13 Einführung in die Wahrscheinlichkeitstheorie

13.1 Modelle, Daten und die Rolle der Stochastik

Physik lebt vom Wechselspiel zweier Richtungen: Auf der einen Seite steht ein **Modell** unserer Welt (ein physikalisches Gesetz mit einer Handvoll Parametern), auf der anderen Seite stehen **Daten**: Messwerte aus dem Labor oder von Teleskopen. Wie hängen die beiden zusammen? Genau diese Frage beantwortet die **Stochastik**, die Lehre vom Rechnen mit zufällig erscheinenden Phänomenen. Sie zerfällt in zwei komplementäre Disziplinen, siehe Abbildung 13.1:

- Die **Wahrscheinlichkeitstheorie** denkt *vorwärts*: Ist das Modell gegeben, sagt sie vorher, welche Daten wir mit welcher Wahrscheinlichkeit beobachten sollten.
- Die **Statistik** denkt *rückwärts*: Sind die Daten gegeben, versucht sie herauszufinden, welches Modell – bzw. welche Parameter eines Modells – am besten zu ihnen passt.

Abbildung 13.1: Modell und Daten stehen über die Stochastik in Verbindung: Die Wahrscheinlichkeitstheorie schließt vorwärts vom Modell auf mögliche Daten, die Statistik rückwärts von den Daten auf das zugrundeliegende Modell. Dass die „Daten“-Kopie rechts deutlich unordentlicher ausfällt als das „Modell“ links, ist Absicht: Ein Modell ist ein sauberes, idealisiertes Abbild der Welt, während echte Messdaten immer mit Rauschen, Unsicherheiten und Unvollkommenheiten behaftet sind.

Um diesen Rückschluss (Statistik) überhaupt führen zu können, müssen wir zunächst verstehen, was eine Wahrscheinlichkeit überhaupt *ist* und welche Rechenregeln für sie gelten. Das ist das Ziel dieses Kapitels; in Kapitel 14 bauen wir darauf auf.

13.2 Ein unfairer Würfel

Als Einstieg betrachten wir einen sechsseitigen Würfel (Abbildung 13.2), von dem wir aus vielen vorherigen Würfen folgende Beobachtungen kennen:

- Eine 1, 2 oder 3 wird in sieben von zehn Würfen geworfen.
- Die Augenzahlen 4, 5 und 6 werden alle gleich wahrscheinlich geworfen.
- Eine 1 oder 2 wird in sechs von zehn Würfen geworfen.
- Eine 1 wird doppelt so oft geworfen wie eine 2.

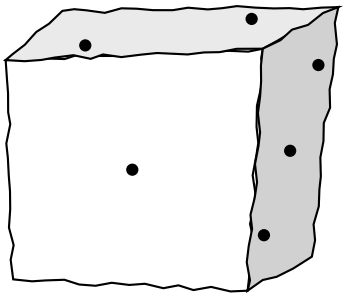


Abbildung 13.2: Ein *unfairer* Würfel: Schon die schiefen, ungleichmäßigen Seitenflächen verraten, dass hier nicht alles mit rechten Dingen zugeht. Sichtbar sind die Augenzahlen 1 (vorne), 2 (oben) und 3 (rechts) – genau die Gruppe, die laut Beobachtung in 7 von 10 Würfen fällt. Die Rückseiten 4, 5 und 6 sind dagegen gleichverteilt.

Aufgabe 13.1

Wie modellieren wir einen solchen Würfel mathematisch?

Es hilft, nicht über einzelne Würfe, sondern über *Ereignisse* nachzudenken – also über Teilmengen der möglichen Ausgänge wie $\{1, 2, 3\}$ oder $\{4\}$.

13.3 Wahrscheinlichkeitsräume

Definition 13.1

Ein **Wahrscheinlichkeitsraum** ist ein Tripel (Ω, Σ, P) mit:

- der **Grundmenge** Ω, der Menge aller möglichen Ausgänge;
- dem **Ereignissystem** Σ, einer Menge von Teilmengen von Ω (den *Ereignissen*); im diskreten Fall wählt man meist die volle **Potenzmenge** $\Sigma = \mathcal{P}(\Omega) := \{U \mid U \subset \Omega\}$;
- dem **Wahrscheinlichkeitsmaß** $P : \Sigma \to [0, 1]$, das jedem Ereignis eine Wahrscheinlichkeit zuordnet und die folgenden (Kolmogorov-)Axiome erfüllt:

$$P(E) \geq 0 \quad \forall E \in \Sigma\,, \tag{13.1}$$

$$P(\Omega) = 1\,, \tag{13.2}$$

$$P(A \cup B) = P(A) + P(B) \quad \text{für } A, B \in \Sigma \text{ mit } A \cap B = \emptyset\,. \tag{13.3}$$

Musterlösung 13.1

Für unseren Würfel ist $\Omega = \{1, 2, 3, 4, 5, 6\}$ und $\Sigma = \mathcal{P}(\Omega)$. Aus den Beobachtungen oben wissen wir:

$$P(\{1, 2, 3\}) = \frac{7}{10}\,, \qquad P(\{4\}) = P(\{5\}) = P(\{6\})\,, \tag{13.4}$$

$$P(\{1, 2\}) = \frac{6}{10}\,, \qquad P(\{1\}) = 2\,P(\{2\})\,. \tag{13.5}$$

Mit Axiom (13.3) (angewandt auf die disjunkten Ereignisse $\{4\}, \{5\}, \{6\}$) und (13.2) folgt $P(\{4\}) + P(\{5\}) + P(\{6\}) = 1 - \frac{7}{10} = \frac{3}{10}$, also $P(\{4\}) = P(\{5\}) = P(\{6\}) = \frac{1}{10}$.

Da $\{1,2,3\} = \{1,2\} \cup \{3\}$ eine disjunkte Vereinigung ist, liefert Axiom (13.3) außerdem

$$P(\{3\}) = P(\{1,2,3\}) - P(\{1,2\}) = \frac{7}{10} - \frac{6}{10} = \frac{1}{10}\,. \tag{13.6}$$

Ebenso ist $\{1,2\} = \{1\} \cup \{2\}$ disjunkt, also $P(\{1\}) + P(\{2\}) = P(\{1,2\}) = \frac{6}{10}$. Setzt man $P(\{1\}) = 2P(\{2\})$ ein, folgt $3P(\{2\}) = \frac{6}{10}$, also

$$P(\{2\}) = \frac{2}{10} = \frac{1}{5}\,, \qquad P(\{1\}) = 2P(\{2\}) = \frac{4}{10} = \frac{2}{5}\,. \tag{13.7}$$

Damit ist unser unfairer Würfel bereits vollständig modelliert – ganz ohne den Begriff der bedingten Wahrscheinlichkeit, den wir erst in Abschnitt 13.4 einführen und direkt an diesem Würfel illustrieren:

Augenzahl k	1	2	3	4	5	6
$P(\{k\})$	0.4	0.2	0.1	0.1	0.1	0.1

(Probe: Die sechs Werte summieren sich, wie es Axiom (13.2) verlangt, zu 1.)

13.4 Bedingte Wahrscheinlichkeit und der Satz von Bayes

Definition 13.2

Für zwei Ereignisse $A, B \in \Sigma$ mit $P(B) > 0$ ist die **bedingte Wahrscheinlichkeit** von A gegeben B definiert als

$$P(A \mid B) := \frac{P(A \cap B)}{P(B)}\,. \tag{13.8}$$

Als Beispiel nutzen wir unseren inzwischen (Abschnitt 13.3) vollständig bekannten Würfel: Wie wahrscheinlich ist eine 1, wenn wir schon wissen, dass eine 1, 2 oder 3 gewürfelt wurde? Mit Gleichung (13.8) und $\{1\} \cap \{1,2,3\} = \{1\}$ folgt

$$P(\{1\} \mid \{1,2,3\}) = \frac{P(\{1\})}{P(\{1,2,3\})} = \frac{0.4}{0.7} = \frac{4}{7}\,. \tag{13.9}$$

Diese bedingte Wahrscheinlichkeit ist *größer* als die unbedingte $P(\{1\}) = 0.4$: Das Wissen, dass eine 1, 2 oder 3 gefallen ist, schließt die Möglichkeiten $4, 5, 6$ (zusammen 0.3 Wahrscheinlichkeit) aus und macht eine 1 dadurch relativ wahrscheinlicher. Genauso findet man $P(\{2\} \mid \{1,2,3\}) = 0.2/0.7 = 2/7$ und $P(\{3\} \mid \{1,2,3\}) = 0.1/0.7 = 1/7$ – die drei bedingten Wahrscheinlichkeiten summieren sich korrekt zu 1, wie es sein muss, wenn wir uns auf das Ereignis $\{1,2,3\}$ beschränken.

13.5 Aufgaben zur bedingten Wahrscheinlichkeit und dem Satz von Bayes

Aufgabe 13.2

a) Zeige, dass für zwei Ereignisse $A, B \in \Sigma$

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)} \tag{13.10}$$

gilt. Diese Beziehung heißt Satz von Bayes und ist wichtig in unserem Kurs.

b) Findet eine intuitive Erklärung für den Satz von Bayes, die eure Eltern verstehen würden.

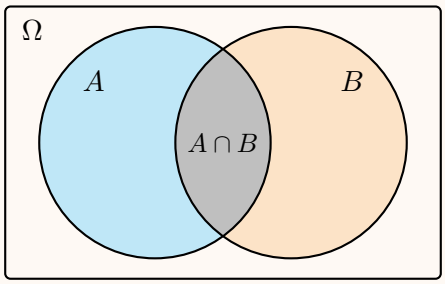
c) Sei $A \in \Sigma$ und A^c das *Gegenereignis* zu A, d.h. die Menge $A^c \in \Sigma$ mit $A \cap A^c = \emptyset$ und $A \cup A^c = \Omega$. Zeige, dass für ein beliebiges Ereignis $B \in \Sigma$

$$P(B) = P(B \mid A) \cdot P(A) + P(B \mid A^c) \cdot P(A^c) \tag{13.11}$$

gilt (man spricht auch vom *Gesetz der totalen Wahrscheinlichkeit*). Versucht, auch diesen Zusammenhang intuitiv zu verstehen.

Musterlösung 13.2

Zur Veranschaulichung der beteiligten Mengen:



a) Für $P(A), P(B) > 0$ liefert Definition 13.2 zwei Ausdrücke für dieselbe Schnittmenge (die grau schattierte Fläche $A \cap B$ oben):

$$P(A \cap B) = P(A \mid B)\, P(B)\,, \qquad P(A \cap B) = P(B \mid A)\, P(A)\,, \tag{13.12}$$

wobei wir im zweiten Schritt $A \cap B = B \cap A$ benutzt haben. Gleichsetzen und Division durch $P(B)$ liefert die Behauptung.
b) Eine mögliche Formulierung: Der Satz von Bayes sagt, wie du eine Vermutung über eine *Ursache* anpassen solltest, nachdem du eine *Beobachtung* gemacht hast. Dabei kombinierst du zwei Dinge: wie plausibel die Ursache *vorher* schon war (Prior $P(A)$) und wie gut die Beobachtung zu dieser Ursache passt (Likelihood $P(B \mid A)$). Beispiel: Du kommst morgens runter und der Rasen ist nass (Beobachtung B). Hat es geregnet (Ursache A) oder lief der Rasensprenger? Wenn es in deiner Gegend selten regnet ($P(A)$ klein), aber Regen fast immer nassen Rasen bedeutet ($P(B \mid A)$ groß), reicht die nasse Wiese allein noch nicht, um sicher auf Regen zu schließen – du musst beides gegeneinander abwägen. Genau das tut der Satz von Bayes rechnerisch.
c) Da $A \cup A^c = \Omega$, gilt für jedes $B \in \Sigma$: $B = B \cap \Omega = B \cap (A \cup A^c) = (B \cap A) \cup (B \cap A^c)$. Diese Vereinigung ist *disjunkt*, denn $(B \cap A) \cap (B \cap A^c) \subset A \cap A^c = \emptyset$. Mit dem dritten Kolmogorov-Axiom (13.3) folgt

$$P(B) = P(B \cap A) + P(B \cap A^c)\,. \tag{13.13}$$

Wendet man auf beide Summanden die Definition der bedingten Wahrscheinlichkeit an ($P(B \cap A) = P(B \mid A)P(A)$ und analog für A^c), folgt die Behauptung. Intuitiv: Jedes Ereignis B tritt entweder zusammen mit A oder zusammen mit A^c ein – ein Drittes gibt es nicht. Die Gesamtwahrscheinlichkeit von B ist deshalb der gewichtete Mittelwert seiner Wahrscheinlichkeit in diesen beiden Fällen, gewichtet mit der Wahrscheinlichkeit des jeweiligen Falls.

Der Satz von Bayes ist einer der folgenreichsten Sätze der Wahrscheinlichkeitstheorie:

Satz von Bayes

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}\,. \tag{13.14}$$

Eine bedingte Wahrscheinlichkeit lässt sich damit „umdrehen“: Aus $P(B \mid A)$ wird $P(A \mid B)$. In der Sprache der Statistik nennt man $P(A)$ den **Prior** (das Vorwissen über A, bevor wir B beobachtet haben), $P(B \mid A)$ die **Likelihood** (wie wahrscheinlich die Beobachtung B unter der Annahme A ist), $P(A \mid B)$ den **Posterior** (unser aktualisiertes Wissen über A, nachdem wir B beobachtet haben) und $P(B)$ die **Evidenz** (eine Normierungskonstante).

Diesen Begriffen begegnen wir in Kapitel 16 wieder, wenn wir Bayessche Inferenz an einem konkreten physikalischen System durchführen.

Aufgabe 13.3

Ein medizinischer Test wird eingesetzt, um eine seltene Krankheit in der Bevölkerung zu erkennen. Die Krankheit tritt bei 0.2 % der Bevölkerung auf. Der Test erkennt erkrankte Personen mit einer Wahrscheinlichkeit von 99 %. Gesunde Personen werden mit einer Wahrscheinlichkeit von 2 % fälschlicherweise als krank markiert.

a) Definiere einen geeigneten Wahrscheinlichkeitsraum für das „Zufallsexperiment“, dass eine Person auf die Krankheit getestet wird. Lege dazu Grundmenge sowie geeignete Ereignisse für den Gesundheitszustand und das Testergebnis fest und gib die bekannten Wahrscheinlichkeiten an.

b) Die Person wird positiv getestet. Berechne die Wahrscheinlichkeit, dass diese Person tatsächlich an der Krankheit leidet.

Hinweis: Beide Gleichungen aus Aufgabe 13.2 sind für die Berechnung hilfreich.

Musterlösung 13.3

a) Grundmenge $\Omega = \{(\text{krank}, +), (\text{krank}, -), (\text{gesund}, +), (\text{gesund}, -)\}$, mit $\Sigma = \mathcal{P}(\Omega)$. Wir betrachten die Ereignisse

$$K = \text{„Person ist krank“}\,, \qquad T = \text{„Test ist positiv“}\,, \tag{13.15}$$

mit Gegenereignissen K^c (gesund) und T^c (negativ). Gegeben sind:

$$P(K) = 0.002\,, \qquad P(T \mid K) = 0.99\,, \qquad P(T \mid K^c) = 0.02\,. \tag{13.16}$$

b) Gesucht ist $P(K \mid T)$. Mit dem Satz von Bayes (Aufgabe 13.2a):

$$P(K \mid T) = \frac{P(T \mid K)\, P(K)}{P(T)}\,. \tag{13.17}$$

Den Nenner $P(T)$ erhalten wir mit dem Gesetz der totalen Wahrscheinlichkeit (Aufgabe 13.2c), angewandt auf die Partition $\{K, K^c\}$:

$$P(T) = P(T \mid K)\, P(K) + P(T \mid K^c)\, P(K^c) = 0.99 \cdot 0.002 + 0.02 \cdot 0.998 \tag{13.18}$$

$$= 0.00198 + 0.01996 = 0.02194\,. \tag{13.19}$$

Damit

$$P(K \mid T) = \frac{0.00198}{0.02194} \approx 0.090\,, \tag{13.20}$$

also nur etwa 9 %! Obwohl der Test zu 99 % zuverlässig ist, liegt bei einem positiven Ergebnis in neun von zehn Fällen *keine* Erkrankung vor. Der Grund: Die Krankheit ist so selten ($P(K) = 0.2\,\%$), dass die (wenigen) falsch-positiven Ergebnisse unter den vielen Gesunden die (wenigen) richtig-positiven Ergebnisse unter den wenigen Kranken zahlenmäßig überwiegen. Diese „Basisraten-Falle“ ist einer der wichtigsten Gründe, warum man Testergebnisse nie isoliert, sondern immer zusammen mit der Prävalenz $P(K)$ interpretieren sollte.

Aufgabe 13.4

In einer Spielshow stehen drei verschlossene Türen zur Auswahl. Hinter einer Tür befindet sich ein Mittel für Weltfrieden, hinter den beiden anderen jeweils ein feucht-schwitziger Händedruck. Eine Kandidatin wählt zunächst zufällig eine Tür. Anschließend öffnet der Moderator eine der beiden übrigen Türen, hinter der sich sicher der feucht-schwitzige Händedruck befindet. Danach bietet er der Kandidatin an, bei ihrer ursprünglichen Wahl zu bleiben oder zur anderen, noch geschlossenen Tür zu wechseln.

a) Definiere einen geeigneten Wahrscheinlichkeitsraum. Beschreibe die Grundmenge und geeignete Ereignisse, die den Ort des Wundermittels und die erste Wahl der Kandidatin berücksichtigen.

b) Sollte die Kandidatin die Tür wechseln oder bei ihrer ursprünglichen Wahl bleiben? Berechne die jeweiligen Wahrscheinlichkeiten.

Musterlösung 13.4

a) Sei $D \in \{1,2,3\}$ die Tür mit dem Wundermittel und $C \in \{1,2,3\}$ die erste Wahl der Kandidatin. Beide sind unabhängig und gleichverteilt, also

$$\Omega = \{1,2,3\} \times \{1,2,3\}\,, \qquad P((d,c)) = \frac{1}{9} \quad \forall (d,c) \in \Omega\,. \tag{13.21}$$

Das Ereignis „Treffer“ $T = \{(d,c) \in \Omega \mid d = c\}$ (die erste Wahl ist bereits richtig) hat 3 der 9 gleichwahrscheinlichen Elementarereignisse, also $P(T) = \frac{1}{3}$ und $P(T^c) = \frac{2}{3}$.
b) Der Moderator öffnet *immer* eine Tür mit Händedruck, die weder die Preistür D noch die gewählte Tür C ist – das ist ihm garantiert möglich, egal was D und C sind. Entscheidend ist: Ob Wechseln zum Gewinn führt, hängt nur davon ab, ob die *erste* Wahl schon richtig war.

- Tritt T ein (Wahrscheinlichkeit $\frac{1}{3}$): Die erste Wahl war schon die Preistür. Die einzige noch geschlossene Tür nach dem Öffnen des Moderators enthält den Händedruck – Wechseln verliert, Bleiben gewinnt.
- Tritt T^c ein (Wahrscheinlichkeit $\frac{2}{3}$): Die erste Wahl war falsch. Der Moderator kann dann gar nicht anders, als die *einzige* verbleibende Tür mit Händedruck zu öffnen – die Preistür bleibt zwangsläufig geschlossen. Wechseln gewinnt, Bleiben verliert.

Also gilt $P(\text{Gewinn beim Wechseln}) = P(T^c) = \frac{2}{3}$ und $P(\text{Gewinn beim Bleiben}) = P(T) = \frac{1}{3}$. Die Kandidatin sollte **wechseln**: Das verdoppelt ihre Gewinnchance von 33 % auf 67 %.

Intuitiv: Die erste Wahl ist zu $\frac{2}{3}$ falsch; durch das gezielte Öffnen einer Verliererтür „bündelt“ der Moderator diese $\frac{2}{3}$ Wahrscheinlichkeit auf die verbleibende Tür.

13.6 Häufige diskrete Wahrscheinlichkeitsverteilungen

Neben unserem Würfel mit endlich vielen Ausgängen gibt es auch für diskrete Zufallsvariablen mit *unbeschränkt* vielen möglichen Werten (etwa eine Anzahl $k \in \{0, 1, 2, \ldots\}$) etablierte Standardverteilungen. Zwei der wichtigsten sind die Poisson- und die Binomialverteilung.

13.6.1 Binomialverteilung

Bei n unabhängigen Versuchen mit Erfolgswahrscheinlichkeit p folgt die Anzahl der Erfolge k der **Binomialverteilung**:

$$B(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}, \quad \langle k \rangle = np, \quad \sigma_k^2 = np(1-p)\,. \tag{13.22}$$

Sie beschreibt z.B. die Anzahl der Erfolge bei wiederholten Münzwürfen oder allgemein bei Experimenten mit nur zwei möglichen Ausgängen („Erfolg“ oder „Misserfolg“); wir begegnen ihr in Aufgabe 13.6 bei einer unfairen Münze wieder. Dabei ist das Symbol $\binom{n}{k} = n!/(k!(n-k)!)$ der sog. Binomialkoeffizient. Die Schreibweise $n! = n \times (n-1) \times \cdots \times 1$ steht für die sog. Fakultät von n.

13.6.2 Poissonverteilung

Im Grenzfall $n \to \infty$, $p \to 0$ bei festem $\lambda = np$ geht die Binomialverteilung in die **Poissonverteilung** über:

$$P(k; \lambda) = \frac{\lambda^k}{k!}\, \mathrm{e}^{-\lambda}, \quad \langle k \rangle = \lambda, \quad \sigma_k^2 = \lambda\,. \tag{13.23}$$

Sie beschreibt seltene Ereignisse mit konstanter Rate (z.B. radioaktiver Zerfall, Photonendetektion) und hat die bemerkenswerte Eigenschaft $\sigma_k = \sqrt{\lambda}$: die relative Unsicherheit $\sigma_k/\langle k \rangle = 1/\sqrt{\lambda}$ geht für viele Ereignisse gegen null. Abbildung 13.3 zeigt beide Verteilungen im Vergleich.

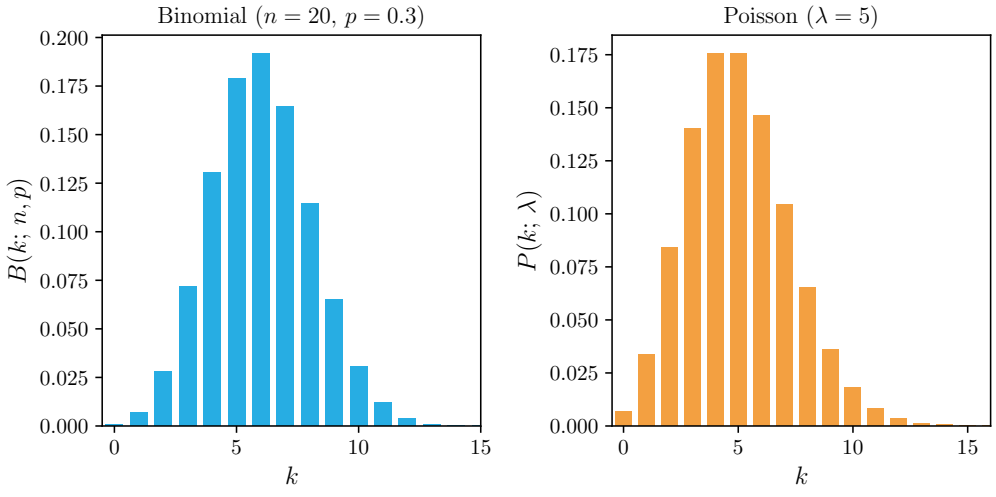


Abbildung 13.3: *Links:* Binomialverteilung $B(k;\, n = 20,\, p = 0.3)$. *Rechts:* Poissonverteilung $P(k;\, \lambda = 5)$ – ihr Grenzfall für seltene Ereignisse mit konstanter Rate.

Bevor wir die Poissonverteilung an einem simulierten Datensatz überprüfen, brauchen wir zwei Kennzahlen, mit denen man einen Datensatz $\{x_1, \ldots, x_N\}$ zusammenfasst: den **Stichprobenmittelwert**

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{13.24}$$

und die **Stichprobenstandardabweichung**

$$s = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2}\,. \tag{13.25}$$

Beide schätzen aus endlich vielen Beobachtungen den (unbekannten) Erwartungswert $\langle x \rangle$ bzw. die Streuung σ_x der zugrundeliegenden Verteilung – mehr dazu in Kapitel 14.

13.7 Statistische Modelle

Bisher hatten wir ein Wahrscheinlichkeitsmaß P, das vollständig feststand. In der Praxis kennen wir P aber selten exakt – wir kennen höchstens seine *Form*, mit ein paar offenen Parametern, die wir aus Daten bestimmen wollen. Das führt zum zentralen Objekt der Statistik:

Definition 13.3

Ein Tripel $(\Omega, \Sigma, (P_\theta)_{\theta \in \Theta})$ heißt **statistisches Modell**, wobei Θ der Raum aller möglichen Parameter ist und P_θ für jedes $\theta \in \Theta$ ein Wahrscheinlichkeitsmaß auf (Ω, Σ) ist.

Beispiele:

i) Eine Münze mit $P_p(\text{„Kopf“}) = p$: Grundmenge $\Omega = \{\text{Kopf}, \text{Zahl}\}$, Parameterraum $\Theta = [0, 1]$. Für $p = \frac{1}{2}$ ist die Münze fair, sonst nicht.

ii) Die Anzahl k der Photonen, die ein Detektor in einem festen Zeitintervall registriert, folgt für viele unabhängige, seltene Ereignisse einer Poissonverteilung $P_\lambda(k)$ mit unbekannter Rate $\lambda > 0$ (Parameterraum $\Theta = (0, \infty)$), vgl. Abschnitt 13.6.

iii) Die Körpergröße x der Teilnehmenden in diesem Raum folgt näherungsweise einer Normalverteilung $P_{(\mu,\sigma)}(x) = \mathcal{N}(x;\, \mu, \sigma^2)$ (vgl. Abschnitt 13.9) mit unbekanntem Mittelwert μ und unbekannter Streuung σ. Hier ist $\theta = (\mu, \sigma)$ selbst schon ein Parameter*vektor* und $\Theta = \mathbb{R} \times (0, \infty)$ entsprechend zweidimensional – Θ muss also kein einfaches Intervall sein, sondern kann jede beliebige Dimension haben.

Das Ziel der Statistik ist es nun, ausgehend von beobachteten Daten $x_1, \ldots, x_N$ auf den „richtigen“ (oder zumindest plausibelsten) Parameter θ zurückzuschließen – also den Pfeil in Abbildung 13.1 rückwärts zu gehen. Mit den Werkzeugen dieses Kurses im Gepäck wenden wir uns dieser Aufgabe in Kapitel 14 zu.

13.8 Aufgaben zu Verteilungen und statistischen Modellen

Aufgabe 13.5

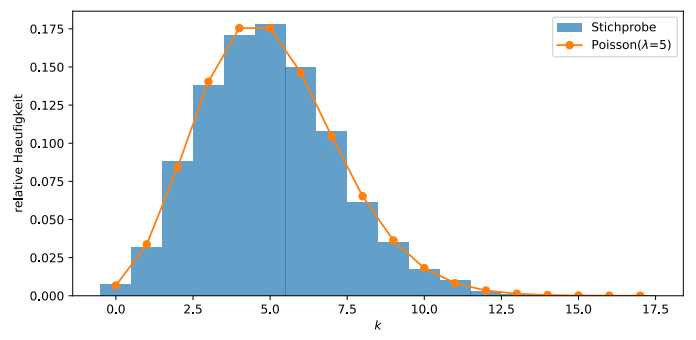
In dieser Aufgabe erkunden wir die Poissonverteilung mit Python. Du darfst in dieser Aufgabe (und auch allen folgenden Aufgaben) natürlich gerne auch das Internet zu Rate ziehen, wenn du noch Fragen hast, wie man in `Python` programmiert. (Simon und Carlo machen das auch so.)

a) Erzeuge mit `numpy.random.poisson(λ=5, size=10000)` einen Datensatz von 10 000 Poisson-Zufallszahlen. Plotte das Histogramm (normiert, `density=True`) und vergleiche es mit der theoretischen Poisson-Verteilung $P(k;\lambda)$ aus dem Text.

b) Berechne für diesen Datensatz den Stichprobenmittelwert $\bar{k}$ und die Stichprobenstandardabweichung s. Vergleiche mit den theoretischen Werten $\langle k\rangle = \lambda$ und $\sigma_k = \sqrt{\lambda}$.

Musterlösung 13.5

```
import numpy as np
import matplotlib.pyplot as plt
from math import factorial, exp

def poisson_wahrscheinlichkeit(k, lam):
    return lam**k / factorial(k) * exp(-lam)

# Teil a): Poisson-Histogramm
lam = 5
data = np.random.poisson(lam, size=10000)
k = np.arange(0, data.max() + 1)
plt.figure(figsize=(8, 4))
plt.hist(data, bins=np.arange(-0.5, data.max() + 1.5, 1), density=True,
         label='Stichprobe', alpha=0.7)
plt.plot(k, [poisson_wahrscheinlichkeit(ki, lam) for ki in k], 'o-',
         label=rf'Poisson($\lambda$={lam})')
plt.xlabel('$k$'); plt.ylabel('relative Haeufigkeit')
plt.legend(); plt.tight_layout(); plt.show()

# Teil b): Kenngroessen
def stichprobenmittelwert(x):
    return sum(x) / len(x)

def stichprobenstandardabweichung(x):
    xbar = stichprobenmittelwert(x)
    return (sum((xi - xbar)**2 for xi in x) / len(x))**0.5

# (np.mean(data) und np.std(data) wuerden hier genauso funktionieren.)
print(f'Mittelwert: {stichprobenmittelwert(data):.3f}  (theoretisch: {lam})')
print(f'Std:        {stichprobenstandardabweichung(data):.3f}   (theoretisch:
↪  {np.sqrt(lam):.3f})')
```



Aufgabe 13.6

Wir haben eine exotische Münze, die mit Wahrscheinlichkeit p Kopf zeigt. Die Münze wird 14-mal geworfen; dabei werden 10-mal Kopf und 4-mal Zahl beobachtet (Daten D). Wir interessieren uns für die Wahrscheinlichkeit, dass die nächsten beiden Würfe Kopf zeigen.

a) Beschreibe das statistische Modell für dieses Experiment.

b) Der sogenannte **Maximum-Likelihood-Schätzer** aus der frequentistischen Statistik schätzt die unbekannte Wahrscheinlichkeit p durch

$$\hat{p} = \frac{\text{Anzahl Kopf}}{\text{Anzahl Würfe}}\,. \tag{13.26}$$

Berechne den Schätzwert $\hat{p}$ und bestimme damit die Wahrscheinlichkeit, dass die nächsten beiden Würfe Kopf zeigen.

c) Schreibe in `Python` eine Funktion, die für ein gegebenes $p = 5/7$ aufeinander folgende 14 Würfe simuliert. Werte diese Funktion 10 000-mal aus und mache eine Liste der Male, in denen genau 10-mal Kopf und 4-mal Zahl beobachtet wurde. Welche frequentistische Wahrscheinlichkeit hat also dieses Ergebnis (10-mal Kopf und 4-mal Zahl in 14 Würfen)?

Hinweis: Die Funktion `np.random.binomial(n, p)` zieht dir direkt eine solche Stichprobe. Google gerne mal, wie man diese importiert und wofür n dabei steht.

d) Bei n unabhängigen Versuchen mit Erfolgswahrscheinlichkeit p folgt die Anzahl der Erfolge k der Verteilung

$$B(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{13.27}$$

Schreibe in `Python` eine Funktion, die die Wahrscheinlichkeit für k-mal Kopf bei n Würfen mit vorgegebenem p berechnet. Berechne damit für

$$p = 0.2,\ 0.4,\ 0.6,\ 0.8,\ 0.9$$

jeweils die Wahrscheinlichkeit, bei 14 Würfen genau 10-mal Kopf und 4-mal Zahl zu beobachten.

Hinweis: Hier ist `np.random.binomial` *nicht* das richtige Werkzeug – das liefert dir nur eine zufällige Stichprobe, aber keine Wahrscheinlichkeit. Berechne stattdessen $n!$ selbst (oder mit `math.factorial`) und setze damit den Binomialkoeffizienten und $B(k; n, p)$ direkt aus der obigen Formel zusammen.

e) Plotte die Ergebnisse der Aufgaben c) und d) gemeinsam und vergleiche sie miteinander. Da c) bisher nur eine einzige frequentistische Wahrscheinlichkeit bei $p = \hat{p}$ liefert, musst du die Simulation aus c) dafür zunächst erweitern: Wiederhole sie für jeden der fünf Werte $p = 0.2, 0.4, 0.6, 0.8, 0.9$ aus Teil d). Trage anschließend sowohl die theoretischen (Teil d) als auch die simulierten (Teil c) Wahrscheinlichkeiten für $P(\text{10-mal Kopf})$ über p auf.

f) (*Bonus*) In der Bayesschen Statistik wird p als zufällige Größe modelliert. Wir nehmen zunächst an, dass alle Werte zwischen 0 und 1 gleich wahrscheinlich sind (*a-priori*-Verteilung). Versuche, die *a-posteriori*-Verteilung (numerisch oder auf dem Papier) zu

berechnen. Wie groß ist die Wahrscheinlichkeit, dass die nächsten beiden Würfe Kopf zeigen, wenn du diese Verteilung zur Berechnung verwendest? Unterscheidet sich das von dem frequentistischen Ergebnis? Basierend auf deiner Analyse: Solltest du Geld darauf verwetten, dass die nächsten beiden Würfe Kopf zeigen?
Hinweis: Erwartungswerte werden wie folgt berechnet:

$$\langle f(p) \rangle = \int_0^1 f(p)\, P(p \mid D)\, \mathrm{d}p\,. \tag{13.28}$$

Wenn du noch nicht gelernt hast, wie man das relevante Integral löst, recherchiere gerne im Internet, wie das numerisch in `Python` geht.

Musterlösung 13.6

a) Die interessierende Zufallsgröße ist die Anzahl k der Köpfe bei $n = 14$ unabhängigen Würfen mit Kopfwahrscheinlichkeit p; sie kann jeden Wert zwischen 0 und 14 annehmen. Das statistische Modell (Definition 13.3, Abschnitt 13.7) ist also $(\Omega, \Sigma, (P_p)_{p\in\Theta})$ mit Grundmenge $\Omega = \{0, 1, \ldots, 14\}$, Ereignissystem $\Sigma = \mathcal{P}(\Omega)$ und Parameterraum $\Theta = [0, 1]$; welche konkrete Wahrscheinlichkeit P_p jedem k zuordnet, leiten wir in Teil d) her.
b) $\hat{p} = 10/14 = 5/7 \approx 0.714$. Unter der Annahme $p = \hat{p}$ (und Unabhängigkeit der Würfe) ist die Wahrscheinlichkeit zweier weiterer Köpfe $\hat{p}^2 = (5/7)^2 = 25/49 \approx 0.510$.
c)

```
import numpy as np

def muenzwuerfe_simulieren(p, n=14):
    """Simuliert n aufeinanderfolgende Muenzwuerfe mit Kopfwahrscheinlichkeit
    ↪  p; gibt Anzahl Kopf zurueck."""
    return np.random.binomial(n, p)

p_hat = 5 / 7  # Ergebnis aus Teil b)
ergebnisse = [muenzwuerfe_simulieren(p_hat) for _ in range(10000)]
treffer = sum(1 for k in ergebnisse if k == 10)
print(f'Frequentistische Wahrscheinlichkeit: {treffer / 10000:.3f}
↪  ({treffer}/10000 Treffer)')
```

Wir simulieren also nicht mit einem beliebigen p, sondern gezielt mit unserem Schätzwert $\hat{p} = 5/7$ aus Teil b). In einem Durchlauf ergaben sich z.B. 2342 von 10 000 Wiederholungen mit genau 10-mal Kopf, also eine frequentistische Wahrscheinlichkeit von ≈ 0.234 (der genaue Wert schwankt von Durchlauf zu Durchlauf leicht). Das Ergebnis wird uns in Teil d) als Vergleichswert dienen.
d) Binomialkoeffizient und Fakultät lassen sich direkt aus ihrer in der Aufgabenstellung gegebenen Definition heraus programmieren:

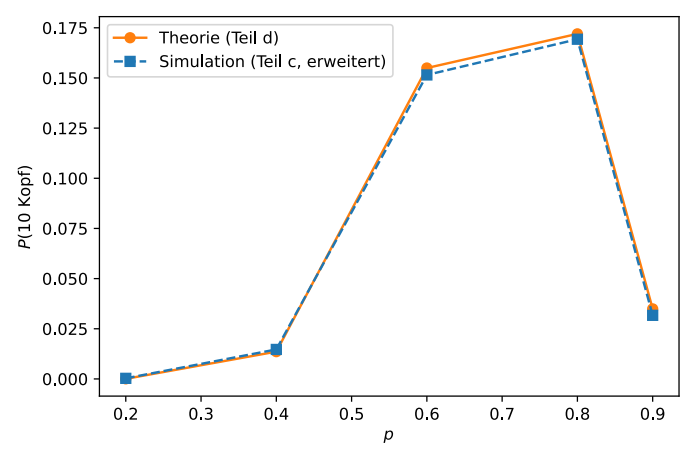
```
from math import factorial

def binomialkoeffizient(n, k):
    return factorial(n) // (factorial(k) * factorial(n - k))

def binomial_wahrscheinlichkeit(k, n, p):
    return binomialkoeffizient(n, k) * p**k * (1 - p)**(n - k)

for p in [0.2, 0.4, 0.6, 0.8, 0.9]:
    print(f'p={p}: P(10 Kopf) = {binomial_wahrscheinlichkeit(10, 14, p):.3g}')
```

p	0.2	0.4	0.6	0.8	0.9
$P(D \mid p)$	4.2×10^{-5}	1.4×10^{-2}	0.155	0.172	0.035

Die Wahrscheinlichkeit steigt also zunächst mit p, erreicht aber *nicht* ihr Maximum bei $p = 0.9$, sondern fällt dorthin schon wieder deutlich ab: Bei $p = 0.8$ ist sie mit 0.172 noch größer als bei $p = 0.9$ mit nur 0.035. Das passt genau zum beobachteten Anteil $10/14 \approx 0.71$, der zwischen 0.6 und 0.8 liegt – die Wahrscheinlichkeit hat dort ihr Maximum und nimmt in *beide* Richtungen wieder ab. Zur Probe werten wir dieselbe Funktion bei $p = \hat{p} = 5/7$ aus Teil b) aus: `binomial_wahrscheinlichkeit(10, 14, 5/7)` ≈ 0.231 – in guter Übereinstimmung mit der frequentistischen Schätzung ≈ 0.234 aus Teil c)!

e) Um die Ergebnisse aus c) und d) auf einem gemeinsamen Plot vergleichen zu können, müssen sie dieselbe Größe als Funktion derselben Variable darstellen. Wir ergänzen dazu die Simulation aus c), die bisher nur für den einen Wert $p = \hat{p}$ lief, um dieselben fünf Werte $p = 0.2, 0.4, 0.6, 0.8, 0.9$ wie in Teil d):

```
import matplotlib.pyplot as plt

p_werte = [0.2, 0.4, 0.6, 0.8, 0.9]
frequentistisch = []
for p in p_werte:
    ergebnisse = [muenzwuerfe_simulieren(p) for _ in range(10000)]
    treffer = sum(1 for k in ergebnisse if k == 10)
    frequentistisch.append(treffer / 10000)

theoretisch = [binomial_wahrscheinlichkeit(10, 14, p) for p in p_werte]

plt.plot(p_werte, theoretisch, 'o-', label='Theorie (Teil d)')
plt.plot(p_werte, frequentistisch, 's--', label='Simulation (Teil c,
↪  erweitert)')
plt.xlabel('$p$'); plt.ylabel(r'$P(10\ \mathrm{Kopf})$')
plt.legend(); plt.tight_layout(); plt.show()
```



Simulation und Theorie stimmen für alle fünf Werte von p sehr gut überein – ein weiterer Beleg dafür, dass die in Teil d) hergeleitete Formel tatsächlich dieselbe Verteilung beschreibt, die wir in c) simuliert haben. Der Wert $p = 0.9$ macht dabei besonders deutlich, dass die Wahrscheinlichkeit tatsächlich ein *Maximum* bei $p \approx 10/14$ hat und keineswegs mit p immer weiter ansteigt.

f) *(Bonus)* Prior $\pi(p) = 1$ für $p \in [0, 1]$ (vgl. die Gleichverteilung $\mathcal{U}(0, 1)$ in Abschnitt 13.9, der Standard-uninformative Prior). Die Likelihood ist die Binomialverteilung aus Teil d),

$P(D \mid p) = \binom{14}{10} p^{10}(1-p)^4$. Mit dem Satz von Bayes:

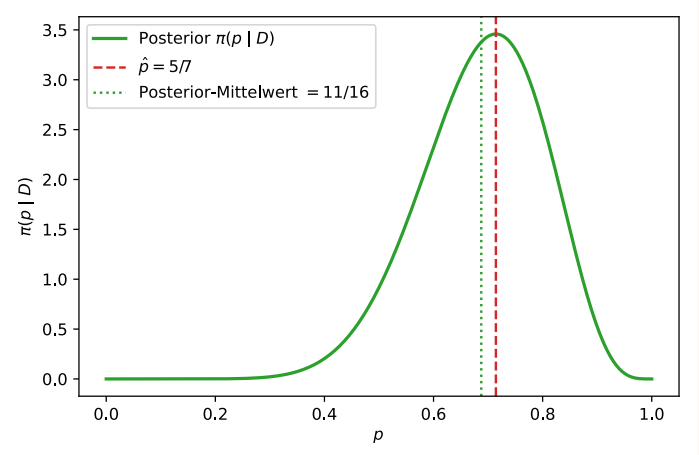
$$\pi(p \mid D) = \frac{P(D \mid p)\,\pi(p)}{P(D)} = \frac{\binom{14}{10} p^{10}(1-p)^4 \cdot 1}{P(D)} \propto p^{10}(1-p)^4\,, \tag{13.29}$$

da weder $\binom{14}{10}$ noch $P(D)$ von p abhängen und deshalb in der Proportionalitätskonstante verschwinden. Die fehlende Normierungskonstante ist das Integral $Z = \int_0^1 p^{10}(1-p)^4\,\mathrm{d}p$; damit ist $\pi(p \mid D) = p^{10}(1-p)^4/Z$. Statt diese Größe von Hand auszurechnen, werten wir $p^{10}(1-p)^4$ auf einem feinen Gitter aus und integrieren numerisch (z.B. mit der Trapezregel). Auf demselben Weg erhalten wir sofort auch die eigentlich gesuchte Größe: die Wahrscheinlichkeit zweier weiterer Köpfe ist ja gerade der Erwartungswert $\langle p^2 \rangle$ von $P(HH \mid p) = p^2$ unter der Posterior-Verteilung,

$$P(HH \mid D) = \langle p^2 \rangle_{\pi(p|D)} = \int_0^1 p^2\,\pi(p \mid D)\,\mathrm{d}p\,. \tag{13.30}$$

```
import numpy as np

p_grid = np.linspace(0, 1, 1000)
unnormiert = p_grid**10 * (1 - p_grid)**4

Z = np.trapezoid(unnormiert, p_grid)
posterior = unnormiert / Z

p_mittelwert = np.trapezoid(p_grid * posterior, p_grid)
p2_erwartung = np.trapezoid(p_grid**2 * posterior, p_grid)

print(f'Z = {Z:.4g}, <p> = {p_mittelwert:.4f}, <p^2> = {p2_erwartung:.4f}')
```

Das liefert $Z \approx 6.66 \times 10^{-5}$, $\langle p \rangle \approx 0.6875$ und $\langle p^2 \rangle \approx 0.4853$.



Der Posterior-Mittelwert $\langle p \rangle \approx 0.6875$ liegt etwas unter $\hat{p} = 5/7 \approx 0.714$: Der uninformative Prior zieht das Ergebnis leicht in Richtung seines eigenen Mittelwerts $\langle p \rangle = \frac{1}{2}$ („Shrinkage“), ein Effekt, der mit mehr Daten immer kleiner wird. Genau denselben Effekt sehen wir bei der eigentlich gesuchten Antwort: Die Bayessche Wahrscheinlichkeit für zwei weitere Köpfe, $P(HH \mid D) = \langle p^2 \rangle \approx 0.485$, liegt ebenfalls minimal unter der naiven frequentistischen Schätzung $\hat{p}^2 \approx 0.510$ aus Teil b) – auch hier zieht der prior die Wahrscheinlichkeit etwas herunter. Im Grenzwert vieler “Messungen” verschwindet die Abhängigkeit vom Prior und die frequentistische und die Bayessche Aussagen nähern sich einander an: Die Daten siegen am Ende über unsere subjektiven Vorannahmen.