Mini-Skript zur Vorlesung Wahrscheinlichkeitsrechnung und Statistik (D-INFK)

Peter Bühlmann

Herbstsemester 2009

1 Der Begriff der Wahrscheinlichkeit

Stochastik befasst sich mit **Zufallsexperimenten**: deren Ergebnisse sind unter "gleichen Versuchsbedingungen" verschieden.

Beispiele:

- Kartenziehen, Würfeln, Roulette
- Simulation
- komplexe Phänomene (zumindest approximativ): Börse, Data-Mining, Genetik, Wetter

Ergebnisse von Zufallsexperimenten werden in **Ereignisse** zusammengefasst.

- *Ereignisraum* (Grundraum) Ω: Menge aller möglichen Ergebnisse des Zufallsexperiments
- *Elementarereignisse* ω: Elemente von Ω, also die möglichen Ergebnisse des Zufallsexperiments
- *Ereignis*: Teilmenge von Ω
- Operationen der Mengenlehre haben natürliche Interpretation in der Sprache der Ereignisse
  Bsp: $A \cap B$ bedeutet "A **und** B"; $A \cup B$ bedeutet "A **oder** B" ("oder" zu verstehen als "und/oder")

Das Vorgehen der Stochastik zur Lösung eines Problems kann in drei Schritte unterteilt werden:

1. Man bestimmt die Wahrscheinlichkeiten gewisser Ereignisse A_i. Dabei sind Expertenwissen, Daten und Plausibilitäten wichtig.
2. Man berechnet aus den Wahrscheinlichkeiten $P[A_i]$ die Wahrscheinlichkeiten von gewissen anderen Ereignissen B_j gemäss den Gesetzen der Wahrscheinlichkeitstheorie (oft vereinfachend unter Unabhängigkeitsannahme).

3. Man interpretiert die Wahrscheinlichkeiten $P[B_j]$ im Hinblick auf die Problemstellung.

Das **Bestimmen von Wahrscheinlichkeiten** (siehe Schritt 1) wird oft konkreter formalisiert.

Beispiel: Kombinatorische Abzählung bei endlichem Ω. Die Wahrscheinlichkeit von einem Ereignis A ist gegeben durch

$$P[A] = |A|/|\Omega| \quad (= \text{Anzahl günstige Fälle/Anzahl mögliche Fälle}).$$

Dies ist das **Laplace-Modell**. Dem Laplace-Modell liegt die **uniforme Verteilung** von Elementarereignissen ω zugrunde:

$$P[\omega] = 1/|\Omega| \text{ für alle } \omega \in \Omega.$$

Andere Wahrscheinlichkeitsverteilungen werden mit Hilfe des Konzepts von **Zufallsvariablen** (siehe Kapitel 2) eingeführt. Es sei aber bereits hier festgehalten: die Stochastik geht weit über das Laplace-Modell hinaus. Für viele Anwendungen ist das Laplace-Modell ungeeignet.

1.1 Rechenregeln für Wahrscheinlichkeiten

Die drei grundlegenden Regeln (Axiome) sind

1. $P[A] \geq 0$: Wahrscheinlichkeiten sind immer nicht-negativ.
2. $P[\Omega] = 1$: sicheres Ereignis Ω hat Wahrscheinlichkeit eins.
3. $P[A \cup B] = P[A] + P[B]$ für alle Ereignisse $A,\ B$, die sich gegenseitig ausschliessen (d.h. $A \cap B = \emptyset$).

Weitere Regeln werden daraus abgeleitet, z.B.

$$\begin{aligned} &P[A^c] = 1 - P[A], \\ &P[A \cup B] = P[A] + P[B] - P[A \cap B], \\ &P[A_1 \cup \ldots \cup A_n] \leq P[A_1] + \ldots + P[A_n]. \end{aligned}$$

1.2 Unabhängigkeit von Ereignissen

Wenn zwischen zwei Ereignissen A und B kein kausaler Zusammenhang besteht (d.h. es gibt keine gemeinsamen Ursachen oder Ausschliessungen), dann soll gelten

$$P[A \cap B] = P[A]P[B]. \tag{1.1}$$

Ereignisse A und B heissen (stochastisch) unabhängig, falls (1.1) gilt. Bei n Ereignissen $A_1, \ldots A_n$ soll Unabhängigkeit bedeuten, dass für jedes $k \leq n$ und jedes $1 \leq i_1 < \ldots < i_k \leq n$ gilt

$$P[A_{i_1} \cap \ldots \cap A_{i_k}] = P[A_{i_1}] \cdots P[A_{i_k}].$$

1.3 Interpretation von Wahrscheinlichkeiten

Die beiden wichtigsten Interpretationen sind die "Idealisierung der relativen Häufigkeiten bei vielen unabhängigen Wiederholungen" (frequentistisch) und das (subjektive) "Mass für den Glauben, dass ein Ereignis eintreten wird" (Bayes'sch).

Frequentistisch: betrachte die relative Häufigkeit eines Ereignis A in n unabhängigen Wiederholungen desselben Experiments,

$$f_n[A] = (\text{Anzahl Auftreten des Ereignis } A \text{ in } n \text{ Experimenten})/n.$$

Dieses Mass $f_n[\cdot]$ basiert auf **Daten** oder **Beobachtungen**. Falls n gross ist, so gilt

$$f_n[A] \to P[A] \ (n \to \infty),$$

wobei $P[A]$ ein Mass in einem **Modell** ist (keine Experimente oder Daten). Die Statistik befasst sich grösstenteils wie man von Daten auf ein Modell (induktiv) schliessen kann, siehe später.

2 Zufallsvariable und Wahrscheinlichkeitsverteilung

Ergebnisse eines Versuchs sind oft Zahlen (Messungen).

2.1 Definition einer Zufallsvariablen

Eine **Zufallsvariable** X ist ein Zufallsexperiment mit möglichen Werten in $\mathbb{R}$, bzw. in einer Teilmenge von $\mathbb{R}$, z.B. $\mathbb{N}_0 = \{0, 1, \ldots\}$. Deren Wert ist im voraus nicht bekannt ist, sondern hängt vom Ergebnis eines Zufallsexperiments ab. Mathematisch heisst das:

$$X : \Omega \to \mathbb{R}, \ \omega \mapsto X(w).$$

Wenn das Ergebnis ω herauskommt, nimmt die Zufallsvariable den Wert $X(\omega)$ an.

Beispiel: Wert einer zufällig gezogenen Jasskarte.
$\Omega = \{Jasskarten\}$; ein $\omega \in \Omega$ ist z.B. ein Schilten-As;
Zufallsvariable

$$\begin{aligned} X: \quad & \text{As irgendeiner Farbe} \mapsto 11 \\ & \text{König irgendeiner Farbe} \mapsto 4 \\ & \ldots \\ & \text{"Brettchen" irgendeiner Farbe} \mapsto 0. \end{aligned}$$

2.2 Wahrscheinlichkeitsverteilung auf $\mathbb{R}$

Eine Zufallsvariable X legt eine Wahrscheinlichkeit Q auf $\mathbb{R}$ fest, die sogenannte **Verteilung** von X:

$$Q[B] = P[\{\omega; X(\omega) \in B\}] = P[X \in B] \quad (B \subseteq \mathbb{R}).$$

(Wir könnten auch direkt den Grundraum $\Omega = \mathbb{R}$ wählen und dann eine Wahrscheinlichkeit $Q[\cdot]$ für geeignete Ereignisse $B \subseteq \mathbb{R}$ definieren).

Beispiel: Wert einer zufällig gezogenen Jasskarte (Fortsetzung).
In obigem Beispiel ist bespielsweise

$$Q[11] = P[\text{As irgendeiner Farbe}] = 4/36.$$

Die **kumulative Verteilungsfunktion** ist definiert als

$$F(b) = P[X \leq b] = Q[(-\infty, b]].$$

Sie enthält dieselbe Information wie die Verteilung $Q[\cdot]$, ist aber einfacher darzustellen.

2.3 Diskrete und stetige Zufallsvariablen

Eine Zufallsvariable X heisst **diskret**, falls die Menge W der möglichen Werte von X endlich oder abzählbar ist. Zum Beispiel $W = \{0, 1, 2, \ldots, 100\}$ oder $W = \mathbb{N}_0 = \{0, 1, 2, \ldots\}$. Die Verteilung einer diskreten Zufallsvariablen ist festgelegt durch die Angabe der sogenannten **Wahrscheinlichkeitsfunktion**:

$$p(x_i) = P[X = x_i] \text{ für alle } x_i \in W.$$

Offensichtlich ist die kumulative Verteilungsfunktion eine Treppenfunktion mit Sprüngen an den Stellen x_i und Sprunghöhen $p(x_i)$; also nicht stetig. Ferner gilt $Q[B] = \sum_{x_i \in B} p(x_i)$.

Eine Zufallsvariable X heisst **stetig** falls die Menge der möglichen Werte W ein Intervall enthält. Zum Beispiel $W = [0, 1]$ oder $W = \mathbb{R}$.

2.4 Erwartungswert und Varianz

Eine Verteilung einer Zufallsvariablen X kann durch mindestens zwei Kennzahlen zusammengefasst werden, eine für die Lage und eine für die Streuung. Die gebräuchlichsten Kennzahlen sind der **Erwartungswert** $E(X) = \mu_X$ für die Lage und die **Standardabweichung** σ_X für die Streuung.

Der Erwartungswert einer transformierten diskreten Zufallsvariablen $Y = g(X)$ ist definiert durch

$$E[g(X)] = \sum_{x_i \in W} g(x_i) p(x_i).$$

Für $g(x) = x$ erhält man die Definition von $E[X]$.

Die Standardabweichung ist die Wurzel aus der **Varianz**, $\sigma_X = \sqrt{Var(X)}$, wobei

$$Var(X) = \sigma_X^2 = E[(X - E(X))^2] = \sum_{x_i \in W} (x_i - \mu_X)^2 p(x_i).$$

(Die letzte Gleichheitsrelation oben gilt nur für den Fall einer diskreten Zufallsvariablen).

Die folgenden **Rechenregeln** erweisen sich oft als nützlich:

$$\begin{aligned} &E[a + bX] = a + bE[X], \ a, b \in \mathbb{R}, \\ &Var(X) = E[X^2] - (E[X])^2, \\ &Var(a + bX) = b^2 Var(X). \end{aligned}$$

In der frequentistischen Interpretation ist der Erwartungswert eine Idealisierung des arithmetischen Mittels der Werte einer Zufallsvariablen bei vielen Wiederholungen. Also: $E[X]$ ist eine Kennzahl im Modell der Wahrscheinlichkeitstheorie.

2.5 Die wichtigsten diskreten Verteilungen

Die **Binomialverteilung** ist die Verteilung der Anzahl "Erfolge" bei n unabhängigen Wiederholungen eines Experiments mit Erfolgswahrscheinlichkeit p. Hier ist $W = \{0, 1, \ldots, n\}$, $p(x) = \binom{n}{x} p^x (1-p)^{n-x}$, $E[X] = np$ und $\sigma_X = \sqrt{np(1-p)}$.

Die **Poissonverteilung** ist eine Approximation der Binomialverteilung für grosses n und kleines p, mit $np = \lambda$. Hier ist $W = \{0, 1, \ldots\}$, $p(x) = e^{-\lambda}\lambda^x/x!$, $E[X] = \lambda$ und $\sigma_X = \sqrt{\lambda}$. Die Anzahl Ausfälle einer Komponente oder eines Systems in einem Intervall der Länge t ist oft in erster Näherung Poisson-verteilt mit Parameter λt.

Die **geometrische Verteilung** ist die Verteilung der Anzahl Wiederholungen bis ein Ereignis mit Wahrscheinlichhkeit p eintritt. Hier ist $W = \{1, 2, \ldots\}$, $p(x) = p(1-p)^{x-1}$, $E[X] = 1/p$ und $\sigma_X = \sqrt{1-p}/p$.

3 Stetige Wahrscheinlichkeitsverteilung

Bei einer stetigen Zufallsvariablen X ist $P[X = x] = 0$ für jedes feste x. Wir betrachten nur Fälle, wo $P[x \leq X \leq x + h]$ für kleine h ungefähr proportional zu h ist. Die Proportionalitätskonstante heisst die **Dichte** f von X.

3.1 Wahrscheinlichkeitsdichte

Die Dichte von einer stetigen Verteilung P ist definiert als

$$f(x) = \lim_{h \downarrow 0} \frac{P[x \leq X \leq x + h]}{h}.$$

Zwischen der Dichte f und der kumulativen Verteilungsfunktion F bestehen die folgenden Beziehungen,

$$f(x) = F'(x), \quad F(x) = \int_{-\infty}^{x} f(u)du.$$

Erwartungswert und Varianz berechnen sich gemäss

$$E[X] = \mu_X = \int_{-\infty}^{\infty} x f(x)dx, \quad Var(X) = \sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x)dx.$$

und es gelten die gleichen Rechenregeln wie im diskreten Fall (Kapitel 2.4). Andere Kennzahlen beruhen auf den **Quantilen** $q(\alpha)$ $(0 < \alpha < 1)$,

$$P[X \leq q(\alpha)] = \alpha \Leftrightarrow F(q(\alpha)) = \alpha \Leftrightarrow q(\alpha) = F^{-1}(\alpha).$$

Das Quantil zu $\alpha = 1/2$ heisst der **Median**.

3.2 Die wichtigsten stetigen Verteilungen

Die **uniforme Verteilung** tritt auf bei Rundungsfehlern und als Formalisierung der völligen "Ignoranz". Hier ist $W = [a, b]$, $f(x) = 1/(b-a)$ für $a \leq x \leq b$, $E[X] = (a+b)/2$ und $\sigma_X = (b-a)/\sqrt{12}$.

Die **Exponentialverteilung** ist das einfachste Modell für Wartezeiten auf Ausfälle und eine stetige Version der geometrischen Verteilung. Hier ist $W = [0, \infty)$, $f(x) = \lambda e^{-\lambda x}$ für $x > 0$, $F(x) = 1 - e^{-\lambda x}$ und $E[X] = \sigma_X = 1/\lambda$. Wenn die Zeiten zwischen den Ausfällen eines Systems Exponential(λ)-verteilt sind, dann ist die Anzahl Ausfälle in einem Intervall der Länge t Poisson(λt)-verteilt.

Die **Normal-** oder **Gauss-Verteilung** ist die häufigste Verteilung für Messwerte. Hier ist $W = \mathbb{R}$, $f(x) = 1/(\sigma\sqrt{2\pi}) \exp(-\frac{(x-\mu)^2}{2\sigma^2})$, $E[X] = \mu$ und $\sigma_X = \sigma$. Die Verteilungsfunktion F ist nicht geschlossen darstellbar, aber es gilt $F(x) = \Phi((x-\mu)/\sigma)$ und Φ ist tabelliert.

3.3 Transformationen

Bei stetigen Verteilungen spielen Transformationen $Y = g(X)$ eine wichtige Rolle. Falls g **linear** ist: $g(x) = a + bx$ mit $b > 0$, dann gilt $E[Y] = a + bE[X]$, $\sigma_Y = b\sigma_X$, $F_Y(x) = F_X((x-a)/b)$ und $f_Y(x) = f_X((x-a)/b)/b$. Durch Skalenänderungen kann man also alle Exponentialverteilungen ineinander überführen, und ebenso durch lineare Transformationen alle Normalverteilungen ineinander.

Für beliebiges g gilt $E[Y] = E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$.

Wenn $X \sim \mathcal{N}(\mu, \sigma^2)$ normalverteilt ist, dann heisst $Y = e^X$ **lognormal-verteilt**. Es gilt z.B. $E(Y) = \exp(\mu + \sigma^2/2)$.

Wenn U uniform auf $[0, 1]$ verteilt ist und F eine beliebige kumulative Verteilungsfunktion, dann ist die Verteilungsfunktion von $X = F^{-1}(U)$ gleich F. Dies ist ein wichtiges Faktum um Verteilungen, respektive Realisierungen von Zufallsvariablen, zu **simulieren**:

1. Erzeuge Realisation u von uniform verteilter Zufallsvariable $U \sim \text{Unif}([0, 1])$. Dies wird mittels einem "Standard-Paket" gemacht.
2. Berechne $x = F^{-1}(u)$. Gemäss obigem Faktum ist dann x eine Realisation einer Zufallsvariablen X mit kumulativer Verteilungsfunktion F.

(Diese Methode ist nicht immer rechentechnisch effizient).

4 Mehrere Zufallsvariablen und Funktionen davon

Das Ziel ist hier, Genaueres über den Unterschied von $P[A]$ und der relativen Häufigkeit $f_n[A]$ von A, respektive von $E[X]$ und dem arithmetischen Mittel, bei n Wiederholungen zu sagen.

4.1 Die i.i.d. Annahme

Dabei müssen wir präzisieren was eine "Wiederholung" ist. Die n-fache Wiederholung eines Zufallsexperimentes ist selber wieder ein Zufallsexperiment. Wenn A ein Ereignis im ursprünglichen Experiment ist, bezeichnen wir mit A_i das Ereignis "A tritt bei der i-ten Wiederholung ein". Dann ist es sinnvoll, folgendes anzunehmen:

- $A_1, \ldots A_n$ sind unabhängig: Unabhängigkeit der Ereignisse

- $P[A_1] = \ldots = P[A_n] = P[A]$: gleiche Wahrscheinlichkeiten

Ebenso, wenn X die ursprüngliche Zufallsvariable ist, dann soll X_i die Zufallsvariable der i-ten Wiederholung bezeichnen. Die **i.i.d.** Annahme verlangt folgendes:

- $X_1, \ldots, X_n$ sind **unabhängig**
- alle X_i haben **dieselbe Verteilung**

Die Abkürzung "i.i.d." kommt vom Englischen: **i**ndependent and **i**dentically **d**istributed.

Unabhängigkeit von Zufallsvariablen heisst, dass zum Beispiel $P[X_i \in A \text{ und } X_j \in B] = P[X_i \in A]P[X_j \in B]$ für alle $i \neq j$ und für alle $A \subseteq \mathbb{R}$, $B \subseteq \mathbb{R}$, und analog für Trippel etc.

Die i.i.d. Annahme ist ein "Postulat", welches in der Praxis in vielen Fällen vernünftig scheint. Die Annahme bringt erhebliche Vereinfachungen um mit mehreren Zufallsvariablen zu rechnen.

4.2 Funktionen von Zufallsvariablen

Ausgehend von $X_1, \ldots, X_n$ kann man neue Zufallsvariablen $Y = g(X_1, \ldots, X_n)$ bilden. Hier betrachten wir die wichtigen Spezialfälle Summe $S_n = X_1 + \ldots + X_n$ und arithmetisches Mittel $\overline{X}_n = S_n/n$. Wir nehmen stets an, dass $X_1, \ldots X_n$ i.i.d. sind.

Wenn $X_i = 1$ falls ein bestimmtes Ereignis bei der i-ten Wiederholung eintritt und $X_i = 0$ sonst, dann ist $\overline{X}_n$ nichts anderes als die relative Häufigkeit dieses Ereignisses. Die Verteilung von S_n ist im allgemeinen schwierig exakt zu bestimmen, mit den folgenden Ausnahmen:

1. Wenn $X_i \in \{0, 1\}$ wie oben, dann ist $S_n \sim$ Binomial(n, p) mit $p = P[X_i = 1]$.
2. Wenn $X_i \sim$ Poisson(λ), dann ist $S_n \sim$ Poisson$(n\lambda)$.
3. Wenn $X_i \sim \mathcal{N}(\mu, \sigma^2)$, dann ist $S_n \sim \mathcal{N}(n\mu, n\sigma^2)$.

Einfacher sind die Berechnungen von Erwartungswert, Varianz und Standardabweichung. Allgemein gilt:

$$\begin{array}{lll} E[S_n] = nE[X_i], & Var(S_n) = nVar(X_i), & \sigma_{S_n} = \sqrt{n}\sigma_{X_i}, \\ E[\overline{X}_n] = E[X_i], & Var(\overline{X}_n) = Var(X_i)/n, & \sigma_{\overline{X}_n} = \sigma_{X_i}/\sqrt{n}. \end{array}$$

Die Streuung der Summe wächst also, aber langsamer als die Anzahl Beobachtungen, während die Streuung des arithmetischen Mittels abnimmt, aber ebenfalls langsamer als die Anzahl Beobachtungen. Um die Genauigkeit des arithmetischen Mittels zu verdoppeln (d.h die Standardabweichung zu halbieren), braucht man viermal soviele Beobachtungen. Die zufälligen Abweichungen von $\overline{X}_n$ zum Erwartungswert $E[X]$ kompensieren sich in dem Sinne, dass $\sigma_{\overline{X}_n}$ abnimmt mit der Ordnung $1/\sqrt{n}$ wenn n wächst.

4.3 Das Gesetz der Grossen Zahlen und der Zentrale Grenzwertsatz

Von den obigen Formeln über Erwartungswert und Varianz wissen wir, dass:

- $E[\overline{X}_n] = E[X_i]$: das heisst $\overline{X}_n$ hat denselben Erwartungswert wie die einzelnen Variablen X_i.
- $\mathrm{Var}(\overline{X}_n) \to 0$ $(n \to \infty)$: das heisst, $\overline{X}_n$ besitzt keine Variabilität mehr im Limes.

Diese beiden Punkte implizieren den folgenden Satz.

Gesetz der Grossen Zahlen (GGZ)
Seien $X_1, \ldots, X_n$ i.i.d. mit Erwartungswert μ. Dann

$$\overline{X}_n \to \mu \ (n \to \infty).$$

Als Spezialfall davon gilt:

$$f_n[A] \to P[A] \ (n \to \infty).$$

(Der Begriff der Konvergenz muss für Zufallsvariablen geeignet definiert werden).

Zur Berechnung der genäherten **Verteilung** von S_n und $\overline{X}_n$ (dies ist ein bedeutend präziseres Resultat als das GGZ) stützt man sich auf den folgenden berühmten Satz.

Zentraler Grenzwertsatz (ZGS)
Seien $X_1, \ldots, X_n$ i.i.d. mit Erwartungswert μ und Varianz σ^2. Dann,

$$S_n \approx \mathcal{N}(n\mu, n\sigma^2) \text{ für grosse } n,$$
$$\overline{X}_n \approx \mathcal{N}(\mu, \sigma^2/n) \text{ für grosse } n.$$

Wie gut diese Approximationen für ein gegebenes n sind, hängt von den Eigenschaften der Verteilung der X_i ab. Mit der sogenannten **Chebychev-Ungleichung**

$$P[|\overline{X}_n - \mu| > c] \leq \sigma^2/(nc^2)$$

ist man stets auf der sicheren Seite. Dafür ist diese aber meistens ziemlich grob.

Immer wenn eine Zufallsvariable als eine Summe von vielen kleinen Effekten aufgefasst werden kann, ist sie wegen des Zentralen Grenzwertsatzes in erster Näherung normalverteilt. Das wichtigste Beispiel dafür sind Messfehler. Wenn sich die Effekte eher multiplizieren als addieren, kommt man zur lognormal-Verteilung (Beispiel Teilchengrössen).

5 Gemeinsame und bedingte Wahrscheinlichkeiten

Oft besteht ein Zufallsexperiment aus verschiedenen Stufen, und man erfährt das Resultat auch entsprechend diesen Stufen. Im einfachsten Fall erfährt man in der ersten Stufe, ob ein bestimmtes Ereignis B eingetreten ist oder nicht, und in der zweiten Stufe erfährt man, welches Ergebnis ω eingetreten ist.

5.1 Bedingte Wahrscheinlichkeit

Im allgemeinen wird die Information aus der ersten Stufe die Unsicherheit über die zweite Stufe verändern. Diese modifizierte Unsicherheit wird gemessen durch die bedingte Wahrscheinlichkeit von A gegeben B bzw. B^c.

Die bedingte Wahrscheinlichkeit von A gegeben B ist definiert als

$$P[A|B] = P[A \cap B]/P[B] \quad \text{(analog für } P[A|B^c]).$$

Dass diese Definition sinnvoll ist, kann man anhand der Entsprechung von Wahrscheinlichkeiten und relativen Häufigkeiten sehen. Insbesondere gilt:

$$P[A|B] = P[A|B^c] = P[A], \text{ falls } A \text{ und } B \text{ unabhängig sind.}$$

5.2 Satz der totalen Wahrscheinlichkeit und Satz von Bayes

Die obige Definition kann man aber auch als $P[A \cap B] = P[A|B]P[B]$ lesen, d.h. $P[A \cap B]$ ist bestimmt durch $P[A|B]$ und $P[B]$. In vielen Anwendungen wird dieser Weg beschritten. Man legt die Wahrscheinlichkeiten für die erste Stufe $P[B]$ und die bedingten Wahrscheinlichkeiten $P[A|B]$ und $P[A|B^c]$ für die zweite Stufe gegeben die erste fest (aufgrund von Daten, Plausibilität und subjektiven Einschätzungen). Dann lassen sich die übrigen Wahrscheinlichkeiten berechnen. Es gilt zum Beispiel der folgende Satz:

Satz der totalen Wahrscheinlichkeit (I)

$$P[A] = P[A \cap B] + P[A \cap B^c] = P[A|B]P[B] + P[A|B^c]P[B^c].$$

Dieses Vorgehen wird besonders anschaulich, wenn man das Experiment als Baum darstellt. Wenn man dagegen von den Wahrscheinlichkeiten der Durchschnitte ausgeht, wählt man besser eine Matrixdarstellung.

Wenn die einzelnen Stufen komplizierter sind, geht alles analog. Betrachte den Fall mit k Ereignissen auf der ersten Stufe $B_1, \ldots B_k$, wobei $B_i \cap B_j = \emptyset$ falls $i \neq j$ und $B_1 \cup \ldots \cup B_k = \Omega$.

Satz der totalen Wahrscheinlichkeit (II)

$$P[A] = \sum_{i=1}^{k} P[A|B_i]P[B_i].$$

In manchen Situationen erhält man die Information über die verschiedenen Stufen aber nicht in der ursprünglichen Reihenfolge, d.h. man kennt zuerst das Ergebnis der zweiten Stufe, weiss also z.B. dass A eingetreten ist. In einem solchen Fall will man die bedingten Wahrscheinlichkeiten der ersten Stufe gegeben die zweite Stufe $P[B_i|A]$ berechnen. Das Ergebnis liefert der folgende Satz:

Satz von Bayes

$$P[B_i|A] = \frac{P[A|B_i]P[B_i]}{P[A|B_1]P[B_1] + \cdots + P[A|B_k]P[B_k]}.$$

Oft ist das numerische Resultat einer solchen Berechnung stark verschieden von dem, was man naiverweise erwartet. Der Satz von Bayes ist vor allem in der subjektiven Wahrscheinlichkeitstheorie sehr wichtig: Wenn man für die verschiedenen Möglichkeiten $B_1, \dots B_k$ subjektive Wahrscheinlichkeiten festlegt und danach erfährt, dass A eingetreten ist, dann muss man die subjektiven Wahrscheinlichkeiten gemäss diesem Satz modifizieren.

5.3 Gemeinsame und bedingte diskrete Verteilungen

Die beiden, obig beschriebenen Stufen können auch durch Zufallsvariablen X und Y gegeben sein. Dann nennt man $P[X = x_i]$ und $P[Y = y_j]$ die **Randverteilungen**, $P[X = x_i, Y = y_j]$ die **gemeinsame Verteilung** (das Komma steht für "und") und $P[Y = y_j | X = x_i]$ die **bedingte Verteilung**.

Bei mehr als zwei Stufen geht alles analog. Die Bäume werden einfach länger und die Matrizen werden zu Feldern in höheren Dimensionen. Das Ganze wird aber sehr rasch unübersichtlich, und es gibt sehr viele Wahrscheinlichkeiten, die man zu Beginn festlegen muss. Eine wesentliche Vereinfachung erhält man, wenn man annimmt, dass jede Stufe zwar von der unmittelbar vorangehenden, aber nicht von den weiter zurückliegenden Stufen abhängt. Das führt auf die sogenannten **Markovketten**, deren Verhalten durch eine Startverteilung und eine Übergangsmatrix gegeben ist.

6 Gemeinsame und bedingte stetige Verteilungen

Bei zwei oder mehreren stetigen Zufallsvariablen kann die gemeinsame und bedingte Verteilung nicht mehr mit Bäumen oder Matrizen dargestellt werden wie in Kapitel 5.

6.1 Gemeinsame Dichte

Die gemeinsame Dichte $f_{X,Y}(\cdot, \cdot)$ von zwei stetigen Zufallsvariablen X und Y ist gegeben, in "Ingenieurnotation", durch

$$P[x \le X \le x + dx,\ y \le Y \le y + dy] = f_{X,Y}(x, y)dxdy.$$

(Die Darstellung als Ableitung einer geeigneten kumulativen Verteilungsfunktion ist nicht sehr instruktiv).

Daraus kann man allgemein Wahrscheinlichkeiten durch Integration berechnen:

$$P[(X, Y) \in A] = \int\int_A f_{X,Y}(x, y)dxdy \quad (A \subseteq \mathbb{R}^2).$$

6.2 Randdichte und bedingte Dichte

Aus der gemeinsamen Dichte erhält man insbesondere die Randdichte von X, bzw. Y

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy, \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dx.$$

Für die bedingte Verteilung von Y gegeben $X = x$ wird die bedingte Dichte benützt:

$$f_Y(y|X = x) = f_{X,Y}(x, y)/f_X(x).$$

Aus den obigen Definitionen ist klar, dass alle wahrscheinlichkeitstheoretischen Aspekte von 2 Zufallsvariablen X und Y durch deren gemeinsame Dichte $f_{X,Y}(x,y)$ vollständig bestimmt sind.

Die Unabhängigkeit von X und Y kann charakterisiert (definiert) werden als

$$f_{X,Y}(x,y) = f_X(x) f_Y(y) \text{ für alle } x,y \in \mathbb{R}^2. \tag{6.2}$$

In diesem Fall genügt das Konzept von 1-dimensionalen Dichten: die gemeinsame Dichte kann dann sehr einfach mittels Multiplikation berechnet werden.

6.3 Erwartungswert bei mehreren Zufallsvariablen

Der Erwartungswert macht nur Sinn für eine $\mathbb{R}$-wertige Grösse (oder Teilmenge von $\mathbb{R}$).

Den Erwartungswert einer transformierten Zufallsvariable $Z = g(X,Y)$ mit $g : \mathbb{R}^2 \to \$$ können wir berechnen als

$$E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y) dx dy.$$

(Im diskreten Fall lautet die entsprechende Formel:
$E[g(X,Y)] = \sum_i \sum_j g(x_i, y_j) P[X = x_i, Y = y_j]$).

Der Erwartungswert von der einen Zufallsvariablen Y gegen $X = x$ ist gegeben durch

$$E[Y|X = x] = \int_{-\infty}^{\infty} y f_Y(y|X = x) dy.$$

6.4 Kovarianz und Korrelation

Da die gemeinsame Verteilung von abhängigen Zufallsvariablen i.A. kompliziert ist, begnügt man sich oft mit einer **vereinfachenden** Kennzahl zur Beschreibung der Abhängigkeit. Die Kovarianz und Korrelation zwischen X und Y sind wie folgt definiert:

$$Cov(X,Y) = E[(X - \mu_X)(Y - \mu_Y)] \quad \text{(Kovarianz)}$$
$$Corr(X,Y) = \rho_{XY} = Cov(X,Y)/(\sigma_X \sigma_Y) \quad \text{(Korrelation)}.$$

Es gelten die folgenden **Rechenregeln**:
$E[X+Y] = E[X] + E[Y]$ für beliebige, auch abhängige Zufallsvariablen;
$Cov(X,Y) = E[XY] - E[X]E[Y]$;
$Cov(X,Y) = 0$ falls X und Y unabhängig sind;
$Cov(a+bX, c+dY) = bdCov(X,Y)$, $Corr(a+bX, c+dY) = sign(b)sign(d)Corr(X,Y)$;
$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$.

Die Korrelation misst Stärke und Richtung der **linearen Abhängigkeit** zwischen X und Y. Es gilt

$$Corr(X,Y) = +1 \text{ genau dann wenn } Y = a + bX \text{ für ein } a \in \mathbb{R} \text{ und ein } b > 0,$$
$$Corr(X,Y) = -1 \text{ genau dann wenn } Y = a + bX \text{ für ein } a \in \mathbb{R} \text{ und ein } b < 0.$$

Überdies gilt:

$$X \text{ und } Y \text{ unabhängig} \quad \Longrightarrow \quad Corr(X,Y) = 0. \tag{6.3}$$

Die Umkehrung gilt i.A. nicht. Ein Spezialfall, wo auch die Umkehrung gilt, wird in Kapitel 6.6 diskutiert.

6.5 Lineare Prognose

Bei der linearen Prognose von Y gestützt auf X macht man den Ansatz $\hat{Y} = a + bX$ und bestimmt die Koeffizienten so, dass der mittlere quadratische Prognosefehler $E[(Y - \hat{Y})^2]$ minimal wird. Man erhält

$$\hat{Y} = \mu_Y + \frac{Cov(X,Y)}{Var(X)}(X - \mu_X), \quad E[(Y - \hat{Y})^2] = (1 - \rho_{XY}^2)Var(Y).$$

6.6 Zwei-dimensionale Normalverteilung

Die wichtigste zweidimensionale Verteilung ist die Normalverteilung mit Erwartungswerten (μ_X, μ_Y) und Kovarianzmatrix Σ wobei $\Sigma_{11} = Var(X)$, $\Sigma_{22} = Var(Y)$ und $\Sigma_{12} = \Sigma_{21} = Cov(X,Y)$. Sie hat die Dichte

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{det(\Sigma)}} \exp\left(-\frac{1}{2}(x - \mu_X, y - \mu_Y)\Sigma^{-1}\begin{pmatrix} x - \mu_X \\ y - \mu_Y \end{pmatrix}\right).$$

Wir sehen von dieser Formel: wenn $Cov(X,Y) = 0$ wird Σ eine Diagonalmatrix und man kann nachrechnen dass dann die Bedingung (6.2) gilt. Das heisst: im Falle der zweidimensionalen Normalverteilung gilt auch die Umkehrung von (6.3). Zudem: die Rand- und bedingten Verteilungen sind wieder (1-dimensional) normal.

6.7 Mehr als 2 Zufallsvariablen

Alle diese Begriffe und Definitionen lassen sich natürlich auf mehr als zwei Zufallsvariablen verallgemeinern. Die Formeln sehen im wesentlichen gleich aus, vor allem wenn man die Sprache der Linearen Algebra verwendet.

Ausblick: Wenn man eine dynamische Grösse während eines Zeitintervalls misst, erhält man einen stochastischen Prozess $\{X(t);\ t \in [a,b]\}$. Die linearen Abhängigkeiten zwischen den Werten zu verschiedenen Zeitpunkten werden dann durch die sogenannte **Autokovarianzfunktion** beschrieben.

7 Deskriptive Statistik

In der *Statistik* will man aus beobachteten Daten Schlüsse ziehen. Meist nimmt man an, dass die Daten Realisierungen von Zufallsvariablen sind (siehe Kap. 8.1), deren Verteilung man aufgrund der Daten bestimmen möchte. Als ersten Schritt geht es aber zunächst einmal darum, die vorhandenen Daten übersichtlich darzustellen und zusammenzufassen. Dies ist das Thema der *beschreibenden* oder *deskriptiven Statistik*.

7.1 Kennzahlen

Für die numerische Zusammenfassung von Daten gibt es diverse Kennzahlen.

Das *arithmetische Mittel* ist

$$\overline{x} = \frac{1}{n}(x_1 + \ldots + x_n)$$

als Kennzahl für die Lage der Daten. Die *empirische Standardabweichung* ist die Wurzel aus der *empirischen Varianz*

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

als Kennzahl für die Streuung der Daten. (Eine Begründung für den Nenner $n-1$ statt n folgt später).

Um weitere Kennzahlen zu definieren, führen wir die geordneten Werte

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

Das *empirische α-Quantil* $(0 < \alpha < 1)$ ist

$$x_{(k)},\ k \text{ die kleinste ganze Zahl} > \alpha n.$$

Wenn αn eine ganze Zahl ist, nimmt man $\frac{1}{2}(x_{(\alpha n)} + x_{(\alpha n+1)})$. Der *empirische Median* ist das 50%-Quantil und ist eine Kennzahl für die Lage. Die *Quartilsdifferenz* ist das empirische 75%-Quantil minus empirisches 25%-Quantil und ist eine Kennzahl für die Streuung.

Einen ganz anderen Aspekt erfasst man, wenn man die Werte gegen den Beobachtungszeitpunkt aufträgt. Damit kann man Trends und andere Arten von systematischen Veränderungen in der Zeit erkennen.

7.2 Histogramm und Boxplot

Wenn man n Werte $x_1, \ldots, x_n$ einer Variablen hat, dann gibt es als grafische Darstellungen das *Histogramm*, den *Boxplot* und die empirische *kumulative Verteilungsfunktion*.

Beim Histogramm bilden wir Klassen $(c_{k-1}, c_k]$ und berechnen die Häufigkeiten h_k= Anzahl Werte in diesem Intervall. Dann trägt man über den Klassen Balken auf, deren Höhe *proportional* ist zu $h_k/(c_k - c_{k-1})$ ist.

Beim Boxplot hat man ein Rechteck, das vom 25%- und vom 75%-Quantil begrenzt ist, und Linien, die von diesem Rechteck bis zum kleinsten- bzw. grössten "normalen" Wert gehen (per Definition ist ein normaler Wert höchstens 1.5 mal die Quartilsdifferenz von einem der beiden Quartile entfernt). Zusätzlich gibt man noch Ausreisser durch Sterne und den Median durch einen Strich an. Der Boxplot ist vor allem dann geeignet, wenn man die Verteilungen einer Variablen in verschiedenen Gruppen (die im allgemeinen verschiedenen Versuchsbedingungen entsprechen) vergleichen will.

Die empirische Verteilungsfunktion ist eine Treppenfunktion, die an den Stellen $x_{(i)}$ von $(i-1)/n$ auf i/n springt. Für eine glattere Version verbindet man die Punkte $(x_{(i)}, (i-0.5)/n)$ durch Strecken.

7.3 Normal- und QQ-Plot

Der Normal- und QQ-Plot ("Quantil-Quantil-Plot") sind oft viel geeignetere grafische Mittel als die empirische kumulative Verteilungsfunktion.

Die empirischen Quantile für $\alpha_k = (k-0.5)/n$ sind gerade die geordneten Beobachtungen $x_{(k)}$ $(x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)})$. Der QQ-Plot trägt die Punkte $F^{-1}(\alpha_k)$ (die theoretischen Quantile einer kumulativen Verteilung F) gegen $x_{(k)}$ (die empirischen Quantile) auf.

Wir interpretieren die Daten $x_1, \ldots, x_n$ als Realisierungen von $X_1, \ldots, X_n$ i.i.d $\sim \tilde{F}$, siehe Kap. 8.1 Falls nun die wahre Verteilung $\tilde{F}$ mit der gewählten Verteilung F im QQ-Plot übereinstimmt, so liefert der QQ-Plot approximativ eine Gerade durch Null mit Steigung 45 Grad. Man kann also Abweichungen der Daten von einer gewählten Modell-Verteilung so grafisch überprüfen.

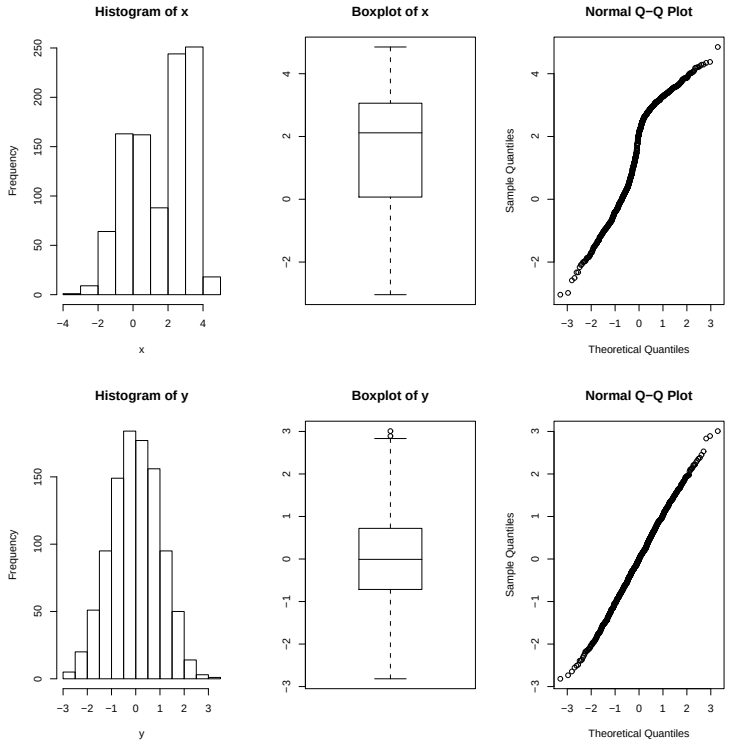

Abbildung 7.1: Histogramm, Boxplot und Normal-Plot für 2 Datensätze der Stichprobengrösse 1000.

Der Normal-Plot ist ein QQ-Plot wo die Modell-Verteilung F die Standard-Normalverteilung $\mathcal{N}(0,1)$ ist. Es gilt dann das folgende: wenn die wahre Verteilung $\tilde{F}$ eine Normalverteilung $\mathcal{N}(\mu, \sigma^2)$ ist, so liefert der Normal-Plot approximativ eine Gerade, welche jedoch im allgemeinen nicht durch Null und nicht Steigung 45 Grad hat. Das heisst: der Normal-Plot liefert eine gute Überprüfung für irgendeine Normalverteilung, auch wenn die Modell-Verteilung als Standard-Normal gewählt wird.

Im Normal- und QQ-Plot kann man insbesondere sehen, ob eine Transformation der Daten angebracht ist, oder ob es Ausreisser gibt, die man besonders behandeln sollte.

Im Grunde genommen ist der QQ-Plot bereits mehr als bloss deskriptive Statistik: es ist ein grafisches Werkzeug um eventuelle Abweichungen von einem Modell festzustellen: vergleiche mit dem Formalismus des statistischen Tests in Kap. 8.3.

8 Schliessende Statistik: Konzepte und erste Anwendungen für diskrete Zufallsvariablen

8.1 Daten als Realisierungen von Zufallsvariablen

In der schliessenden Statistik wollen wir anhand von Daten (Beobachtungen) Aussagen über ein Wahrscheinlichkeitsmodell machen. Dass man dies tun kann ist zunächst erstaunlich: man benützt die induktive Logik um probabilistische Aussagen (d.h. Aussagen, welche mit typischerweise hoher Wahrscheinlichkeit gelten) zu machen.

Grundlegend für die schliessende Statistik ist die Annahme, dass Daten Realisierungen von Zufallsvariablen sind. Das heisst: eine Beobachtung (oder "Messung") x ist entstanden in dem ein $\omega \in \Omega$ zufällig gezogen wurde, so dass die Zufallsvariable X den Wert $X(\omega) = x$ annimmt. Bei mehreren Daten geht alles analog: n Beobachtungen $x_1, \ldots, x_n$ werden aufgefasst als Realisierungen von Zufallsvariablen $X_1, \ldots, X_n$, welche die Werte $X_i = x_i$ $(i = 1, \ldots, n)$ angenommen haben.

8.2 Erste Konzepte

Wir betrachten folgende Situation. Gegeben ist eine Beobachtung x (eine Realisierung) einer Binomial(n, p)- oder einer Poisson(λ)-verteilten Zufallsvariablen X: z.B. Anzahl Ausfälle bei n Wiederholungen oder während einer Beobachtungsdauer t, wobei dann $\lambda = \mu t$ und μ die erwartete Anzahl Ausfälle pro Zeiteinheit ist. Wir möchten daraus Rückschlüsse auf den unbekannten Parameter p bzw. λ ziehen. Genauer geht es um folgende drei Fragestellungen:

- Welches ist der plausibelste Wert des unbekannten Parameters (*Punktschätzung*)?
- Ist ein bestimmter vorgegebener Parameterwert p_0, bzw. λ_0 (z.B. ein Sollwert) mit der Beobachtung verträglich (*Test*) ?
- Was ist der Bereich von plausiblen Parameterwerten (*Vertrauensintervall*) ?

Die Punktschätzer sind hier wie folgt: plausibel sind $\hat{p} = X/n$ bzw. $\hat{\lambda} = X$ bzw. $\hat{\mu} = X/t$. Die Schätzer sind also wiederum Zufallsvariablen und im allgemeinen nicht gleich dem unbekannten Wert (darum die Bezeichnung mit dem Hut). Die realisierte Schätzung ist dann $\hat{p} = x/n$ bzw. $\hat{\lambda} = x$ bzw. $\hat{\mu} = x/t$, d.h. man ersetzt X durch dessen Realisierung x. Diese realisierten Grössen können dann berechnet werden (mittels der realisierten Beobachtung).

Bemerkung: Die Notation $\hat{p}$, $\hat{\lambda}$ und $\hat{\mu}$ unterscheidet leider nicht zwischen dem Schätzer als Funktion von Zufallsvariable(n) und dessen realisierter Wert, welches eine numerische Zahl ist.

8.3 Das Testproblem

Beim Testproblem beschränken wir uns hier zur Vereinfachung der Notation auf die Binomialverteilung.

Wir legen eine *Nullhypothese* für einen Parameterwert fest

$$H_0: \quad p = p_0.$$

Man überlegt sich dann anhand der Problemstellung, welche *Alternative* geeignet ist

$$\begin{aligned} H_A: \quad & p \neq p_0 \text{ (zwei-seitig)} \\ & p > p_0 \text{ (ein-seitig nach oben)} \\ & p < p_0 \text{ (ein-seitig nach unten).} \end{aligned}$$

Wenn wir nur an Abweichungen nach oben interessiert sind, d.h. $H_A : p > p_0$, dann lehnen wir die Nullhypothese $H_0 : p = p_0$ ab, falls $x \geq c$. Das ist qualitativ betrachtet plausibel: die quantitative Wahl von c wird wie folgt gemacht. Wir nehmen einmal an, dass die Nullhypothese stimmt. Dann ist die Wahrscheinlichkeit, die Nullhypothese fälschlicherweise abzulehnen (d.h. ein *Fehler 1. Art*)

$$P_{p_0}[X \geq c] = \sum_{k=c}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k}.$$

Wir sollten also c nicht zu klein wählen. Umgekehrt möchten wir aber auch c nicht zu gross wählen, weil wir sonst zu häufig einen *Fehler 2. Art* begehen: kein Verwerfen der Nullhypothese H_0, obwohl sie falsch ist. Man schliesst einen Kompromiss, indem man das kleinste $c = c(\alpha)$ nimmt, so dass

$$P_{p_0}[X \geq c] \leq \alpha.$$

Dabei ist α eine im voraus festgelegte (kleine) Zahl, das sogenannte *Signifikanzniveau*. Obige (Un-)Gleichung besagt, dass die Wahrscheinlichkeit eines Fehlers 1. Art mit dem Signifikanzniveau α kontrolliert ist. Die Wahrscheinlichkeit für einen Fehler 2. Art ist nicht explizit kontrolliert, deswegen, weil man nur einen – und hier wählt man den schlimmeren Fehler 1. Art – direkt kontrollieren kann. Nach all diesen Überlegungen kommt man zum Rezept, dass H_0 verworfen wird, falls $x \geq c_\alpha$.

Im Fall, wo man nach Abweichungen nach unten interessiert ist, d.h. $H_A : p < p_0$, geht alles analog. Bei zwei-seitiger Alternative $H_A : p \neq p_0$, verwerfen wir die Nullhypothese $H_0 : p = p_0$, wenn $x \leq c_1$ oder $x \geq c_2$. Hier wählt man c_1 möglichst gross und c_2 möglichst klein unter den Einschränkungen dass

$$\sum_{k=0}^{c_1} \binom{n}{k} p_0^k (1-p_0)^{n-k} \leq \alpha/2, \quad \sum_{k=c_2}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k} \leq \alpha/2.$$

Offensichtlich gilt $x \geq c(\alpha)$ (für die ein-seitige Alternative $H_A : p > p_0$) genau dann, wenn der sogenannte *P-Wert*

$$P_{p_0}[X \geq x]$$

kleiner als α ist. Dieser P-Wert wird von vielen Computer-Paketen geliefert.

8.3.1 Zusammenfassung eine statistischen Tests

Die Durchführung eines statistischen Tests kann, zumindest teilweise, "rezeptartig" erfolgen.

1. Lege Nullhypothese $H_0: \ \theta = \theta_0$ fest. (θ bezeichnet hier allgemein einen Parameter in einem wahrscheinlichkeitstheoretischen Modell).

2. Anhand der Problemstellung, spezifiziere vernünftige Alternative $H_A: \ \theta \neq \theta_0$ (zweiseitig) oder $H_A: \ \theta > \theta_0$ (ein-seitig nach oben) oder $H_A: \ \theta < \theta_0$ (ein-seitig nach unten).

3. Wähle Signifikanzniveau α, z.B. $\alpha = 0.05$ oder 0.01.

4. Konstruiere Verwerfungsbereich für H_0, so dass

$$P_{\theta_0}[\text{Fehler 1. Art}] \leq \alpha.$$

5. Erst jetzt: betrachte ob die Beobachtung x (oder eine Funktion von mehreren Beobachtungen) in den Verwerfungsbereich fällt: falls ja, so verwerfe H_0 (die Alternative ist dann "signifikant"). Falls x nicht in den Verwerfungsbereich fällt, so belassen wir H_0 (was noch lange nicht heisst, dass deswegen H_0 statistisch bewiesen ist).

Viele Computer-Pakete liefern Punkt 4 insofern, dass der P-Wert gegeben wird. Man entscheidet dann in Punkt 5 so, dass H_0 verworfen wird, falls der P-Wert kleiner als α ist.

8.4 Vertrauensintervalle

Ein Vertrauensintervall I zum Niveau $1-\alpha$ (oft auch *Konfidenzintervall* genannt) besteht aus allen Parameterwerten, die im Sinne eines statistischen Tests zum Signifikanzniveau α mit der Beobachtung verträglich sind (üblicherweise nimmt man den zweiseitigen Test). Mathematisch heisst das:

$$I = \{\theta_0; \text{ Nullhypothese } H_0: \ \theta = \theta_0 \text{ wird belassen}\}.$$

Diese Beziehung stellt eine Dualität zwischen Tests und Vertrauensintervall dar.

Die Berechnung kann grafisch, oder mit einer Tabelle ,oder basierend auf der Normalapproximation erfolgen. Letztere ergibt

$$\frac{x}{n} \pm \Phi^{-1}(1-\frac{\alpha}{2})\sqrt{\frac{x}{n}(1-\frac{x}{n})\frac{1}{n}} \text{ ist Vertrauensintervall für } p, \text{ falls } X \sim \text{Binom}(n,p),$$
$$x \pm \Phi^{-1}(1-\frac{\alpha}{2})\sqrt{x} \text{ ist Vertrauensintervall für } \lambda, \text{ falls } X \sim \text{Poisson}(\lambda).$$

Das Vertrauensintervall ist zufällig: es fängt den unbekannten wahren Parameter mit Wahrscheinlichkeit $1-\alpha$ ein.

8.5 Mehrere Beobachtungen

Wenn man n Beobachtungen hat, geht man oft zu den Summen über: $x_1 + \ldots + x_n$. Für die zugehörigen Summen von Zufallsvariablen, welche als i.i.d. angenommen werden, kennen wir dann die Verteilung der Summen approximativ (ZGS) oder auch in einigen Fällen exakt. Damit lassen sich Verwerfungsbereiche und Konfidenzintervalle konstruieren.

9 Statistik bei normalverteilten Daten

Wir betrachten folgende Situation. Gegeben sind n Beobachtungen (Realisierungen) $x_1, \ldots, x_n$ von Zufallsvariablen $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$. Typischerweise sind dies n Messungen einer unbekannten Grösse μ, und σ gibt an, wie genau die Messungen sind. Die Annahme der Normalverteilung wird meist mit dem Zentralen Grenzwertsatz begründet. Ausserdem nehmen wir noch an, dass es keine Beeinflussungen zwischen den einzelnen Beobachtungen gibt, so dass die Unabhängigkeit der Zufallsvariablen gerechtfertigt erscheint.

Wir möchten aus $x_1, \ldots, x_n$ Rückschlüsse auf die unbekannten Parameter μ und σ ziehen. Wie zuvor geht es um die drei Fragestellungen Punktschätzung, Test für einen vorgegebenen Wert (Sollwert) und Vertrauensintervall (Bereich von plausiblen Werten). Weil μ meist von grösserem Interesse ist als σ, behandeln wir die letzten beiden Fragestellungen nur für μ.

9.1 Schätzungen

Die Punktschätzungen sind:

$$\hat{\mu} = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \hat{\sigma}^2 = S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2.$$

Diese Schätzungen sind Funktionen von Zufallsvariablen, also wieder zufällig und im allgemeinen verschieden vom unbekannten wahren Wert. Die realisierten Werte dieser Schätzung (den Wert den man berechnen kann) erhält man indem die Zufallsvariable X_i mit deren realisiertem Wert x_i ersetzt wird.

Der Erwartungswert der Schätzer ist

$$E(\hat{\mu}) = \mu$$
$$E(\hat{\sigma}^2) = \sigma^2.$$

(Dies ist der Grund für den Nenner $n-1$).

9.2 Testen

Beim Testen einer Nullhypothese $H_0 : \mu = \mu_0$ stellt sich wieder die Frage, wie die Alternative H_A aussieht.

Zur Vereinfachung nehmen wir zuerst an, dass σ bekannt ist. Dann lehnen wir bei zwei-seitiger Alternative $H_A : \mu \neq \mu_0$ die Nullhypothese $H_0 : \mu = \mu_0$ ab, falls

$$|\overline{X}_n - \mu_0| > \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1 - \frac{\alpha}{2}). \tag{9.4}$$

Analog ist der Verwerfungsbereich für ein-seitige Alternative $H_A : \mu > \mu_0$: $\overline{X}_n - \mu_0 > \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\alpha)$ und mit der Relation "<", falls $H_A : \mu < \mu_0$. Der Test mit der Entscheidungsregel in (9.4) heisst *z-Test*. Die Begründung für (9.4) ist wie folgt:

$$\overline{X}_n \sim \mathcal{N}(\mu_o, \sigma^2/n) \text{ unter der Nullhypothese } H_0.$$

Somit folgt mit einfacher Umformung, dass

$$P_{\mu_0}[|\overline{X}_n - \mu_0| > \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\frac{\alpha}{2})] = \alpha,$$

das heisst, die Wahrscheinlichkeit eines Fehlers 1. Art ist gerade gleich dem Signifikanzniveau α.

Kenntnis von σ, welche für den z-Test benötigt wird, ist in der Praxis meist unrealistisch. Wenn wir σ nicht kennen, ersetzen wir es durch den Schätzer S_n. Um die zufälligen Abweichungen vom wahren σ zu berücksichtigen, müssen wir dann jedoch die Grenzen etwas grösser machen. Man kann zeigen, dass das Niveau α eingehalten wird (d.h. die Wahrscheinlichkeit eines Fehlers 1. Art ist gleich α), wenn wir statt der Normalverteilung die sogenannte *t-Verteilung* mit $n-1$ *Freiheitsgraden* verwenden. Die Quantile $t(m; 1-\alpha)$ dieser Verteilung sind tabelliert für häufig gebrauchte α's, oder sie können mittels Computer numerisch berechnet werden. Die Entscheidungregel bei zwei-seitiger Alternative (um H_0 zu verwerfen) lautet dann anstelle von (9.4),

$$|\overline{X}_n - \mu_0| > \frac{S}{\sqrt{n}}t(n-1; 1-\alpha/2).$$

Und analog für ein-seitige Alternativen. Dieser Test heisst *t-Test*.

9.3 Vertrauensintervall

Das Vertrauensintervall zum Niveau $1-\alpha$ besteht aus allen Parameterwerten μ_0 , bei denen der t-Test (üblicherweise zweiseitig) zum Niveau α nicht verwirft:

$$[\overline{X}_n - \frac{S_n}{\sqrt{n}}t(n-1; 1-\frac{\alpha}{2}), \overline{X}_n + \frac{S_n}{\sqrt{n}}t(n-1; 1-\frac{\alpha}{2})]. \tag{9.5}$$

Man kann mit Formel (9.5) nachprüfen, dass

$$P[\mu \in I] = 1-\alpha,$$

wobei I das Vertrauensintervall in (9.5) ist. Das heisst, dass das *zufällige* Intervall I den unbekannten wahren Parameter μ mit Wahrscheinlichkeit $1-\alpha$ überdeckt. Der Beweis dafür ist wie folgt:

$$\begin{aligned} P[\mu \in I] &= P[-\frac{S_n}{\sqrt{n}}t(n-1; 1-\alpha/2) \le \overline{X}_n - \mu \le \frac{S_n}{\sqrt{n}}t(n-1; 1-\alpha/2)] \\ &= P[\underbrace{-t(n-1; 1-\alpha/2)}_{=t(n-1;\alpha/2)} \le \sqrt{n}\frac{\overline{X}_n - \mu}{S_n} \le t(n-1; 1-\alpha/2)] \\ &= 1-\alpha/2-\alpha/2 = 1-\alpha. \end{aligned}$$

Ausblick auf eine verwandte Methode in der Praxis: in der *statistischen Qualitätskontrolle* wird in regelmässigen Abständen eine kleine Stichprobe vom Umfang n aus dem Produktionsprozess gezogen, die Zielgrösse gemessen, gemittelt und gegen die Zeit aufgetragen. Fällt ein Mittelwert ausserhalb der Kontrollgrenzen "Sollwert $\pm 3\sigma/\sqrt{n}$" oder sind 9 aufeinanderfolgende Mittelwerte alle grösser oder alle kleiner als der Sollwert, dann ist der Produktionsprozess ausser Kontrolle.

10 Punktschätzungen: allgemeine Methoden

Wir betrachten folgende Situation: Gegeben sind n Beobachtungen $x_1, \dots x_n$, die wir als Realisierungen von n i.i.d. Zufallsvariablen $X_1, \dots, X_n$ ansehen, siehe Kap. 8.1. Die Verteilung von X_i sei bekannt bis auf einen unbekannten Parameter θ. Dabei kann θ auch mehrere Komponenten haben und ist dann ein Parametervektor. Bei der Normalverteilung ist z.B. $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+$. In diesem allgemeinen Rahmen stellen wir zwei Schätzmethoden vor.

10.1 Momentenmethode

Die Momentenmethode nimmt an, dass wir den unbekannten Parameter θ ausdrücken können mithilfe der Momente $\mu_k = E[X^k]$ $(1 \leq k \leq p)$, d.h.

$$\theta_j = g_j(\mu_1, \dots \mu_p) \quad (j = 1, \dots r),$$

wobei r die Dimension des Parametervektors ist. Der Momentenschätzer ersetzt nun die wahren μ_k durch deren empirische Analoga:

$$\hat{\theta}_j = g_j(\hat{\mu}_1, \dots, \hat{\mu}_p) \quad (j = 1, \dots, r),$$
$$\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k.$$

Der Momentenschätzer ist einfach, aber nicht immer die optimale (im Sinne einer zu definierenden besten Genauigkeit für den unbekannten Parameter) Methode. Überdies ist der Momentenschätzer nicht eindeutig wie das folgende Beispiel zeigt.

Beispiel: $X_1, \dots, X_n$ i.i.d. $\sim$ Poisson(λ). Es gilt $E[X_i] = \lambda$. Also können wir $g_1(\cdot)$ als die Identität wählen und erhalten den Momentenschätzer

$$\hat{\lambda} = \overline{X}_n = n^{-1} \sum_{i=1}^{n} X_i.$$

Es gilt aber auch: $Var(X_i) = \lambda$. Wir können also auch

$$g_1(\mu_1, \mu_2) = \mu_2 - \mu_1^2$$

wählen und erhalten so einen anderen Momentenschätzer

$$\hat{\lambda} = n^{-1} \sum_{i=1}^{n} X_i^2 - (\overline{X}_n)^2 = \frac{n-1}{n} S_n^2.$$

In diesem Beispiel zieht man $\hat{\lambda} = \overline{X}_n$ vor, denn dies ist auch der sogenannte Maximum-Likelihood Schätzer, welcher im Allgemeinen genauer ist.

10.2 Maximum-likelihood Schätzer

Die Maximum-Likelihood Methode nimmt als Schätzung denjenigen Parameterwert θ, der die sogenannte log-Likelihoodfunktion

$$\ell(\theta) = \begin{cases} \sum_{i=1}^n \log p_\theta(X_i) & \text{für diskrete } X_i \\ \sum_{i=1}^n \log f_\theta(X_i) & \text{für stetige } X_i \end{cases}$$

maximiert. Die Maximum-Likelihood-Schätzung ist meist genauer, und es gibt einfache Verfahren, um auch ein Vertrauensintervalle für θ zu konstruieren und Hypothesen der Form $\theta = \theta_0$ zu testen (siehe Lehrbücher).

Beispiel (Fortsetzung): $X_1, \ldots, X_n$ i.i.d. $\sim$ Poisson(λ). Die Punktwahrscheinlichkeiten sind dann

$$p_\lambda(x) = \frac{\lambda^x}{x!} e^{-\lambda}.$$

Die log-Likelihoodfunktion ist somit

$$\begin{aligned} \ell(\lambda) &= \sum_{i=1}^n (X_i \log(\lambda) - \log(X_i!) - \lambda) \\ &= \sum_{i=1}^n (X_i \log(\lambda) - \lambda) - C, \quad C = \sum_{i=1}^n \log(X_i!). \end{aligned}$$

Beachte dass die Konstante C keinen Einfluss auf die Maximierung von $\ell(\lambda)$ hat. Leitet man $\ell(\lambda)$ ab und setzt $\ell'(\lambda) = 0$, so erhält man die Lösung des Maximierungs-Problems und somit den Maximum-Likelihood Schätzer

$$\hat{\lambda} = \overline{X}_n = n^{-1} \sum_{i=1}^n X_i.$$

Dies ist derselbe Schätzer wie bei der Momentenmethode wo g_1 die Identität ist.

11 Vergleich zweier Stichproben

Wichtige Anwendungen der Statistik liegen im Vergleich verschiedener Versuchsbedingungen, oder allgemeiner bei der Bestimmung der Auswirkung verschiedener erklärender Variablen auf eine Zielgrösse. Als einfachsten Fall behandeln wir jetzt den Vergleich zweier Methoden (Gruppen, Versuchsbedingungen, Behandlungen) hinsichtlich des Erwartungswertes.

11.1 Gepaarte und ungepaarte Stichproben

In allen Anwendungen ist neben der Auswertung auch die korrekte Planung des Versuches wichtig. Man muss sicherstellen, dass eventuelle Unterschiede tatsächlich durch die verschiedenen Methoden und nicht durch eine andere Störgrösse verursacht sind. Die beiden wichtigsten Prinzipien dazu sind *Blockbildung* und *Randomisierung*.

Randomisierung bedeutet hier, dass man die Reihenfolge der Versuche und die Zuordnung von Versuchseinheit zu Versuchsbedingung zufällig wählt: man hat dann Beobachtungen (realisierte Zufallsvariablen)

$$x_1, x_2, \ldots, x_n \text{ unter Versuchsbedingung 1,}$$
$$y_1, y_2, \ldots, y_m \text{ unter Versuchsbedingung 2.}$$

Im Allgemeinen ist $m \neq n$, aber nicht notwendigerweise. Bei solch zufälliger Zuordnung von verschiedenen Versuchseinheiten zu zwei verschiedenen Versuchsbedingungen spricht man von einer *ungepaarten Stichprobe*.

Beispiel: zufällige Zuordnung von 100 Testpatienten zu Gruppe der Grösse 60 mit Medikamenten-Behandlung und zu anderer Gruppe der Grösse 40 mit Placebo-Behandlung.

Andererseits liegt eine *gepaarte Stichprobe* vor, wenn beide Versuchsbedingungen an derselben Versuchseinheit eingesetzt werden:

$$x_1, \ldots, x_n \text{ unter Versuchsbedingung 1,}$$
$$y_1, \ldots, y_n \text{ unter Versuchsbedingung 2.}$$

Notwendigerweise ist dann: die Stichprobengrösse n ist für beide Versuchsbedingungen dieselbe.

Beispiel: Vergleich zweier Reifentypen, wo bei jedem Testfahrzeug und jedem Fahrer beide Reifentypen verwendet werden.

11.2 Gepaarte Vergleiche

Bei der Analyse von gepaarten Vergleichen arbeitet man stets mit den Differenzen innerhalb der Paare,

$$u_i = x_i - y_i \quad (i = 1, \ldots, n),$$

welche wir als Realisierungen von i.i.d. Zufallsvariablen $U_1, \ldots, U_n$ auffassen. Kein Unterschied zwischen den beiden Versuchsbedingungen heisst dann einfach $E[U_i] = 0$. Dies kann man formal testen mit der Nullhypothese $H_0 : E[U_i] = 0$ und mit der zwei-seitigen (oder auch ein-seitigen) Alternative $H_A : E[U_i] \neq 0$. Die folgenden Tests bieten sich dazu an:

1. der t-Test, siehe Kap. 9.2;
2. der sogenannte *Vorzeichen-Test*, falls die Normalverteilung nicht gerechtfertigt scheint: betrachte Anzahl positiver U_i und benütze die Methoden für die Binomialverteilung um die Nullhypothese $H_0 : p = p_0 = 0.5$ zu testen, siehe Kap. 8.3;
3. der sogenannte *Wilcoxon-Test*, siehe unten.

Der Wilcoxon-Test ist ein Kompromiss, der weniger voraussetzt als der t-Test und die Information der Daten besser ausnützt als der Vorzeichen-Test. Dazu bildet man die Ränge der Differenzen bezüglich des Absolutwertes: $\text{Rang}(|U_i|) = k$ heisst, dass $|U_i|$ den k-ten kleinsten Wert hat unter $|U_1|, \ldots, |U_n|$. Wenn einzelne $|U_i|$ zusammenfallen, teilt man die

Ränge auf durch Mittelung. Ausserdem sei noch V_i der Indikator dafür, ob U_i positiv ist, d.h. $V_i = 1$ falls $U_i > 0$ ist und $V_i = 0$ sonst. Dann verwirft man die Nullhypothese, falls

$$W = \sum_{i=1}^{n} \text{Rang}(|U_i|)V_i$$

zu gross oder zu klein oder beides ist (je nach Spezifikation der Alternative). Die Schranken für zu gross oder zu klein entnimmt man aus Tabellen oder Statistikpaketen für den Computer. Man kann zeigen, dass dieser Test das Niveau exakt einhält (d.h. die Wahrscheinlichkeit für einen Fehler 1. Art ist gleich α), wenn die U_i i.i.d. sind und eine um 0 symmetrische Dichte haben. Beim t-Test wird das Niveau zwar auch ungefähr eingehalten bei vielen nichtnormalen Verteilungen (wegen dem ZGS), aber unter Umständen ist die Wahrscheinlichkeit eines Fehlers 2. Art beim t-Test *viel grösser* als beim Wilcoxon-Test.

In der Praxis ist der Wilcoxon-Test allermeist dem t- oder Vorzeichen-Test vorzuziehen. Nur falls die Daten sehr gut mit einer Normalverteilung beschrieben werden ist der t-Test für gute Datenanalyse "vollumfänglich tauglich": diese Annahme oder Bedingung kann man z.B. mit dem Normal-Plot (siehe Kap. 7.3) grafisch überprüfen.

11.3 Zwei-Stichproben Tests

Wie bereits beschrieben gibt es Fälle (ungepaarte Stichproben), wo man keine Paare bilden kann. Dann hat man i.i.d Zufallsvariablen $X_1, \dots X_n$ für die eine Versuchsbedingung und $Y_1, \dots Y_m$ für die andere, und man nimmt an, dass alle Zufallsvariablen unabhängig sind. Die effektiv gemachten Beobachtungen sind wie üblich als Realisierungen von diesen Zufallsvariablen zu interpretieren. Das einfachste Problem lässt sich unter folgender Annahme lösen:

$$\begin{aligned} X_i &\sim \mathcal{N}(\mu_X, \sigma^2) \quad (i = 1, \dots, n), \\ Y_i &\sim \mathcal{N}(\mu_Y, \sigma^2) \quad (i = 1, \dots, m). \end{aligned}$$

Der *Zwei-Stichproben t-Test* verwirft dann die Nullhypothese $H_0 : \mu_X = \mu_Y$, falls

$$\begin{aligned} &\frac{|\overline{X}_n - \overline{Y}_m|}{S_{pool}\sqrt{1/n + 1/m}} > t(n + m - 2; 1 - \frac{\alpha}{2}) \text{ bei Alternative } H_A : \mu_X \neq \mu_Y, \\ &\frac{\overline{X}_n - \overline{Y}_m}{S_{pool}\sqrt{1/n + 1/m}} > t(n + m - 2; 1 - \alpha) \text{ bei Alternative } H_A : \mu_X > \mu_Y. \quad (11.6) \end{aligned}$$

Dabei ist

$$S^2_{pool} = \frac{1}{n + m - 2}\left(\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 + \sum_{i=1}^{m}(Y_i - \overline{Y}_m)^2\right)$$

die gepoolte Schätzung für die gemeinsame Varianz σ^2. Die Wahl des Nenners in (11.6) ergibt sich aus $\text{Var}(\overline{X}_n - \overline{Y}_m) = \sigma^2(\frac{1}{n} + \frac{1}{m})$.

Die Verallgemeinerungen des Zwei-Stichproben t-Tests bei ungleichen Varianzen $\sigma^2_X \neq \sigma^2_Y$ ist in der Literatur zu finden (z.B. Rice, Kap. 11.2, Example C) auf S. 395). Ebenfalls in der Literatur zu finden ist der Zwei-Stichproben Wilcoxon-Test, welcher ein für die Praxis sehr guter Test für ungepaarte Stichproben ist (z.B. Rice, Kap. 11.2.3).

Anhang

Die wichtigsten 1-dimensionalen Verteilungen

Verteilung	$p(x)$, bzw. $f(x)$	Wertebereich	$E(X)$	$\mathrm{Var}(X)$
Binomial(n,p)	$\binom{n}{x}p^x(1-p)^{n-x}$	$\{0,1,\dots n\}$	np	$np(1-p)$
Geometrisch(p)	$p(1-p)^{x-1}$	$\{1,2,\dots\}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson(λ)	$e^{-\lambda}\frac{\lambda^x}{x!}$	$\{0,1,\dots\}$	λ	λ
Uniform(a,b)	$\frac{1}{b-a}$	$[a,b]$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential(λ)	$\lambda e^{-\lambda x}$	$\mathbb{R}^+ = \{x \in \mathbb{R}; x > 0\}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma(α,λ)	$\frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}$	$\mathbb{R}^+ = \{x \in \mathbb{R}; x > 0\}$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$
Normal(μ,σ^2)	$\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$	$\mathbb{R}$	μ	σ^2