel wie "clever vermuten" oder "intelligent vermuten". Es steht also zu vermuten, dass Stochastik sich mit der Analyse von Phänomenen befasst, deren Ausgang zumindest dem Beobachter nicht gewiss ist, der vom Zufall abhängig ist.

Insgesamt zerfällt die Stochastik in zwei große Teilgebiete, die Wahrscheinlichkeitstheorie und die Statistik, von denen sich die letzte wieder in die beschreibende und die schließende Statistik gliedert. Die Rolle von Wahrscheinlichkeitstheorie und Statistik sind hierbei gewissermaßen komplementär (oder "dual" wie der Mathematiker sagt): Während für ein typisches Problem der Wahrscheinlichkeitstheorie die Situation einschließlich des zugrunde liegenden Zufallsmechanismus vollständig beschrieben ist und die Aufgabe darin liegt, das langfristige Verhalten einer großen Stichprobe, die nach diesem Modell gezogen wird, vorherzusagen, ist das Problem in der Statistik gerade umgekehrt. Hier ist eine (hoffentlich große) Stichprobe gegeben und die Aufgabe besteht darin zu entscheiden, welches der zugrunde liegende Zufallsmechanismus ist.

Beiden Fragestellungen ist offenbar eine Modellierung des Zufalls und somit der Begriff der Wahrscheinlichkeit zu eigen. Interessanterweise ist dabei nicht vollständig klar, was "Wahrscheinlichkeit" im philosophischen Sinne eigentlich bedeuten soll (wir werden in Kürze sehen, dass diese Frage mathematisch weniger virulent ist: Wir geben eine Definition, die für unsere mathematische Arbeit vollständig ausreichend ist, das Grundproblem dabei aber unberücksichtigt lässt). Der intuitive Begriff der Wahrscheinlichkeit, der besagt, dass für ein Ereignis mit Wahrscheinlichkeit $1/2$ etwa gilt, dass in den allermeisten Versuchsreihen auf lange Sicht das Ereignis in der Hälfte aller Fälle auftauchen wird und in der anderen Hälfte nicht, ist zwar unter den richtigen Voraussetzungen mathematisch wahr (wir werden dieses Phänomen unter den Namen "Gesetz der großen Zahlen" später kennenlernen), taugt aber nicht für eine mathematische Definition: Wie sollte man bei einer Rechnung "für die allermeisten Versuchsreihen" und "auf lange Sicht" formalisieren? Und wie ließe sich mit einer solchen Formalisierung rechnen? Umgekehrt ist es auch schwer zu sagen, was die Wahrscheinlichkeit eines Individualereignisses sein soll. Wenn der Wetterbericht meldet, dass die Regenwahrscheinlichkeit morgen bei 80 % liegt, was genau meint er damit? Der morgige Tag wird nur einmal eintreten und an ihm wird es regnen oder nicht. Wir haben also keine Möglichkeit, unsere relative Häufigkeitsinterpretation der Wahrscheinlichkeit sinnvoll anzuwenden. Umgekehrt werden wir übermorgen wissen, ob es morgen geregnet haben wird – das Ereignis "morgen wird es regnen" ändert somit seine Warhscheinlichkeit; dies ist etwas, was im Rahmen der üblichen Wahrscheinlichkeitstheorie eher unerwünscht ist (auch Naturgesetze sollten ja mit der zeitlichen Veränderung unterliegen).

Wie wir im Laufe dieses Kurses kennenlernen werden, sind diese eher metaphysischen Betrachtungen für die mathematische Theorie der Wahrscheinlichkeit nicht von Belang. Ob etwas tatsächlich zufällig ist, wie sich der Zufallsmechanismus in eine Situation "einschleicht", oder ob Wahrscheinlichkeit nur das eigene Unwissen modelliert (wie beispiels-

weise in verschiedenen statistischen Fragestellungen) oder dem Ereignis inherent ist (wie das heutzutage für die Quantenmechanik angenommen wird).

Insgesamt gliedert sich der vorliegende Kurs in drei Teile. Der erste, kürzere Teil ist der deskriptiven Statistik gewidmet. Hier geht es vorrangig um die Frage, wie man erhobene Daten auf sinnvolle und ansprechende Weise darstellen kann. Hierbei geben heutzutage auch viele Computerprogramme Hilfestellungen. Trotzdem oder gerade deshalb finden sich gegenwärtig auch in beinahe jeder Publikation mit statistischen Auswertungen absichtliche oder unabsichtliche Fehler. In einem gesonderten Abschnitt werden wir auf solche Fehlerquellen eingehen.

Im zweiten Teil werden wir uns der Wahrscheinlichkeitsrechnung widmen. Hierbei gehen wir vor allem auf die sogenannte diskrete Wahrscheinlichkeitstheorie ein, also diejenige, bei der der Grundraum höchstens abzählbar ist. Eine naheliegende Wahrscheinlichkeitsannahme dort, die sogenannte Laplace-Wahrscheinlichkeit, bringt es mit sich, dass wir die Mächtigkeit gewisser Teilmengen abzählen müssen. Dies gehört eigentlich in eine andere mathematische Teildisziplin, die Kombinatorik; wir werden uns aber im Rahmen dieser Vorlesung dennoch damit beschäftigen, um bestimmte interessante Wahrscheinlichkeiten berechnen zu können.

Der dritte Teil dieser Vorlesung ist der sogenannten schließenden Statistik gewidmet. Dabei geht es darum, aufgrund einer Beobachtung (d. h. einer Stichprobe von möglichst großem Umfang) auf das zugrunde liegende Modell zu schließen. Diese Situation tritt in der Praxis häufig auf. Beispielsweise möchte ein Meinungsforschungsinstitut vor einer Wahl das Wahlergebnis möglichst gut vorhersagen. Da man aber schlecht alle Wahlberechtigten fragen kann, muss man sich auf eine Stichprobe beschränken. Wir werden im dritten Kapitel auf die Möglichkeiten und Risiken einer Schlussfolgerung aufgrund eines solchen Datensatzes eingehen.

2 Beschreibende Statistik

In diesem Kapitel geht es um die Darstellung der Daten aus einer statistischen Analyse. Hierbei können wir uns nicht damit befassen, wie diese Daten erhoben werden, obschon schon bei diesem Prozess viele Fehler auftreten können. Beispielsweise können die Messinstrumente eines Physikers systematisch falsch justiert sein oder ein Soziologe kann Fragen stellen, die ihm keine richtigen Ergebnisse liefern können. Ein bekanntes Beispiel für die letzte Situation ist ein Fragebogen zum Sexualverhalten heterosexueller Mitteleuropäer. Die Frage nach der Anzahl verschiedener Sexualpartner im Leben ergab, dass Frauen durchschnittlich drei verschiedene Sexualpartner in ihrem Leben haben, während Männer deren sieben haben (?!).

Unsere Arbeit als Statistiker beginnt in dem Moment, in dem wir den Datensatz zur Verfügung gestellt bekommen (dies schließt allerdings nicht aus, dass der Statistiker dem "Experimentator" beim sinnvollen Zusammenstellen des Experiments hilfreich zur Seite steht – diese Forschungsrichtung läuft unter dem Schlagwort "Experimental Design"). Als erstes klassifizieren wir hierbei die eintreffenden Daten. Dabei nehmen wir stets an, dass unsere *Untersuchungseinheit* ω aus einer *Grundgesamtheit* Ω zufällig entnommen ist. Was Ω ist, wird dabei durch unsere Untersuchung bestimmt:

Beispiel 2.1

- *Untersuchen wir die sozialen Verhältnisse der Einwohner Deutschlands, so besteht unsere Grundgesamtheit Ω aus allen Einwohnern Deutschlands; eine Untersuchungseinheit $\omega \in \Omega$ ist dann ein Bürger Deutschlands.*
- *Wollen wir zur Konzeption der Klausur "Stochastik für Studierende des Lehramts GHR" eine Untersuchung über die Studierenden dieser Vorlesung durchführen, so besteht die Grundgesamtheit Ω aus allen Studierenden dieser Vorlesung. $\omega \in \Omega$ ist ein zufällig ausgewählter Studierender dieser Vorlesung.*

Nun besteht das Erheben einer Statistik ja nicht allein im zufälligen Auswählen eines $\omega \in \Omega$, sondern auch darin, Daten dieses ω zu erfassen. Mit anderen Worten erfassen wir *Merkmale*. Die Funktion, die ω diese Merkmalsausprägung x zuordnet, bezeichnen wir oft mit X. Es ist also

$$X : \Omega \to S,$$

wobei S die Menge aller möglichen Merkmalsausprägungen bezeichnet. Hat ω die Merkmalsausprägung x, so schreiben wir auch

$$X(\omega) = x.$$

Beispiel 2.2

- *Wollen wir die Altersverteilung in Deutschland feststellen, so wäre Ω gerade die Menge aller Einwohner Deutschlands, $S = \mathbb{N}_0$ und*

$$X : \Omega \to \mathbb{N}_0$$

würde $\omega \in \Omega$ sein Alter zuweisen.

- *Sind wir an der Augenfarbe der Studierenden dieser Vorlesung interessiert, so ist*

$$\begin{aligned} \Omega &= \{\omega : \omega \textit{ ist Studierende(r) dieser Vorlesung}\} \\ S &= \{\textit{blau, grün, braun, grau,} \ldots\} \end{aligned}$$

  und $X : \Omega \to S$ *weist* ω *seine Augenfarbe zu.*

Es sei noch bemerkt, dass in vielen Untersuchungen gleich mehrere Merkmale gleichzeitig erhoben werden: $X(\omega)$ kann also durchaus aus $(X_1(\omega), \ldots, X_d(\omega))$ bestehen.

2.1 Datentypen

Wie schon Beispiel 2.2 zeigt, gibt es verschiedene Datentypen: Das Alter eines Menschen ist durch eine Zahl beschreibbar und mit diesen Zahlen lässt sich auch sinnvoll rechnen; beispielsweise hat jemand, der 50 Jahre alt ist, doppelt so lange gelebt wie jemand, der 25 Jahre alt ist. Die Augenfarbe eines Menschen hingegen ist keine Zahl; man kann ihr eine Zahl zuordnen, etwa

$$\text{blau} \mathrel{\hat{=}} 1, \quad \text{grün} \mathrel{\hat{=}} 2, \ldots.$$

Allerdings kann man mit den derart entstehenden Zahlen nicht sinnvoll rechnen, grüne Augen sind nicht etwa doppelt so gut wie blaue Augen. Dementsprechend unterscheiden wir in *qualitative Merkmale* und *quantitative Merkmale*.

Qualitative Merkmale sind solche, die sich durch ihre verschiedenen Ausprägungen charakterisieren lassen. Diese Ausprägungen sind i. a. keine Zahlen, ihnen können aber Zahlen zugeordnet werden, mit denen sich dann aber nicht sinnvoll rechnen lässt.

Quantitative Merkmale sind messbar und können durch Zahlen (mit denen dann auch gerechnet werden kann) erfasst werden.

Beispiel 2.3 *Qualitative Merkmale sind u. a.: Augenfarbe, Geschlecht, Wohnort, mathematische Kenntnisse, Schulnoten(!), ...*
Quantitative Merkmale sind hingegen: Schuhgrößen, Semesterzahl, Alter einer Person, Körpergröße, ...

Hierzu noch zwei Anmerkungen: Zum einen soll bemerkt werden, dass Schulnoten in der Tat qualitative Merkmale sind – sie sind zwar durch Zahlen bezeichnet, aber jemand, der eine 4 schreibt, ist ja nicht etwa genau so gut wie jemand, der zwei 2en geschrieben hat. Dies impliziert aber – wir werden später noch einmal darauf zurückkommen – dass beispielsweise die übliche Mittelwertbildung bei Zensuren nicht zulässig ist. Zum zweiten gibt es noch eine weitere Einteilung von quantitativen Größen in diskrete und stetige Größen. Hierbei nehmen diskrete Größen ihre Werte in einer höchstens abzählbaren Menge an, etwa $\mathbb{N}$, während stetige Größen überabzählbar viele Werte annehmen können, beispielsweise ist ihr Wertebereich ein Intervall $[a, b] \neq \emptyset$ oder $\mathbb{R}$. Beispiele für diskrete Größen

sind Schuhgrößen oder Semesterzahl, während stetige Größen Alter oder Körpergröße sind. Da man aber in der Praxis auch bei stetigen Größen nur selten beliebige Präzision bei deren Angabe vorfindet (wer gibt schon seine Größe als einen Dezimalbruch der Form 1,817253649... m an?), werden wir auf diese Unterscheidung verzichten.

Dagegen ist eine weitere Unterscheidung der qualitativen Merkmale wichtig. Selbst wenn man mit qualitativen Größen nicht rechnen kann, so gibt es doch solche darunter, die eine Rangordnung erlauben und andere, wo dies nicht der Fall ist. Beispielsweise ist bei Schulnoten eine 1 besser als eine 2, die ist besser als eine 3 usw. Hingegen ist es bei Augenfarben unsinnig darüber zu sprechen, dass beispielsweise blau besser oder schlechter ist als grün. Entsprechend unterscheiden wir die folgenden Skalen:

Nominalskala: Die Ausprägungen nominal skalierter Merkmale können nicht geordnet werden. Der einzig mögliche Vergleich ist der Test auf Gleichheit der Merksmalsausprägungen zweier Untersuchungsgrößen.

Ordinal- oder Rangskala: Die Merkmalsausprägungen können geordnet werden. Eine Interpretation der Rangordnung ist möglich, eine Interpretation der Abstände ist aber nicht möglich.

Metrische Skala: Unter den Merkmalsausprägungen gibt es eine Rangordnung. Zusätzlich können auch die Abstände zwischen den Merkmalsausprägungen gemessen und interpretiert werden. Metrisch skalierte Merkmale können noch weiter unterteilt werden in

- Intervallskala: Es sind nur Differenzenbildungen zwischen den Merkmalsausprägungen zulässig; diese erlauben die Abstände zu bestimmen.
- Verhältnisskala: Es existiert zudem ein natürlicher Nullpunkt. Quotientenbildung ist zulässig, Verhältnisse sind sinnvoll interpretierbar.
- Absolutskala: Zusätzlich zur Verhältnisskala kommt eine natürliche Einheit hinzu.

Beispiel 2.4 *1. Das Merkmal "Farbe" ist nomimal skaliert (wie schon besprochen).*

2. *Das Merkmal "Schulnote" ist ordinal skaliert.*
3. *Das Merkmal "Temperatur" ist metrisch skaliert. Der Unterschied zwischen 30 Grad Celsius und 20° C ist derselbe wie der zwischen 20° C und 10° C. Es ist aber Unsinn zu sagen, 20° C sei doppelt so warm wie 10° C.*
4. *Das Merkmal "Geschwindigkeit" ist ebenfalls metrisch skaliert. Zusätzlich zu Geschwindigkeitsdifferenzen lassen sich auch Geschwindigkeitsverhältnisse messen: 30 km/h ist doppelt so schnell wie 15 km/h. 0 km/h ist ein natürlicher Nullpunkt.*
5. *Das Merkmal "Semesterzahl" ist ebenfalls metrisch skaliert. Da man es am besten in natürlichen Zahlen misst, liegt es auf einer Absolutskala.*

2.2 Datenpräsentation

In diesem Kapitel wollen wir uns damit befassen, wie wir den erhobenen Datensatz am besten darstellen. Das einfachste Instrument hierbei ist die *Urliste*. Die Urliste enthält alle erhobenen Daten.

Beispiel 2.5 *Beim Weitsprung einer 5. Klasse wurden von jedem Teilnehmer das Geschlecht und die Weite erfasst. Die entsprechende Urliste hat die folgende Gestalt:*

$$\begin{aligned}\mathbb{L} \;=\; &\{(w;3,51),(w;2,75),(w;3,06),(m;4,37),(m;3,52),(w;3,99),(m;3,12),(w;2,47),\\ &(m;3,38),(w;3,90),(w;2,98),(m;2,81),(m;4.07),(w;4,01),(m;3,74),(m;3,56)\}.\end{aligned}$$

Schon dieses Beispiel macht Vor- und Nachteile der Urliste deutlich. Einerseits ist die Urliste die Quelle, die die Gesamtheit der erhobenen Daten am besten erfasst – es sind einfach alle erhobenen Daten in ihr enthalten. Andererseits verliert man bei größer werdenden Datenmengen schnell den Überblick; wer könnte beispielsweise anhand der Liste $\mathbb{L}$ auf den ersten Blick sagen, welche Schulnoten man für welche Leistung geben wollte/sollte?

Formen, die Daten graphisch besser zusammenzufassen, sind

- *Stabdiagramm (nominale Daten) und Histogramm (metrische Daten)*
  Stabdiagramm und Histogramm sind zwei gängige Arten (eindimensionale) Daten darzustellen. Hierzu werden die Daten in k Klassen eingeteilt und deren absolute Häufigkeiten werden zu einem Diagramm über den Klassen abgetragen. Dies ergibt das Stabdiagramm. Hierbei sollte die Klasseneinteilung möglichst äquidistant sein (nicht zu verwechseln mit Klassen gleicher Stichprobengröße). Die Länge der "Stäbe" im Stabdiagramm ist also proportional zu den absoluten (oder relativen) Häufigkeiten der Klassen. Für das Histogramm trägt man über den Klassen Rechtecke ab, die *flächenproportional* zu den absoluten (oder relativen) Häufigkeiten der Klassen sind.

  Beispiel 2.6 *In Beispiel 2.5 wählen wir zur Darstellung der "Weite" der einzelnen Teilnehmer die 4 Klassen*

  $$\begin{aligned}L_1 &:= \{x : 2,00 \le x \le 3,00\}\\ L_2 &:= \{x : 3,01 \le x \le 3,50\}\\ L_3 &:= \{x : 3,51 \le x \le 4,00\} \quad \textit{und}\\ L_4 &:= \{x : 4,01 \le x \le 4,50\}.\end{aligned}$$

Das Stabdiagramm hätte dann folgendes Aussehen:

Das Histogramm sieht folgendermaßen aus:

Bei beiden Arten der Darstellung erhebt sich die Frage, wieviele Klassen man wählen soll: Wählt man zuviele, so hat man nur wenig gegenüber der Urliste gewonnen. Wählt man zu wenige, so verliert man im Extremfall alle Information (wenn man nur eine Klasse konstruiert). Gängig ist die Sturge'sche Regel, die vorschlägt, dass

$$\text{für sehr große } n \;\; k \approx \log_2 n \quad \text{bzw.} \quad k \approx \sqrt{n} \text{ für kleinere } n$$

gelten soll, wobei n der Stichprobenumfang und k die Anzahl der Klassen ist. An diese Regel haben wir uns in Beispiel 2.6 gehalten (denn $4 = \log_2 16$).

- *Kreisdiagramm*

  Das Kreisdiagramm eignet sich besonders zur Darstellung relativer Häufigkeiten und Prozentsätze bei einer kleinen Anzahl von Klassen: Hierzu wird der Kreis derart in Segmente unterteilt, dass die Flächen proportional sind zu den relativen Häufigkeiten der Klassen.

Beispiel 2.7 *Mit den Daten der Liste* $\mathbb{L}$ *aus Beispiel 2.5 und den Klassen* $L_1, \ldots, L_4$

aus Beispiel 2.6 ergibt sich folgendes Kreisdiagramm:

Ein weiteres Mittel, um unsere Daten graphisch aufzubereiten, können wir erst besprechen, wenn wir Begriffe wie Median, Quartil, Interquartilabstand, usw. geklärt haben. Dies geschieht in den folgenden beiden Abschnitten.

2.3 Lagemaße

Im folgenden werden wir nicht versuchen, die Daten graphisch darzustellen, sondern ihnen Kennzahlen zuzuordnen, die sie möglichst gut beschreiben sollen. Die Lagemaße sind dabei solche Kennzahlen, die beschreiben, wo sich die Daten im Parameterbereich aufhalten. Was dies genau bedeutet, wird deutlicher, wenn wir sie einzeln studieren. Wir gehen hierbei immer von Daten (einer Stichprobe) $(X_i)_{i=1,\ldots,n}$ aus. Der wohl bekannteste Lageparameter ist der folgende:

Empirischer Mittelwert
Der empirische Mittelwert der Daten $x_i,\ i = 1, \ldots, n$, ist definiert als

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

Diese Größe ist sicher den meisten bekannt. Dennoch sei hier eine kleine Zwischenbemerkung eingeschoben, die oft übersehen wird. Die Bildung von $\bar{x}$ erfordert nämlich, dass man die Größen $x_1, \ldots, x_n$ addieren darf. Dazu müssen diese auf einer *metrischen* Skala liegen. Strenggenommen ist die Durchschnittsbildung $\bar{x}$ also beispielsweise für Schulnoten unsinnig (was niemanden davon abhält, dies dennoch zu praktizieren).

Beispiel 2.8 *Für die Stichprobe 1, 3, 5, 7, 9 ist* $\bar{x} = \frac{1}{5}(1 + 3 + 5 + 7 + 9) = 5$. *Für die Stichprobe 1, 3, 5, 5, 7, 9 ist* $\bar{x}$ *auch gleich 5, während für die Stichprobe 1, 3, 5, 7, 90*

$$\bar{x} = 21,2$$

gilt. Diese Stichproben werden uns noch weiter begleiten.

Der arithmetische Mittelwert oder empirische Mittelwert ist vielleicht die bekannteste und gebräuchlichste Art mit Daten umzugehen. Ein Grund hierfür mag sein, dass die Summe der Abweichungen von $\bar{x}$ aller Daten x_i gleich null, also der Mittelpunkt (Schwerpunkt) der Werte x_i ist. In der Tat

$$\sum_{i=1}^{n}(x_i - \bar{x}) = (\sum_{i=1}^{n} x_i) - n\bar{x} = n\bar{x} - n\bar{x} = 0.$$

Wenn wir über weitere Größen nachdenken wollen, die die Lage der Daten $x_1, \ldots, x_n$ darstellen sollen, müssen wir uns Gedanken machen, was denn wünschenswerte Eigenschaften solcher Größen sein sollten. Eine wichtige Forderung an Lageparameter der Verteilung eines Merkmals ist die sogenannte *Translationsinvarianz*. Dies soll bedeuten, dass für eine lineare Transformation

$$y = L(x) = a + bx$$

der Daten, die aus den Daten $x_1, \ldots, x_n$ die Daten

$$y_i = a + bx_i$$

macht, der Lageparameter λ_x sich zu λ_y transformieren lässt, wobei

$$\lambda_y = a + b\lambda_x$$

gilt. Wir wollen uns an einem Beispiel für den empirischen Mittelwert von der Nützlichkeit dieses Konzepts überzeugen.

Beispiel 2.9 *Wir messen täglich die Mittagstemperatur in °C und ermitteln daraus einen Jahresdurchschnittswert in °C. Messen wir stattdessen die Temperatur in °F, so ist der Zusammenhang zwischen den Celsiustemperaturen (x_i) und den Fahrenheittemperaturen (y_i)*

$$y_i = 32 + 1,8x_i.$$

Natürlich sollte dann auch für die Durchschnittstemperaturen in °C $(\bar{x})$ und in °F $(\bar{y})$ gelten:

$$\bar{y} = 32 + 1,8\bar{x}.$$

Dies gilt in der Tat:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(32 + 1,8x_i) = 32 + 1,8\frac{1}{n}\sum_{i=1}^{n} x_i = 32 + 1,8\bar{x}.$$

Dies zeigt, dass der empirische Mittelwert translationsinvariant ist.

Die folgenden beiden Lagemaße haben diese Eigenschaft ebenfalls:

Median
Der Median oder Zentralwert ist grob gesagt der mittlere Wert der geordneten Stichprobe. Mit anderen Worten ordnen wir die Stichprobe $x_1, \ldots, x_n$ zu $x_{(1)}, \ldots, x_{(n)}$, so dass

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

gilt. Der Median, den wir mit $\tilde{x}$ (manchmal auch mit $\tilde{x}_{0.5}$) bezeichnet werden, ist dann ein Wert, so dass höchstens 50 % der Stichprobenwerte kleiner sind als $\tilde{x}$ und höchstens 50 % der Stichprobenwerte größer sind als $\tilde{x}$. Man beachte, dass für diese Definition die Daten zumindest auf einer Ordinalskala liegen sollten. Eine Möglichkeit bei metrischen Daten $\tilde{x}$ zu wählen ist die folgende

$$\tilde{x} = \tilde{x}_{0.5} = \begin{cases} x_{(\frac{n+1}{2})}, & \text{falls } n \text{ ungerade} \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}), & \text{falls } n \text{ gerade} \end{cases} .$$

Beispiel 2.10 *In den Stichproben aus Beispiel 2.8 ist für alle Stichproben 5 der Median. Für die Stichprobe 1, 3, 4, 5, 6, 9 ist jeder Wert im Intervall* $[4, 5]$ *ein Median, z. B.* $4\frac{1}{2}$.

Man kann sich nun fragen, was die Vor- bzw. Nachteile des Medians gegenüber dem empirischen Mittelwert sind. Der eine Vorteil des Medians wurde eben schon angesprochen: er ist auch sinnvoll für Daten auf einer ordinalen Skala (beispielsweise Schulnoten) anwendbar, für die die Bildung des arithmetischen Mittels nicht notwendig sinnvolle Resultate liefert. Um einen weiteren Vorteil des Medians kennenzulernen, studieren wir das folgende

Beispiel 2.11 *Wir betrachten die folgende (fiktive) Einkommensverteilung in einem erdölfördernden Land. Dort haben 1 000 Erdölarbeiter einen Monatslohn von 1 000 $, während der Ölscheich selbst 1 000 000 $ pro Monat verdient. Das "Durchschnittseinkommen", also der empirische Mittelwert der Daten, 1 998,02 $ ist dann wenig aussagekräftig: er ist beinahe das Doppelte der Einkünfte der Erdölarbeiter, aber nur ein winziger Bruchteil der Einkünfte des Scheichs. Der Median $ 1 000 repräsentiert hingegen das Einkommen von 99,9 % der Bevölkerung exakt.*

Der Median ist somit viel toleranter gegenüber Ausreißern als der empirische Mittelwert. Letzterer hat hingegen den nicht zu unterschätzenden Vorteil, schon bei kleinen Stichprobengrößen dichter am Mittelwert der Gesamtpopulation zu liegen als der Median. Er hat also bessere Konvergenzeigenschaften.

Ein dritter Lageparameter für die "mittlere Lage" der Stichprobe ist der

Modalwert
Der Modalwert ist der häufigste Stichprobenwert. Er hat den Vorteil, für jede Art von Daten ein sinnvolles Ergebnis zu liefern, allerdings muss dieses nicht unbedingt aussagekräftig sein.

Beispiel 2.12 *In den Stichproben 1, 3, 5, 7, 9 und 1, 3, 5, 7, 90 ist jeder Wert Modalwert (alle Werte kommen gleich häufig vor), für die Stichprobe 1, 3, 5, 5, 7, 9 ist 5 der Modalwert.*

Man kann sich natürlich fragen, ob man nicht nur über das mittlere Stichprobenverhalten sondern auch über die "Ränder" etwas aussagen kann. Ein Versuch, dies zu tun, ist in Anlehnung an den Median definiert. Er führt zu den Begrifffen

Quartile, Quantile, Percentile
Für $0 \leq \alpha \leq 1$ ist das α-Quantil $\tilde{x}_\alpha$ dadurch definiert, dass für eine Stichprobe vom Umfang n höchstens $n\alpha$ Werte kleiner sind als $\tilde{x}_\alpha$ und höchstens $n(1-\alpha)$ Werte größer sind als $\tilde{x}_\alpha$. Die α-Percentile sind die Werte für $\alpha\cdot$ 10 % und die 25 %- und 75 %-Quantile heißen auch unteres bzw. oberes Quartil. Offenbar ist das 50 %-Quantil der Median.

Beispiel 2.13 *Bei der Stichprobe*

$$1, 3, 4, 4, 5, 5, 6, 8, 9, 10$$

ist $\tilde{x}_{0,5} = 5$ *der Median,* $\tilde{x}_{0,25} = 4$ *(unteres Quartil) und* $\tilde{x}_{0,75} = 8$ *(oberes Quartil) sind die Quartile.*

2.4 Streumaße

Lagemaße allein sind zur Beschreibung von Daten nicht ausreichend. Dazu das folgende Beispiel.

Beispiel 2.14 *Zwei Zulieferfirmen für einen Automobilkonzern sollen Türen von exakt 1,00 m Breite liefern. Zulieferer A liefert auch genau solche Türen, während Zulieferer B Türen liefert, die zur Hälfte 0,95 m breit sind und zur Hälfte 1,05 m. Beide Lieferanten liefern Türen mit einer durchschnittlichen Breite von 1,00 m, aber während die Türen von Zulieferer A in Ordnung sind, sind die von Zulieferer B komplett unbrauchbar.*

Wir benötigen also eine Kennzahl, die angibt, wie aussagekräftig ein Lageparameter ist. Wir werden in der Folge solche Streumaße vorstellen:

Spannweite
Die Spannweite r einer Stichrpobe ist definiert als der Abstand zwischen größtem Stichprobenwert $x_{(n)}$ und kleinstem Stichprobenwert $x_{(1)}$

$$r := x_{(n)} - x_{(1)}.$$

Beispiel 2.15 *In der Stichprobe aus Beispiel 2.13 ist die Spannweite*

$$r = 10 - 1 = 9.$$

Quartilabstand
Die Spannweite einer Stichprobe kann durch einen Ausreißer sehr groß werden. Dadurch,

dass ein solcher Ausreißer auch noch durch einen Messfehler oder einen Zahlendreher produziert werden kann, ist die Spannweite sehr fehleranfällig. Eine Möglichkeit, dies zu korrigieren, ist der Quartilsabstand

$$d_Q := \tilde{x}_{0,75} - \tilde{x}_{0,25}.$$

d_Q definiert den zentralen Bereich einer Verteilung, in dem 50 % der Werte liegen.

Beispiel 2.16 *In der Stichprobe aus Beispiel 2.13 ist*

$$d_Q := 8 - 4 = 4.$$

Mittlere absolute Abweichung von Median und arithmetischem Mittel
Größen, die eine durchschnittliche Abweichung von einem Lageparameter angeben, lassen sich als Streumaße verwenden. Je nachdem, ob man den Median oder das arithmetische Mittel als geeignete Lageparameter wählt, bestimmt man das Streumaß in Bezug auf $\tilde{x}_{0,5}$ oder $\bar{x}$. Die einfachsten entsprechenden Streumaße sind die mittleren absoluten Abweichungen bei metrischen Daten

$$\tilde{d} := \frac{1}{n}\sum_{i=1}^{n} |x_i - \tilde{x}_{0,5}| \quad \text{und}$$
$$\bar{d} := \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|.$$

Beispiel 2.17 *In der Stichprobe aus Beispiel 2.13 war $\tilde{x}_{0,5} = 5$ der Median, also ist*

$$\tilde{d} = \frac{1}{10}\sum_{i=1}^{10} |x_i - 5| = \frac{21}{10} = 2,1.$$

Man berechnet, dass für diese Stichprobe gilt, dass $\bar{x} = 5,5$ der Mittelwert ist. Also ist

$$\bar{d} = \frac{1}{10}\sum_{i=1}^{10} |x_i - 5,5| = 2,2.$$

Das wohl bekannteste Streumaß ist durch die folgenden Größen dargestellt:

Varianz und Standardabweichung
Varianz und Standardabweichung sind Streumaße um den Lageparameter $\bar{x}$. Die Varianz misst den mittleren quadratischen Abstand zum empirischen Mittelwert $\bar{x}$:

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

Mit ein wenig Rechenaufwand kommen wir zu einer äquivalenten Formel

Proposition 2.18 *Es gilt der Verschiebungssatz*

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

Beweis: Es gilt

$$\begin{aligned} s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}(\sum_{i=1}^{n} x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\bar{x}\frac{1}{n}\sum_{i=1}^{n} x_i + \bar{x}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\bar{x}^2 + \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2. \end{aligned}$$

□

Beispiel 2.19 *Fünf* $(n = 5)$ *Studierende, die in Münster und Umgebung wohnen, messen an einem Montagmorgen die Temperaturwerte* x_i *und benutzen die folgende Arbeitstabelle zur Berechnung der Varianz:*

i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	x_i^2
1	5	-4	16	25
2	7	-2	4	49
3	9	0	0	81
4	11	2	4	121
5	13	4	16	169
	$\bar{x} = 9$	$\sum(x_i - \bar{x})^2 = 40$		$\sum x_i^2 = 445$

Mit der Definition der Varianz ergibt sich

$$s^2 = \frac{1}{5}\sum_{i=1}^{5}(x_i - \bar{x})^2 = \frac{40}{5} = 8.$$

Analog ergibt sich mit dem Verschiebungssatz

$$s^2 = \frac{1}{5}\sum_{i=1}^{5} x_i^2 - \bar{x}^2 = \frac{445}{5} - 81 = 89 - 81 = 8.$$

Ebenfalls gebräuchlich als Streumaß ist die Standardabweichung. Sie ist definiert als die Wurzel der Varianz:

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2}.$$

Beispiel 2.20 *Für die Daten aus Beispiel 2.19 gilt*

$$s = \sqrt{s^2} = \sqrt{8} = 2\sqrt{2}.$$

Bemerkung 2.21 *In der beschreibenden Statistik ist die obige Definition von Varianz und Standardabweichung gebräuchlich. In der induktiven Statistik, die wir später kennenlernen werden, gibt es gute Gründe (Stichwort "Erwartungstreue", d. h. wir wollen die "wahre Varianz" so gut wie möglich schätzen)*

$$\tilde{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \textit{und}$$

$$\tilde{s} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

zu definieren. Einige Statistikprogramme arbeiten auch mit $\tilde{s}^2$ und $\tilde{s}$ statt mit s^2 und s, andere lassen die Wahl. Offenbar verschwindet der Unterschied zwischen s^2 und $\tilde{s}^2$ für größere n.

2.5 Die empirische Verteilungsfunktion

Die obigen Parameter versuchen, die Stichprobe mit wenigen Kenngrößen (Lageparameter, Streumaße) zu erfassen. Dies führt einerseits zu einer sehr handlichen Beschreibung der Daten, birgt aber auch die Gefahr der übermäßigen Vereinfachung durch Datenverlust. Eine Möglichkeit, die Stichprobe ohne Datenverlust kompakt zu beschreiben, ist die empirische Verteilungsfunktion. Dies ist eine Funktion

$$F : \mathbb{R} \to [0, 1],$$

die an den Datenpunkten um $\frac{1}{n}$ springt (wobei n der Stichprobenumfang ist), entsprechend mehr bei mehrfachen Datenpunkten. Sie ist definiert als

$$F(x) = \frac{1}{n} \cdot |\{i : x_i \leq x\}|,$$

wobei $|\cdot|$ die Anzahl der Elemente der Menge zwischen den Betragsstrichen bezeichnet.

Beispiel 2.22 *Für die Stichprobe -1,1,3,3,5 ist die empirische Verteilungsfunktion F_1 die Funktion, die monoton von 0 (bei $-\infty$) nach 1 (bei $+\infty$) wächst und dabei an den Werten*

-1,1 und 5 um jeweils $\frac{1}{5}$ und an die Stelle 3 um $\frac{2}{5}$ springt.

2.6 Semigraphische Darstellung von Daten – der Boxplot

Wir sind nun in der Lage, eine weitere, quasi semigraphische Darstellung der Daten vorzustellen, den Boxplot. Er enthält als wesentliche Information über die Stichprobe: den Median, die Quartile, den Interquartilabstand sowie die außen liegenden Daten. Genauer verfährt man wie folgt: Man zeichnet eine Box (Kasten) vom unteren Quartil $\tilde{x}_{0,25}$ zum oberen Quartil $\tilde{x}_{0,75}$. In der Mitte dieser Box kennzeichnet man den Median durch eine waagerechte Linie. Vom oberen und vom unteren Rand der Box zeichnet man je eine Strecke der 1,5-fachen Länge der Box. Daten außerhalb werden einzeln markiert und zwar als $*$, falls sie zwischen 1,5 – 3 Boxlängen vom unteren bzw. oberen Boxrand entfernt sind und ansonsten als 0.

Ein typischer Boxplot, z. B. für die Körpergröße einer Population, sieht folgendermaßen aus:

2.7 Fehler und Fallen in der beschreibenden Statistik

Die beschreibende Statistik bietet neben der Möglichkeit einer kompakten Beschreibung der Daten auch die Gefahr, die Daten so zu präsentieren, dass diese falsch verstanden werden können oder sollen. Möglichkeiten hierfür sind u. a.
– Stauchung oder Streckung von Achsen

– Abschneiden von Achsen
– mehrdimensionale Piktogramme
– die Interpretation von Korrelation als Kausalität
– verkehrte Wahl der Bezugsgröße
– u.v.m.

Beispiele werden in der Vorlesung gegeben.

3 Einführung in die Wahrscheinlichkeitsrechnung

3.1 Axiomatische Grundlagen

Obschon die Beschäftigung mit Wahrscheinlichkeit und ihren Gesetzmäßigkeiten mehr als 300 Jahre alt ist – das erste wahrscheinlichkeitstheoretische Gesetz, J. Bernoullis Gesetz der großen Zahlen, stammt aus dem Jahre 1705 und wurde posthum 1713 publiziert – war die Frage einer genauen Axiomatisierung der Wahrscheinlichkeitstheorie lange offen. Hilbert stellte bei seinem berühmt gewordenen Vortrag 1900 die Axiomatisierung der Wahrscheinlichkeitstheorie und Physik(!) als das sechste seiner 23 offenen Probleme. Dieses wurde 1933 von A. N. Kolmogorov gelöst. Sein Ansatz baut nicht auf die intuitive Vorstellung von Wahrscheinlichkeiten als Limes relativer Häufigkeiten auf, sondern stützt sich auf das (damals noch recht junge) Gebiet der Maßtheorie. Wir werden diese Axiomatisierung nicht ganz im Detail nachvollziehen (können), weil uns dazu der Begriff des Maßes bzw. des Maßintegrals fehlt. Einige Elemente seiner Theorie aber können wir leicht übertragen. Zunächst bezeichne Ω die Menge aller möglichen Versuchsausgänge eines Zufallsexperiments.

Beispiel 3.1 *1. Besteht das Zufallsexperiment aus dem einmaligen Werfen einer Münze, so ist*

$$\Omega = \{\textit{Kopf, Zahl}\}.$$

2. Besteht das Experiment aus dem einmaligen Werfen eines Würfels, so ist

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

In der Potenzmenge

$$\mathcal{P}\Omega := \{A \subseteq \Omega\}$$

wollen wir nun die Mengen auszeichnen, denen wir eine Wahrscheinlichkeit zuweisen wollen. Hierbei sind die folgenden Regeln naheliegend:

- Die Wahrscheinlichkeit von Ω kann man immer messen.
- Kann man die Wahrscheinlichkeit von $A \in \Omega$ messen, so auch die von $A^c := \Omega \backslash A$.
- Kann man die Wahrscheinlichkeit von $A_1, A_2, \ldots$ usw. messen, so auch die von $\bigcup_{i=1}^{\infty} A_i$.

Wir wollen ein System $\mathcal{A}$ von Teilmengen von Ω als *σ-Algebra* bezeichnen, wenn gilt:

- $\Omega \in \mathcal{A}$,
- $A \in \mathcal{A} \;\Rightarrow\; A^c \in \mathcal{A}$,

- $A_1, A_2, \ldots \in \mathcal{A} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$.

Beispiel 3.2 *1. Über der Menge*

$$\Omega = \{\textit{Kopf, Zahl}\} \stackrel{\wedge}{=} \{0, 1\}$$

(wenn wir z. B. Kopf=1 und Zahl=0 setzen) gibt es genau zwei σ-Algebren:

$$\begin{aligned} \mathcal{A}_1 &= \{\emptyset, \Omega\} \quad \textit{und} \\ \mathcal{A}_2 &= \mathcal{P}\Omega = \{\emptyset, \{0\}, \{1\}, \Omega\}. \end{aligned}$$

Dies folgt, da jede σ-Algebra die Menge Ω enthalten muss und somit auch $\emptyset$. Weiter impliziert $\{1\} \in \mathcal{A}_2$ sofort auch $\{0\} \in \mathcal{A}_2$.

2. Über jeder beliebigen Menge Ω sind stets

$$\mathcal{A}_1 = \{\emptyset, \Omega\} \quad \textit{und} \quad \mathcal{A}_2 = \mathcal{P}\Omega$$

σ-Algebren. Der Grund dafür, dass man nicht immer $\mathcal{P}\Omega$ als σ-Algebra nimmt und damit eine Wahrscheinlichkeit auf Ω definiert (das wäre an sich sehr praktisch, denn so könnte man sicher sein, jeder Teilmenge von Ω eine Wahrscheinlichkeit zuweisen zu können) ist der, dass sich auf $\mathcal{P}\Omega$ nicht immer eine Wahrscheinlichkeit mit allen gewünschten Eigenschaften definieren lässt (beispielsweise kann man auf

$$\Omega = [0, 1] \quad \textit{und} \quad \mathcal{A} = \mathcal{P}\Omega$$

keine Wahrscheinlichkeit definieren, die einem Intervall seine Länge zuweist). Dies kann und soll uns aber in der Folge nicht weiter interessieren. Für endliche oder abzählbar unendliche Mengen können wir getrost stets die Potenzmenge $\mathcal{P}\Omega$ als σ-Algebra verwenden.

Bei der Definition von Wahrscheinlichkeit spielt schließlich die Intuition eine große Rolle. Ein "üblicher" Begriff von Wahrscheinlichkeit würde die Wahrscheinlichkeit eines Ereignisses A als Limes der relativen Häufigkeiten des Auftretens von A bei n unabhängigen Versuchen definieren, wenn n gegen unendlich geht. Diese Definition birgt einige Nachteile (beispielsweise werden die relativen Häufigkeiten des Auftretens von A typischerweise von den Versuchen abhängen, man kann z. B. ja sowohl K, K, Z, K als auch Z, Z, K, K als Ergebnis eines vierfachen Münzwurfs erhalten); allerdings wäre es sicherlich nützlich, wenn man relative Häufigkeiten als Wahrscheinlichkeiten auffassen könnte. Die folgende Definition ist also gewissermaßen den relativen Häufigkeiten abgeschaut:

Definition 3.3 *Es sei Ω eine Menge, $\Omega \neq \emptyset$, und $\mathcal{A}$ eine σ-Algebra über Ω. Eine Wahrscheinlichkeit $\mathbb{P}$ ist eine Funktion*

$$\mathbb{P} : \mathcal{A} \to [0, 1]$$

mit

- $\mathbb{P}(\Omega) = 1$.
- *Sind $A_1, A_2, \ldots$ paarweise disjunkt, d. h. $A_i \cap A_j = \emptyset$ für $i \neq j$, so gilt:*

$$\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) \quad (\sigma\text{-Additivität})$$

Diese Regeln implizieren insbesondere, dass

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A) \tag{3.1}$$

für alle $A \in \mathcal{A}$ gilt, also auch

$$\mathbb{P}(\emptyset) = 0. \tag{3.2}$$

Der Grund hierfür ist, dass A und A^c disjunkt sind und $A \cup A^c = \Omega$ gilt. Also ist

$$1 = \mathbb{P}(\Omega) = \mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c),$$

also $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$; setzt man für $A = \Omega$ ein, so ergibt sich (3.2).

Beispiel 3.4 *Den fairen Münzwurf über $\Omega = \{0, 1\}$, $\mathcal{A} = \mathcal{P}\Omega$ modelliert man mit*

$$\mathbb{P}(\{0\}) = \mathbb{P}(\{1\}) = \frac{1}{2}.$$

Den fairen Würfelwurf auf $\Omega = \{1, \ldots, 6\}$, $\mathcal{A} = \mathcal{P}\Omega$, modelliert man mit

$$\mathbb{P}(\{1\}) = \ldots = \mathbb{P}(\{6\}) = \frac{1}{6}.$$

Im Beispiel 3.4 haben wir die Wahrscheinlichkeit $\mathbb{P}$ gerade dadurch festgelegt, dass wir für jedes $\omega \in \Omega$ angeben, was dessen Wahrscheinlichkeit sein soll. So lange Ω höchstens abzählbar ist, geht dies ganz allgemein.

Satz 3.5 *Ist Ω höchstens abzählbar, $\mathcal{A} = \mathcal{P}\Omega$ und $(p_\omega)_{\omega \in \Omega}$ eine Folge nicht negativer Zahlen mit*

$$\sum_{\omega \in \Omega} p_\omega = 1,$$

dann ist durch

$$\mathbb{P}(A) = \sum_{\omega \in A} p_\omega \tag{3.3}$$

für $A \subseteq \Omega$ eindeutig eine Wahrscheinlichkeit auf $\mathcal{A}$ festgelegt.

Beweis: Die Eindeutigkeit ist klar, da man positive Summen beliebig umordnen darf. $\mathbb{P}$, wie in (3.3) definiert, ist auch eine Wahrscheinlichkeit, denn es ist $0 \leq \mathbb{P}(A) \leq 1$ für alle, $A \subset \Omega$,

$$\mathbb{P}(\Omega) = \sum_{\omega \in \Omega} p_\omega = 1,$$

und sind $A_1, A_2, \ldots$ paarweise disjunkt, so ist

$$\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{\omega \in \bigcup_{i=1}^{\infty} A_i} p_\omega = \sum_{i=1}^{\infty} \sum_{\omega \in A_i} p_\omega = \sum_{i=1}^{\infty} \mathbb{P}(A_i),$$

da $\omega \in \bigcup_{i=1}^{\infty} A_i$ impliziert, dass $\omega \in A_i$ für genau ein i gilt. □

Wir werden in der Folge immer mit einem sogenannten *Wahrscheinlichkeitsraum* $(\Omega, \mathcal{A}, \mathbb{P})$ arbeiten (Ω eine Menge, $\mathcal{A}$ eine σ-Algebra über Ω und $\mathbb{P}$ eine Wahrscheinlichkeit über $\mathcal{A}$). Dabei nennen wir Ω den Grundraum oder auch das *sichere Ereignis*, $\omega \in \Omega$ heißt *Ergebnis*, $\{\omega\}$ *Elementarereignis*, $A \in \mathcal{A}$ heißt *Ereignis* und $\mathbb{P}$ heißt *Wahrscheinlichkeit*.

3.2 Endliche Wahrscheinlichkeitsräume und mehrstufige Zufallsexperimente

In diesem Kapitel wollen wir uns Zufallsexperimenten zuwenden, bei denen der Grundraum nicht nur abzählbar, sondern sogar endlich ist. Beispiele hierfür haben wir im vorhergehenden Kapitel mit dem fairen Münzwurf und dem fairen Würfeln schon kennengelernt. Diese beiden Beispiele (siehe Beispiel 3.4) haben noch eine weitere Eigenschaft, die im Falle endlich vieler möglicher Versuchsausgänge (d. h. Ω ist eine endliche Menge) oft anzutreffen ist: Alle Elementarereignisse $\omega \in \Omega$ haben dieselbe Wahrscheinlichkeit. Solche Experimente heißen auch Laplace-Experimente.

Definition 3.6 *Sei Ω eine endliche Menge und $\mathcal{A} = \mathcal{P}\Omega$. Eine Wahrscheinlichkeit $\mathbb{P}$ über $\mathcal{A}$ heißt Laplace-Wahrscheinlichkeit (und $(\Omega, \mathcal{A}, \mathbb{P})$ Laplace-Experiment), wenn gilt*

$$\mathbb{P}(\{\omega\}) = \frac{1}{|\Omega|} \quad \textit{für alle} \quad \omega \in \Omega.$$

Hierbei bezeichnet $|\Omega|$ die Anzahl der Elemente in Ω. Für ein Laplace-Experiment gilt offenbar

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} \quad \textit{für alle} \quad A \subseteq \Omega.$$

Man nennt die Elemente in A auch die "günstigen Fälle", daher spricht man auch in Laplace-Experimenten davon, dass $\mathbb{P}(A)$ die Anzahl der günstigen Fälle geteilt durch die Anzahl der möglichen Fälle ist.

In Laplace-Experimenten sind alle Versuchsausgänge definitionsgemäß gleichwahrscheinlich. Wir sprechen oft auch von einem "zufälligen" oder "rein zufälligen" Experiment.

Beispiel 3.7 *Eine Urne enthält 100 Kugeln, die mit 00, 01, ..., 99 nummeriert sind. Sei X die erste und Y die zweite Ziffer einer rein zufällig gezogenen Kugel. Dann berechnet sich beispielsweise $\mathbb{P}(X = 3)$, $\mathbb{P}(X \neq Y)$ oder $\mathbb{P}(X + Y \neq 8)$ folgendermaßen:*

$$\Omega = \{00, \ldots, 99\},\ \mathcal{A} = \mathcal{P}\Omega \quad \textit{und} \quad \mathbb{P}(\{\omega\}) = \frac{1}{100}$$

für alle $\omega \in \Omega$.

Das Ereignis

$$\{X = 3\} = \{30, 31, \ldots, 39\}.$$

Also ist $|\{X = 3\}| = 10$ und

$$\mathbb{P}(X = 3) = \frac{|\{X = 3\}|}{100} = \frac{10}{100} = \frac{1}{10}.$$

$$\{X \neq Y\} = \{X = Y\}^c = \{00, 11, 22, \ldots, 99\}^c.$$

Also ist $|\{X \neq Y\}| = 90$ und

$$\mathbb{P}(X \neq Y) = \frac{90}{100} = \frac{9}{10}.$$

Schließlich ist

$$\{X + Y \neq 8\} = \{X + Y = 8\}^c = \{08, 17, \ldots, 71, 80\}^c.$$

Da $|\{X + Y = 8\}| = 9$, folgt $|\{X + Y \neq 8\}| = 91$ und somit

$$\mathbb{P}(X + Y \neq 8) = \frac{91}{100}.$$

Beispiel 3.8 *Ein Klub hat 100 Mitglieder. Eine Umfrage ergab die folgende Statistik:*

	Männer	*Frauen*
sind keine Vegetarier	*48*	*12*
sind Vegetarier	*16*	*24*

Ω lässt sich z. B. folgendermaßen beschreiben:

$$\Omega = \{(i, g, v) |\, i \in \{1, \ldots, 100\}, g \in \{\textit{Mann,Frau}\},\ v \in \{\textit{Vegetarier, Nichtvegetarier}\}\},$$

wobei $i = 1, \ldots, 100$ die Nummer des Mitglieds ist, g sein Geschlecht und v bezeichnet, ob er/sie Vegetarier ist. Dann berechnet sich z. B. die Wahrscheinlichkeit dafür, dass eine

zufällig gezogene Person Vegetarier ist, mittels $\mathcal{A} = \mathcal{P}\Omega$ *und* $\mathbb{P}(\{\omega\}) = \frac{1}{100}$ *für alle* $\omega \in \Omega$ *als*

$$\mathbb{P}(\{\omega | \omega \text{ ist Vegetarier}\}) = \frac{16+24}{100} = \frac{2}{5},$$

da 16 Männer und 24 Frauen Vegetarier sind. Wissen wir von der Person schon, dass die eine Frau ist, so ist die Wahrscheinlichkeit, dass sie auch Vegetarierin ist, $= \frac{24}{36} = \frac{2}{3}$. *Dies gehört aber eigentlich in den Themenkreis der bedingten Wahrscheinlichkeit und soll später untersucht werden.*

Bisher haben wir sogenannte *einstufige* Zufallsversuche betrachtet. Eine Münze wurde einmal geworfen, mit einem Würfel wurde einmal gewürfelt. Oft finden aber mehrere solcher Prozesse hintereinander statt, man spricht dann von *mehrstufigen* Experimenten. Der Ablauf eines solchen Experiments lässt sich in Form eines Baumdiagramms veranschaulichen. Wir wollen dies anhand eines Beispiels kennenlernen:

Beispiel 3.9 *In einer Urne befinden sich 5 Kugeln, 3 tragen ein "a" und 2 ein "n". Es wird ein Buchstabe gezogen, notiert und dann wieder zurückgelegt. Dann wird das Experiment wiederholt. Beide Stufen des Versuchs sind gewissermaßen Kopien voneinander, man spricht daher auch von unabhängigen und identisch verteilten Versuchsdurchführungen – aber dies soll später besprochen werden. Für dieses zweistufige Experiment ist der Grundraum*

$$\Omega = \{(a,a),(a,n),(n,a),(n,n)\}$$

und $\mathcal{A} = \mathcal{P}\Omega$. *Um die Wahrscheinlichkeiten für die Elemente aus* Ω *zu bestimmen, denken wir uns die Kugeln in der Urne zusätzlich nummeriert: 1 – 3 steht für a, 4 und 5 für n. Dann gibt es insgesamt 25 Arten, die Kugeln der Urne durch zweifaches Ziehen zu kombinieren; es sind nämlich alle Kombinationen* (i,j) *mit* $1 \leq i \leq 5$ *und* $1 \leq j \leq 5$. *Von diesen gibt es neun, die zur Kombination* $(a,a) \in \Omega$ *führen, nämliche alle Paare* (i,j) *mit* $1 \leq i \leq 3$ *und* $1 \leq j \leq 3$. *Sechs Kombinationen führen zu* $(a,n) \in \Omega$, *nämlich alle* (i,j) *mit* $1 \leq i \leq 3$ *und* $j = 4$ *oder 5. Ähnlich gibt es sechs Kombinationen, die zu* $(n,a) \in \Omega$ *führen, nämlich alle* (i,j) *mit* $i = 4,5$ *und* $1 \leq j \leq 3$. *Schließlich führen die vier Kombinationen* (i,j) *mit* $i = 4,5$ *und* $j = 4,5$ *zu* $(n,n) \in \Omega$. *Da alle Kugeln die gleiche Wahrscheinlichkeit haben, gezogen zu werden, haben auch alle* (i,j), $1 \leq i \leq 5$, $1 \leq j \leq 5$ *dieselbe Wahrscheinlichkeit. Es gibt 25 solcher Paare, also hat jedes* (i,j) *die Wahrscheinlichkeit* $\frac{1}{25}$, *gezogen zu werden. Somit ist im obigen* Ω *die Wahrscheinlichkeit für* (a,a) $\frac{9}{25}$, *die Wahrscheinlichkeit für* (a,n) *und* (n,a) $\frac{6}{25}$ *und die Wahrscheinlichkeit für* (n,n) $\frac{4}{25}$.

Eine andere Art dies zu verstehen ist die folgende: Wir beschreiben den Versuch durch

folgendes Baumdiagramm:

Man bekommt also die Wahrscheinlichkeit der Kombinationen (a,a), (a,n), (n,a) *und* (n,n) *durch Multiplikation der Wahrscheinlichkeiten entlang des Pfades.*

Dies führt zu folgender Definition:

Definition 3.10 *In einem mehrstufigen Zufallsexperiment ist die Wahrscheinlichkeit eines* $\{\omega\}$ *das Produkt der Wahrscheinlichkeiten entlang des Pfades, das diesem* $\omega \in \Omega$ *entspricht. Die Wahrscheinlichkeit eines Ereignisses* A *ist die Summe der Wahrscheinlichkeiten aller Pfade* $\{\omega\}$*, die zu* A *gehören, d. h. aller* $\omega \in A$*.*

Wir wollen dieses Prinzip nun in verschiedenen Beispielen kennenlernen. Das erste ist eine kleine Variation des Beispiels 3.9:

Beispiel 3.11 *Die Situation sei die gleiche wie in Beispiel 3.9 mit dem einzigen Unterschied, dass die einmal gezogenen Buchstaben nicht zurückgelegt werden. Ensprechend ändert sich das Baumdiagramm:*

Somit ändern sich die Wahrscheinlichkeiten für die Elemente in

$$\Omega = \{(a,a),(a,n),n,a),(n,n)\}$$

zu

$$\mathbb{P}(\{(a,a)\}) = \frac{3}{10} = \mathbb{P}(\{(a,n)\}) = \mathbb{P}(\{(n,a)\}) \quad \textit{und} \quad \mathbb{P}(\{(n,n)\}) = \frac{1}{10}.$$

Beispiel 3.12 *In einem Boot sitzen neun Passagiere, davon sind vier Schmuggler und fünf ehrliche Personen. Ein Zollbeamter wählt drei Personen zur Kontrolle aus: Alle drei sind Schmuggler. Wie groß ist die Wahrscheinlichkeit, dass dies reiner Zufall ist?*

$$\Omega = \{(i,j,k) : i,j,k \in \{s,e\}\},$$

wobei s für Schmuggler und e für ehrlich steht. Der Pfad, der dem Ereignis (s,s,s) entspricht, sieht nun so aus:

$\frac{4}{9}$ s $\frac{3}{8}$ s $\frac{2}{7}$ s

Also hat $(s,s,s) \in \Omega$ die Wahrscheinlichkeit

$$\mathbb{P}((s,s,s)) = \frac{4}{9} \cdot \frac{3}{8} \cdot \frac{2}{7} = \frac{1}{21},$$

d. h. weniger als 5 %; vermutlich hatte der Zollbeamte also eine gewisse Menschenkenntnis.

Beispiel 3.13 *In einem Schubfach befinden sich 4 schwarze, 6 rote und 2 weiße Socken. Im Dunkeln wählt jemand (ein Mathematiker?) zwei Socken rein zufällig. Wie groß ist die Wahrscheinlichkeit, dass die beiden Socken die gleiche Farbe haben?*

Es ist

$$\Omega = \{(i,j) | i,j \in \{s,r,w\}\}.$$

Nun führen drei Pfade zu einem Paar gleicher Farbe:

Also ist

$$\mathbb{P}(\{(i,j) : i = j\}) = \frac{1}{3} \cdot \frac{3}{11} + \frac{1}{2} \cdot \frac{5}{11} + \frac{1}{6} \cdot \frac{1}{11} = \frac{1}{3}.$$

Beispiel 3.14 *Jedesmal, wenn Prof. L. 7 Personen beisammen sieht, wettet er 100:1, dass darunter mindestens zwei Personen am gleichen Wochentag geboren sind. Ist dies eine günstige Wette?*

Die Wahrscheinlichkeit, dass Prof. L. die Wette verliert, lässt sich wie folgt berechnen:

$$\Omega = \{(i_1, \ldots, i_7) : i_j \in \{1 \ldots 7\},\ j = 1 \ldots 7\},$$

wobei i, j für den Wochentag steht, an dem die Person j geboren ist. Die Wahrscheinlichkeit, dass Prof. L. die Wette verliert entspricht nun dem folgenden Pfad:

Die Wahrscheinlichkeit für diesen Pfad ist

$$\frac{7 \cdot 6 \ldots 2 \cdot 1}{7^7} = \frac{7!}{7^7} = \frac{6!}{7^6} = 0,00612,$$

die Wette ist also günstig für den Prof.

Beispiel 3.15 *In M. gibt es nur zwei Arten Wetter: Nass (N) und Trocken (T). Ist es heute nass, so ist es mit Wahrscheinlichkeit $\frac{5}{6}$ auch morgen nass und mit Wahrscheilichkeit $\frac{1}{6}$ trocken. Ist es heute trocken, so ist es morgen mit Wahrscheinlichkeit $\frac{3}{10}$ auch trocken und mit Wahrscheinlichkeit $\frac{7}{10}$ nass. [Frage: Welche Stadt ist M.?] Heute ist es trocken: Mit welcher Wahrscheinlichkeit ist es übermorgen auch trocken?*

Offenbar gibt es zwei Pfade, dass es übermorgen trocken ist:

Die Wahrscheinlichkeit, dass es übermorgen trocken ist, berechnet sich also so:

$$\frac{3}{10} \cdot \frac{3}{10} + \frac{7}{10} \cdot \frac{1}{6} = \frac{9}{100} + \frac{7}{60} = \frac{27}{300} + \frac{36}{300} = \frac{52}{300} = \frac{13}{75}.$$

3.3 Unabhängigkeit und bedingte Wahrscheinlichkeit I

Zur Motivation begeben wir uns noch einmal zurück zu Beispiel 3.9 und 3.11, also den Beispielen, die in mehrstufige Zufallsexperimente eingeführt haben. Wir wollen erkennen, dass es sich bei der Multiplikationsregel um eine Form der *Unabhängigkeit* bzw. *bedingten*

Unabhängigkeit handelt, wobei wir den letzten Begriff nicht definieren, wohl aber kurz anreißen. In der Tat ist ja in Beispiel 3.9 durch das Zurücklegen der gezogenen Buchstaben die Grundsituation vor jedem Zug dieselbe. Da zudem die Buchstaben "gedächtnislos" sind, d. h. die Wahrscheinlichkeit ein a oder ein n zu ziehen nicht davon abhängt, ob ich zuvor ein a oder ein n gezogen habe, kann man guten Gewissens von der Unabhängigkeit der beiden Züge in Beispiel 3.9 sprechen. Die Situation in Beispiel 3.11 ist dagegen eine andere. Hier hängt die Wahrscheinlichkeit im zweiten Zug ein a oder ein n zu ziehen, sehr wohl vom ersten Zug ab, denn hier sind dann entsprechend weniger a's oder n's in der Urne, die beiden Züge sind also nicht unabhängig.

Wir wollen den Begriff der Unabhängigkeit nun mathematisch fassen. Dazu müssen wir zunächst den Begriff der "bedingten Wahrscheinlichkeit" definieren. Hierzu sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $A, B \in \mathcal{A}$ seien zwei Ereignisse. Nehmen wir nun an, jemand mit hellseherischen Fähigkeiten sagte uns, dass das Ereignis B eintritt. Was ist nun – bedingt darauf, dass wir wissen, dass B eintritt – die Wahrscheinlichkeit, dass auch A eintritt? Diese Wahrscheinlichkeit wollen wir mit $\mathbb{P}(A|B)$ bezeichnen. Da wir schon wissen, dass B eintritt, kommen für das Eintreten von A nur die $\omega \in \Omega$ in Frage, die sowohl in A ($\omega \in A$) als auch in B ($\omega \in B$) sind, also die ω mit

$$\omega \in A \cap B.$$

$\mathbb{P}(A|B)$ muss also proportional sein zu $\mathbb{P}(A \cap B)$. Allerdings ist $\mathbb{P}(A \cap B)$ in A keine Wahrscheinlichkeit mehr, denn für $A = \Omega$ gilt

$$\mathbb{P}(\Omega \cap B) = \mathbb{P}(B),$$

und dies ist in der Regel nicht 1. Um also eine Wahrscheinlichkeit zu erhalten, *teilen* wir $\mathbb{P}(A \cap B)$ durch $\mathbb{P}(B)$ (vorausgesetzt das ist nicht null, was wir aber getrost fordern dürfen, denn die Berechnung der Wahrscheinlichkeit eines Ereignisses A, vorausgesetzt ein Ereignis B tritt ein, dass aber sowieso nie eintritt, führt zu offensichtlichen logischen Problemen). Dies ergibt die folgende

Definition 3.16 *Es sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $A, B \in \mathcal{A}$ mit $\mathbb{P}(B) \neq 0$. Als die bedingte Wahrscheinlichkeit von A gegeben B definieren wir*

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

Beispiel 3.17 *Wir modellieren den einfachen fairen Würfelwurf mit $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{A} = \mathcal{P}\Omega$ und*

$$\mathbb{P}(\{\omega\}) = \frac{1}{6} \quad \textit{für alle} \quad \omega \in \Omega.$$

Es seien A und B die Ereignisse

$$\begin{aligned} A &= \textit{"Die Augenzahl ist durch 3 teilbar"}, \\ B &= \textit{"Die Augenzahl ist gerade"}, \end{aligned}$$

also

$$A = \{3, 6\} \quad \textit{und} \quad B = \{2, 4, 6\}.$$

Dann ist

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\{6\})}{\mathbb{P}(\{2,4,6\})} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}.$$

Mit Kenntnis von Definition 3.16 ist es nun einfach, den Begriff der Unabhängigkeit zu definieren. Es liegt nahe, zwei Ereignisse A und B, die den Voraussetzungen von Definition 3.16 genügen, unabhängig zu nennen, wenn die Kenntnis des Eintretens von B die Wahrscheinlichkeit des Eintretens von A nicht ändert, d. h. wenn

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

gilt. Mit anderen Worten soll gelten

$$\begin{aligned} & \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \mathbb{P}(A) \\ \Leftrightarrow \quad & \mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B). \end{aligned}$$

Die letzte Zeile hat hierbei den Vorteil, dass sie symmetrisch in A und B ist, dass sie auch für $P(B) = 0$ sinnvoll und zweckmäßig ist, und dass deshalb auf die Voraussetzung $\mathbb{P}(B) > 0$ verzichtet werden kann.

Definition 3.18 *Es sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, $A, B \in \mathcal{A}$ heißen unabhängig, falls*

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

gilt.

Beispiel 3.19 *Wir wollen nun zur Situation von Beispiel 3.9 zurückkehren und rechtfertigen, dass unsere Intuition dort nicht fehlgeleitet war und der erste und zweite Zug dort tatsächlich unabhängig sind. Hier war ja*

$$\begin{aligned} & \Omega = \{(a,a), (a,n), (n,a), (n,n)\} \\ & \mathcal{A} = \mathcal{P}\Omega \qquad \textit{und} \\ & \mathbb{P}(\{(a,a)\}) = \frac{3}{5} \cdot \frac{3}{5} = \frac{9}{25}, \ \mathbb{P}(\{(a,n)\}) = \frac{3}{5} \cdot \frac{2}{5} = \frac{6}{25}, \\ & \mathbb{P}(\{(n,a)\}) = \frac{2}{5} \cdot \frac{3}{5} = \frac{6}{25}, \ \mathbb{P}(\{(n,n)\}) = \frac{2}{5} \cdot \frac{2}{5} = \frac{4}{25}. \end{aligned}$$

Es sei nun

$$\begin{aligned} A &= \text{“}a \textit{ bei der ersten Ziehung}\text{”} \\ B &= \text{“}n \textit{ bei der zweiten Ziehung}\text{”}, \end{aligned}$$

also

$$A = \{(a,n),(a,a)\} \quad \textit{und} \quad B = \{(a,n),(n,n)\}.$$

Dann sind A und B unabhängig. In der Tat gilt ja

$$A \cap B = \{(a,n)\}$$

und weiter

$$\begin{aligned} \mathbb{P}(A) &= \frac{9}{25} + \frac{6}{25} = \frac{3}{5} \\ \mathbb{P}(B) &= \frac{6}{25} + \frac{4}{25} = \frac{2}{5}. \end{aligned}$$

Also

$$\mathbb{P}(A \cap B) = \frac{6}{25} = \frac{3}{5} \cdot \frac{2}{5} = \mathbb{P}(A)\mathbb{P}(B).$$

Dass Unabhängigkeit im stochastischen Sinn durchaus nicht immer etwas mit unserer Intuition, dass Unabhängigkeit von "unabhängigen" (d. h. verschiedenen, gegenseitig unbeeinflussten) Versuchen stammt, zu tun haben muss, zeigt ein weiterer Blick auf Beispiel 3.17.

Beispiel 3.20 *Die Ereignisse*

$$\begin{aligned} A &= \textit{"Die Augenzahl ist durch 3 teilbar"} \qquad \textit{und} \\ B &= \textit{"Die Augenzahl ist gerade"} \end{aligned}$$

sind beim einfachen fairen Würfeln unabhängig (obwohl beide Ereignisse sich auf denselben Wurf beziehen!). In der Tat haben wir ja in Beispiel 3.17 berechnet, dass

$$\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{1}{3}$$

gilt. Außerdem gilt auch $\mathbb{P}(A) = \frac{1}{3}$, also

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

Man kann sich nun fragen, wie sich der Begriff der Unabhängigkeit auf mehrere Ereignisse überträgt. Dass man für Ereignisse $A_1, \ldots, A_n$ nicht einfach ihre Unabhängigkeit mit dem Bestehen der Identität

$$\mathbb{P}(A_1 \cap \ldots \cap A_n) = \mathbb{P}(A_1) \ldots \mathbb{P}(A_n)$$

gleichsetzt, hat den Grund darin, dass man Unabhängigkeit zu einer *vererbbaren* Eigenschaft machen möchte, d. h. man möchte, dass mit $A_1, \ldots, A_n$ auch jede Teilfamilie von $A_1, \ldots, A_n$ unabhängig ist. Dies wäre mit einer solchen Definition nicht gewährleistet. Daher definiert man direkt:

Definition 3.21 *Es sei* $(\Omega, \mathcal{A}, \mathbb{P})$ *ein Wahrscheinlichkeitsraum und* $A_1, \ldots, A_n \in \mathcal{A}$, $n \in \mathbb{N},\ n \geq 2$. $A_1, \ldots, A_n$ *heißen unabhängig, falls für jedes* $1 \leq m \leq n$ *und alle* $i_1, \ldots, i_m$ *gilt*

$$\mathbb{P}(A_{i_1} \cap \ldots \cap A_{i_m}) = \mathbb{P}(A_{i_1}) \ldots \mathbb{P}(A_{i_m}).$$

Man sieht unmittelbar, dass diese Definition dem oben formulierten Vererbbarkeitsgrundsatz genügt. Der Zusammenhang zwischen der Unabhängigkeit zum einen und den im vorhergehenden Kapitel formulierten Pfadregeln zum anderen besteht nun darin, dass diese Pfadregeln für unabhängige Versuche unmittelbar evident sind. In der Tat wollen wir bei n Zufallsexperimenten, $n \in \mathbb{N},\ n \geq 2$, davon sprechen, dass diese unabhängig sind, falls alle Ereignisse $A_1, \ldots, A_n$, wobei $A_j,\ j = 1, \ldots, n$, ein beliebiges Ereignis ist, das sich aber *nur auf das j'te Experiment beziehen darf*, unabhängig sind. Damit folgt

Proposition 3.22 *Sind die Experimente in einem mehrstufigen Zufallsversuch unabhängig, so gilt die Pfadregel, Definition 3.10.*

Beweis: Dies folgt in der Tat unmittelbar aus der Definition der unabhängigen Experimente bzw. der Definition der Unabhängigkeit von n Ereignissen, Definition 3.21. □

Allerdings haben wir in Definition 3.10 die Pfadregel nicht nur für unabhängige, sondern für beliebige Ereignisse formuliert. Wieso ist diese richtig? [Die Frage ist formal falsch, denn eine Definition kann natürlich nicht richtig oder falsch sein, die Frage ist vielmehr, wieso Definition 3.10 nicht im Widerspruch zur in diesem Kapitel entwickelten Theorie steht.] Die Antwort erhält man, indem man die Formel für die bedingte Wahrscheinlichkeit umstellt. Es ist ja

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

Analog erhält man für drei Ereignisse $A, B, C \in \mathcal{A}$ durch zweimaliges Anwenden dieser Regel:

$$\begin{aligned} \mathbb{P}(A \cap B \cap C) &= \mathbb{P}(A \cap B|C) \cdot \mathbb{P}(C) \\ &= \mathbb{P}(A|B \cap C) \cdot \mathbb{P}(B \cap C|C) \cdot \mathbb{P}(C) \\ &= \mathbb{P}(A|B \cap C) \cdot \mathbb{P}(B|C) \cdot \mathbb{P}(C) \end{aligned}$$

(dies lässt sich auch beweisen, indem man einfach die Definitionen der entsprechenden bedingten Wahrscheinlichkeiten einsetzt).

Ebenso zeigt man die folgende Produktformel für bedingte Wahrscheinlichkeiten von Ereignissen $A_1, \ldots, A_n,\ n \in \mathbb{N},\ n \geq 2$:

$$\mathbb{P}(A_1 \cap \ldots \cap A_n) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2|A_1) \cdot \mathbb{P}(A_3|A_2 \cap A_1) \ldots \mathbb{P}(A_n|A_1 \cap \ldots \cap A_{n-1}). \tag{3.4}$$

Hat man nun ein mehrstufiges Zufallsexperiment mit n Stufen, und $A_1, \ldots, A_n$ sind Ereignisse, wobei sich das Ereignis A_j nur auf das j-te Experiment bezieht, so ist (3.4) nichts anderes als die Pfadregel Definition 3.10, die wir somit auf dem Hintergrund der bedingten Wahrscheinlichkeit gerechtfertigt haben.

3.4 Kombinatorik I – Produkt- und Summenregel

Wir kehren hier zur Definition eines Laplace-Experiments, Definition 3.6, zurück. In solchen Experimenten über einem *endlichen* Zustandsraum Ω gilt für alle $A \subseteq \Omega$

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

Offenbar ist für die Bestimmung der Wahrscheinlichkeit eines Ereignisses A ein Zählproblem zu lösen. Wer sich jemals ein wenig mit Kombinatorik beschäftigt hat, weiß, dass Zählen in der Tat ein Problem sein kann. Wir werden in diesem Abschnitt zwei einfache Zählregeln formulieren und ihre Konsequenzen für Stichprobenumfänge aufzeigen. Es sei jedoch hier bemerkt, dass die Kombinatorik, also die Bestimmung der Größen von endlichen Mengen, das Auffinden von Zählprinzipien, eine Teildisziplin der (diskreten) Mathematik ist, dass sie also nicht Teildisziplin der Wahrscheinlichkeitstheorie ist, und dass es extrem harte kombinatorische Probleme gibt, die bis zum heutigen Tage nicht gelöst sind. Die erste ist die *Summenregel*:

Regel 3.23 *Dazu sei Ω eine endliche Menge und $A_1, \ldots, A_r \subset \Omega$ seien paarweise disjunkt, d. h.*

$$A_i \cap A_j = \emptyset \quad \textit{für alle} \quad i \neq j.$$

Es sei

$$A := A_1 \cup A_2 \cup \ldots \cup A_r.$$

Dann gilt

$$|A| = |A_1| + \ldots + |A_r| = \sum_{j=1}^{r} |A_j|.$$

Dies lässt sich (lustigerweise) mit der Definition der Laplace-Wahrscheinlichkeit beweisen. Wegen der 3. Eigenschaft von Wahrscheinlichkeiten gilt ja für $A, B \in \mathcal{A}$

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

(denn $\mathbb{P}(A \cup B) = \mathbb{P}(A \cup B \backslash A) = \mathbb{P}(A) + \mathbb{P}(B \backslash A)$). Also gilt für $A \cap B = \emptyset$

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

und allgemein für paarweise disjunkte $A_1, \ldots, A_n$

$$\mathbb{P}(A) = \mathbb{P}(\bigcup_{j=1}^{n} A_j) = \sum_{j=1}^{n} \mathbb{P}(A_j).$$

Setzt man für $\mathbb{P}(A_j)$ die Laplace-Wahrscheinlichkeit $\frac{|A_j|}{|\Omega|}$ ein, so erhält man für paarweise disjunkte $A_1, \ldots, A_n$

$$\frac{|A|}{|\Omega|} = \frac{|\bigcup_{j=1}^{n} A_j|}{|\Omega|} = \sum_{j=1}^{n} \frac{|A_j|}{|\Omega|}.$$

Gilt ferner $\bigcup_{j=1}^{n} A_j = \Omega$, so ist die linke Seite 1, also

$$1 = \sum_{j=1}^{n} \frac{|A_j|}{|\Omega|}.$$

Dies ist gleichbedeutend mit

$$|\Omega| = \sum_{j=1}^{n} |A_j|.$$

Das zweite Zählprinzip ist die Produktregel. Wir formulieren sie so:

Regel 3.24 *Ein Versuch bestehe aus r Stufen, wobei der Ausgang einer Stufe keinen Einfluss auf die Ergebnisse späterer Stufen habe. Haben die einzelnen Versuche $n_1, n_2, \ldots, n_r$ Ausfälle, dann hat der Gesamtversuch*

$$n_1 \cdot n_2 \cdot \ldots \cdot n_{r-1} \cdot n_r$$

mögliche Ausgänge.

Dies kann man (wenn man möchte) ähnlich wie die Summenregel aus den Prinzipien der Wahrscheinlichkeit ableiten, indem man die Unabhängigkeit benutzt. Wir überlassen dies dem Leser/der Leserin und zeigen die Nützlichkeit dieser Regeln anhand einiger Beispiele.

Beispiel 3.25 *Eine Münze mit den Seiten 0 und 1 werde 6 mal geworfen. Wieviele mögliche Ausgänge gibt es? Da jeder Münzwurf zwei mögliche Ausgänge hat und die Münzwürfe sich nicht gegenseitig beeinflussen, gibt es insgesamt*

$$2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 = 2^6 = 64$$

verschiedene mögliche Versuchsausgänge. Allgemein gibt es im n-fachen Münzwurf 2^n mögliche Versuchsausgänge.

Beispiel 3.26 *In einer Urne sind 5 Kugeln mit den Nummern 1 bis 5. Hiervon werden drei Kugeln nacheinander ohne Zurücklegen gezogen. Wieviele Ausfälle gibt es? Dieser Versuch hat drei Stufen. In der ersten Stufe können 5 Kugeln gezogen werden, in der zweiten 4, in der dritten 3. Der Ausgang einer Stufe beeinflusst die Möglichkeiten in der nächsten Stufe (eine einmal gezogene Kugel kann nicht nochmals gezogen werden), nicht aber ihre Anzahl (die Anzahl der restlichen Kugeln ist stets dieselbe). Also folgt mit der Produktregel, dass der Gesamtversuch*

$$5 \cdot 4 \cdot 3 = 60$$

mögliche Ausfälle hat.

Beispiel 3.27 *Frau M. hat 6 Hüte, 5 Blusen und 5 Röcke, jeweils in den Farben schwarz und weiß und zwar 3 schwarze Hüte (und 3 weiße), 2 schwarze Röcke (und 3 weiße) und 3 schwarze Blusen (und 2 weiße). Gibt es keine Restriktion, so gibt es nach der Produktregel* $6 \cdot 5 \cdot 5 = 150$ *Kombinationsmöglichkeiten. Nun trägt Frau M. aber höchstens ein schwarzes Kleidungsstück. Die Produktregel ist nicht mehr direkt anwendbar. Jede der Kombinationsmöglichkeiten www, sww, wsw, wws bestimmt eine Teilmenge der möglichen Kombinationsmöglichkeiten. Nach der Produktregel haben diese* $3 \cdot 3 \cdot 2$, $3 \cdot 3 \cdot 2$, $3 \cdot 2 \cdot 2$ *und* $3 \cdot 3 \cdot 3$ *mögliche Kombinationen. Da die Teilmengen disjunkt sind, gibt es nach der Summenformel insgesamt*

$$3 \cdot 3 \cdot 2 + 3 \cdot 3 \cdot 2 + 3 \cdot 2 \cdot 2 + 3 \cdot 3 \cdot 3 = 75$$

mögliche Kleidungsmöglichkeiten für Frau M.

3.5 Kombinatorik II – Stichproben

Wir werden nun die oben gelernten Regeln auf die Bestimmung der Anzahl von Stichprobenmengen anwenden. Hierbei spielen zwei Aspekte eine Rolle: Die Auswahl der Teilmenge und die Frage, ob wir die Anordnung der Elemente berücksichtigen wollen. Je nachdem, ob wir der Reihenfolge der Stichprobenelemente eine Bedeutung beimessen oder nicht, sprechen wir von *geordneten* bzw. *ungeordneten* Stichproben.

Die Ausgangssituation in diesem Kapitel wird stets sein, dass wir eine Grundgesamtheit Ω vom Umfang n haben (z. B. eine Urne mit n Kugeln, die wir mit $1, \ldots, n$ nummerieren), aus der wir s mal mit und mal ohne Zurücklegen ziehen.

Zunächst behandeln wir

Geordnete Stichproben mit Zurücklegen
Hier zieht man also eine von n Kugeln zufällig, notiert ihre Nummer und legt sie wieder zurück. Dies wird s-mal wiederholt. Man hat also einen Versuch mit s Stufen, von denen keine die anderen beeinflusst und jede n mögliche Ausgänge hat. Nach der Produktregel gibt es also n^s mögliche Versuchsausgänge des Gesamtversuchs.

Bemerkung 3.28

1. *Bei einer geordneten Stichprobe zählt also die Reihenfolge der Stichprobenelemente. Ein gutes Beispiel hierfür sind Wörter. Wenn ich aus einem Alphabet mit Zurücklegen die Buchstaben B, E, I und L ziehe, so können diese je nach Reihenfolge die Wörter*

   $$BEIL,\ BLEI,\ LEIB \quad oder \quad LIEB$$

   ergeben. Diese haben offensichtlich unterschiedliche Bedeutungen.

2. *Wenn man die Kugeln im obigen Experiment rein zufällig, d. h. gemäß einer Laplace-Wahrscheinlichkeit, zieht, so hat jeder der n^s möglichen Versuchsausgänge die Wahrscheinlichkeit $\frac{1}{n^s}$.*

3. *Es gibt eine andere (eine "duale") Darstellung desselben Experiments. Statt s mal aus n Kugeln zu ziehen, werfen wir s Kugeln rein zufällig in n Urnen. Genauer seien s unterscheidbare (z. B. nummerierte) Kugeln gegeben und n unterscheidbare Urnen. Wir legen nun die s Kugeln in die n Urnen. Jede Urne darf beliebig viele Kugeln aufnehmen. Da jede der s Kugeln in eine von n Urnen fällt, können wir das Ergebnis unseres Experiments durch einen Vektor der Länge s*

$$(a_1, \ldots, a_s) \in \{1, \ldots, n\}^s$$

notieren. Dabei beschreibt a_j die Nummer der Urne, in der die j-te Kugel gelandet ist. Da jedes a_j einen Wert zwischen 1 und n annehmen kann, gibt es insgesamt

$$\underbrace{n \cdot n \cdot \ldots \cdot n}_{s\text{-mal}} = n^s$$

Möglichkeiten.

Wir wenden uns nun einigen wichtigen Beispielen zu:

Beispiel 3.29 *Es sei $A = \{1, \ldots, s\}$ und $B = \{1, \ldots, n\}$. Wie viele Funktionen von A nach B gibt es?*

$$f : A \to B$$

muss jedem Element aus A ein Element aus B zuordnen. f transportiert also sozusagen jedes Element $j \in A$, $j = 1, \ldots, s$, in eine der n Urnen. Es gibt somit n^s solcher Funktionen.

Beispiel 3.30 *Die Anzahl der Teilmengen einer n-elementigen Menge (inklusive der Menge selbst und der leeren Menge) lässt sich mit Hilfe von Beispiel 3.29 berechnen. Dieses Beispiel ist von großer Wichtigkeit. Hierzu sei*

$$\Omega = \{\omega_1, \ldots, \omega_n\}$$

eine n-elementige Menge. Wir können nun eine Teilmenge A von Ω durch eine Abbildung f_A beschreiben

$$\begin{aligned} f_A : \quad & \Omega \to \{0, 1\} \\ & \omega \mapsto \begin{cases} 0 & \omega \notin A \\ 1 & \omega \in A \end{cases} . \end{aligned}$$

Durch diese Beschreibung sehen wir, dass es ebenso viele Teilmengen von Ω gibt, wie es Funktionen von Ω nach $\{0, 1\}$ gibt. Die Anzahl dieser Funktionen haben wir aber in Beispiel 3.29 bestimmt. Da es n Elemente in Ω und 2 Elemente in $\{0, 1\}$ gibt, gibt es 2^n solcher Funktionen, also auch 2^n Teilmengen von Ω.

Beispiel 3.31 *Wie viele Tippreihen im Toto gibt es? Dazu muss man wissen, dass für das gewöhnliche Fußballtoto, die sogenannte 11er-Wette, die Resultate von 11 Spielen zu tippen sind und zwar jeweils Sieg der Heimmannschaft ("1"), Unentschieden ("0") oder Sieg der Gastmannschaft ("2"). Das Ausfüllen eines Tippzettels ist also ein Versuch mit 11 Stufen und 3 möglichen Ausgängen pro Stufe. Nach dem oben Überlegten gibt es also*

$$3^{11} = 177\ 147$$

Tippreihen. Man kann sich auch fragen, wie viele davon komplett falsch sind. Es gibt pro Spiel 2 falsche Tippergebnisse, also insgesamt

$$2^{11} = 2\ 048$$

komplett falsche verschiedene Tippzettel. Dies ist verglichen mit den 177 147 insgesamt möglichen Tippreihen sehr wenig (ca. 1,1 %), so dass Sie mit großer Wahrscheinlichkeit wenigstens ein Spiel richtig vorhersagen, selbst wenn Sie gar keine Ahnung von Fußball haben.

Geordnete Stichproben ohne Zurücklegen
Auch hier ziehen wir wieder aus einer Urne mit n Kugeln mit den Nummern $1, \ldots, n$ s Kugeln, diesmal jedoch *ohne die Kugeln zurückzulegen.* Man notiert ihre Nummern in der Reihenfolge, in der sie erscheinen. Es gibt also insgesamt s Stufen in diesem Versuch. Anders als im ersten Fall sind diese aber keine Kopien voneinander, denn die Kugeln, die in den vorherigen Ziehungen gezogen wurden, beeinflussen das Ergebnis der nachfolgenden Ziehungen. Es sind also der Reihe nach

$$n, n-1, n-2, \ldots, n-(s-1)$$

Ergebnisse in den aufeinanderfolgenden Stufen möglich. Nach der Produktregel folgt, dass das Gesamtexperiment

$$(n)_s := n \cdot (n-1) \ldots (n-s+1)$$

mögliche Ausgänge hat. Ist $s = n$, d. h. wird die Urne vollständig leer gemacht, so gibt es

$$n! := (n)_n = n \cdot (n-1) \ldots 3 \cdot 2 \cdot 1$$

mögliche Versuchsausgänge. Insbesondere ergibt sich:
Eine Menge mit n Elementen lässt sich auf $n!$ verschiedene Arten anordnen.

Üblicherweise setzt man $0! = 1$. $n!$ ist eine äußerst schnell wachsende Funktion in n (schneller als exponentiell!). Das führt dazu, dass auch für moderat große n, $n!$ nicht mehr gut zu berechnen ist (schon 70! sprengt die Exponentialanzeige eines handelsüblichen Taschenrechners). Für größere n benutzt man daher die Approximation der *Stirlingformel*

$$n! \sim (\frac{n}{e})^n \sqrt{2\pi n}.$$

Hierbei schreiben wir für zwei Folgen $(a_n)_n$ und $(b_n)_n$, $a_n \sim b_n$, falls

$$\frac{a_n}{b_n} \to 1 \quad \text{für} \quad n \to \infty$$

gilt und bemerken, dass dies nicht impliziert, dass $a_n - b_n$ gegen 0 konvergiert (für die Stirlingformel ist dies sogar definitiv falsch).

Bemerkung 3.32 *1. Bei einer zufälligen Ziehung hat somit jede Stichprobe die Wahrscheinlichkeit*

$$\frac{1}{(n)_s} = \frac{(n-s)!}{n!}.$$

2. Auch hier gibt es eine duale, manchmal bequemere Formulierung des Problems: Es sollen s unterscheidbare Kugeln auf n unterscheidbare Urnen verteilt werden und es sei $s \leq n$. Es gelte aber das Ausschlussprinzip: Pro Urne darf man höchstens eine Kugel haben (dieses Prinzip ist in der Tat in manchen physikalischen Modellen sehr wichtig, siehe z. B. Fermi-Dirac-Statistiken). Wieviele Verteilungen gibt es? Da man die erste, zweite, ..., s-te Kugeln auf $n, n-1, \ldots, (n-s+1)$ Arten unterbringen kann, gibt es

$$(n_s) = n \cdot (n-1) \ldots (n-s+1)$$

verschiedene Ausfälle.

Beispiel 3.33 *Wie viele injektive Abbildungen gibt es von $A = \{1, \ldots, s\}$ nach $B = \{1, \ldots, n\}$? Wenn $s > n$ ist, ist klar, dass es gar keine injektiven Abbildungen von A nach B geben kann. Ist $s \leq n$, so lässt sich ein injektives*

$$f : A \to B$$

gerade als die Funktion auffassen, die jede der s Kugeln in eine Urne schickt, so dass nicht zwei Kugeln in derselben Urne landen. Nach der obigen Bemerkung muss es also $(n)_s$ viele solcher injektiven Abbildungen geben.

Beispiel 3.34 *Auf wie viele Arten können 8 Türme so auf ein Schachbrett gestellt werden, dass sie sich gegenseitig nicht schlagen können? Dazu muss man vom Schach nur wissen, dass Türme vertikal und horizontal ziehen können und dass ein Schachbrett 8×8 Felder hat. Folglich muss also in jeder Reihe und Spalte ein Turm stehen. Wir positionieren die Türme spaltenweise. Für den Turm in der ersten Spalte gibt es noch 8 Möglichkeiten (8 Reihen, in die ich ihn stellen kann), für den zweiten 7 usw. Also gibt es insgesamt*

$$8! = 40\ 320$$

mögliche Anordnungen. Ähnlich überlegt man sich, dass es auf einem $n \times n$-Brett $n!$ Möglichkeiten gibt um n Türme so zu positionieren, dass sie sich nicht schlagen können.

Beispiel 3.35 *Eine Sekretärin tippt n Briefe und adressiert n Umschläge. Dann steckt sie die Briefe in die Umschläge, ohne auf die Adressen zu schauen (das macht eine gute Sekretärin natürlich üblicherweise nicht; so ein Verhalten ist eher einem schusseligen Mathematikprofessor zuzutrauen). Wie groß ist die Wahrscheinlichkeit, dass mindestens ein Brief in den richtigen Umschlag kommt?*

Wir lösen das komplementäre Problem, nämlich die Bestimmung der Wahrscheinlichkeit, dass keiner der Briefe richtig adressiert wird. Mathematisch kann man dieses Problem so

formulieren: Wieviele fixpunktfreie Permutationen der Menge $\{1,\dots,n\}$ *gibt es? Hierbei ist eine Permutation eine injektive Abbildung*

$$\sigma : \{1,\dots,n\} \to \{1,\dots,n\}$$

und ein Fixpunkt einer solchen Permutation ist ein $j \in \{1,\dots,n\}$*, für welches* $\sigma(j) = j$ *gilt. Eine Permutation heißt fixpunktfrei, falls es keine Fixpunkte gibt. Es sei* a_n *die Anzahl der fixpunktfreien Permutationen von* $\{1,\dots,n\}$*. Offenbar gilt* $a_1 = 0$ *(ein Element steht notwendig auf seinem Platz) und* $a_2 = 1$*, denn* $\sigma(1) = 2,\ \sigma(2) = 1$ *ist die einzige fixpunktfreie Permutation der Menge* $\{1,2\}$*. Allgemein überlegt man sich folgendes: Sei* σ *eine fixpunktfreie Permutation von* $\{1,\dots,n\}$*. Somit ist* $\sigma(1) \neq 1$*. Wir nehmen an,* $\sigma(1) = 2$ *und multiplizieren das Ergebnis dann mit* $(n-1)$*, denn* $\sigma(1)$ *kann auf die* $(n-1)$ *Plätze* $2,3,\dots,n$ *verteilt sein. Ist nun gleichzeitig* $\sigma(2) = 1$*, so müssen nur noch die Elemente* $3,\dots,n$ *fixpunktfrei permutiert werden. Hierfür gibt es* a_{n-2} *Möglichkeiten. Ist* $\sigma(2) \neq 1$*, so benötigen wir eine Permutation, die 2 nicht auf 1, 3 nicht auf 3, 4 nicht auf 4 ... und n nicht auf n wirft. Davon gibt es* a_{n-1} *viele. Also insgesamt die Rekursionsformel*

$$a_n = (n-1)(a_{n-1} + a_{n-2}).$$

Hiermit lassen sich aus $a_1 = 0$ *und* $a_2 = 1$ *sukzessive die* a_n *berechnen; z. B. ist*

$$a_3 = 2,\ a_4 = 9 \quad \textit{und} \quad a_5 = 44.$$

Die Frage ist, ob sich auch eine geschlossene Formel für die a_n *herleiten lässt. Dies geschieht so: Setzt man die Rekursion für* a_n *ein, so ergibt sich*

$$a_n - na_{n-1} = -[a_{n-1} - (n-1)a_{n-2}]. \tag{3.5}$$

Setzt man

$$d_n := a_n - na_{n-1},\ n \geq 2,$$

so ergibt (3.5)

$$d_n = -d_{n-1},$$

also

$$d_n = (-1)d_{n-1} = (-1)^2 d_{n-2} = \dots = (-1)^{n-2} d_2.$$

Daher ist

$$d_n = (-1)^n \quad \textit{und} \quad a_n - na_{n-1} = (-1)^n.$$

Daraus folgt

$$a_n = na_{n-1} + (-1)^n.$$

Dividiert man dies durch $n!$*, so erhält man*

$$\frac{a_n}{n!} = \frac{a_{n-1}}{(n-1)!} + \frac{(-1)^n}{n!}. \tag{3.6}$$

Nun lässt sich der Term $\frac{a_{n-1}}{(n-1)!}$ *auf der rechten Seite von (3.6) ebenso entwickeln*

$$\frac{a_{n-1}}{(n-1)!} = \frac{a_{n-2}}{(n-2)!} + \frac{(-1)^{n-1}}{(n-1)!}$$

usw. Dies ergibt

$$\frac{a_n}{n!} = \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} + \ldots + \frac{(-1)^n}{n!}.$$

Da zudem $0! = 1!$ *ist, also auch* $\frac{1}{0!} = \frac{1}{1!}$ *gilt, folgt*

$$a_n = n!(\frac{1}{0!} - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \ldots + \frac{(-1)^n}{n!}).$$

Der Klammerausdruck auf der rechten Seite konvergiert gegen $\frac{1}{e}$*; somit ist für große* n

$$\frac{a_n}{n!} \sim \frac{1}{e} \approx 0,3679.$$

$\frac{a_n}{n!}$ *ist aber gerade die Wahrscheinlichkeit, rein zufällig eine fixpunktfreie Permutation auszuwählen. Umgekehrt ist die Wahrscheinlichkeit für mindestens einen Fixpunkt*

$$1 - \frac{a_n}{n!} \approx 1 - \frac{1}{e} \approx 0,6321,$$

also beinahe $\frac{2}{3}$*. Die (von* n *abhängige) Wahrscheinlichkeit, dass die Sekretärin also mindestens einen Brief korrekt adressiert, ist daher also ungefähr 0,6321 (was unabhängig ist von* n*).*

Ungeordnete Stichproben ohne Zurücklegen

Nicht immer spielt bei einer Stichprobe die Reihenfolge der Stichprobenelemente eine Rolle. Beispielsweise ist es beim Lotto (was einem Ziehen der Kugeln *ohne* Zurücklegen entspricht) irrelevant, die Kugeln in der richtigen Reihenfolge vorherzusagen; wichtig ist nur, dass man die auftretenden Kugeln überhaupt vorhersagt. Gegeben sei also eine Menge von n Elementen. Daraus werde (quasi in einem Griff) eine Menge von s Elementen gezogen. Wir wollen die Anzahl der möglichen Ergebnisse mit $\binom{n}{s}$ bezeichnen. Offenbar ist $\binom{n}{s}$ die Anzahl der s-elementigen Teilmengen von $\{1, \ldots, n\}$. $\binom{n}{s}$ lässt sich relativ einfach bestimmen. Dazu erinnern wir uns, dass es

$$(n)_s = \frac{n!}{(n-s)!}$$

mögliche geordnete Stichproben aus $\{1, \ldots, n\}$ gibt. Da man jede dieser Stichproben auf $s!$ Arten permutieren kann und dabei dieselbe ungeordnete Stichprobe erhält, gilt

$$\binom{n}{s} = \frac{(n)_s}{s!} = \frac{n!}{(n-s)!s!} = \binom{n}{n-s}.$$

Bemerkung 3.36 *1. Es ist* $\binom{n}{0} = 1$*, denn es gibt nur eine 0-elementige Teilmenge jeder Menge, die leere Menge.*

2. *Man stelle sich vor, es soll aus einer n-elementigen Population ein Ausschuss der Größe s gebildet werden, von denen einer der Vorsitzende ist. Dies kann entweder so geschehen, dass man erst den Vorsitzenden wählt – dafür gibt es $\binom{n}{1}$ Möglichkeiten – und dann die $s-1$ restlichen Mitglieder (das kann auf $\binom{n-1}{s-1}$ Arten geschehen), oder man wählt erst die s Mitglieder – das geht auf $\binom{n}{s}$ Möglichkeiten – und die wählen einen Vorsitzenden (d. h. $s = \binom{s}{1}$ Möglichkeiten). Da beides dieselbe Anzahl möglicher Konstellationen liefern muss, gilt*

$$n\binom{n-1}{s-1} = s(\frac{n}{s}) \quad oder \quad \binom{n}{s} = \frac{n}{s}\binom{n-1}{s-1}.$$

3. *Wir hatten gezeigt, dass die Menge $\{1,\ldots,n\}$ 2^n Teilmengen hat. Es gibt weiter $\binom{n}{0}, \binom{n}{1}, \ldots, \binom{n}{n}$ Teilmengen der Größe $0, 1, \ldots, n$. Also gilt*

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

Dies lässt sich aus dem Binomischen Lehrsatz ableiten. Es ist ja

$$\sum_{k=0}^{n} \binom{n}{k} = \sum_{k=0}^{n} \binom{n}{k} \cdot 1^k \cdot 1^{n-k} \underset{\text{Binom. Lehrsatz}}{=} (1+1)^n = 2^n.$$

Beispiel 3.37 *Das Lottoproblem: Eine Urne enthalte 49 Kugeln mit den Nummern 1 bis 49. Sechs davon seien rot, die restlichen weiß. Es wird eine Stichprobe von 6 Kugeln entnommen. Es gibt – da es nicht auf die Reihenfolge der entnommenen Kugeln ankommen soll – insgesamt*

$$\binom{49}{6} = \frac{49 \cdot 48 \cdot 47 \cdot 46 \cdot 45 \cdot 44}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6} = 13\ 983\ 816$$

mögliche Stichproben. Bei rein zufälliger Auswahl hat also jede Stichprobe eine Wahrscheinlichkeit von 1:13 983 816 gezogen zu werden. Wir fragen nun nach der Wahrscheinlichkeit, dass eine Stichprobe (die rein zufällig gezogen ist), $0, 1, \ldots, 6$ rote Kugeln enthält. Diese Wahrscheinlichkeit muss nach dem soeben Überlegten

$$\mathbb{P}(\text{“}s \text{ rote Kugeln}'') = \frac{f(s)}{\binom{49}{6}}$$

sein, wobei $f(s)$ die Anzahl der Stichproben mit s roten Kugeln beschreibt. $f(s)$ lässt sich nun wie folgt berechnen: Man muss genau s rote Kugeln ziehen, dafür gibt es $\binom{6}{s}$ Möglichkeiten. Aus den 43 anderen Kugeln müssen $6-s$ Kugeln gezogen werden, hierfür gibt es $\binom{43}{6-s}$ Möglichkeiten. Nach der Produktregel ist also

$$f(s) = \binom{6}{s}\binom{43}{6-s}.$$

Beispiel 3.38 *Auf wie viele Arten kann man 3 A's, 4 B's und 5 C's anordnen? (Hierbei ist angenommen, dass sich gleiche Buchstaben untereinander nicht unterscheiden.) Insgesamt kann man die 12 Buchstaben auf 12! Arten permutieren. Nun kann man die Plätze, auf denen die A's bzw. B's bzw. C's stehen, permutieren, ohne dass dies etwas an der entstehenden Konfiguration ändert. Dies geht auf 3!4!5! (Produktregel!) verschiedene Arten. Also gibt es*

$$\frac{12!}{3!4!5!} = 27\ 720$$

mögliche Konfigurationen.

Ähnlich berechnet sich die Anzahl der möglichen Permutationen von a 0en und b 1en $(a, b \in \mathbb{N})$*. Insgesamt hat man also* $a + b$ *Elemente zu permutieren. Eine Permutation ist eindeutig festgelegt, wenn man die Positionen der 0en kennt. Diese kann man offenbar auf* $\binom{a+b}{a}$ *Weisen wählen. Da die Permutation ebenso entsteht, wenn man die Position der 1en kennt und es dafür* $\binom{a+b}{b}$ *Möglichkeiten gibt, folgt: Man kann a 0en und b 1en auf*

$$\binom{a+b}{a} = \binom{a+b}{b}$$

Arten anordnen.

Beispiel 3.39 *Was ist die Wahrscheinlichkeit, beim 10-fachen Münzwurf genau 5 mal "Kopf" zu werfen, wenn die Münze fair ist? Wir formalisieren das Experiment folgendermaßen (das sollte man an sich immer machen):*

$$\begin{aligned}
&\Omega = \{0,1\}^{10} = \{(\omega_1, \ldots, \omega_{10}) : \omega_i \in \{0,1\}\} \\
&\mathcal{A} = \mathcal{P}\Omega \\
&\mathbb{P}(\{\omega\}) = \frac{1}{|\Omega|} = \frac{1}{2^{10}} \quad \textit{für alle} \quad \omega \in \Omega.
\end{aligned}$$

Die uns interessierende Menge ist

$$A = \{\omega \in \Omega : \omega_1 + \ldots + \omega_{10} = 5\}.$$

Nach Beispiel 3.37 hat A gerade $\binom{10}{5}$ *Elemente. Also ist*

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{\binom{10}{5}}{2^{10}} = \frac{252}{1\ 024} \approx \frac{1}{4}.$$

Wir kommen nun zur letzten (und schwierigsten) Möglichkeit:

Ungeordnete Stichproben mit Zurücklegen
Aus einem (unbegrenzten) Vorrat von 4 verschiedenen Sorten von Dingen a, b, c, d soll eine ungeordnete Stichprobe von 7 Objekten ausgewählt werden. Auf wie viele Arten geht das? Offenbar lässt sich dieselbe Frage auch so stellen: Aus der Menge $\{a, b, c, d\}$ soll mit Zurücklegen eine Stichprobe vom Umfang 7 entnommen werden. Auf wie viele Arten

geht das? Das duale (und somit wieder äquivalente) Problem lautet: 7 ununterscheidbare Kugeln sollen auf 4 Urnen a, b, c, d verteilt werden. Wie viele Möglichkeiten gibt es?

In jeder der drei Fragen kann man eine Stichprobe durch ein Tupel aus 7 0en und 3 Strichen (|) beschreiben in der Form

$$00|000|0|0.$$

Dabei steht in diesem Beispiel die erste Gruppe von 0en für zwei a-Kugeln, die zweite für drei b-Kugeln, die dritte für eine c-Kugel und die vierte für eine d-Kugel. Die Anzahl der möglichen Versuchsausgänge ist also durch die Anzahl der möglichen Permutationen der 7 0en und 3 1en beschrieben. Nach dem, was wir im vorigen Kapitel überlegt haben, gibt es hierfür

$$\binom{7+3}{3} = \binom{10}{3} = \binom{10}{7} = 120$$

Möglichkeiten.

Ersetzt man allgemein 7 durch s und 4 durch n, so erhält man:

Man kann s ununterscheidbare Kugeln auf $\binom{n+s-1}{s}$ Arten in n Urnen legen

und

Eine Menge der Mächtigkeit n enthält $\binom{n+s-1}{s}$
Stichproben vom Umfang s mit Zurücklegen.

Beispiel 3.40 *Wir nehmen für einen Moment an, das beliebte Lotto 6 aus 49 würde mit Zurücklegen gespielt. Wie viele mögliche Ausgänge gäbe es, wenn die Reihenfolge der gezogenen Kugeln nach wie vor keine Rolle spielte? Anwort: Nach dem soeben Gesagten gäbe es*

$$\binom{49+6-1}{6} = \binom{54}{6}$$

Stichproben. Da dies noch größer ist als $\binom{49}{6}$, wäre ein 6er hier noch schwieriger. Genauer wäre die Wahrscheinlichkeit für 6 Richtige dann

$$\frac{1}{\binom{54}{6}} = \frac{1}{25\ 827\ 165}.$$

3.6 Unendliche Wahrscheinlichkeitsräume

Um nicht den Eindruck zu erwecken, es gäbe nur endliche Wahrscheinlichkeitsräume, geben wir ein kurzes Intermezzo über Wahrscheinlichkeiten auf *unendlichen* Mengen Ω. Diese sind in gewissem Sinne oft die eigentlich interessanten, können aber meist nicht

durch so naive Methoden behandelt werden, wie die eben vorgestellten (z. B. ist die einfache Laplace-Verteilung auf solchen Mengen offensichtlich unmöglich). Wir beschäftigen uns hier mit einem kleinen Abriss interessanter Beispiele.

Das erste Beispiel eines unendlichen Wahrscheinlichkeitsraumes ist das einer Wartezeit. Diese kann zumindest prinzipiell unendlich sein.

Beispiel 3.41 *Eine faire Münze wird so lange geworfen, bis zum ersten Mal Kopf erscheint. Hier ist*

$$\Omega = \mathbb{N},\ \mathcal{A} = \mathcal{P}\Omega.$$

Die Elemente aus $\mathbb{N}$ *entsprechen hierbei den folgenden Münzwürfen*

$$\begin{aligned} 1 &\stackrel{\wedge}{=} K \\ 2 &\stackrel{\wedge}{=} Z, K \\ 3 &\stackrel{\wedge}{=} Z, Z, K \\ 4 &\stackrel{\wedge}{=} Z, Z, Z, K \\ &\vdots \end{aligned}$$

Da der einzelne Münzwurf ein Laplace-Experiment ist, und die einzelnen Münzwürfe unabhängig sind, sollten wir

$$\mathbb{P}(\{1\}) = \frac{1}{2},\ \mathbb{P}(\{2\}) = \frac{1}{4}, \ldots$$

und allgemein

$$\mathbb{P}(\{n\}) = 2^{-n}$$

setzen. Das so gewählte $\mathbb{P}$ *ist in der Tat eine Wahrscheinlichkeit auf* $\Omega = \mathbb{N}$*, denn*

$$\mathbb{P}(\Omega) = \sum_{n\in\mathbb{N}} \mathbb{P}(\{n\}) = \sum_{n=1}^{\infty} 2^{-n} = \frac{1}{2}\sum_{n=0}^{\infty} 2^{-n} = \frac{1}{2}\frac{1}{1-\frac{1}{2}} = \frac{1}{2}\cdot 2 = 1.$$

Bezeichnen wir mit X *die Anzahl der benötigten Würfe bis zum ersten Erscheinen von "Kopf", so lassen sich z. B. die Wahrscheinlichkeiten der Ereignisse "*X *ist gerade" oder "*X *ist ungerade" berechnen. In der Tat ist ja*

$$\begin{aligned} &\mathbb{P}(\text{“}X \textit{ ist ungerade}\text{”}) = \mathbb{P}(\{1\}) + \mathbb{P}(\{3\}) + \mathbb{P}(\{5\}) + \ldots \\ = \ &\frac{1}{2} + \frac{1}{8} + \frac{1}{32} + \ldots = \frac{1}{2}\sum_{n=0}^{\infty}(\frac{1}{4})^n = \frac{1}{2}\cdot\frac{1}{1-\frac{1}{4}} = \frac{2}{3}. \end{aligned}$$

Entsprechend gilt

$$\mathbb{P}(\text{“}X \textit{ ist gerade}\text{”}) = 1 - \frac{2}{3} = \frac{1}{3}.$$

Beispiel 3.42 *In diesem Beispiel wird der unendliche Wahrscheinlichkeitsraum aus den Punkten einer Ebene bestehen. Diese sind sogar überabzählbar. Im Casino von Sikinien*

wird das folgende Glücksspiel gespielt: Aus einer Entfernung von 5 Metern wirft ein Spieler einen Kulotnik (eine Münze von 3 cm Durchmesser) auf einen karierten Tisch mit 4 cm × 4 cm Karos. Fällt die Münze ganz in das Innere eines Karos, gewinnt der Spieler 12 Kulotnik (und erhält seine Münze zurück), anderenfalls verliert er die Münze. Was sind seine Gewinnchancen? Die erste wichtige Beobachtung dabei ist, dass die Lage der Münze eindeutig durch die Lage ihres Mittelpunkts x bestimmt ist. Wir wählen als

$$\Omega = \{x : x \in \mathbb{R}^2, \text{ und } x \text{ ist ein Punkt auf dem Tisch}\}$$

und stellen uns hierunter die Orte der Münzmittelpunkte vor. Wenn der Spieler nun kein besonders geübter Werfer ist, so wird die Wahrscheinlichkeit, dass der Münzmittelpunkt in einem Flächensegment A landet, zu der Fläche von A (die wird mit $\lambda(A)$ notiert) proportional sein. Da $\mathbb{P}$ eine Wahrscheinlichkeit sein soll, also $\mathbb{P}(\Omega) = 1$ gilt, setzen wir

$$\mathbb{P}(A) = \lambda(\Omega).$$

Wir müssen allerdings hierfür eigentlich auch eine σ-Algebra auf Ω definieren. Der Versuch,

$$\mathcal{A} = \mathcal{P}\Omega$$

zu setzen schlägt fehl (es gibt Teilmengen von Ω, denen man keine vernünftige Fläche zuordnen kann; das ist überraschend und kann hier nicht bewiesen werden). Wir setzen

$$\mathcal{A} = \{A : A \text{ hat eine Fläche}\}$$

und vertrauen darauf, dass dies erstens sinnvoll und zweites eine σ-Algebra ist. Auf jeden Fall sind alle schönen Mengen, z. B. Kreise, Quadrate, geometrische Figuren, ... in $\mathcal{A}$, denn ihre Fläche können wir angeben. Damit nun die Münze ganz im Inneren eines Karos landet, muss ihr Mittelpunkt im Inneren eines Quadrats von Kantenlänge 1 cm landen. Der Tisch habe n Karos. Somit folgt

$$P(\text{Münze im Karo}) = \frac{\text{günstige Fläche}}{\text{mögliche Fläche}} = \frac{n \cdot 1cm^2}{n \cdot 4^2 cm^2}$$

Also ist die Wahrscheinlichkeit für einen Gewinn 1/16. Wie bei so vielen anderen Glücksspielen gewinnt also auch hier die Bank.

Beispiel 3.43 *Es sei ω das Gewicht eines Neugeborenen. Wir verwenden hier*

$$\Omega = \mathbb{R}^+ = \{x \in \mathbb{R} : x > 0\}.$$

Bei der Wahl einer σ-Algebra $\mathcal{A}$ auf $\mathbb{R}^+$ stoßen wir auf dieselben Probleme wie oben und glauben daher einfach, dass es $\mathcal{A}$ gibt. In diesem Beispiel ist es allerdings unsinnig, Intervallen gleicher Länge die gleiche Wahrscheinlichkeit zuzuordnen, denn erstens wird das nie eine Wahrscheinlichkeitsverteilung auf $\mathbb{R}^+$ und zweitens ist die Wahrscheinlichkeit, dass ein Neugeborenes zwischen 3 und 4 Kilogramm wiegt, viel größer als die, dass

es zwischen 500 Gramm und 1,5 kg wiegt. Man verwendet stattdessen eine sogenannte Dichtefunktion f, d. h. eine (stückweise stetige) Funktion

$$f : \mathbb{R} \to \mathbb{R}$$

mit $f \geq 0$, f ist integrierbar und

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

Die Wahrscheinlichkeit für $a < \omega < b$ $(a, b \in \mathbb{R}^+)$ berechnet sich dann als

$$\mathbb{P}(a < \omega < b) = \int_a^b f(x)dx.$$

In unserem Beispiel ist klar, dass man $f(x) = 0$ für alle $x < 0$ setzen sollte. f kann ansonsten nicht durch Nachdenken bestimmt werden, sondern muss aus Daten geschätzt werden (f kann z. B. auch von Land zu Land oder Region zu Region verschieden sein). Häufig wird für f die Normalverteilung zu den Parametern μ (=Mittelwert) und σ^2 (=Varianz), kurz die $\mathcal{N}(\mu, \sigma^2)$-Verteilung, gewählt. Hier ist

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2},\ \mu \in \mathbb{R},\ \sigma \in \mathbb{R}^+.$$

Dass in der Tat

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

gilt, erfordert dabei etwas mehr Analysis, als die meisten von uns kennen. Für das vorliegende Beispiel ist die Wahl einer Normalverteilung allerdings nur sehr eingeschränkt sinnvoll. Egal, wie man μ und σ^2 wählt, gilt stets $f(x) > 0$ für alle $x \in \mathbb{R}$, d. h. auch negative Geburtsgewichte wären möglich (wenn auch für gewisse Wahlen von μ und σ nur sehr unwahrscheinlich). Das ist zumindest interpretationsbedürftig.

3.7 Zufallsvariablen

Oftmals interessiert bei einem Zufallsexperiment gar nicht der Versuchsausgang selbst, sondern nur ein gewisser Teilaspekt, ähnlich wie bei einem physikalischen Versuch nicht alle prinzipiell erlebbaren Messdaten interessieren, sondern nur einige wenige Aspekte. Dieses Konzentrieren auf das Wesentliche geschieht in der Wahrscheinlichkeitstheorie mit Hilfe von *Zufallsvariablen*.

Definition 3.44 *Eine Zufallsvariable auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ ist eine Funktion*

$$X : \Omega \to \mathbb{R},$$

so dass für jedes Paar $a, b \in \mathbb{R},\ a < b$

$$\{\omega : a < X(\omega) < b\} \in \mathcal{A}$$

gilt.

Bemerkung 3.45 *Die Bedingung in der letzten Definition besagt, dass wir für alle $a < b$ die Wahrscheinlichkeit des Ereignisses $\{a < X(\omega) < b\}$ berechnen können. Dies ist eine wünschenswerte Eigenschaft.*

Beispiel 3.46 *1. Es sei $\Omega = \{1,2,3,4,5,6\}$, $\mathcal{A} = \mathcal{P}\Omega$ und $\mathbb{P}(\{\omega\}) = \frac{1}{6}$ für alle $\omega \in \Omega$. Wir haben also erneut das Modell eines einfachen fairen Würfelwurfs. Interessieren wir uns nur dafür, ob die geworfene Augenzahl gerade oder ungerade ist, kommt die folgende Zufallsvariable in Betracht:*

$$\begin{array}{rcl} X : \Omega & \to & \mathbb{R} \\ \omega & \mapsto & \begin{cases} 1 & \omega = 2,4,6 \\ 0 & \omega = 1,3,5 \end{cases} \end{array} .$$

Dies ist in der Tat eine Zufallsvariable, denn

$$\{\omega : a < X(\omega) < b\} = \begin{cases} \emptyset & b \leq 0 \text{ oder } a \geq 1 \text{ oder } b \leq a \\ \{1,3,5\} & a \leq 0 \text{ und } a < b \leq 1 \\ \{2,4,6\} & a > 0 \text{ und } b \geq 1 \\ \Omega & a \leq 0 \text{ und } b \geq 1 \end{cases} .$$

All diese Mengen sind in $\mathcal{A}$.

2. Eine Zahl zwischen 1 und 16 werde rein zufällig gezogen. Es ist also $\Omega = \{1,\ldots,16\}$, $\mathcal{A} = \mathcal{P}\Omega$ und $\mathbb{P}(\{\omega\}) = \frac{1}{16}$ für alle $\omega \in \Omega$. Die folgende Zufallsvariable beschreibt die Anzahl der 1en in der Dezimaldarstellung der gezogenen Zahl:

$$X(\omega) = \begin{cases} 2 & \omega = 11 \\ 1 & \omega = 1,10,12,13,14,15,16 \\ 0 & \text{sonst} \end{cases} .$$

Das Wichtige (und Faszinierende) an Zufallsvariablen ist, dass sich ihre wichtigsten Eigenschaften aus der sogenannten Verteilung der Zufallsvariablen gewinnen lassen; andererseits gibt es verschiedene Zufallsvariablen, die dieselbe Verteilung besitzen. Eine Zufallsvariable abstrahiert also gewissermaßen vom zugrunde liegenden Wahrscheinlichkeitsraum. Diesen wollenwir uns in der Folge stets noch als endlich oder abzählbar unendlich vorstellen, d. h. als diskret. Genauer definieren wir:

Definition 3.47 *Es sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein diskreter Wahrscheinlichkeitsraum. Die Wahrscheinlichkeit $\mathbb{P}^X = (p_x)_{x\in\mathbb{R}}$ mit*

$$p_x = \begin{cases} \mathbb{P}(X = x) & \text{falls } x \text{ im Bild von } X \text{ ist} \\ 0 & \text{sonst} \end{cases}$$

auf $\mathbb{R}$ heißt <u>Verteilung</u> von X.

Beispiel 3.48 *Die Zufallsvariable X aus Beispiel 3.46.1 und die Zufallsvariable Y auf $\Omega = \{K, Z\}$, $\mathcal{A} = \mathcal{P}\Omega$, $\mathbb{P}(\{K\}) = \mathbb{P}(\{Z\}) = \frac{1}{2}$ mit*

$$Y(\omega) = \begin{cases} 1 & \omega = K \\ 0 & \omega = Z \end{cases}$$

haben dieselbe Verteilung $\mathbb{P}^X = \mathbb{P}^Y$ mit

$x =$	*0*	*1*
p_x	$\frac{1}{2}$	$\frac{1}{2}$

Betrachten wir nun einen Spieler der ein Spiel mit den möglichen Auszahlungen $X(\omega_i)$, $i = 1, \ldots, n$ spielt. $p_1, \ldots, p_n$ seien die Wahrscheinlichkeiten von $\omega_1, \ldots, \omega_n$. Gemäß der Häufigkeitsinterpretation der Wahrscheinlichkeit sollte sein Gewinn in einer sehr großen Serie von n Spielen in etwa

$$X(\omega_1) \cdot p_1 + \ldots + X(\omega_n) \cdot p_n$$

betragen. Sein Durchschnittsgewinn ist dann in etwa

$$X(\omega_1) \cdot p_1 + \ldots + X(\omega_n) \cdot p_n.$$

Dies motiviert die folgende Definition:

Definition 3.49 *Unter dem Erwartungswert der diskreten Zufallsvariable X auf dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ versteht man die Größe*

$$\mathbb{E}X = X(\omega_1) \cdot p_1 + \ldots + X(\omega_n) \cdot p_n,$$

falls $|\Omega| = n$ endlich ist und

$$\mathbb{E}X = \sum_{\omega \in \Omega} X(\omega) p_\omega$$

(wobei $p_\omega = \mathbb{P}(\{\omega\})$), falls Ω unendlich ist und die Reihe

$$\sum_{\omega \in \Omega} |X(\omega)| p_\omega < +\infty$$

konvergiert.

Bevor wir Beispiele für die Berechnung von Erwartungswerten geben, sammeln wir zunächst ein paar Regeln, die dafür nützlich sind.

Satz 3.50 *Ist $\mathbb{P}_X$ die Verteilung von X, so gilt*

$$\mathbb{E}X = \sum_{y \in \mathbb{R}} y \cdot \mathbb{P}_X(y),$$

$\mathbb{E}X$ hängt also nur von der Verteilung von X ab.

Beweis: Es sei stillschweigend angenommen, dass $\mathbb{E}X$ existiert, d. h. dass $\sum_\omega |X(\omega)|p_\omega < +\infty$ ist. Dann kann man die Reihe $\sum_\omega X(\omega)p_\omega$ beliebig umordnen, ohne dass dies ihren Wert ändert: Es gilt also

$$\begin{aligned}\mathbb{E}X &= \sum_\omega X(\omega)p_\omega = \sum_{y\in\mathbb{R}} \sum_{\omega: X(\omega)=y} X(\omega)\cdot p_\omega \\ &= \sum_{y\in\mathbb{R}} y\mathbb{P}(X=y) = \sum_{y\in\mathbb{R}} y\mathbb{P}_X(y).\end{aligned}$$

□

Als zweites beobachten wir, dass der Erwartungswert linear ist.

Satz 3.51 *Es seien $a, b \in \mathbb{R}$ und X und Y Zufallsvariablen auf dem gleichen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ mit existierenden Erwartungswerten. Dann existiert auch der Erwartungswert von $aX + bY$ und es gilt*

$$\mathbb{E}(aX + bY) = a\mathbb{E}X + b\mathbb{E}Y.$$

Beweis: Aufgrund der Dreiecksungleichung gilt

$$\sum_{\omega\in\Omega} |aX(\omega) + bY(\omega)|p_\omega \le |a| \sum_{\omega\in\Omega} |X(\omega)|p_\omega + |b| \sum_{\omega\in\Omega} |Y(\omega)|p_\omega < \infty.$$

Daher existiert $\mathbb{E}(aX + bY)$ und es gilt

$$\begin{aligned}\mathbb{E}(aX + bY) &= \sum_{\omega\in\Omega} (aX(\omega)) + bY(\omega))p_\omega \\ &= a\sum_{\omega\in\Omega} X(\omega)p_\omega + b\sum_{\omega\in\Omega} Y(\omega)p_\omega \\ &= a\mathbb{E}X + b\mathbb{E}Y.\end{aligned}$$

□

Wir sind nun in der Lage, ein paar wichtige Verteilungen und ihre Erwartungswerte kennenzulernen.

Beispiel 3.52 *1. Die Bernoulli-Verteilung zum Parameter $0 < p < 1$:*
Eine Zufallsvariable X heißt Bernoulli-verteilt zum Parameter p, falls

$$\mathbb{P}(X_i = 1) = p = 1 - \mathbb{P}(X_i = 0)$$

gilt. Bezeichnet man beim Münzwurf "Kopf" oder "Erfolg" mit "1" und "Zahl" oder "Misserfolg" mit "0", so ist X gerade der Indikator für einen Erfolg beim 1-fachen Münzwurf. Der Erwartungswert von X berechnet sich als

$$\mathbb{E}X = \sum_{x\in\mathbb{R}} x \cdot p_x = 1 \cdot \mathbb{P}(X = 1) + 0 \cdot \mathbb{P}(X = 0) = p.$$

2. *Die Binomialverteilung zu den Parametern $n \in \mathbb{N}$ und $0 < p < 1$:*
X heißt binomialverteilt zu den Parametern n und p (kurz $B(n,p)$-verteilt), falls X nur die Werte $0, 1, \ldots, n$ annimmt und

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k \in \{0, 1, \ldots, n\}$$

gilt. Dies ergibt in der Tat eine Wahrscheinlichkeitsverteilung auf $\{0, \ldots, n\}$, denn

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

$\mathbb{P}(X = k)$ beschreibt gerade die Wahrscheinlichkeit im n-fachen unabhängigen Münzwurf k Köpfe zu sehen, d. h. k-mal Erfolg zu haben, wenn die Wahrscheinlichkeit für "Kopf" in einem Wurf gerade p ist. In der Tat hat ja jede Versuchsreihe mit k "Köpfen" und $n-k$ "Zahlen" nach der Pfadregel die Wahrscheinlichkeit $p^k(1-p)^{n-k}$. Andererseits kann man die k "Köpfe" gerade auf $\binom{n}{k}$ Arten auf die n Versuche anordnen. Mit der Summenregel folgt die Behauptung. Will man nun den Erwartungswert einer $B(n,p)$-verteilten Zufallsgröße berechnen, so muss man entweder

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

berechnen (das geht, ist aber ein wenig trickreich), oder man benutzt Satz 3.51. Wir schreiben einfach

$$X = X_1 + \ldots + X_n,$$

wobei $X_i = 1$, wenn der i-te Versuch Kopf zeigt und 0 sonst. X ist $B(n,p)$-verteilt, wenn X_i $B(1,p)$-verteilt ist und $\mathbb{E}X$ hängt nur von der Verteilung von X ab. X_i hat nach Teil 1 den Erwartungswert $\mathbb{E}X_i = p$. Also gilt

$$\mathbb{E}X = \mathbb{E}X_1 + \ldots + \mathbb{E}X_n = p + \ldots + p = n \cdot p.$$

3. *Geometrisch verteilte Zufallsgrößen:*
Wir verallgemeinern Beispiel 3.41 ein wenig. Eine Münze mit Erfolgswahrscheinlichkeit p werde so lange geworfen, bis sie erstmalig "Kopf" zeigt. X beschreibe die Anzahl der Versuche, die dafür notwendig sind. Für $X = n$ müssen wir also zunächst $(n-1)$ Mal "Zahl" werfen und dann "Kopf". Analog der Rechnung in Beispiel 3.41 ergibt dies

$$\mathbb{P}(X = n) = (1-p)^{n-1} \cdot p.$$

Dies ist in der Tat eine Wahrscheinlichkeit auf $\mathbb{N}$, denn

$$\sum_{n=1}^{\infty} (1-p)^{n-1} \cdot p = p \sum_{n=0}^{\infty} (1-p)^n = p \cdot \frac{1}{1-(1-p)} = \frac{p}{p} = 1.$$

Der Erwartungswert von X berechnet sich wie folgt:

$$\begin{aligned}\mathbb{E}X &= \sum_{n=0}^{\infty} n \cdot (1-p)^{n-1} \cdot p = p \sum_{n=0}^{\infty} n \cdot (1-p)^{n-1} \\ &= p \cdot \sum_{n=1}^{\infty} n(1-p)^{n-1} = p \cdot \left(\frac{1}{1-(1-p)}\right)^2 = \frac{1}{p}.\end{aligned}$$

Man muss also durchschnittlich $\frac{1}{p}$ lange warten, bis erstmalig "Kopf" fällt, also bei einer fairen Münze ($p = \frac{1}{2}$) z. B. zwei Mal.

Wir werden in einem der nächsten Kapitel sehen, dass der Erwartungswert tatsächlich den Mittelwert von Zufallsvariablen gut beschreibt.

3.8 Unabhängigkeit und bedingte Wahrscheinlichkeit II

Wir wollen in diesem Abschnitt zunächst den Begriff der Unabhängigkeit, den wir schon für Ereignisse kennengelernt haben, auf Zufallsvariablen übertragen. Intuitiv wollen wir wieder Zufallsvariablen unabhängig nennen, falls der Ausgang einer von ihnen nicht vom Ausgang der anderen abhängt. Technisch nennen wir $X_1, \ldots, X_n$ unabhängig, wenn alle Ereignisse, die sich durch $X_1, \ldots, X_n$ beschreiben lassen, unabhängig sind.

Definition 3.53 *Seien $X_1, \ldots, X_n$ Zufallsvariablen, die auf dem gleichen (diskreten) Wahrscheinlichkeitsraum $(\Omega, \mathcal{P}\omega, \mathbb{P})$ definiert sind. $X_1, \ldots, X_n$ heißen (stochastisch) unabhängig, wenn für alle $a_1, \ldots, a_n \in \mathbb{R}$ gilt*

$$\mathbb{P}(X_1 = a_1, \ldots, X_n = a_n) = \mathbb{P}(X_1 = a_1) \cdot \ldots \cdot \mathbb{P}(X_n = a_n).$$

Dies ist genau dann der Fall, wenn für alle Teilmengen $A_1, \ldots, A_n \subseteq \mathbb{R}$ gilt

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \ldots \mathbb{P}(X_n \in A_n).$$

Beispiel 3.54 *Es sei $\Omega = \{(0,0), (0,1), (1,0), (1,1)\}$, $\mathcal{A} = \mathcal{P}\Omega$ und*

$$\mathbb{P}\{(0,0)\} = (1-p)^2,\ \mathbb{P}\{(0,1)\} = \mathbb{P}\{(1,0)\} = p(1-p),\ \mathbb{P}\{(1,1)\} = p^2$$

(also der 2-fache Münzwurf mit Erfolgswahrscheinlichkeit $0 < p < 1$). Sei weiter

$$X(\omega) = \begin{cases} 0 & \omega = (0,0) \text{ oder } \omega = (0,1) \\ 1 & \text{sonst} \end{cases} \quad \text{und}$$

$$Y(\omega) = \begin{cases} 0 & \omega = (0,0) \text{ oder } \omega = (1,0) \\ 1 & \text{sonst} \end{cases}.$$

Dann sind X und Y stochastisch unabhängig. In der Tat gilt z. B.

$$\begin{aligned}\mathbb{P}(X = 0, Y = 1) &= \mathbb{P}(\{(0,1)\}) = (1-p) \cdot p \\ &= \mathbb{P}(\{(0,0),(0,1)\}) \cdot \mathbb{P}(\{(0,1),(1,1)\}) \\ &= \mathbb{P}(X = 0) \cdot \mathbb{P}(Y = 1).\end{aligned}$$

Eine der wichtigsten Eigenschaften stochastisch unabhängiger Zufallsvariablen ist:

Satz 3.55 *Sind X und Y stochastisch unabhängige Zufallsvariablen, so gilt*

$$\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]. \tag{3.7}$$

Beweis: Es gilt

$$\begin{aligned}
\mathbb{E}(X \cdot Y) &= \sum_{\omega \in \Omega} X(\omega)Y(\omega)\mathbb{P}(\{\omega\}) = \sum_{x,y} xy \cdot \mathbb{P}(X = x, Y = y) \\
&= \sum_{x,y} xy \cdot \mathbb{P}(X = x)\mathbb{P}(Y = y) \\
&= \sum_{x} x \cdot \mathbb{P}(X = x) \sum_{y} y \cdot \mathbb{P}(Y = y) = \mathbb{E}X \cdot \mathbb{E}Y.
\end{aligned}$$

□

Man kann sich natürlich fragen, ob (3.7) schon gleichbedeutend ist mit der Unabhängigkeit von X und Y. Dies ist jedoch nicht der Fall:

Beispiel 3.56 *Es sei $\Omega = \{0, 1, -1\}$, $\mathcal{A} = \mathcal{P}\Omega$ und*

$$\mathbb{P}(\{0\}) = \frac{1}{2},\ \mathbb{P}(\{1\}) = \mathbb{P}(\{-1\}) = \frac{1}{4}.$$

Es sei ferner $X(\omega) = \omega$ und $Y(\omega) = |X(\omega)|$. Beachte, dass $\mathbb{E}X = \frac{1}{4} - \frac{1}{4} + 0 = 0$ und dass $X \cdot Y(\omega) = X(\omega)$ ist, also $\mathbb{E}[X \cdot Y] = 0$. Somit gilt (3.7), aber natürlich sind X und Y nicht unabhängig, z. B. ist

$$0 = \mathbb{P}(X = 0, Y = 1) \neq \mathbb{P}(X = 0) \cdot \mathbb{P}(Y = 1) = \frac{1}{4}.$$

Variablen, die (3.7) genügen, sind allerdings in der Analyse von Zufallsvariablen auch sehr wichtig. Sie bekommen einen eigenen Namen:

Definition 3.57 *X, Y seien Zufallsvariablen mit $EX^2 < \infty$, $EY^2 < \infty$. Wenn*

$$\mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] = 0,$$

dann heißen X und Y unkorreliert. Dies ist genau dann der Fall, wenn (3.7) gilt.

Bevor wir uns an die Analyse des Verhaltens unabhängiger und unkorrelierter Zufallsvariablen machen, schieben wir einen kleinen Exkurs über eine der wichtigsten Rechenregeln

für abhängige Ereignisse ein, den Satz von Bayes. Hierzu sei noch einmal an die Definition der bedingten Wahrscheinlichkeit

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

für Ereignisse $A, B \in \mathcal{A}$ mit $\mathbb{P}(B) > 0$ erinnert. Die Grundfrage bei der Bayesschen Formel, die wir nun vorstellen wollen, lautet: Was kann man über $\mathbb{P}(B|A)$ sagen, wenn man $\mathbb{P}(A|B)$ kennt? Hierzu zunächst ein (einigermaßen überraschendes) Beispiel:

Beispiel 3.58 *Die Infektionsrate bei BSE beträgt 0,1 %, d. h. jedes 1 000. Rind ist an BSE erkrankt. Da es sich um eine sehr gefährliche Krankheit handelt, hat man einen sehr sicheren Test entwickelt: Ein erkranktes Rind wird mit 99,9 % Wahrscheinlichkeit positiv getestet, ein gesundes Rind wird mit 95 % Wahrscheinlichkeit negativ, d. h. gesund, getestet. Was ist nun die Wahrscheinlichkeit, dass ein positives Rind auch wirklich an BSE erkrankt ist? Wir führen die folgenden Ereignisse ein:*
$\Omega = \{$*Menge aller Rinder*$\}$, $\mathcal{A} = \mathcal{P}\Omega$, $\mathbb{P}$ *die Gleichverteilung auf* Ω,
$B := \{\omega : \omega$ *ist an BSE erkrankt*$\}$,
$B^c := \{\omega : \omega$ *ist gesund*$\}$,
$\oplus = \{\omega : \omega$ *ist positiv getestet*$\}$,
$\ominus = \{\omega : \omega$ *ist negativ getestet*$\} = \oplus^c$.

Wir fragen nach $\mathbb{P}(B|\oplus)$ *und kennen* $\mathbb{P}(\oplus|B)$, $\mathbb{P}(\oplus|B^c)$ *so wie* $\mathbb{P}(B)$. *Dies genügt, denn*

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

Wegen

$$\oplus = (\oplus \cap B) \cup (\oplus \cap B^c),$$

und da die Vereinigung disjunkt ist, folgt

$$\begin{aligned} \mathbb{P}(\oplus) &= \mathbb{P}(\oplus \cap B) + \mathbb{P}(\oplus \cap B^c) \\ &= \mathbb{P}(\oplus|B) \cdot \mathbb{P}(B) + \mathbb{P}(\oplus|B^c) \cdot \mathbb{P}(B^c). \end{aligned}$$

Eingesetzt ergibt dies

$$\mathbb{P}(B|\oplus) = \frac{\mathbb{P}(\oplus|B) \cdot \mathbb{P}(B)}{\mathbb{P}(\oplus|B) \cdot \mathbb{P}(B) + \mathbb{P}(\oplus|B^c) \cdot \mathbb{P}(B^c)}.$$

Setzt man die konkreten Zahlen ein, erhält man

$$\mathbb{P}(B|\oplus) = \frac{0,999 \cdot 0,001}{0,999 \cdot 0,001 + 0,05 \cdot 0,999} = \frac{1}{1+50} = \frac{1}{51} \approx 0,0196.$$

Die Wahrscheinlichkeit, dass ein positiv getestetes Rind erkrankt ist, ist also etwa 1/51. Obwohl der Test sehr zuverlässig erscheint (siehe die Daten) ist nur 1 von 50 positiv getesteten Rindern wirklich BSE-krank (also nicht gleich notschlachten).

Verfolgt man die Argumentation des Beispiels, so haben wir en passant die folgenden Formeln gezeigt:

Satz 3.59 *Es seien $A, B_1, \ldots, B_n \in \mathcal{A}$ (wobei $(\Omega, \mathcal{A}, \mathbb{P})$ zugrunde liege) mit $\mathbb{P}(B) > 0$ für alle $i = 1, \ldots, n$. Weiter sei $B_i \cap B_j = \emptyset$ für $i \neq j$ und $\Omega = \bigcup_{i=1}^n B_i$. Dann gilt*

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

Dies ist der Satz von der totalen Wahrscheinlichkeit. Ferner gilt für alle A mit $\mathbb{P}(A) > 0$ und alle $k = 1, \ldots, n$

$$\mathbb{P}(B_k|A) = \frac{\mathbb{P}(B_k) \cdot \mathbb{P}(A|B_k)}{\sum_{i=1}^n \mathbb{P}(B_i) \cdot \mathbb{P}(A|B_i)}.$$

Dies ist die Formel von Bayes.

Den Beweis dieses Satzes haben wir (mehr oder weniger) im letzten Beispiel geführt. Wir wollen ihn nicht wiederholen, sondern stattdessen noch ein Beispiel anfügen.

Beispiel 3.60 *Johannes, Lukas und Markus sind zum Tode verurteilt. Einer von ihnen wird ausgelost und begnadigt, sein Name aber wird geheimgehalten. Johannes sagt sich: "Die Chance, dass ich es bin, ist 1/3." Er sagt dem Wächter: "Einer der beiden, Lukas oder Markus, wird sicher hingerichtet; du verrätst mir also nichts, wenn du mir sagst, wer von beiden es ist." Der Wächter antwortet: "Markus wird hingerichtet." Darauf fühlt sich Johannes ermutigt, denn Lukas oder er werden begnadigt und die Wahrscheinlichkeit ist 1/2, dass er es ist. Hat er Recht?*

Dieses schöne Beispiel stellen wir als Übungsaufgabe.

3.9 Varianz und Kovarianz

In Abschnitt 3.7 hatten wir den Erwartungswert einer Zufallsvariablen X kennengelernt. Dies war grob gesprochen eine Beschreibung des mittleren Verhaltens von X. Nun kann dieses mittlere Verhalten durch $\mathbb{E}X$ aber mehr oder weniger gut beschrieben werden: Drei Zufallsvariablen X, Y und Z mit $X \equiv 0$, $\mathbb{P}(Y = +1) = \mathbb{P}(Y = -1) = \frac{1}{2}$ und $\mathbb{P}(Z = 1\ 000) = \mathbb{P}(Z = -1\ 000) = \frac{1}{2}$ haben alle den Erwartungswert 0. Im ersten Fall (für X) ist der Erwartungswert aber identisch gleich dem einzigen Wert von X, im Fall von Z ist jeder Wert von Z um 1 000 vom Erwartungswert entfernt. Diese Problematik führt wie im 2. Kapitel dazu, eine Art "Qualitätsmaß" für den Erwartungswert $\mathbb{E}X$ einzuführen, die Varianz. Die Varianz einer Zufallsvariablen ist die mittlere Abweichung vom Erwartungswert. Hierbei müssen wir uns noch verständigen, wie wir die Abweichung messen wollen: Der naive Ansatz $\mathbb{E}[X - \mathbb{E}X]$ ist wegen

$$\mathbb{E}[X - \mathbb{E}X] = \mathbb{E}X - \mathbb{E}\mathbb{E}X = \mathbb{E}X - \mathbb{E}X = 0$$

für alle X fruchtlos. Man hat sich (aus vielerlei Gründen) auf die quadratische Abweichung geeinigt:

Definition 3.61 $\mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}X)^2]$ *heißt die Varianz einer Zufallsvariablen X. Hierbei wird vorausgesetzt, dass EX endlich ist.*

Wir werden zunächst eine Formel kennenlernen, mit der sich die Varianz einer Zufallsvariablen oft leichter berechnen lässt:

Satz 3.62 *Es gilt der Verschiebungssatz*

$$\mathbb{V}(X) = \mathbb{E}[X^2] - (\mathbb{E}X)^2.$$

Bemerkung 3.63 *Genauer müsste es heißen: Die Varianz ist endlich genau dann, wenn $\mathbb{E}(X^2)$ endlich ist. Das sei hier aber nur ohne Beweis bemerkt.*

Beweis: Es gilt:

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

□

Schon aus der quadratischen Form der Varianz lässt sich erkennen, dass sie nicht linear sein kann. In der Tat gilt:

Satz 3.64 *Es seien $X, X_1, \ldots, X_n, Y$ Zufallsvariablen mit endlichen Varianzen und $a, b \in \mathbb{R}$. Dann gilt*

1. $\mathbb{V}(aX) = a^2\mathbb{V}(X)$.
2. $\mathbb{V}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathbb{V}(X_i) + \sum_{i \neq j} Cov(X_i, X_j)$,

wobei $Cov(X_i, X_j) := \mathbb{E}[(X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j)]$ gesetzt ist.

Beweis:

1. Nach dem Verschiebungssatz gilt

$$\begin{aligned} \mathbb{V}(aX) &= \mathbb{E}[(aX)^2] - (\mathbb{E}(aX))^2 \\ &= a^2\mathbb{E}X^2 - a^2\mathbb{E}X^2 \\ &= a^2(\mathbb{E}X^2 - (\mathbb{E}X)^2) = a^2\mathbb{V}X. \end{aligned}$$

2. Da

$$(\sum_{i=1}^{n} a_i)^2 = \sum_{i=1}^{n} a_i^2 + \sum_{i \neq j} a_i a_j$$

gilt, folgt

$$\begin{aligned}
\mathbb{V}(\sum_{i=1}^{n} X_i) &= \mathbb{E}[(\sum_{i=1}^{n} X_i - \mathbb{E}\sum_{i=1}^{n} X_i)^2] \\
&= \mathbb{E}[(\sum_{i=1}^{n} (X_i - \mathbb{E}X_i))^2] \\
&= \mathbb{E}[\sum_{i=1}^{n} (X_i - \mathbb{E}X_i)^2 + \sum_{i \neq j} (X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j)] \\
&= \sum_{i=1}^{n} \mathbb{E}[(X_i - \mathbb{E}X_i)^2] + \sum_{i \neq j} \mathbb{E}[(X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j)] \\
&= \sum_{i=1}^{n} \mathbb{V}(X_i) + \sum_{i \neq j} Cov(X_i, X_j).
\end{aligned}$$

□

Bemerkung: X, Y heißen unkorreliert, wenn $Cov(X, Y) = 0$ ist. Damit folgt sofort

Korollar 3.65 *Für paarweise unkorrelierte Zufallsvariablen gilt*

$$\mathbb{V}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathbb{V}(X_i). \tag{3.8}$$

Bemerkung 3.66 *Da paarweise unabhängige Zufallsvariablen insbesondere paarweise unkorreliert sind, gilt (3.8) insbesondere für paarweise unabhängige Zufallsvariablen.*

Wir beschließen diesen Abschnitt mit einer der wichtigsten Ungleichungen, der Chebyschev-Ungleichung. Diese wird uns im nächsten Abschnitt über das Gesetz der großen Zahlen noch von großem Nutzen sein.

Satz 3.67 *(Chebyschev-Ungleichung)*
Es sei X eine Zufallsvariable mit endlicher Varianz. Dann gilt für jedes $a > 0$

$$\mathbb{P}(|X - \mathbb{E}X| \geq a) \leq \frac{\mathbb{V}(X)}{a^2}.$$

Beweis: Es gilt

$$\begin{aligned}
\mathbb{V}(X) &= \mathbb{E}((X-\mathbb{E}X)^2) = \sum_x (x-\mathbb{E}X)^2 \cdot \mathbb{P}(X=x) \\
&\geq \sum_{x:|x-\mathbb{E}X|\geq a} (x-\mathbb{E}X)^2 \cdot \mathbb{P}(X=x) \\
&\geq a^2 \sum_{x:|x-\mathbb{E}X|\geq a} \mathbb{P}(X=x) \\
&= a^2 \mathbb{P}(|X-\mathbb{E}X| \geq a).
\end{aligned}$$

Division durch a^2 ergibt das Gewünschte. Hierbei haben wir in der ersten Ungleichung die Positivität von Quadraten ausgenutzt. □

Die Chebyschev-Ungleichung lässt sich verallgemeinern; beispielsweise gilt für jede positive, monoton steigende Funktion $g : \mathbb{R} \to \mathbb{R}$ mit $\mathbb{E}g(X) < +\infty$ die Ungleichung

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}g(X)}{g(a)}.$$

Diese sogenannte Markov-Ungleichung werden wir aber für die Zwecke dieser Vorlesung nicht benötigen und daher auch nicht beweisen.

3.10 Das Gesetz der großen Zahlen

Wenn man jemanden auf der Straße fragt, was denn die Wahrscheinlichkeit eines Ereignisses sei, so wird er oder sie vermutlich ungefähr so antworten: "Führt man häufig dasselbe Experiment durch, bei dem das Ereignis A eintreten kann, so wird sich auf lange Sicht die relative Häufigkeit des Eintretens von A bei einem Wert einpendeln. Dies ist die Wahrscheinlichkeit von A." Wie schon in der Einleitung besprochen, kann man dies nicht bedenkenlos als Definition von "Wahrscheinlichkeit" nehmen. Dieser empirische Sachverhalt muss sich in unserem Modell widerspiegeln (sonst müssten wir unser Modell modifizieren). Dieses Gesetz, das heutzutage unter dem Namen "Gesetz der großen Zahlen" bekannt ist, wurde 1705 von Jakob Bernoulli formuliert und bewiesen und 1713 posthum publiziert. Es ist gewissermaßen das fundamentale Naturgesetz der Wahrscheinlichkeitstheorie, ohne das große Teile der Stochastik nicht möglich wären. Beispielsweise wäre es völlig unsinnig, immer größere Stichproben zur Bestimmung einer Wahrscheinlichkeit zu erheben oder dem Begriff "Wahrscheinlichkeit" könnte kein regulärer Sinn zugeordnet werden. In seiner heutigen Formulierung lautet das Gestz der großen Zahlen so:

Satz 3.68 *(Gesetz der großen Zahlen)*
Es seien $X_1, X_2, \ldots$ paarweise unkorrelierte Zufallsvariablen mit $\mathbb{E}X_i = E$ für alle i und $\mathbb{V}X_i = V < +\infty$ für alle i (wir setzen also voraus, dass Erwartungswert und Varianz der

X_i endlich und gleich sind). Dann gilt für jedes $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}[|\frac{1}{n}\sum_{i=1}^{n} X_i - E| \geq \varepsilon] = 0.$$

Der empirische Mittelwert konvergiert also gegen den Erwartungswert der X_i.

Beweis: Da

$$\mathbb{E}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}X_i = \frac{1}{n}\sum_{i=1}^{n} E = E$$

gilt, folgt aus der Chebyschev-Ungleichung

$$\mathbb{P}[|\frac{1}{n}\sum_{i=1}^{n} X_i - E| > \varepsilon] \leq \frac{1}{\varepsilon^2}\mathbb{V}(\frac{1}{n}\sum_{i=1}^{n} X_i). \tag{3.9}$$

Mit Korollar 3.65 folgt weiter

$$\mathbb{V}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\mathbb{V}(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}(X_i) = \frac{1}{n}\cdot V.$$

Also konvergiert die rechte Seite von (3.9) gegen 0. Da die linke Seite aber nicht-negativ ist, war genau dies zu zeigen. □

Der Zusammenhang zu der eingangs erwähnten Formulierung ist der folgende:

Korollar 3.69 *Ein Experiment werde n-mal unabhängig durchgeführt. Für ein Ereignis A sei X_i der Indikator für A im i-ten Versuch, also 1, falls A im i-ten Versuch eintritt und 0 sonst. Dann gilt*

$$\lim_{n\to\infty} \mathbb{P}[|\frac{1}{n}\sum_{i=1}^{\infty} X_i - \mathbb{P}(A)| \geq \varepsilon] = 0.$$

Die Folge der relativen Häufigkeiten des Eintretens von A konvergiert also gegen die Wahrscheinlichkeit von A.

Beweis: Dies folgt unmittelbar aus dem Gesetz der großen Zahlen, da unabhängige Ereignisse unkorreliert sind und

$$\mathbb{E}X_1 = \mathbb{E}\mathbb{1}_A = 0 \cdot \mathbb{P}(A^c) + 1 \cdot \mathbb{P}(A) = \mathbb{P}(A)$$

gilt. □

Ein paar Bemerkungen sind hier wichtig:

Bemerkung 3.70 *1. Satz 3.68 ist das sogenannte schwache Gesetz der großen Zahlen. Es gibt auch das starke Gesetz der großen Zahlen. Dieses besagt, dass*

$$\mathbb{P}(\frac{1}{n}\sum_{i=1}^{n} X_i \to E) = 1 \tag{3.10}$$

gilt. Es setzt die Existenz unendlich vieler unabhängiger Zufallsvariablen voraus (aber das haben wir nonchalant auch in Satz 3.68 vorausgesetzt, ohne dass es dort notwendig wäre). Die in Satz 3.67 behauptete Konvergenz ist verschieden von der in (3.10) behaupteten. Wir können diesen Unterschied hier aber ebensowenig zeigen wie das starke Gesetz der großen Zahlen.

2. *Wenden wir Korollar 3.69 z. B. auf den n-fachen Münzwurf an, so behauptet Korollar 3.69, dass die relative Häufigkeit von "Kopf" gegen p, die Wahrscheinlichkeit von "Kopf", konvergiert. Dies bedeutet aber nicht, dass für gerades n und beispielsweise eine faire Münze die Wahrscheinlichkeit zur Zeit n genau n/2 mal "Kopf" gesehen zu haben, gegen 1 geht. Im Gegenteil: Diese Wahrscheinlichkeit ist etwa $\frac{1}{\sqrt{\pi n}}$ und geht daher mit großen n gegen 0.*

3. *Man kann das folgende Spiel betrachten: Tom und Jerry werfen eine faire Münze. Fällt "Kopf", bekommt Tom von Jerry 1 Euro, anderenfalls bekommt Jerry von Tom einen Euro. Das Gestz der großen Zahlen besagt, dass sich Toms Kapitalstand geteilt durch die Anzahl der Spiele auf lange Sicht nahe Null einpendelt. Bedeutet dies auch, dass mit großer Wahrscheinlichkeit beide ungefähr zu 50 % der Zeit in Führung liegen? Nein! Das Arcus-Sinus-Gesetz besagt im Gegenteil, dass mit sehr großer Wahrscheinlichkeit ein Spieler beinahe die ganze Zeit führt. Auch dieses können wir hier nicht behandeln.*

3.11 Weitere Grenzwertsätze

Mit dem Gesetz der großen Zahlen haben wir einen Grenzwertsatz in der Wahrscheinlichkeitstheorie kennengelernt. Solche Grenzwertsätze sind die eigentliche Stärke der Wahrscheinlichkeitstheorie, denn sie besagen, dass es in dem scheinbaren Chaos zufälliger Folgen regelmäßige Strukturen gibt.

Für die konkrete Berechnung von Wahrscheinlichkeiten hilft das Gesetz der großen Zahlen aber wenig. Stellen wir uns vor, wir werfen 1 000 mal eine Münze mit Erfolgswahrscheinlichkeit $p = \frac{1}{2}$. Dann wissen wir, dass die Anzahl der gefallenen Köpfe nahe bei 500 liegen sollte, d. h. das Ereignis, mehr als 900 mal "Kopf" zu sehen, sollte eine kleine Wahrscheinlichkeit haben. Aber mehr wissen wir nicht. Die beiden folgenden Grenzwertsätze geben nähere Auskunft über die Wahrscheinlichkeit solcher Ereignisse. Der erste davon, der Poissonsche Grenzwertsatz, beschäftigt sich dabei mit der Wahrscheinlichkeit *seltener Ereignisse*. Genauer werden wir eine große Anzahl von unabhängigen Erfolg-Misserfolg-Experimenten durchführen (also z. B. Münzwürfe, die hierfür aber nur ein Beispiel sind), wobei jeder Versuch nur eine sehr kleine Erfolgswahrscheinlichkeit hat. Wir fragen uns

nach der Wahrscheinlichkeit für k Erfolge im Gesamtexperiment. In dieser Situation ist die sogenannte Poisson-Verteilung von Interesse.

Definition 3.71 *Eine Wahrscheinlichkeitsverteilung π_λ auf $\mathbb{N}_0$ heißt Poisson-Verteilung zum Parameter $\lambda > 0$, falls gilt*

$$\pi_\lambda(n) = \frac{\lambda^n}{n!} e^{-\lambda}$$

für alle $n \in \mathbb{N}_0$.

Satz 3.72 *π_λ ist in der Tat eine Wahrscheinlichkeitsverteilung. Hat eine Zufallsvariable mit Werten in $\mathbb{N}_0$ die Verteilung π_λ, so gilt*

$$\mathbb{E}X = \lambda \quad \textit{und} \quad \mathbb{V}X = \lambda.$$

Beweis: Da $\pi_\lambda(n) \geq 0$ gilt, ist nur noch

$$\sum_{n \in \mathbb{N}_0} \pi_\lambda(n) = 1$$

zu klären. Dies aber folgt wegen

$$\sum_{n \in \mathbb{N}_0} \pi_\lambda(n) = \sum_{n \in \mathbb{N}_0} \frac{\lambda^n}{n!} e^{-\lambda} = e^{-\lambda} \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} = e^{-\lambda} e^{\lambda} = 1$$

aus der Exponentialreihendarstellung von e^λ. Der Erwartungswert einer π_λ-verteilten Zufallsvariablen berechnet sich als

$$\sum_{n=0}^{\infty} n \frac{\lambda^n}{n!} e^{-\lambda} = e^{-\lambda} \lambda \cdot \sum_{n=1}^{\infty} \frac{\lambda^{n-1}}{(n-1)!} = e^{-\lambda} \cdot \lambda \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} = e^{-\lambda} \cdot \lambda e^{\lambda} = \lambda.$$

Für die Berechnung der Varianz berechnen wir zunächst

$$\begin{aligned} \mathbb{E}X^2 &= \sum_{n=0}^{\infty} n^2 \frac{\lambda^n}{n!} e^{-\lambda} = \sum_{n=1}^{\infty} n(n-1) \frac{\lambda^n}{n!} e^{-\lambda} + \sum_{n=1}^{\infty} n \frac{\lambda^n}{n!} e^{-\lambda} \\ &= \lambda^2 \sum_{n=2}^{\infty} \frac{\lambda^{(n-2)}}{(n-2)!} e^{-\lambda} + \mathbb{E}X = \lambda^2 e^{-\lambda} \sum_{n=2}^{\infty} \frac{\lambda^{n-2}}{(n-2)!} + \lambda \\ &= \lambda^2 e^{-\lambda} e^{\lambda} + \lambda = \lambda^2 + \lambda. \end{aligned}$$

Mit Satz 3.62 folgt

$$\mathbb{V}X = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

□

Wir kommen nun zum angekündigten Poissonschen Grenzwertsatz (der im übrigen mit Siméon Denise Poisson (1782 – 1840) entdeckt wurde und nichts mit Fisch zu tun hat).

Satz 3.73 *(Poissonscher Grenzwertsatz)*
Es seien $X_1, \ldots, X_n$ unabhängige Bernoulli-Variable mit $\mathbb{P}(X_i = 1) = 1 - \mathbb{P}(X_i = 0) = p$ und p sei so klein, dass $np = \lambda$ gelte. Dann gilt

$$\lim_{n\to\infty} \mathbb{P}(\sum_{i=1}^{n} X_i = k) = \pi_\lambda(k),$$

wobei $\pi_\lambda(k)$ die Wahrscheinlichkeit von k unter einer Poisson-Verteilung zum Parameter λ ist. Die Wahrscheinlichkeit für k Erfolge konvergiert also gegen $\pi_\lambda(k)$.

Beweis: Es gilt

$$\begin{aligned}
\mathbb{P}(\sum_{i=1}^{n} X_i = k) &= \binom{n}{k} p^k (1-p)^{n-k} \\
&= \frac{1}{k!} np \cdot ((n-1)\cdot p) \ldots ((n-k+1)\cdot p)(1 - \frac{\lambda}{n})^{n-k} \\
&\approx \frac{1}{k!}\lambda^k (1 - \frac{\lambda}{n})^n (1 - \frac{\lambda}{n})^k \\
&\approx \frac{1}{k!}\lambda^k e^{-\lambda}.
\end{aligned}$$

Hierbei haben wir die Folgendarstellung der Exponentialfunktion

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

verwandt. □

Wir wollen diesen Grenzwertsatz in Aktion erleben:

Beispiel 3.74 *Auf einer Buchseite befinden sich 1000 Buchstaben, jeder davon hat eine Wahrscheinlichkeit von 2 °/₀₀, ein Druckfehler zu sein, und die Fehler seien unabhängig. Die Anzahl der Druckfehler pro Seite folgt somit einer $B(1000,\ 0,002)$-Verteilung. Diese wird hervorragend durch eine π_2-Verteilung angenähert. Die Wahrscheinlichkeit, dass sich auf einer Seite höchstens 1 Druckfehler befindet, ist also ziemlich genau*

$$\pi_2(0) + \pi_2(1) = e^{-2} + \frac{e^{-2}2^1}{1!} = e^{-2}(1+2) = 3e^{-2} \simeq 0,406.$$

Beispiel 3.75 *In einem Jahr feiert ein großer Betrieb sein 100jähriges Bestehen. Die Direktion beschließt, allen Kindern von Betriebsangehörigen, die am 100. Geburtstag geboren werden, ein Sparkonto mit 5 000 Euro anzulegen. Bekannt ist, dass etwa 730 Kinder von Betriebsangehörigen pro Jahr geboren werden, also im Mittel zwei pro Tag. Man hat deshalb 10 000 Euro Auslagen zu erwarten. Um sicher zu gehen, werden für diese Zwecke 25 000 Euro eingeplant. Wie groß ist die Wahrscheinlichkeit, dass das Geld nicht reicht?*

Hier ist $n = 730$ *und* $p = \frac{1}{365}$ *(die Wahrscheinlichkeit, dass eines der 730 Kinder am 100. Geburtstag geboren wird). Also ist*

$$\lambda = n \cdot p = 2.$$

Ist X *die Anzahl der Kinder, die am 100. Geburtstag geboren werden, so ist*

$$\mathbb{P}(X \geq 6) = 1 - \mathbb{P}(X \leq 5) \approx 1 - e^{-2}\frac{109}{15} = 0,0166.$$

Die Wahrscheinlichkeit ist also sehr klein. Interessanterweise beruht das Beispiel (mit etwas anderen Zahlen, die Verhältnisse aber stimmen) auf einer wahren Begebenheit. Am 100. Geburtstag wurden 36(!) Kinder geboren.

Wir wollen uns nun einem zweiten Grenzwertsatz zuwenden. Dieser stellt zusammen mit dem Gesetz der großen Zahlen so etwas wie das erste und zweite Gebot der Wahrscheinlichkeitstheorie dar. Sein Name "Zentraler Grenzwertsatz" stammt jedoch nicht von seiner zentralen Rolle in der Wahrscheinlichkeitstheorie, sondern davon, dass er erlaubt, die Wahrscheinlichkeit zentraler Ereignisse, also solcher, die in der Nähe des Erwartungswertes stattfinden, zu berechnen.

Satz 3.76 *(Zentraler Grenzwertsatz)*
Es sei $X_1, X_2, \ldots$ *eine Folge von Zufallsvariablen, die unabhängig sind und alle dieselbe Verteilung besitzen. Es sei*

$$\begin{aligned} \mathbb{E}X_i &= \mu \qquad \textit{und} \\ \mathbb{V}X_i &= \sigma^2 > 0 \quad \textit{für alle } i \textit{in} \mathbb{N}. \end{aligned}$$

Dann gilt für alle $a < b \in \mathbb{R}$

$$\lim_{n\to\infty} \mathbb{P}(a \leq \frac{\sum_{i=1}^{n}(X_i - \mu)}{\sqrt{n\sigma^2}} \leq b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx. \tag{3.11}$$

Bemerkung 3.77 *Die Verteilung auf der rechten Seite hatten wir schon in einem vorhergehenden Abschnitt als die* $\mathcal{N}(0,1)$*-Verteilung (Standardnormalverteilung) kennengelernt. Die sogenannte Standardisierung, also das Abziehen von* μ *und das Teilen durch* $\sqrt{n\sigma^2}$ *in (3.11) wird verständlich, wenn man weiß, dass eine Zufallsvariable* Y *die* $\mathcal{N}(0,1)$*-verteilt ist, d. h. für die gilt*

$$\mathbb{P}(a \leq Y \leq b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx \quad \textit{für alle} \quad a < b,$$

Erwartungswert 0 und Varianz 1 hat, also

$$\mathbb{E}Y = 0 \quad \mathbb{V}Y = 1.$$

Genau dies wird mit der Standardisierung auch für $\sum_{i=1}^{n} X_i$ *erreicht.*

Der Beweis von Satz 3.76 ist nicht ganz einfach und kann hier nicht gegeben werden. Wir wollen zunächst seine Bedeutung für den n-fachen Münzwurf kennenlernen und dann Beispiele studieren.

Korollar 3.78 *Seien $X_1, X_2, \ldots$ unabhängige Bernoulli-Variable mit*

$$\mathbb{P}(X_i = 1) = p = 1 - \mathbb{P}(X_i = 0).$$

Dann gilt für alle $a < b$:

$$\lim_{n\to\infty} \mathbb{P}(a \leq \frac{\sum_{i=1}^n X_i - np}{\sqrt{np(1-p)}} \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

Beweis: Dies folgt sofort aus Satz 3.77, denn

$$\mathbb{E}(\sum_{i=1}^n X_i) = np \quad \text{und} \quad \mathbb{V}(\sum_{i=1}^n X_i) = np(1-p).$$

□

Bevor wir uns Beispielen zuwenden, sollte noch bemerkt werden, dass das Integral auf der rechten Seite von (3.11) nicht elementar berechnet werden kann. Es ist aber vielerorts tabelliert.

Beispiel 3.79 *Wie groß ist die Wahrscheinlichkeit, dass bei 6 000-fachem Würfeln eines fairen Würfels die "6" mindestens 1 100 mal fällt? Man kann dies mit viel Geduld und einem sehr guten Computerprogramm direkt berechnen. Die gesuchte Wahrscheinlichkeit ist einfach*

$$\sum_{k=1100}^{6000} B(k; 6000, \frac{1}{6}).$$

Dies zu berechnen ist aber mühselig. Mit Hilfe von Korollar 3.78 kommt man auf

$$\mathbb{P}(\sum_{k=1}^{6000} X_k \geq 1100) = \mathbb{P}\left(\frac{\sum_{k=1}^{6000} X_k - np}{\sqrt{np(1-p)}} \geq \frac{1100 - np}{\sqrt{np(1-p)}}\right).$$

Wegen $p = \frac{1}{6}$ (Wahrscheinlichkeit für eine "6") und $n = 6000$, erhält man

$$\begin{aligned} \mathbb{P}(\sum_{k=1}^{6000} X_k \geq 1100) &= \mathbb{P}\left(\frac{\sum_{k=1}^{6000} X_k - 1000}{\frac{100}{6}\sqrt{3}} \geq \frac{100}{\frac{100}{6}\sqrt{3}}\right) \\ = \int_{2\sqrt{3}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx &= 1 - \Phi(2\sqrt{3}). \end{aligned}$$

Hierbei haben wir

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

gesetzt und $\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1$ *ausgenutzt. Den Wert* $\Phi(2\sqrt{3}) = \Phi(3,464)$ *kann man in einer Tabelle nachschauen und erhält dann*

$$\mathbb{P}(\sum_{k=1}^{6000} X_k \geq 1100) \approx 0,00028.$$

Die Wahrscheinlichkeit für mehr als 1 100 "6"en ist also extrem gering.

Beispiel 3.80 *Berechnung einer Stichprobengröße.*
Man will den Anteil der Raucher p einer sehr großen Bevölkerung schätzen und zwar auf 2 % genau. Will man p mit 100 %iger Sicherheit auf 2 % genau schätzen, so muss man fast die ganze Bevölkerung befragen. Ist man dagegen mit 95 %iger Sicherheit (oder einer anderen kleiner als 100 %iger) zufrieden, so genügt eine geringere Stichprobe, die wir uns der Einfachheit halber als "mit Zurücklegen" gezogen vorstellen. Wie groß muss diese Stichprobe nun sein? Eine Stichprobe vom Umfang n enthalte S_n *Raucher. Es liegt nahe (und wir werden das im nächsten Kapitel auch rechtfertigen), als Schätzer für p*

$$\hat{p} = \frac{S_n}{n}$$

zu verwenden. Ein Blick in die Tabelle der Standardnormalverteilung zeigt, dass

$$\mathbb{P}(|\frac{S_n - np}{\sqrt{np1-p}}| \leq 2) = \mathbb{P}(-2 \leq \frac{S_n - np}{\sqrt{np1-p}} \leq 2) \approx \Phi(2) - \Phi(-2) = 1 - 2\Phi(-2) \approx 0,954.$$

Das Ereignis $|\frac{S_n - np}{\sqrt{np1-p}}| \leq 2$ *tritt also mit etwas mehr als 95 %iger Wahrscheinlichkeit ein. Nun ist*

$$|\frac{S_n - np}{\sqrt{np(1-p)}}| \leq 2 \quad \textit{gleichbedeutend mit} \quad |\frac{S_n}{n} - p| \leq 2\sqrt{\frac{p(1-p)}{n}}.$$

Somit gilt auch

$$\mathbb{P}(|\hat{p} - p| \leq 2\sqrt{\frac{p(1-p)}{n}}) \approx 0,95.$$

Wir wollen, dass $|\hat{p} - p| \leq 0,02$ *mit 95 %iger Sicherheit ist und setzen deshalb*

$$2\sqrt{\frac{p(1-p)}{n}} = 0,02 \quad \textit{d. h.} \quad n = 10\ 000p(1-p).$$

Dummerweise wollen wir p ja gerade schätzen, d. h. wir wissen nicht, wie groß $p(1-p)$ *ist. Eine schnelle Überlegung lehrt uns aber, dass* $p(1-p)$ *eine Parabel in p ist, die ihren Scheitel in* $p = \frac{1}{2}$ *hat. Also gilt stets*

$$p(1-p) \leq \frac{1}{4}.$$

Somit genügen 10 000 $p(1-p) \leq 2\ 500$ *Probanden, um den Anteil der Raucher in der Bevölkerung mit 95 % Sicherheit auf 2 % Genauigkeit zu schätzen. Solche Ergebnisse lösen bei vielen Geistes- und Gesundheitswissenschaftlern oftmals Bestürzung aus, weil typische Stichprobengrößen dort um Größenordnungen kleiner sind.*

4 Beurteilende Statistik

Dieses ist das zweite Kapitel in diesem Skript, das die Überschrift "Statistik" trägt, doch im Gegensatz zu Kapitel 2, in dem es nur darum ging, die Daten ansprechend und charakteristisch aufzubereiten, wollen wir hier Schlüsse aus unseren Daten ziehen. Genauer versuchen wir gewissermaßen, die inverse Aufgabe der Wahrscheinlichkeitstheorie zu lösen: Wir versuchen nun von den Daten auf die zugrunde liegende Verteilung zu schließen. Dieses Problem existiert in verschiedenen Schwierigkeitsgraden. Das schwierigste und realistischste davon, dass wir nämlich nichts oder so gut wie nichts a priori über die Verteilung der Daten wissen, können wir hier noch nicht einmal ansatzweise behandeln. Dies wäre die sogenannte nicht-parametrische Situation.

In der parametrischen Situation wissen wir schon von vornherein, dass die Verteilung der beobachteten Daten aus einer Klasse stammt, die wir mit einem endlich-dimensionalen Vektor beschreiben können. Diesen Fall wollen wir genauer anschauen. Wir werden hierbei archetypisch auch nur eine Situation genauer studieren, den n-fachen Münzwurf: Es seien also n unabhängige Beobachtungen $x_1, \dots, x_n$ gegeben, die alle vom Wurf einer Münze stammen. Diese habe Erfolgswahrscheinlichkeit p, also

$$\mathbb{P}(X_i = 1) = p = 1 - \mathbb{P}(X_i = 0), \tag{4.1}$$

und p sei dabei unbekannt. Die Familie der zugrunde liegenden Wahrscheinlichkeiten ist also $(\mathbb{P}_p)_{p\in[0,1]}$, p ist also der (eindimensionale) Parameter, um unsere Wahrscheinlichkeiten zu parametrisieren. Ziel ist es nun, mehr Informationen über p zu sammeln. Genauer wollen wir die folgenden Problemkreise untersuchen:
1. Schätze p.
2. Teste Hypothesen über p.

Prinzipiell kennt die beurteilende Statistik auch noch eine dritte Fragestellung, nämlich Bereiche anzugeben, in denen sich p mit großer Wahrscheinlichkeit befindet, sogenannte Konfidenzintervalle. Diese sollen hier aber nicht untersucht werden.

Für den Rest des Kapitels sei also $\Omega = \{0, 1\}$, $\mathcal{A} = \mathcal{P}\Omega$ und $\mathbb{P}_p$ die in (4.1) beschriebene Klasse von Verteilungen auf Ω. Weiter seien unabhängige Zufallsvariablen $X_1, \dots, X_n$, die nach $\mathbb{P}_p$ verteilt sind, gegeben. Insbesondere gilt also

$$\begin{aligned}\mathbb{P}_p(X_1 = x_1, \dots, X_n = x_n) &= \mathbb{P}_p(X_1 = x_1) \cdot \ldots \cdot \mathbb{P}_p(X_n = x_n) = p^k \cdot (1-p)^{n-k} \\ &\quad \text{für} \quad \sum_{i=1}^{n} x_i = k.\end{aligned}$$

4.1 Das Schätzproblem

In diesem Kapitel sollen Mechanismen beschrieben werden, die uns eine möglichst gute Schätzung von p erlauben. Eine solche Schätzung sollte natürlich tunlichst von den

gemachten Beobachtungen $X_1 = x_1, \ldots, X_n = x_n$ abhängen. Wir definieren die Schätzfunktion deshalb in folgender Weise:

Definition 4.1 *Eine Schätzfunktion bzw. ein Schätzer für p ist eine Funktion*

$$\begin{aligned} \hat{p} : \mathbb{R}^n &\to \mathbb{R} \\ (x_1, \ldots, x_n) &\mapsto \hat{p}(x_1, \ldots, x_n). \end{aligned}$$

Bemerkung 4.2

1. *Die Idee bei der obigen Definition ist die, dass die $x_1, \ldots, x_n$ die Ausgänge der Zufallsvariablen $X_1, \ldots, X_n$ sind. $\hat{p}$ ist also ein zufälliger Wert, wobei der Zufall daher kommt, dass $x_1, \ldots, x_n$ zufällige Werte sind.*

2. *Die einfachste Schätzfunktion ist stets zu vermuten, dass die Münze fair ist, d. h. man wählt*
$$g(x_1, \ldots, x_n) = \frac{1}{2}$$
für alle $x_1, \ldots, x_n \in \{0, 1\}^n$.

3. *Die bekannteste Schätzfunktion für p ist das arithmetische Mittel (das sicher beinahe jeder nennen wird, den man nach einer Schätzung für p fragt), also*
$$\hat{p}(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{4.2}$$
Der Rest dieses Abschnitts beschäftigt sich damit, den naiven Schätzer in (4.2) zu rechtfertigen.

Dies geschieht so, dass wir uns zunächst fragen, was ein sinnvolles Prinzip wäre, um einen Schätzer zu konstruieren. Dieses Prinzip ist das sogenannte Maximum-Likelihood-Prinzip. Ausgangspunkt dieses Prinzips ist die Häufigkeitsinterpretation der Wahrscheinlichkeit. Danach werden wir eher ein wahrscheinliches als ein unwahrscheinliches Ereignis beobachten. Das Maximum-Likelihood-Prinzip ist es nun, einen Schätzer für p so zu konstruieren, dass unter dem geschätzten Wert für p die Beobachtung, die wir gemacht haben, maximale Wahrscheinlichkeit hat. Dieser Schätzer heißt Maximum-Likelihood-Schätzer.

Definition 4.3 *Ein Maximum-Likelihood-Schätzer für p ist jedes $\hat{p}$ mit*

$$\mathbb{P}_{\hat{p}}(x_1, \ldots, x_n) = \max_{p \in [0,1]} \mathbb{P}_p(x_1, \ldots, x_n).$$

Um dieses Konzept besser zu verstehen, berechnen wir einfach den Maximum-Likelihood-Schätzer in unserer Situation.

Satz 4.4 *Sei $0 \leq p \leq 1$. In der zu analysierenden Situation ist Maximum-Likelihood-Schätzer der naive Schätzer*

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Beweis: Eine Beobachtung $x = (x_1, \ldots, x_n)$ mit $k := \sum_{i=1}^{n} x_i$ 1en und $(n-k)$ 0en hat die Wahrscheinlichkeit

$$P_p(\{x\}) = \mathbb{P}_p(\{(x_1, \ldots, x_n)\}) = p^k (1-p)^{n-k}.$$

Diese wollen wir in p maximieren, d. h. wir suchen das Maximum der Funktion

$$p \mapsto L_x(p) := p^k (1-p)^{n-k}, \quad k := \sum_{i=1}^{n} x_i.$$

Diese Funktion ist nicht so einfach zu maximieren, jedenfalls ist

$$p \mapsto \log L_x(p)$$

leichter zu maximieren. Das ist auch völlig ausreichend, denn der natürliche Logarithmus ist wie jeder andere Logarithmus eine monotone Funktion, d. h. ein Maximum von $\log L_x(\cdot)$ ist auch ein Maximum von L_x. Nun ist

$$\mathcal{L}_x(p) := \log L_x(p) = k \log p + (n-k) \log(1-p).$$

Weiter ist

$$\begin{aligned} \frac{d}{dp} \mathcal{L}_x(p) &= \frac{k}{p} - \frac{n-k}{1-p} \qquad \text{und} \\ \frac{d^2}{dp^2} \mathcal{L}_x(p) &= -\frac{k}{p^2} - \frac{n-k}{(1-p)^2} < 0. \end{aligned}$$

Eine Nullstelle der ersten Ableitung wird also in der Tat ein lokales Maximum von $\mathcal{L}_x(\cdot)$ sein. Nun ist

$$\begin{aligned} & \frac{k}{p} - \frac{n-k}{1-p} = 0 \\ \Leftrightarrow \quad & k - kp = np - kp \\ \Leftrightarrow \quad & p = \frac{k}{n}. \end{aligned}$$

Man vergewissert sich zudem, dass für $0 < k < n$ $\mathbb{P}_1(\{x\})$ und $\mathbb{P}_0(\{x\})$ echt kleiner sind als $\mathbb{P}_{\hat{p}}(\{x\})$. Also ist (man erinnere sich, dass $k = \sum_{i=1}^{n} x_i$ war)

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$$

in der Tat ein Maximum-Likelihood-Schätzer. □

Die Maximum-Likelihood-Methode ist *die* Methode schlechthin, um gute Schätzer zu konstruieren und dies gilt für eine sehr große Klasse von Problemen. Für den Fall der Binomialverteilung haben wir gesehen, dass der naive Schätzer (das arithmetische Mittel) der Maximum-Likelihood-Schätzer ist. Dies bedeutet, dass er aus einem uns vernünftig erscheinenden Prinzip konstruiert werden kann, aber es bedeutet natürlich nicht, dass es auch per se ein guter Schätzer sein muss. Wir wollen hier noch zwei Qualitätsmerkmale für den Maximum-Likelihood-Schätzer überprüfen:

Das erste geht davon aus, dass wir vielleicht in jeder einzelnen Schätzung, d. h. bei jeder einzelnen Beobachtung $(x_1, \ldots, x_n)$ einen Schätzfehler haben, dass wir aber wenigstens im Mittel den richtigen Wert für p schätzen wollen (im Durchschnitt über die $(x_1, \ldots, x_n)$ nicht über p).

Definition 4.5 *Ein Schätzer $\bar{p} = \bar{p}(X_1, \ldots, X_n)$ für p heißt erwartungstreu, wenn für alle $p \in [0, 1]$*

$$\mathbb{E}_p \bar{p} = p$$

gilt. Dabei besagt der untere Index p, dass wir den Erwartungswert bei zugrunde liegendem p bilden.

Es stellt sich heraus, dass unser naiver Schätzer $\hat{p} = \frac{1}{n} \sum_{i=1}^{n} x_i$ erwartungstreu ist:

Satz 4.6 *Der naive Schätzer*

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

ist erwartungstreu (wir schreiben große X_i um zu zeigen, dass $\hat{p}$ eine Zufallsvariable ist).

Beweis: Aus den Eigenschaften des Erwartungswerts folgt

$$\begin{aligned} \mathbb{E}_p \hat{p} &= \mathbb{E}_p \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_p X_i \\ &= \frac{1}{n} \sum_{i=1}^{n} p = \frac{1}{n} \cdot np = p. \end{aligned}$$

$\hat{p}$ ist also in der Tat erwartungstreu. □

Das zweite Qualitätsmerkmal geht von der Überlegung aus, dass für endliches n jeder Schätzer fehlerbehaftet sein mag, dass wir uns aber wünschen, dass für immer größere Stichproben die Qualität unseres Schätzers immer besser wird, und der Schätzer schließlich gegen den zu schätzenden Wert konvergiert.

Definition 4.7 *Ein Schätzer $\bar{p} = \bar{p}(X_1, \ldots, X_n)$ heißt konsistent, wenn für alle $\varepsilon > 0$ gilt*

$$\lim_{n\to\infty} \mathbb{P}_p(|\bar{p}(X_1, \ldots, X_n) - p| \geq \varepsilon) = 0$$

für alle $p \in [0, 1]$.

Auch diese Eigenschaft hat unser naiver Schätzer:

Satz 4.8 *Der naive Schätzer $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ ist für die Münzwurf-Situation konsistent.*

Beweis: Es gilt für alle $\varepsilon > 0$

$$\mathbb{P}_p(|\hat{p} - p| \geq \varepsilon) = \mathbb{P}_p(|\frac{1}{n}\sum_{i=1}^{n} X_i - p| \geq \varepsilon)$$

und daher folgt die Behauptung aus dem schwachen Gesetz der großen Zahlen für den Münzwurf. □

Der naive Schätzer wird für uns noch von großem Nutzen sein, wenn wir im nächsten Abschnitt Hypothesen über das zugrunde liegende p testen wollen.

4.2 Testtheorie im Münzwurf

Im vorhergehenden Abschnitt hatten wir gesehen, dass ein sehr guter Schätzer für das unbekannte p der empirische Mittelwert, d. h. der naive Schätzer, ist. In der Klasse der erwartungstreuen Schätzer ist er sogar in gewissem Sinne der beste (dies werden wir hier aber weder spezifizieren noch beweisen). Ist nun aber unser p (der Erfolgsparameter) eine irrationale Zahl, so werden wir p mit keiner Schätzung unseres naiven Schätzers richtig schätzen: egal, wie groß n ist und welche Stichprobe $x_1, \ldots, x_n$ wir beobachten, stets wird $\frac{1}{n}\sum_{i=1}^{n} x_i$ ein Bruch sein und damit verschieden von p. Wir wollen in diesem Abschnitt bescheidener sein und sehen, ob wir zumindest Aussagen über p mit großer Wahrscheinlichkeit als richtig oder falsch nachweisen können.

Wir wollen dies zunächst an einem Beispiel kennenlernen:

Beispiel 4.9 Wir wollen uns der Frage zuwenden, ob ein neugeborenes Küken Körner erkennen kann, oder ob es dies durch Erfahrung lernen muss. Hierzu führen wir das folgende Experiment durch: Sobald ein Küken geschlüpft ist, werden ihm falsche Körner aus Papier vorgesetzt. Die Hälfte sind kleinere Kreise, die andere Hälfte kleine Dreiecke von gleicher Fläche. Nun beobachten wir das Küken 16 mal beim Picken. Davon picke es X Kreise. Wir haben nun zwei Hypothesen:

H_0: Das Verhalten des Kükens ist völlig indifferent gegenüber der Form der Körner, d. h. es pickt runde Körner und eckige Körner mit gleicher Wahrscheinlichkeit.

H_1: Das Küken bevorzugt runde Körner.

Dass das Küken dreieckige Körner lieber mag, scheint uns unwahrscheinlich. Man kann dieses Experiment dadurch mathematisieren, dass man sich die gepickten Körner als eine Folge von 16 0en und 1en vorstellt. Dabei schreiben wir 1, falls das Küken ein rundes Korn pickt und 0 für ein dreieckiges Korn. Wir nehmen an, dass diese 0en und 1en aus stochastisch unabhängigen Experimenten stammen (wir sind also nun in der Situation des Münzwurfs), und wir versuchen aufgrund der Stichprobe zwischen den beiden Hypothesen

$$H_0 : p = \frac{1}{2} \qquad \text{und} \qquad H_1 : p > \frac{1}{2}$$

zu entscheiden. Egal, wie wir uns entscheiden und welche Entscheidungsregel wir verwenden, können wir zwei Fehler begehen: Den Fehler erster Art: H_0 ist wahr und wird verworfen. Den Fehler zweiter Art: H_0 ist falsch und wird angenommen. Offensichtlich ist es unmöglich, beide Fehler gleichzeitig in den Griff zu bekommen. Versucht man beispielsweise den Fehler erster Art zu eliminieren, so wird man H_0 stets annehmen müssen. Dies aber maximiert den Fehler zweiter Art.

In der statistischen Theorie und Praxis hat sich das folgende Verfahren eingebürgert: Man sucht nach einem Test, bei dem der Fehler 1. Art kleiner ist als ein vorgegebenes Signifikanzniveau α. Typische Werte für α sind 5 %, 2,5 %, 1 %, 0,5 % oder 0,1 %. Wir wollen unsere Hypothesen auf das Niveau α=5 % testen (weil wir es nicht so tragisch finden, wenn wir uns irren – je gravierender ein Irrtum ist, desto kleiner müssen wir α wählen). Das folgende Testverfahren liegt nahe: Wir schätzen p aus den Daten mit Hilfe des naiven Schätzers $\hat{p}$. Wenn $\hat{p}$ so aussieht, wie wir es unter H_0 vermuten würden (also z. B. nicht zu sehr von $1/2$ entfernt liegt), so nehmen wir H_0 an, sonst H_1. Dabei müssen wir, da wir den Fehler 1. Art klein halten wollen, H_0 auch dann annehmen, wenn $\hat{p}$ sogar größer ist als $1/2$, aber nicht zu groß. Genauer sieht das Testverfahren so aus:

1. Schätze p durch $\hat{p}$
2. Akzeptiere H_0, falls $\hat{p} \leq \Gamma = \Gamma(\alpha)$
3. Verwerfe H_0, falls $\hat{p} > \Gamma = \Gamma(\alpha)$.

Das Intervall $]\Gamma(\alpha), 1]$ heißt dabei Ablehnungsbereich für diesen Test. Die Schranke für $\Gamma = \Gamma(\alpha)$ berechnet sich dabei so, dass unter H_0

$$\mathbb{P}(\hat{p} > \Gamma) \leq \alpha, \qquad \text{also} \qquad \mathbb{P}_{1/2}(\hat{p} > \Gamma) \leq \alpha$$

gilt. In unserem Fall wollen wir also

$$\mathbb{P}_{1/2}(\hat{p} > \Gamma) \leq 0,05$$

sicherstellen. Wegen $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ ist dies gleichbedeutend mit

$$\begin{aligned} \mathbb{P}_{1/2}(\sum_{i=1}^{n} X_i > n\Gamma) &\leq 0,05, \quad \text{d. h.} \\ \sum_{k=n\Gamma}^{n} b(k, 16, \frac{1}{2}) &\leq 0,05. \end{aligned}$$

Berechnet man die Werte der Binomialverteilung $b(k, 16, \frac{1}{2})$, sieht man, dass

$$\sum_{k=12}^{16} b(k, 16, \frac{1}{2}) \cong 0,0384 < 0,05,$$

während $\sum_{k=11}^{16} b(k, 16, \frac{1}{2}) > 0,05$ gilt. Wir wählen also

$$n\Gamma = 12, \quad \text{d. h.} \quad \Gamma = \frac{12}{16} = \frac{3}{4}.$$

Ist $\hat{p} \leq 0,75$, werden wir H_0 annehmen, ansonsten verwerfen.

Dieses Beispiel erlaubt es uns auch schon, die Regeln für sogenannte einseitige Testprobleme aufzustellen: Zu testen sei

$$\begin{array}{llll} H_0 : p \leq p_0 & \text{gegen} & H_1 : p > p_0 & \text{bzw.} \\ H_0 : p < p_0 & \text{gegen} & H_1 : p \geq p_0 & \text{oder} \\ H_0 : p = p_0 & \text{gegen} & H_1 : p > p_0 & \end{array}$$

auf dem Niveau α (d. h. die Wahrscheinlichkeit für einen Fehler 1. Art darf höchstens α sein). Dann befolge man folgendes Verfahren:

1. Schätze p durch $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ aus den Daten.
2. Bestimme $\Gamma = \Gamma(\alpha)$ möglichst klein, so dass

   $$\mathbb{P}_{p_0}(\hat{p} > \Gamma) \leq \alpha$$

   und dabei Γ möglichst klein.
3. Falls $\tilde{p} \leq \Gamma$, nehme H_0 an, ansonsten verwerfe H_0.

Bemerkung 4.10 *1. Dieses Testverfahren ist insofern optimal, dass es beispielsweise für $H_0 : p = p_0$ gegen $H_1 : p = p_1,\ p_1 > p_0$ bei festgehaltenem Fehler 1. Art den kleinsten Fehler 2. Art liefert. Dies ist der Inhalt des sogenannten Neyman-Pearson-Lemmas, das wir hier nicht beweisen können.*

2. *Dass wir auch für $H : p < p_0$ bzw. $H_0 : p \leq p_0$ nur*

$$\mathbb{P}_{p_0}(\hat{p} > \Gamma) \leq \alpha$$

kontrollieren müssen, liegt daran, dass

$$p \mapsto \mathbb{P}_p(\hat{p} > \Gamma)$$

eine monotone Funktion ist.

3. *Analog kann man $H_0 : p = p_0$ gegen $H_1 : p < p_0$ und ähnliche Hypothesen testen. Das Verfahren ist dann:*
   1. *Schätze $\hat{p}$ aus den Daten.*
   2. *Finde $\Gamma = \Gamma(\alpha)$ möglichst groß, so dass $\mathbb{P}_{p_0}(\hat{p} < \Gamma) \leq \alpha$.*
   3. *Akzeptiere H_0, falls $\hat{p} \geq \Gamma$, ansonsten verwerfe H_0.*

Beispiel 4.11 *Zwei Spieler A und B würfeln. Dabei behauptet B, dass der Würfel gezinkt sei und weniger 6en würfelt als normalerweise. A hingegen behauptet, der Würfel sei fair. Es werden 10 Probewürfe gemacht, dabei fällt eine 6. Kann man die Hypothese*

$$H_0 : \text{ Der Würfel ist fair } \quad (p = \frac{1}{6})$$
$$\text{gegen} \qquad H_1 : p < \frac{1}{6}$$

auf dem Niveau $\alpha = 0,05$ verwerfen? Hierbei ist p die unbekannte Wahrscheinlichkeit, eine 6 zu würfeln. Wir tabellieren die Wahrscheinlichkeiten $\mathbb{P}_{1/6}(X = k)$, die für uns relevant sind (X sei die Anzahl von 6en in 10 Würfen)

k	$\mathbb{P}_{1/6}(X = k)$	$\mathbb{P}_{1/6}(X \leq k)$
0	0,161	0,161

Damit sehen wir bereits, dass wir mit 10 Würfen die Hypothese H_0 nie verwerfen können. Wir akzeptieren also H_0.
(Aufgabe: Wie groß muss die Anzahl der Würfe mindestens sein, damit man einen nicht leeren Ablehnungsbereich hat?)

Neben den bisher behandelten einseitigen Tests gibt es auch zweiseitige Tests. Diese sind von der Form

$$H_0 : p = p_0 \qquad \text{gegen} \qquad H_1 : p \neq p_0.$$

Es liegt auf der Hand, den Test folgendermaßen zu gestalten:

1. Schätze p durch $\hat{p} = \frac{1}{n} \sum_{i=1}^{n} x_i$.

2. Bestimme $\Gamma = \Gamma(\alpha)$ möglichst klein mit

$$\mathbb{P}_{p_0}(\hat{p} \notin [p_0 - \Gamma, p_0 + \Gamma]) \leq \alpha.$$

3. Akzeptiere H_0, falls $\hat{p} \in [p_0 - \Gamma, p_0 + \Gamma]$, ansonsten verwerfe H_0.

Wir wollen dies an einem Beispiel kennenlernen.

Beispiel 4.12 *Sind Ratten farbenblind? Wir wollen klären, ob Ratten eine der Farben rot oder grün vorziehen. Hierfür planen wir den folgenden Versuch: Ratten werden durch einen Gang geschickt, der sich in zwei Gänge verzweigt, die grün bzw. rot gestrichen sind. Je nach dem Ausgang des Experiments entscheiden wir uns für*

H_0 : *Die Ratten entscheiden sich für die beiden Gänge mit gleicher Wahrscheinlichkeit* $(p = \frac{1}{2})$ *oder*

H_1 : *Die Ratten bevorzugen einen der Gänge, d. h. eine der Farben* $(p \neq \frac{1}{2})$.

Hierbei steht p für die Wahrscheinlichkeit, z. B. den roten Gang zu betreten. Wir wenden nun den oben beschriebenen Weg an: Es stehen uns 10 Ratten zur Verfügung, d. h. wir können eine Stichprobe vom Umfang $n = 10$ *erheben. Wir suchen nun unser* Γ *und damit das kritische Gebiet. Dazu müssen wir* Γ *so wählen, dass*

$$\mathbb{P}_{1/2}(\frac{1}{n}S_n \notin [\frac{1}{2} - \Gamma, \frac{1}{2} + \Gamma)) \leq \alpha, \quad S_n := \sum_{i=1}^{n} X_i,$$

und Γ *dabei möglichst klein. Wir fertigen die folgende Tabelle an:*

k	$\mathbb{P}(S_n = k)$
0	$\frac{1}{1024}$
10	$\frac{1}{1024}$
1	$\frac{10}{1024}$
9	$\frac{10}{24}$
2	$\frac{45}{1024}$
8	$\frac{45}{1024}$

Wir sehen also

$$\mathbb{P}_{1/2}(|S_n - 5| \geq 4) = \frac{22}{1024} < 0,05$$

und

$$\mathbb{P}_{1/2}(|S_n - 5| \geq 3) > 0,05.$$

Wir gehen also folgendermaßen vor: Gehen die Ratten 9 oder 10 mal in denselben Gang, so werden wir die Hypothese H_0*, dass Ratten farbenblind sind, verwerfen, anderenfalls werden wir sie auf dem Niveau* $\alpha = 0,05$ *akzeptieren.*

Bemerkung 4.13 *1. Bei einem einseitigen Test, beispielsweise der Form*

$$H_0 : p < p_0 \qquad gegen \qquad H_1 : p > p_0$$

ist prinzipiell auch ein Vertauschen der Hypothesen H_0 und H_1 möglich. Dabei tauscht man auch die Fehler 1. Art und 2. Art, man sichert somit einen anderen Fehler ab. Ein solches Vertauschen ist bei einem zweiseitigen Test

$$H_0 : p = p_0 \qquad gegen \qquad H_1 : p \neq p_0$$

nicht möglich.

2. Man sollte beachten, dass die Verwerfungsbereiche von ein- und zweiseitigen Tests verschieden sind. Eine einseitige Hypothese kann verworfen werden, während die zweiseitige angenommen wird. Weiß man z. B. in Beispiel 4.13, dass Ratten – wenn überhaupt – rot bevorzugen, d. h. testet man einseitig

$$H_0 : p = \frac{1}{2} \qquad gegen \qquad H_1 : p > \frac{1}{2},$$

so sieht man, dass der Ablehnungsbereich für α=10 % so konstruiert wird, dass man H_0 bei 8, 9 oder 10 Entscheidungen für den grünen Gang H_0 ablehnt, während man dies im Originalproblem nur bei 9 oder 10 Entscheidungen für "grün" täte. Das ist gewissermaßen paradox: Obwohl man eine stärkere Evidenz für H_0 hat (denn $p < \frac{1}{2}$ ist ja a priori ausgeschlossen im modifizierten Probem), lehnt man H_0 im modifizierten Problem eher ab (nämlich schon bei 8 grünen Versuchen).

Für große Versuchsumfänge n können wir wieder den Zentralen Grenzwertsatz verwenden, den wir schon in Kapitel 3 kennengelernt haben.

Beispiel 4.14 *Eine Münze wird 1 000 mal geworfen. Es soll wieder*

$$H_0 : p = \frac{1}{2} \qquad gegen \qquad H_1 : p \neq \frac{1}{2}$$

getestet werden, also fragen wir, ob die Münze fair ist. Das Testniveau betrage α=5 % und wir beobachten 550 mal Kopf. Können wir H_0 bestätigen oder verwerfen? Wir konstruieren den Ablehnungsbereich. Gesucht ist Γ, so dass

$$\mathbb{P}_{1/2}(|S_{1000} - 500| \geq \Gamma) \leq \alpha$$

und dabei Γ möglichst klein. Es ist nach dem Zentralen Grenzwertsatz:

$$\mathbb{P}_{1/2}(|S_{1000} - 500| \geq \Gamma) = \mathbb{P}_{1/2}(|\frac{S_{1000} - 500}{\sqrt{1000 \cdot \frac{1}{4}}}| \geq \frac{2\Gamma}{\sqrt{1000}}) \approx 2 - 2\Phi(\frac{2\Gamma}{\sqrt{1000}}).$$

Dabei ist $\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$ gesetzt. Es soll also

$$2 - 2\Phi(\frac{2\Gamma}{\sqrt{1000}}) = 0,05$$

sein und deshalb

$$\Phi(\frac{2\Gamma}{\sqrt{1000}}) = 0,975.$$

Aus einer $\mathcal{N}(0,1)$*-Tafel entnehmen wir, dass dazu*

$$\frac{2\Gamma}{\sqrt{1000}} \simeq 1,96$$

sein muss, also

$$\Gamma \cong 31.$$

Wir werden H_0 *also annehmen, wenn wir zwischen 469 und 531 mal Kopf sehen und ansonsten verwerfen. In diesem Fall (550 mal Kopf) würden wir* H_0 *also verwerfen. Ist stattdessen* $\alpha = 0,001$ *vorgegeben, so führt dieselbe Rechnung zu*

$$\begin{aligned} 2 - 2\Phi(\frac{2\Gamma}{\sqrt{1000}}) &= 0,001, \qquad \textit{d. h.} \\ \Phi(\frac{2\Gamma}{\sqrt{1000}}) &= 0,9995, \qquad \textit{also} \\ \frac{2\Gamma}{\sqrt{1000}} &\approx 3,3, \qquad \textit{d. h.} \\ \Gamma &\cong 52. \end{aligned}$$

In diesem Falle würde man H_0 *annehmen.*