Statistik

D-UWIS & D-ERDW & D-AGRL

Sommersemester 2005

basierend auf Skript von H.-R. Künsch

Peter Bühlmann
Seminar für Statistik
ETH Zürich

Inhaltsverzeichnis

Aufbau der Vorlesung

Die Vorlesung behandelt zuerst die Wahrscheinlichkeitsrechnung und Statistik für diskrete Variablen, welche Werte zum Beispiel in $\{0, 1\}$, in $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$ oder in $\mathbb{Z} = \{\ldots, -1, 0, 1, \ldots\}$ annehmen.

Danach werden die erarbeiteten Konzepte auf stetige Variablen übertragen, mit Wertebereichen, mit Wertebereichen zum Beispiel in $\mathbb{R}$ oder $[0, 1]$. Deshalb ist der Aufbau leicht repetitiv, was sich aber in vorigen Jahren gut bewährt hat.

Schlussendlich wird auf komplexere Modellierung anhand der multiplen Regressions-Analyse eingegangen.

Kapitel 1

Einführung (Stahel, Kap. 1)

Die Bedeutung der Statistik liegt, für viele Wissenschaften, in der Fähigkeit

verallgemeinernde Schlüsse von Daten ("Stichproben") auf zukünftige Daten oder umfassendere Populationen zu machen.

Insbesondere wird dabei berücksichtigt, dass

alle Daten gewissen Schwankungen unterworfen sind.

Um dies zu quantifizieren benützt man

Modelle und Gesetze der Wahrscheinlichkeitstheorie.

Der Zufall gehorcht gewissen Gesetzen aus der Wahrscheinlichkeitstheorie, welche genauso zuverlässig sind wie andere Naturgesetze. Ob die Welt tatsächlich zufällig ist, oder ob der Zufall nur eine Terminologie ist für all diejenigen deterministischen Faktoren, die wir unmöglich alle in Betracht ziehen können, ist für viele Betrachtungen unwesentlich.

Kapitel 2

Modelle für Zähldaten

2.1 Einleitung (Stahel, Kap. 4.1)

Ein **Wahrscheinlichkeitsmodell** beschreibt, welche Ergebnisse eines Experiments möglich sind und welche Chancen die verschiedenen Ergebnisse haben. Wir behandeln hier diskrete Wahrscheinlichkeitsmodelle, bei denen die möglichen Ergebnisse endlich oder "abzählbar" (z.B. in $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$) sind. Ein Wahrscheinlichkeitsmodell erlaubt zudem mittels Simulation andere mögliche Daten zu erzeugen und so zu differenzieren, welche Schwankungen noch plausibel sind und welche nicht.

Die Experimente sind hier immer als **Zufallsexperiment** zu verstehen:

> Zufallsexperiment =
>
> Experiment, dessen Ausgang auch nicht von einem Orakel exakt vorhersagbar ist

2.2 Diskrete Wahrscheinlichkeit (Stahel, Kap. 4.2, 4.6)

Für die Beschreibung von Zufallsexperimenten benützen wir ein Wahrscheinlichkeitsmodell. Dieses besteht aus den folgenden Komponenten:

- **Grundraum** Ω
- **Elementarereignis** ω
- **Wahrscheinlichkeit** P

Der Grundraum und die Elementarereignisse sind wie folgt:

$$\Omega = \{\underbrace{\text{mögliche Elementarereignisse } \omega}_{\text{mögliche Ausgänge/Resultate}}\}$$

Beispiel: 2-maliges Werfen einer Münze
$\Omega = \{KK, KZ, ZK, ZZ\}$ wobei K = "Kopf" und Z = "Zahl" bezeichnet
Elementarereignis: zum Beispiel $\omega = KZ$

Unter einem **Ereignis** A versteht man eine Teilmenge von Ω:

$$\text{Ereignis A} \subset \Omega$$

Beispiel (Forts.): $A = \{\text{genau 1-mal Kopf}\} = \{KZ, ZK\}$.

Die Operationen der Mengenlehre (Komplement, Vereinigung, Durchschnitt) haben eine natürliche Interpretation in der Sprache der Ereignisse.

$$\begin{aligned} A \cup B &\Leftrightarrow \text{ A \textbf{oder} B, wobei das "oder" nicht-exklusiv ist ("oder/und")} \\ A \cap B &\Leftrightarrow \text{ A \textbf{und} B} \\ A^c &\Leftrightarrow \text{ \textbf{nicht} A} \end{aligned}$$

Beispiel: $A =$ morgen scheint die Sonne, $B =$ morgen regnet es.
$A \cup B$ bedeutet: morgen scheint die Sonne oder morgen regnet es (und dies kann auch bedeuten, dass morgen die Sonne scheint und dass es morgen regnet); $A \cap B$ bedeutet: morgen scheint die Sonne und morgen regnet es; A^c bedeutet: morgen scheint die Sonne nicht.

Eine **Wahrscheinlichkeit** ordnet jedem Ereignis A eine Wahrscheinlichkeit $P(A)$ zu. Dabei sind die folgenden drei grundlegenden Regeln (Axiome von Kolmogorov) erfüllt:

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

Weitere Regeln können daraus abgeleitet werden, z.B.

$$P(A^c) = 1 - P(A),$$
$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Im diskreten Fall ist eine Wahrscheinlichkeit festgelegt durch die Wahrscheinlichkeiten der Elementarereignisse $P(\{\omega\})$:

$$P(A) = \sum_{\omega \in A} P(\{\omega\}).$$

Beispiel (Forts.) Für $A =$ genau einmal Kopf $= \{KZ, ZK\}$ hat man
$P(A) = P(KZ) + P(ZK) = 1/4 + 1/4 = 1/2$.

Im Wesentlichen werden in der Wahrscheinlichkeitstheorie die Wahrscheinlichkeiten gewisser Ereignisse A festgelegt (auf Grund von Plausibilitäten, Symmetrieüberlegungen, wissenschaftlichen Theorien, Fachwissen und Daten) und daraus die Wahrscheinlichkeiten von gewissen anderen Ereignissen B aus den obigen Gesetzen hergeleitet.
(Die Statistik geht umgekehrt vor: aus Daten, d.h. aus der Information, dass gewisse Ereignisse eingetreten sind, versucht man Rückschlüsse auf ein unbekanntes Wahrscheinlichkeitsmodell (unbekannte Wahrscheinlichkeiten) zu machen).

Interpretation von Wahrscheinlichkeiten:
• Idealisierung der relativen Häufigkeiten bei vielen unabhängigen Wiederholungen (**frequentistisch**)
• Mass für den Glauben, dass ein Ereignis eintreten wird" (**Bayes'sch**)

2.2.1 Der Begriff der Unabhängigkeit

Wenn man die Wahrscheinlichkeiten $P(A)$ und $P(B)$ kennt, lässt sich im Allgemeinen daraus nicht $P(A \cap B)$ berechnen.

Wenn zwischen den Ereignissen A und B kein kausaler Zusammenhang besteht (d.h. es gibt keine gemeinsamen Ursachen oder Ausschliessungen), dann definiert man:

$$\text{A und B heissen (stochastisch) unabhängig } \Leftrightarrow P(A \cap B) = P(A)P(B).$$

Im Falle von unabhängigen Ereignissen, vereinfachen sich viele Situation: insbesondere sehen wir, dass in diesem Fall $P(A \cap B)$ aus $P(A)$ und $P(B)$ berechnet werden kann. In der Praxis wird oftmals aufgrund von Plausibilitätsüberlegungen **deklariert**, dass zwei Ereignisse unabhängig sind.

Bei mehreren Ereignissen $A_1, \ldots A_n$ bedeutet Unabhängigkeit, dass zum Beispiel

$$\begin{aligned} &P(A_1 \cap A_2) = P(A_1)P(A_2), \\ &P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3). \end{aligned}$$

Die allgemeine Formel für unabhängige Ereignisse lautet:

$P(A_{i_1} \cap \ldots \cap A_{i_k}) = P(A_{i_1}) \cdots P(A_{i_k})$ für jedes $k \leq n$ und jedes $1 \leq i_1 < \ldots < i_k \leq n$.

2.3 Zufallsvariable (Stahel, Kap. 4.3, 4.4)

Oft sind mit einem Zufallsexperiment Zahlenwerte verknüpft, d.h. zu jedem Elementarereignis (Ergebnis) ω gehört ein Zahlenwert $X(\omega) = x$.

Beispiel: Wert einer gezogenen Jass-Karte

$$\begin{array}{rcl} \omega = \text{ As} & \mapsto & X(\omega) = 11 \\ \omega = \text{ König} & \mapsto & X(\omega) = 4 \\ \vdots & & \vdots \\ \omega = \text{ Sechs} & \mapsto & X(\omega) = 0 \end{array}$$

In obigen Beispiel ist also $X(\cdot)$ eine Funktion. Allgemein definiert man:

Eine **Zufallsvariable** X ist eine **Funktion**:

$$\begin{aligned} X: \quad & \Omega \to \mathbb{R} \\ & \omega \mapsto X(\omega) \end{aligned}$$

Die Funktion $X(\cdot)$ ist nicht zufällig; das Argument ω hingegen schon.

Die Notation X (oder auch $Y, Z, \ldots$) ist eher unüblich für die Bezeichung einer Funktion: wir werden aber sehen, dass man mit Zufallsvariablen manchmal ähnlich rechnen kann wie mit gewöhnlichen Variablen x(oder auch $y, z, \ldots$).

Je nach Ausgang des Zufallsexperiments (d.h. von ω) erhalten wir einen anderen Wert $x = X(\omega)$: der Wert x ist **eine Realisierung** der Zufallsvariablen X. Eine Realisierung einer Zufallsvariablen ist also das Ergebnis eins Zufallsexperiments (welches mit einer Zahl beschrieben werden kann).

Eine Zufallsvariable X heisst *diskret*, falls der Wertebereich $W = W_X$ (Menge der möglichen Werte von X) diskret ist, d.h. abzählbar (man kann die möglichen Werte durchnumerieren). Zum Beispiel ist $W = \{0, 1, \ldots, 10\}$ endlich und deshalb diskret; $W = \mathbb{N}_0 = \{0, 1, 2, \ldots\}$ ist zwar unendlich, aber trotzdem diskret; $W = \mathbb{R}$ ist nicht diskret (sondern kontinuierlich). Wir werden in diesem Kapitel bloss diskrete Zufallsvariablen genauer diskutieren.

Wahrscheinlichkeitsverteilung einer Zufallsvariablen

Die Werte einer Zufallsvariablen X (die möglichen Realisationen von X) treten mit gewissen Wahrscheinlichkeiten auf. Diese sind wie folgt definiert:

$$\begin{aligned} &\text{Wahrscheinlichkeit, dass } X \text{ den Wert } x \text{ annimmt} \\ &= P(X = x) = P(\{\omega;\ X(\omega) = x\}) \\ &= \sum_{\omega; X(\omega) = x} P(\omega). \end{aligned}$$

Beispiel (Forts): $X =$ Wert einer gezogenen Jass-Karte

$$\begin{aligned} &\text{Wahrscheinlichkeit für Zahl } 4 = P(X = 4) \\ = \;& P(\{\omega;\ \omega = \text{ ein König}\}) \\ = \;& P(\text{Eicheln-König}) + P(\text{Rosen-König}) + P(\text{Schellen-König}) + P(\text{Schilten-König}) \\ = \;& 4/36 = 1/9. \end{aligned}$$

Die "Liste" von $P(X = x)$ für alle möglichen Werte x heisst (diskrete) **(Wahrscheinlichkeits-) Verteilung** der (diskreten) Zufallsvariablen X. Zu einer Zufallsvariablen X gehört immer eine (Wahrscheinlichkeits-) Verteilung und umgekehrt:

$$\text{Zufallsvariable } X \;\Leftrightarrow\; \text{(Wahrscheinlichkeits-) Verteilung}$$

Es gilt immer

$$\sum_{\text{alle möglichen } x} P(X = x) = 1.$$

Beispiel (Forts): $X =$ Wert einer gezogenen Jass-Karte
Die Wahrscheinlichkeitsverteilung von X ist

$$\begin{aligned} P(X = 11) &= 1/9 \\ P(X = 10) &= 1/9 \\ P(X = 4) &= 1/9 \\ P(X = 3) &= 1/9 \\ P(X = 2) &= 1/9 \\ P(X = 0) &= 4/9 \end{aligned}$$

Wenn man nur an der Zufallsvariablen X interessiert ist, kann man den zu Grunde liegenden Raum Ω vergessen, man braucht nur die Verteilung von X.

2.4 Binomialverteilung (Stahel Kap. 5.1)

Wir betrachten die Situation wo es um das Messen der Anzahl Erfolge (oder Misserfolge) geht. Solche Anwendungen treten z.B. auf bei der Qualitätskontrolle, Erfolg/Misserfolg bei Behandlungen (medizinisch, biologisch) oder auch bei Glücksspielen.

Beispiel: Werfen einer Münze
Es wird eine Münze geworfen, welche dann zufällig entweder auf Kopf (K) oder Zahl (Z) fällt.
Betrachte die Zufallsvariable X mit Werten in $W = \{0, 1\}$, welche folgendes beschreibt:

$$\begin{aligned} X &= 0 \qquad \text{falls Münze auf Z fällt,} \\ X &= 1 \qquad \text{falls Münze auf K fällt.} \end{aligned}$$

Die Wahrscheinlichkeitsverteilung von X kann durch einen einzelnen Parameter π beschrieben werden:

$$P(X = 1) = \pi, \quad P(X = 0) = 1 - \pi, \quad 0 \leq \pi \leq 1.$$

Fall die Münze fair ist, so ist $\pi = 1/2$.

Bernoulli(π)-Verteilung:

Eine Zufallsvariable X mit Werten in $W = \{0, 1\}$ heisst Bernoulli(π)-verteilt, falls
$$P(X = 1) = \pi, \quad P(X = 0) = 1 - \pi, \quad 0 \leq \pi \leq 1.$$

Die Bernoulli-Verteilung ist eine triviale Mathematisierung um das Eintreffen oder Nicht-Eintreffen eines Ereignis zu beschreiben.

Beispiel (Forts.): n-maliges Werfen einer Münze
Wir betrachten $X =$ Anzahl K bei n unabhängigen Münzwürfen. Es ist klar, dass der Wertebereich von X die Menge $W = \{0, 1, \ldots, n\}$ ist. X lässt sich auch darstellen als Summe von unabhängigen Bernoulli-Zufallsvariablen

$$\begin{aligned} X &= \sum_{i=1}^{n} X_i, \\ X_i &= \begin{cases} 1 & \text{falls K im i-ten Wurf} \\ 0 & \text{falls Z im i-ten Wurf.} \end{cases} \end{aligned}$$

Die Verteilung von X im obigen Beispiel lässt sich analytisch ausrechnen. Falls $X_1, \ldots, X_n$ alle unabhängig und Bernoulli(π)-verteilt sind, so gilt zum Beispiel:

$$\begin{aligned} P(X = 0) &= P(\text{alle } X_1 = \ldots = X_n = 0) = (1 - \pi)^n, \\ P(X = 1) &= P(\text{ein } X_i = 1 \text{ und alle anderen } X_j = 0) = \binom{n}{1} \pi (1 - \pi)^{n-1}. \end{aligned}$$

Allgemein gilt die Formel der Binomial-Verteilung.

Binomial(n, π)-Verteilung:

Eine Zufallsvariable X mit Werten in $W = \{0, 1, \ldots, n\}$ heisst Binomial(n, π)-verteilt, falls

$$P(X = x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}, \ x = 0, 1, \ldots, n$$

wobei $0 \leq \pi \leq 1$ der Erfolgsparameter der Verteilung ist.

(Dabei ist $\binom{n}{x}$ der Binomialkoeffizient, der angibt, auf wie viele Arten man x Erfolge und $n - x$ Misserfolge anordnen kann).

Wie bereits in obigem Beispiel motiviert, beschreibt X die Anzahl Erfolge/Misserfolge (Eintreten eines bestimmten Ereignis) bei n **unabhängigen** Versuchen. Das Prädikat "unabhängig" ist wesentlich für die Korrektheit der Binomialverteilung.

Beispiel: Spermasexing (Tages-Anzeiger 6.12.2000)
Geschlechts-Beeinflussung von Kuhkälbern mit einer Methode, die Spermasexing genannt wird: deren Ziel ist, ein weibliches Kalb zu züchten. In einem Testlauf wurden zwölf Kühe mit Spermien besamt, die optisch nach dem Y-Chromosom sortiert wurden (d.h. mit der Spermasexing-Methode). Da die Methode nicht hundertprozentig sicher ist, können wir das als Zufallsexperiment auffassen. Sei $X =$ Anzahl weiblicher gezüchteter Kuhkälber. Eine vernünftiges Modell ist dann:

$$X \sim \text{ Binomial}(12, \pi),$$

wobei π unbekannt ist. Effektiv beobachtet wurden $x = 11$ weiblich gezüchtete Kuhkälber: d.h. $X = x = 11$ wurde tatsächlich **realisiert**.

Eigenschaften der Binomialverteilung (siehe Abb. 2.1): $P(X = x)$ ist maximal wenn x gleich dem ganzzahligen Teil von $(n + 1)\pi$ ist, und auf beiden Seiten von diesem Wert nehmen die Wahrscheinlichkeiten monoton ab. Wenn $n\pi(1 - \pi)$ nicht allzu klein ist, ist die Verteilung glockenförmig.

2.5 Kennzahlen einer Verteilung (Stahel Kap. 5.3)

Eine beliebige (diskrete) Verteilung kann vereinfachend zusammengefasst werden durch 2 Kennzahlen, den **Erwartungswert** $\mathcal{E}(X)$ und die **Varianz** $\text{Var}(X)$ (oder **Standardabweichung** $\sigma(X) = \sqrt{\text{Var}(X)}$).

Der Erwartungswert beschreibt die mittlere Lage der Verteilung und ist wie folgt definiert:

$$\mathcal{E}(X) = \sum_{x \in W_x} x P(X = x), \quad W_x = \text{ Wertebereich von } X.$$

Die Varianz beschreibt die Variabilität der Verteilung, und die Standardabweichung deren Streuung:

$$\text{Var}(X) \sum_{x \in W_x} (x - \mathcal{E}(X))^2 P(X = x)$$
$$\sigma(X) = \sqrt{\text{Var}(X)}.$$

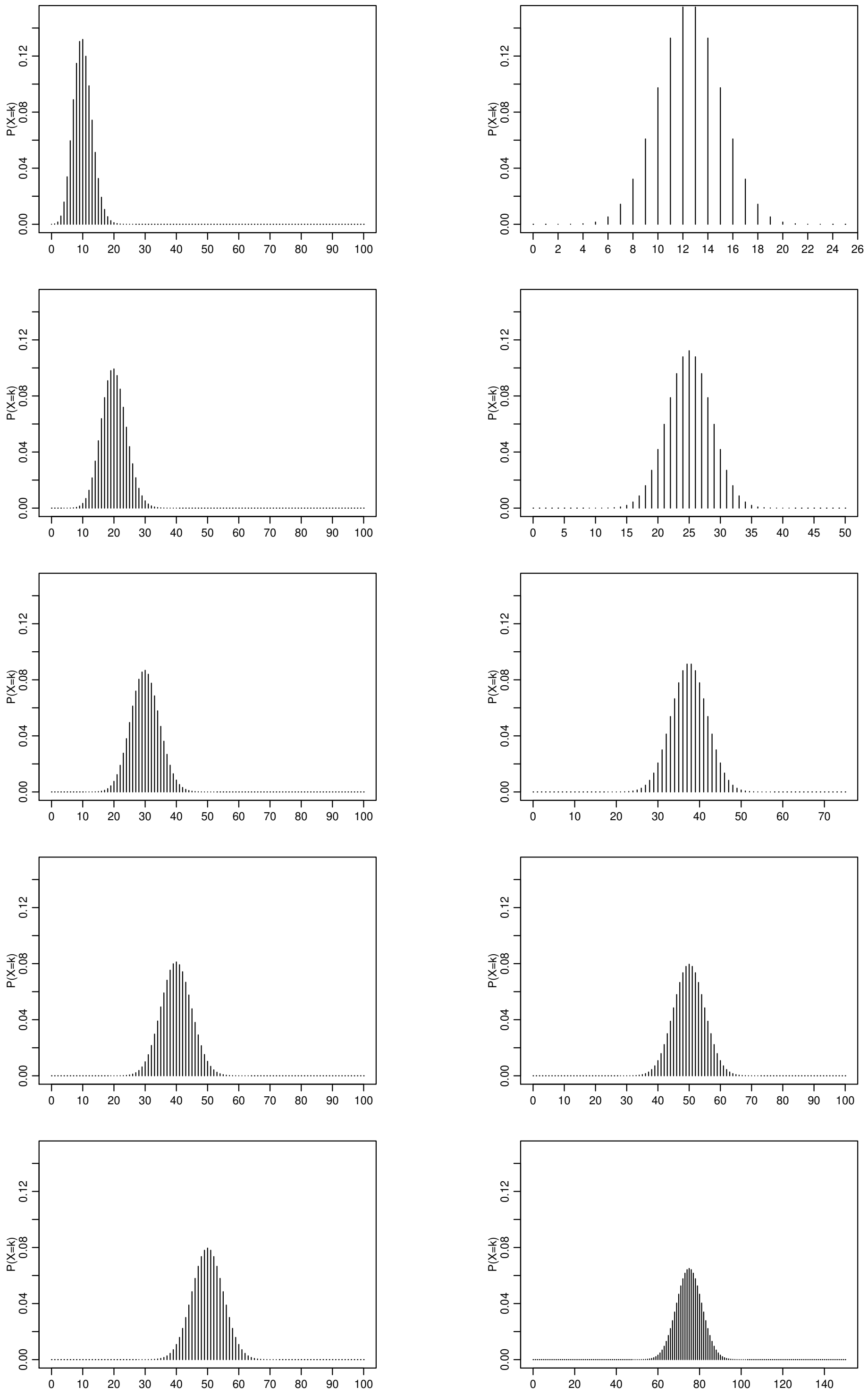

Abbildung 2.1: Die Binomialwahrscheinlichkeiten $P(X = x)$ als Funktion von x für verschiedene n's und π's. Links ist $n = 100$ und $\pi = 0.1, 0.2, 0.3, 0.4, 0.5$ und rechts ist $\pi = 0.5$ und $n = 25, 50, 75, 100, 150$.

Die Standardabweichung hat dieselbe Dimension wie die Messdaten: wird z.B. X in Metern (m) gemessen, so besitzt $\mathrm{Var}(X)$ die Dimension Quadratmeter (m^2) und $\sigma(X)$ wiederum Meter (m).

Beispiel: Sei $X \sim$ Bernoulli(π).
Dann:

$$
\begin{aligned}
\mathcal{E}(X) &= 0 \cdot P(X=0) + 1 \cdot P(X=1) = \pi, \\
\mathrm{Var}(X) &= (0-\mathcal{E}(X))^2 P(X=0) + (1-\mathcal{E}(X))^2 P(X=1) = \pi^2(1-\pi) + (1-\pi)^2\pi \\
&= \pi(1-\pi), \\
\sigma(X) &= \sqrt{\pi(1-\pi)}.
\end{aligned}
$$

Allgemeiner gilt für die Binomial-Verteilung (beachte, dass Bernoulli(π) = Binomial(1,π)):

$$
\begin{aligned}
&X \sim \text{ Bernoulli}(n,\pi), \\
&\mathcal{E}(X) = n\pi, \quad \mathrm{Var}(X) = n\pi(1-\pi), \quad \sigma(X) = \sqrt{n\pi(1-\pi)}.
\end{aligned}
$$

2.5.1 Kumulative Verteilungsfunktion

Manchmal ist es geeigneter, wenn man statt der "Liste" $P(X=x)$ (für alle x) die sogenannte **kumulative Verteilungsfunktion** angibt:

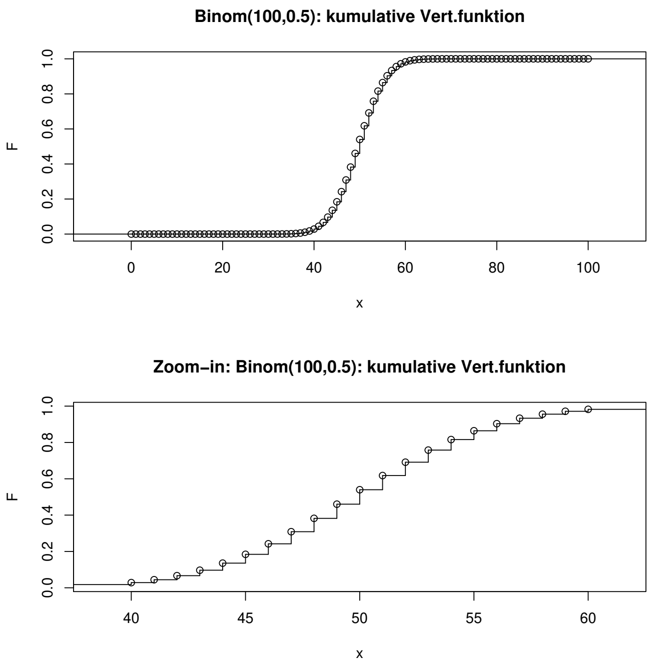

Abbildung 2.2: Kumulative Verteilungsfunktion $F(\cdot)$ für $X \sim$ Binomial(100,0.5). Unten: zoom-in für die Werte $x \in [40, 60]$.

$$F(x) = P(X \le x) = \sum_{k \le x} P(X = k).$$

Die Funktion $F(\cdot)$ ist monoton wachsend (allerdings nicht strikt monoton) und es gilt:

$$F(-\infty) = 0, \quad F(+\infty) = 1.$$

Siehe auch Abbildung 2.2. Die Kenntnis der "Liste" $P(X = x)$ (für alle x) ist äquivalent zur Kenntnis der Funktion $F(\cdot)$, denn aus dem einen kann man das andere berechnen. Zum Beispiel gilt für X mit Wertebereich $W_X = \{0, 1, \ldots, n\}$: $P(X = x) = F(x) - F(x-1)$ $(x = 1, 2 \ldots, n)$ und $P(X = 0) = F(0)$.

2.6 Poissonverteilung (Stahel Kap. 5.2)

Der Wertebereich der Binomial(n, π)-Verteilung ist $W = \{0, 1, \ldots, n\}$. Falls eine Zufallsvariable nicht im vornherein einen beschränkten Wertebereich hat, so bietet sich für Zähldaten die Poisson-Verteilung an.

Eine Zufallsvariable X mit Werten in $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$ heisst Poisson(λ)-verteilt, falls

$$P(X = x) = \exp(-\lambda)\frac{\lambda^x}{x!} \quad (x = 0, 1, 2, \ldots)$$

wobei $\lambda > 0$ ein Parameter der Verteilung ist.

Die Poisson-Verteilung ist die Standardverteilung für unbeschränkte **Zähldaten**.

Beispiele: Die Poisson(λ)-Verteilung kann bei folgenden Anwendungen gebraucht werden um die Verteilung einer Zufallsvariablen X zu modellieren:
$X =$ Anzahl Schadenmeldungen eines Versicherten pro Jahr
$X =$ Anzahl spontaner Ereignisse in einer Nervenzelle während einer Sekunde
(via Transmitterfreisetzung an einer Synapse)

Die Kennzahlen sind wie folgt: für $X \sim$ Poisson(λ):

$$\mathcal{E}(X) = \lambda, \quad \mathrm{Var}(X) = \lambda, \quad \sigma(X) = \sqrt{\lambda}.$$

2.6.1 Poisson-Approximation der Binomial-Verteilung

Betrachte $X \sim$ Binomial(n, π) und $Y \sim$ Poisson(λ). Falls n gross und π klein mit $\lambda = n\pi$, dann:

$$P(X = x) = \binom{n}{x}\pi^x(1-\pi)^{n-x} \approx P(Y = x) = \exp(-\lambda)\frac{\lambda^x}{x!} \ (x = 0, 1, \ldots, n).$$

Das heisst: für grosse n und kleine π: Binomial$(n, \pi) \approx$ Poisson(λ) für $\lambda = n\pi$. Mit anderen Worten: die Poisson-Verteilung kann interpretiert werden als Verteilung für **seltene Ereignisse bei vielen unabhängigen Versuchen** (selten für einen einzelnen Fall, die Gesamt-Anzahl kann trotzdem gross sein).

Kapitel 3

Statistik für Zähldaten

3.1 Drei Grundfragestellungen der Statistik (Stahel Kap. 7.1)

Die Statistik befasst sich mit dem Schliessen von einer oder mehreren Beobachtungen auf einen (oder mehrere) Parameter in einem wahrscheinlichkeitstheoretischen Modell.

Beispiel (Forts.): Sei $x = 11$ die effektive Anzahl weiblicher gezüchteter Kuhkälber beim Spermasexing (vgl. Kapitel 2.4). Wir fassen $x = 11$ auf als **Realisierung** der Zufallsvariablen $X \sim$ Binom$(12, \pi)$. Wir möchten jetzt Schlüsse ziehen von der Beobachtung $x = 11$ auf den unbekannten Parameter π.

1. Grundfragestellung: Welches ist der zu den Beobachtungen plausibelste Parameterwert? Die Antwort auf diese 1. Grundfrage heisst **(Punkt-)Schätzung**.

2. Grundfragestellung: Sind die Beobachtungen kompatibel (statistisch vereinbar) mit einem vorgegebenen Parameterwert? Die Antwort auf diese 2. Grundfrage heisst **statistischer Test**.

3. Grundfragestellung: Welche Parameterwerte sind mit den Beobachtungen kompatibel (statistisch vereinbar)? Die Antwort auf diese 3. Grundfrage heisst **Konfidenzintervall** oder **Vertrauensintervall**. Das Konfidenzintervall ist allgemeiner und informativer als ein statistischer Test.

Beispiel (Forts.): Im Beispiel von Spermasexing können die Grundfragen so lauten:
1. Welches ist der plausibelste Wert π (zu der Beobachtung $x = 11$)?
2. Ist die Beobachtung $x = 11$ kompatibel mit $\pi = 0.7$?
3. Welcher Bereich (Intervall) für den Parameter π ist mit der Beobachtung $x = 11$ kompatibel?

3.2 Schätzung, statistischer Test und Vertrauensintervall bei Binomial-Verteilung (Stahel Kap. 7.2, 8.2, 9.1, 9.2)

Wir betrachten folgende Situation: gegeben ist eine Beobachtung x, welche als Realisierung von $X \sim$ Binomial(n, π) aufgefasst wird. Wir möchten Schlüsse ziehen über den unbekannten Parameter π.

3.2.1 (Punkt-)Schätzung

Eine Schätzung für π kann pragmatisch hergeleitet werden. Da $\mathcal{E}(X) = n\pi$ (siehe Kapitel 2.5) gilt: $\pi = \mathcal{E}(X)/n$. Der Wert n (Anzahl unabhängiger Versuche) ist als bekannt vorausgesetzt: die einzige Unbekannte ist dann $\mathcal{E}(X)$. Eine pragmatisch motivierte Schätzung ist dann: $\widehat{\mathcal{E}(X)} = x(=$ Beobachtung), d.h. man setzt die Beobachtung gleich dem Erwartungswert. Somit ergibt sich:

$$\hat{\pi} = x/n.$$

3.2.2 Statistischer Test

Beispiel: Es wird 100-mal eine Münze geworfen.
Betrachte $X =$ Anzahl Kopf (K) bei 100 Würfen. Es ist vernünftig, das Modell zu benutzen: $X \sim$ Binomial$(100, \pi)$. Beobachtet (realisiert) wurde $x = 58$. Wir möchten testen, ob die Münze fair ist, d.h. ob $\pi = 1/2$.

Motivation

Im obigen Beispiel stellen wir die folgende Überlegung an. Wir nehmen einmal an, dass die Münze fair ist, d.h. dass $\pi = 1/2$, und berechnen die Wahrscheinlichkeiten für "unplausible" Ereignisse von der Form $\{X \geq c\}$ für "grosse" Werte c. Die Absicht dabei ist zu quantifizieren, ob die beobachtete Anzahl $x = 58$ bereits zu einem "unplausiblem" Ereignis gehört (so dass man schliessen müsste, dass die Münze nicht mehr fair ist, das heisst $\pi > 1/2$). Die folgende Tabelle liefert die Zahlen für $X \sim$ Binomial$(100, 1/2)$.

	$c = 52$	$c = 53$	$c = 54$	$c = 55$	$c = 56$	$c = 57$	$c = 58$	$c = 59$	$c = 60$
$P(X \geq c)$	0.382	0.309	0.242	0.184	0.136	0.097	0.067	0.044	0.028

Typischerweise deklariert man ein Ereignis als "unplausibel" falls dessen Wahrscheinlichkeit weniger oder gleich 5% beträgt. In unserem Beispiel sehen wir, dass die Beobachtung $x = 58$, welche zu dem Ereignis $X \geq 58$ gehört, eine zugehörige Wahrscheinlichkeit von 6.7% hat und deshalb immer noch als plausibel eingestuft wird. Das heisst, dass die Beobachtung $x = 58$ durchaus noch als plausibel bei einer fairen Münze eingestuft werden kann. Hätte man aber 59-mal Kopf beobachtet, so würde man dies als nicht mehr sehr plausibel bei einer fairen Münze einstufen: die zugehörige Wahrscheinlichkeit ist mit 4.4% bereits eher klein. (Natürlich ist die Grenze, welche durch eine Wahrscheinlichkeit von 5% gegeben ist, willkürlich. Später werden wir dies mit dem sogenannten P-Wert charakterisieren, siehe unten.

Formales Vorgehen

Ein statistischer Test für den Parameter π im Modell $X \sim$ Binomial(n, π) ist wie folgt aufgebaut.

1. Spezifiziere die sogenannte **Nullhypothese** H_0:

$$H_0: \ \pi = \pi_0,$$

und (anhand der Problemstellung) eine sogenannte **Alternative** H_A:

$$\begin{aligned} H_A: \quad & \pi \neq \pi_0 \text{ (zwei-seitig)} \\ & \pi > \pi_0 \text{ (ein-seitig nach oben)} \\ & \pi < \pi_0 \text{ (ein-seitig nach unten).} \end{aligned}$$

Beispiel (Forts.): Es wird 100-mal eine Münze geworfen.
Betrachte $X =$ Anzahl Kopf (K) bei 100 Würfen mit dem Modell $X \sim$ Binomial$(100, \pi)$. Wir möchten folgende Frage untersuchen: Ist die Münze fair? Oder ist sie unfair, dass sie eher zu oft auf Kopf (K) landet? Im Formalismus des statistischen Tests heisst das: $H_0: \pi = \pi_0 = 1/2$ und $H_A: \pi > \pi_0 = 1/2$. Insbesondere die Wahl der Alternative muss bezüglich der interessierenden Fragestellung getroffen werden.

2. Lege das sogenannte **Signifikanzniveau** α fest. Typischerweise wählt man $\alpha = 0.05$ (5%) oder auch $\alpha = 0.01$ (1%).

3. Bestimme den sogenannten **Verwerfungsbereich** K. Qualitativ zeigt K in Richtung der Alternative:

$$\begin{aligned} K = [0, c_u] \cup [c_o, n] \quad & \text{falls } H_A: \pi \neq \pi_0, \\ K = [c, n] \quad & \text{falls } H_A: \pi > \pi_0, \\ K = [0, c] \quad & \text{falls } H_A: \pi < \pi_0. \end{aligned}$$

Quantitativ wird K so berechnet, dass

$$P_{H_0}(X \in K) = \underbrace{P_{\pi_0}}_{\text{von Binomial}(n, \pi_0)} (X \in K) \overset{\approx}{\leq} \alpha. \tag{3.1}$$

Beispiel (Forts.): Bei 100-maligem Münzwurf.
$H_0: \pi = 1/2$ (d.h. $\pi_0 = 1/2$) und $H_A: \pi > 1/2$. Für $\alpha = 0.05$ haben wir in der Tabelle oben gesehen, dass $K = [59, 100]$ ist.

4. Erst jetzt betrachte, ob die Beobachtung x in den Verwerfungsbereich K fällt:
falls ja: so verwerfe H_0 (die Alternative ist dann "signifikant")
falls nein: belasse H_0 (was nicht heisst, dass deswegen H_0 statistisch bewiesen ist).
Diese Art der Test-Entscheidung beruht auf dem Widerspruchs-Prinzip: ein statisticher Nachweis/Beweis gelingt bloss dann, wenn die Nullhypothese verworfen werden kann. Diese Art von wissenschaftlicher Evidenz-Findung wurde bereits von Aristoteles in der Antike propagiert.

Beispiel (Forts.): Bei 100-maligem Münzwurf.
Da $x = 58$ effektiv beobachtet wurde: H_0 wird belassen. Das heisst, dass es keine statistische Evidenz gibt (auf dem Signifikanzniveau $\alpha = 0.05$), dass die Münze zu Gunsten von Kopf (K) gefälscht ist.

Beispiel (Forts.): Beim Spermasexing (vgl. Kapitel 2.4) wurden $x = 11$ weibliche Kuhkälber gezüchtet von insgesamt 12 Kälbern. Der Hersteller der Methode behauptet, dass die Erfolgswahrscheinlichkeit grösser als 70% ist. Der Test ist wie folgt.

Modell: $X \sim$ Binomial$(12, \pi)$
$H_0: \ \pi = \pi_0 = 0.7$
$H_A: \ \pi > \pi_0 = 0.7$
Signifikanzniveau: wir wählen $\alpha = 0.05$
Verwerfungsbereich: $P_{\pi=0.7}(X \in K) \overset{\approx}{\leq} 0.05 \leadsto \ K = \{12\}$
Entscheid: H_0 wird belassen, d.h. die Aussage des Herstellers ist nicht signifikant

Fehler 1. und 2. Art

Bei einem statistischen Test treten 2 Arten von Fehlern auf.
Fehler 1. Art: Fälschliches Verwerfen von H_0, obwohl H_0 richtig ist.
Fehler 2. Art: Fälschliches Beibehalten von H_0, obschon die Alternative zutrifft.
Der Fehler 1. Art wird als "schlimmer" betrachtet: er wird direkt kontrolliert mittels der Konstruktion eines Tests: die Formel (3.1) besagt:

$$P(\text{Fehler 1. Art}) = P_{H_0}(X \in K) \overset{\approx}{\leq} \alpha.$$

Das Signifikanzniveau kontrolliert also die Wahrscheinlichkeit für eine Fehler 1. Art. Es gilt aber auch:

$$P(\text{Fehler 2. Art}) \text{ wird grösser falls } \alpha \text{ kleiner gewählt wird.}$$

Die Wahl von α steuert also einen Kompromiss zwischen Fehler 1. und 2. Art. Weil man aber primär einen Fehler 1.Art vermeiden will, wählt man α klein, z.B. $\alpha = 0.05$.

Beispiel (Forts.): Beim Spermasexing nehmen wir einmal an, dass in Tat und Wahrheit der Parameter $\pi = 0.8 \in H_A$ ist (die Spezifikationen des Tests sind wie oben: $H_0: \ \pi = 0.7$, $H_A: \ \pi > 0.7$ und $\alpha = 0.05$). Da der Verwerfungsbereich $K = \{12\}$ ist (siehe oben), gilt dann:

$$P(\text{Test behält } H_0 \text{ bei, obschon } \pi = 0.8) = P_{\pi=0.8}(X \leq 11) = 1 - P_{\pi=0.8}(X = 12) = 0.93.$$

Das heisst, dass ein Fehler 2. Art (unter der Annahme dass $\pi = 0.8$) mit grosser Wahrscheinlichkeit auftritt. Das ist natürlich enttäuschend, aber unvermeidlich bei der kleinen Anzahl von 12 Versuchen. Beachte, dass die Wahrscheinlichkeit für einen Fehler 1. Art $\overset{\approx}{\leq} 0.05$ ist.

Der P-Wert

Die Entscheidung eines Tests mit "Verwerfen" oder "Beibehalten" der Nullhypothese H_0 ist abhängig von der etwas willkürlichen Wahl des Signifikanzniveaus α. Mathematisch bedeutet dies, dass der Verwerfungsbereich $K = K(\alpha)$ abhängig von der Wahl von α ist.

Man kann sich einfach überlegen, dass qualitative Folgendes gilt:

$$\text{Verwerfungsbereich } K = K(\alpha) \text{ wird kleiner mit kleiner werdendem } \alpha,$$

weil bei kleinem α die Wahrscheinlichkeit für einen Fehler 1. Art klein ist (und dies wird natürlich erreicht, wenn es tendenziell schwierig ist eine Nullhypothese H_0 zu verwerfen - sprich Verwerfungsbereich K klein). Umgekehrt gilt natürlich auch, dass $K = K(\alpha)$ grösser wird mit wachsendem α. Dies impliziert: es gibt ein Signifikanzniveau, bei dem die Nullhypothese H_0 "gerade noch" verworfen wird.

Der P-Wert ist definiert als das kleinste Signifikanzniveau,
bei dem die Nullhypothese H_0 (gerade noch) verworfen wird

Der P-Wert kann folgendermassen gerechnet werden: die Beobachtung $X = x$ kommt auf die Grenze des Verwerfungsbereichs $K = K(\text{P-Wert})$ mit Signifikanzniveau = P-Wert zu liegen; siehe auch Abbildung 3.1.

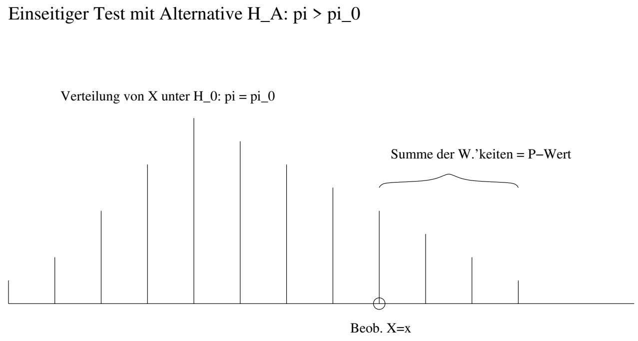

Abbildung 3.1: Schematische Darstellung des P-Werts bei einer einseitigen Alternative $H_A: \ \pi > \pi_0$.

Der P-Wert liefert mehr Information als bloss die Test-Entscheidung bei einem vorbestimmten Signifikanzniveau α (z.B. $\alpha = 0.05$). Insbesondere gilt aufgrund der Definition des P-Werts:

$$\begin{aligned} &\text{verwerfe } H_0 \text{ falls P-Wert} \leq \alpha \\ &\text{belasse } H_0 \text{ falls P-Wert} > \alpha. \end{aligned}$$

Zusätzlich zu dieser Entscheidungsregel quantifiziert der P-Wert wie signifikant eine Alternative ist (d.h. wie gross die Evidenz ist für das Verwerfen von H_0). Sprachlich wird manchmal wie folgt übersetzt:

$$\begin{aligned} \text{P-Wert } &\approx 0.05: \text{ schwach signifikant} \\ \text{P-Wert } &\approx 0.01: \text{ signifikant} \\ \text{P-Wert } &\approx 0.001: \text{ stark signifikant} \\ \text{P-Wert } &\leq 10^{-4}: \text{ äusserst signifikant} \end{aligned}$$

Beispiel (Forts.): Beim Spermasexing betrachten wir die Null-Hypothese $\pi = 0.7$ und die Alternative $\pi > 0.7$. Beobachtet wurde $x = 11$, aufgefasst als Realisierung von $X \sim \text{Binomial}(12, \pi)$. Der P-Wert ist dann:

$$P_{\pi=0.7}(X \geq 11) = P_{\pi=0.7}(X = 11) + P_{\pi=0.7}(X = 12) = 0.085.$$

Wie wir bereits früher gesehen haben, liefert dies kein Verwerfen von H_0 auf dem Signifikanzniveau $\alpha = 0.05$. (Wenn man - aus irgendwelchen Gründen - im voraus das Signifikanzniveau $\alpha = 0.09$ gewählt hätte, so könnte man H_0 auf diesem Signifikanzniveau $\alpha = 0.09$ verwerfen).

3.2.3 Vertrauensintervall

Informativer als ein statistischer Test ist ein sogenanntes Vertrauensintervall (auch Konfidenzintervall genannt). Es liefert eine Antwort auf die 3. Grundfragestellung von Kapitel 3.1: Welche Werte von π sind mit der Beobachtung x kompatibel (statistisch vereinbar).

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

$$I = \{\pi_0;\ \text{Nullhypothese } H_0: \ \pi = \pi_0 \text{ wird belassen}\}. \tag{3.2}$$

Diese Beziehung stellt eine Dualität zwischen Tests und Vertrauensintervall dar.

Die Berechnung kann grafisch, oder mit einer Tabelle erfolgen. Falls n "gross" ist, so kann die sogenannte Normalapproximation (siehe Kap. 4.5) benützt werden. Letztere ergibt folgendes approximatives Konfidenzintervall I zum Niveau $1 - \alpha = 0.95$ für den unbekannten Parameter π:

$$I \approx \frac{x}{n} \pm 1.96\sqrt{\frac{x}{n}(1 - \frac{x}{n})\frac{1}{n}} \tag{3.3}$$

Das Vertrauensintervall $I = I(x)$ hängt von der Beobachtung ab. Wenn man anstelle der Beobachtung die zugehörige Zufallsvariable X einsetzt, so ist $I(X)$ zufällig und hat die Eigenschaft:

$$P(\pi \in I(X)) \overset{\approx}{>} 1 - \alpha.$$

Dies kann so interpretiert werden, dass der wahre Parameter π mit Wahrscheinlichkeit $1 - \alpha$ im Konfidenzintervall I enthalten ist.

Beispiel (Forts.): Beim Spermasexing erhält man für ein zweiseitiges Konfidenzintervall zum Niveau $1 - \alpha = 0.95$ mittels einer Tabelle oder dem Computer für die Berechnung von (3.2):

$$I = (0.615, 0.998)$$

Das heisst, dass der wahre "Zucht"-Parameter π mit einer Wahrscheinlichkeit von 95% in I liegt. Es besteht also auf Grund der kleinen Stichprobe grosse Unsicherheit, wie erfolgreich die Methode bei langfristigem Einsatz tatsächlich sein wird. Die Näherungsformel in (3.3) ist für dieses Beispiel nicht besonders gut, weil $n = 12$ eher klein ist. Man erhält mit (3.3):

$$I \approx (0.760, 1.073)$$

Der rechte Endpunkt ist natürlich zu gross, denn der Parameter π ist ja kleiner oder gleich 1.

3.3 Schätzung, Test und Vertrauensintervall bei Poisson-Verteilung (Stahel, Kap. 7.2, 8.1, 9.1)

Wir betrachten folgende Situation: gegeben ist eine Beobachtung x, welche als Realisierung von $X \sim$ Poisson(λ) aufgefasst wird. Wir möchten Schlüsse ziehen über den unbekannten Parameter λ.

3.3.1 (Punkt-)Schätzung

Da $E(X) = \lambda$ (siehe Kapitel 2.6), und unter Verwendung der pragmatischen Schätzung von $E(X)$ mittels der Beobachtung x, erhält man die Schätzung:

$$\hat{\lambda} = x.$$

3.3.2 Statistischer Test

Ein statistischer Test für den Parameter λ im Modell $X \sim$ Poisson(λ) erfolgt völlig analog zu der Konstruktion bei der Binomial-Verteilung in Kapitel 3.2.2.

1. Spezifiziere die **Nullhypothese** H_0:

$$H_0: \ \lambda = \lambda_0,$$

und (anhand der Problemstellung) eine **Alternative** H_A:

$$\begin{aligned} H_A: \quad & \lambda \neq \lambda_0 \text{ (zwei-seitig)} \\ & \lambda > \lambda_0 \text{ (ein-seitig nach oben)} \\ & \lambda < \lambda_0 \text{ (ein-seitig nach unten).} \end{aligned}$$

2. Lege das **Signifikanzniveau** α fest, zum Beispiel $\alpha = 0.05$.

3. Bestimme den **Verwerfungsbereich** K. Qualitativ zeigt K in Richtung der Alternative:

$$\begin{aligned} K = [0, c_u] \cup [c_o, \infty) \quad & \text{falls } H_A: \ \lambda \neq \lambda_0, \\ K = [c, \infty) \quad & \text{falls } H_A: \ \lambda > \lambda_0, \\ K = [0, c] \quad & \text{falls } H_A: \ \lambda < \lambda_0. \end{aligned}$$

Quantitativ wird K so berechnet, dass

$$P_{H_0}(X \in K) = \underbrace{P_{\lambda_0}}_{\text{von Poisson}(\lambda_0)} (X \in K) \overset{\approx}{\leq} \alpha.$$

4. Erst jetzt betrachte, ob die Beobachtung x in den Verwerfungsbereich K fällt:
falls ja: verwerfe H_0;
falls nein: belasse H_0.

Die Konzepte wie Fehler 1. und 2. Art oder P-Wert sind identisch wie in Kapitel 3.2.2.

3.3.3 Vertrauensintervall

Das Vertrauensintervall I zum Niveau $1 - \alpha$ (siehe auch Kapitel 3.2.3) besteht aus allen Werten λ, die beim zugehörigen Test akzeptiert werden. Manchmal kann auch die folgende Approximation benutzt werden für ein zweiseitiges Konfidenzintervall zum Niveau $1-\alpha = 0.95$:

$$I = I(x) \approx x \pm 1.96\sqrt{x}.$$

Beispiel: Im Jahr 1992 gab es $x = 554$ Tote bei Verkehrsunfällen in der Schweiz. Wir fassen diese Beobachtung auf als eine Realisierung von $X \sim$ Poisson(λ). Die Schätzung ist dann $\hat{\lambda} = 554$ und das approximative Vertrauensintervall ist $I = I(x) \approx (507.9, 600.1)$.

Kapitel 4

Modelle und Statistik für Messdaten

4.1 Einleitung

In vielen Anwendungen hat man es nicht mit Zähl-, sondern mit Messdaten zu tun, bei denen die Werte im Prinzip kontinuierlich sind. Zur Illustration betrachten wir zwei Datensätze. Beim ersten werden zwei Methoden zur Bestimmung der latenten Schmelzwärme von Eis verglichen. Wiederholte Messungen der freigesetzten Wärme beim Uebergang von Eis bei $-0.72°$ C zu Wasser bei $0°$ C ergaben die folgenden Werte (in cal/g):

Methode A	79.98	80.04	80.02	80.04	80.03	80.03	80.04	79.97	80.05	80.03
Methode A	80.02	80.00	80.02							
Methode B	80.02	79.94	79.98	79.97	79.97	80.03	79.95	79.97		

Obwohl die Messungen mit der grösstmöglichen Sorgfalt durchgeführt und alle Störeinflüsse ausgeschaltet wurden, variieren die Messungen von Fall zu Fall. Wir werden diese Variationen innnerhalb der Messreihen als zufällig modellieren, das heisst wir interpretieren diese Werte als Realisierungen von Zufallsvariablen. Wir werden dann die Frage beantworten, ob die Unterschiede zwischen den Methoden ebenfalls als zufällig angesehen werden können, oder ob ein systematischer Unterschied plausibler ist, der auch in der ganzen Population, d.h. in weiteren Messungen, bestehen bleibt. Im letzteren Fall werden wir dann noch zusätzlich angeben, wie gross der systematische Unterschied etwa ist.

Im zweiten Beispiel wurde bei 11 Individuen die Aggregation von Blutplättchen vor und nach dem Rauchen einer Zigarette gemessen. Die folgenden Daten geben den Anteil aggregierter Blutplättchen (in Prozent) nach einer Stimulation an.

Individuum	1	2	3	4	5	6	7	8	9	10	11
Vorher	25	25	27	44	30	67	53	53	52	60	28
Nachher	27	29	37	56	46	82	57	80	61	59	43

Wieder variieren die Werte in einer nicht vorhersehbaren Art. Diesmal handelt es sich jedoch weniger um Messfehler, sondern um Variation zwischen Individuen (vermutlich gäbe es auch noch eine gewisse Variation beim gleichen Individuum, wenn der Test wiederholt würde). Die Aggregation bei diesen 11 Individuen ist meistens, aber nicht immer nach dem Rauchen höher, und die Fragestellung lautet, ob es sich hier um einen zufälligen Effekt handelt, der auf die spezifische Stichprobe beschränkt ist, oder ob man dieses Resultat auf eine grössere Population verallgemeinern kann. Im letzteren Fall möchte man wieder angeben, wie gross die mittlere Zunahme etwa ist.

4.2 Deskriptive Statistik (Stahel, Kap. 2 und 3.1, 3.2)

Bei einer statistischen Analyse ist es wichtig, nicht einfach blind ein Modell anzupassen oder ein statistisches Verfahren anzuwenden. Die Daten sollten immer mit Hilfe von geeigneten grafischen Mitteln dargestellt werden, da man nur auf diese Weise unerwartete Strukturen und Besonderheiten entdecken kann. Kennzahlen können einen Datensatz grob charakterisieren. Im Folgenden werden die Daten mit $x_1, \ldots, x_n$ bezeichnet.

4.2.1 Kennzahlen

Häufig will man die Verteilung der Daten numerisch zusammenfassen. Dazu braucht man mindestens zwei Kenngrössen, eine für die Lage und eine für die Streuung. Die bekanntesten solchen Grössen sind das *arithemtische Mittel* für die Lage,

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

und die *empirische Standardabweichung* für die Streuung,

$$s_x = \sqrt{\text{var}} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2}.$$

(der Nenner $n-1$, anstelle von n, ist mathematisch begründet und hat die Eigenschaft, dass kein "systematischer" Fehler auftritt).

Alternative Kenngrössen sind der *Median* als Lagemass und die *Quartilsdifferenz* als Streuungsmass. Diese werden mit Hilfe von sogenannten Quantilen definiert.

Quantil

Das *empirische α-Quantil* ist anschaulich gesprochen der Wert, bei dem $\alpha \times 100\%$ der Datenpunkte kleiner und $(1-\alpha) \times 100\%$ der Punkte grösser sind.

Zur formalen Definition führen wir die geordneten Werte ein:

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

Das empirische α-Quantil ist dann gleich

$$\begin{array}{ll} \frac{1}{2}(x_{(\alpha n)} + x_{(\alpha n+1)}) & \text{falls } \alpha \cdot n \text{ eine ganze Zahl ist,} \\ x_{(k)} \text{ wobei} k = \text{nächstgrössere ganze Zahl von } \alpha \cdot n; & \text{falls } \alpha \cdot n \text{ keine ganze Zahl ist.} \end{array}$$

Der (empirische) Median ist das empirische 50%-Quantil: d.h., es markiert die "mittlere" Beobachtung und ist also ein Mass für die Lage der Daten.

Die Quartilsdifferenz ist gleich

$$\text{empirisches 75\%-Quantil} - \text{empirisches 25\%-Quantil}$$

und ist ein Streuungsmass für die Daten.

Median und Quartilsdifferenz haben den Vorteil, dass sie robust sind: das heisst, dass sie viel weniger stark durch extreme Beobachtungen beeinflusst werden können als arithmetisches Mittel und Standardabweichung.

Beispiel: Messung der Schmelzwärme von Eis mit Methode A
Aufgrund von $n = 13$ Messungen: das arithmetische Mittel $\overline{x} = 80.02$ und die Standardabweichung $s_x = 0.024$. Ferne, für $n = 13$ ist $0.25n = 3.25$, $0.5n = 6.5$ und $0.75n = 9.75$. Damit ist das 25%-Quantil gleich $x_{(4)} = 80.02$, der Median gleich $x_{(7)} = 80.03$ und das 75%-Quantil gleich $x_{(10)} = 80.04$.

Standardisierung

Durch Verschiebung und Skalierung der Werte kann man erreichen, dass zwei oder mehrere Datensätze die gleiche Lage und Streuung haben. Insbesondere kann man einen Datensatz so standardisieren, dass das arithmetische Mittel gleich Null und die Standardabweichung gleich 1 ist. Dies erreicht man mitels der linear transformierten Variablen

$$z_i = \frac{x_i - \overline{x}}{s_x} \; (i = 1, \ldots, n).$$

Alle Aspekte einer Verteilung, die bei einer Verschiebung oder Skalierung unverändert bleiben, machen die Form der Verteilung aus. Dazu gehört insbesondere die Schiefe (Asymmetrie) der Verteilung, für die es auch Kennzahlen gibt.

4.2.2 Grafische Methoden

Einen Überblick über die auftretenden Werte ergibt das *Histogramm*. Um ein Histogramm zu zeichnen, bildet man Klassen $(c_{k-1}, c_k]$ und berechnet die Häufigkeiten h_k, d.h. die Anzahl Werte in diesem Intervall. Dann trägt man über den Klassen Balken auf, deren Fläche *proportional* zu h_k ist.

Beim Boxplot hat man ein Rechteck, das vom empirischen 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; siehe Abbildung 4.1.

Die *empirische kumulative Verteilungsfunktion* $F_n(\cdot)$ ist eine Treppenfunktion, die links von $x_{(1)}$ gleich null ist und bei jedem $x_{(i)}$ einen Sprung der Höhe $\frac{1}{n}$ hat (falls ein Wert mehrmals vorkommt, ist der Sprung ein Vielfaches von $\frac{1}{n}$). In andern Worten:

$$F_n(x) = \frac{1}{n} \text{Anzahl}\{i \mid x_i \leq x\}.$$

Abbildung 4.2 zeigt die empirische kumulative Verteilungsfunktion für die Messungen der Schmelzwärme von Eis mit Methode A.

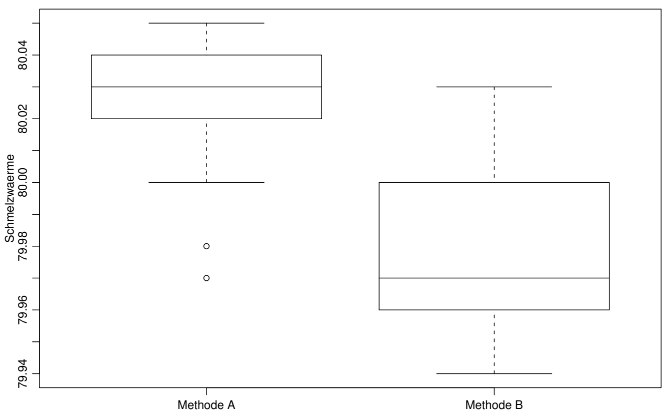

Abbildung 4.1: Boxplots für die zwei Methoden zur Bestimmung der Schmelzwärme von Eis.

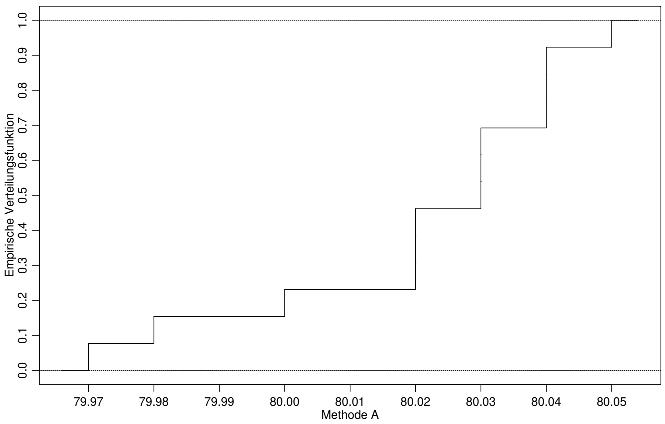

Abbildung 4.2: Empirische kumulative Verteilungsfunktion der Messungen der Schmelzwärme von Eis mit Methode A.

Mehrere Variablen

Wenn wir zwei verschiedene Grössen messen, d.h. wenn die Daten von der Form $(x_1, y_1), \ldots (x_n, y_n)$ sind, interessiert man sich in erster Linie für die Zusammenhänge und Abhängigkeiten zwischen den Variablen. Diese kann man aus dem *Streudiagramm* ersehen, welches die Daten als Punkte in der Ebene darstellt: Die i-te Beobachtung entspricht dem Punkt mit Koordinaten (x_i, y_i). Die Abbildung 4.3 zeigt das Streudiagramm für die Werte "vorher" und "nachher" bei der Blutplättchen-Aggregation. Man sieht einen klaren monotonen Zusammenhang, Individuen haben also eine Tendenz zu starker, bzw. schwacher Aggregation, unabhängig vom Rauchen.

Für die numerische Zusammenfassung der Abhängigkeit ist die **empirische Korrelation** r am gebräuchlichsten:

$$r = \frac{s_{xy}}{s_x s_y}, \quad s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}.$$

Die empirische Korrelation ist eine dimensionslose Zahl zwischen -1 und +1. Das Vorzeichen von r misst die Richtung und der Betrag die Stärke des linearen Zusammenhangs

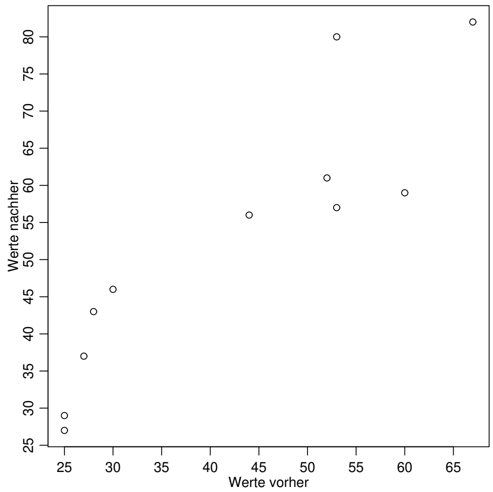

Abbildung 4.3: Streudiagramm der Blutplättchen-Aggregation vor und nach dem Rauchen einer Zigarette.

zwischen den beiden Variablen. Im Fall der Aggregation von Blutplättchen ist die empirische Korrelation gleich 0.9, was den Eindruck vom Streudiagramm bestätigt. Man sollte jedoch nie r berechnen, ohne einen Blick auf das Streudiagramm zu werfen, da ganz verschiedene Strukturen den gleichen Wert von r ergeben können.

4.3 Stetige Zufallsvariablen und Wahrscheinlichkeitsverteilungen (Stahel, Kap. 6.1 – 6.4, 11.2)

Eine Zufallsvariable X heisst stetig, wenn deren Wertebereich W_X kontinuierlich ist; z.B. $W_x = \mathbb{R}$, $\mathbb{R}^+$ oder $[0, 1]$.

In Kapitel 2.3 hatten wir gesehen, dass die Wahrscheinlichkeitsverteilung einer diskreten Zufallsvariablen beschrieben werden kann, indem man die "Punkt"-Wahrscheinlichkeiten $P(X = x)$ für alle möglichen x im Wertebereich angibt. Für eine stetige Zufallsvariable X gilt jedoch:

$$P(X = x) = 0 \text{ für alle } x \in W_X.$$

Dies impliziert, dass wir die Wahrscheinlichkeitsverteilung von X nicht mittels der Angaben von "Punkt"-Wahrscheinlichkeiten beschreiben können.

Die Wahrscheinlichkeitsverteilung einer stetigen Zufallsvariablen X kann jedoch beschrieben werden, indem man die Wahrscheinlichkeiten für alle Intervalle $(a, b]$ $(a < b)$ angibt:

$$P(X \in (a, b]) = P(a < X \leq b).$$

Diese Information ist auch in der kumulativen Verteilungsfunktion $F(x) = P(X \leq x)$ enthalten, denn:

$$P(a < X \leq b) = F(b) - F(a).$$

Zusammenfassend heisst dies, dass die Wahrscheinlichkeitsverteilung einer stetigen Zufallsvariablen X durch die kumulative Verteilungsfunktion beschrieben werden kann.

4.3.1 (Wahrscheinlichkeits-)Dichte

Die Idee einer "Punkt"-Wahrscheinlichkeit $P(X = x)$ kann im Infinitesimalen auch für stetige Zufallsvariablen formuliert werden.

Die (Wahrscheinlichkeits-)Dichte $f(\cdot)$ ist definiert als Ableitung der kumulativen Verteilungsfunktion:

$$f(x) = F'(x).$$

Damit erhalten wir folgende Interpretation:

$$P(x < X \leq x + h) \approx hf(x) \;\; \text{falls } h \text{ klein ist.}$$

Die Begründung dafür ist:

$$P(x < X \leq x + h)/h = (F(x + h) - F(x))/h \approx f(x)$$

wobei die letzte Approximation aus der Definition einer Ableitung folgt.

Weil $F(x) = \int_{-\infty}^{x} f(y)dy$ die Stammfunktion der Dichte ist, gelten die folgenden Eigenschaften:

1. $f(x) \geq 0$ für alle x (da $F(\cdot)$ monoton wachsend ist)
2. $P(a < X \leq b) = \int_a^b f(x)dx$
3. $\int_{-\infty}^{\infty} f(x)dx = 1$ (wegen 2.)

Kennzahlen von stetigen Verteilungen

Der Erwartungswert $\mathcal{E}(X)$ und die Standardabweichung σ_X einer stetigen Zufallsvariablen X haben dieselbe Bedeutung wie im diskreten Fall in Kapitel 2.5; einzig die Berechnungen sehen etwas anders aus. Es gelten:

$$\mathcal{E}(X) = \int_{-\infty}^{\infty} xf(x)dx,$$
$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x - \mathcal{E}(X))^2 f(x)dx, \quad \sigma_X = \sqrt{\mathrm{Var}(X)}.$$

In der frequentistischen Interpretation ist der Erwartungswert eine Idealisierung des arithmetischen Mittels der Werte einer Zufallsvariablen (bei vielen Wiederholungen).

Für den Erwartungswert der transformierten Zufallsvariablen $Y = g(X)$, wobei $g: \mathbb{R} \to \mathbb{R}$ eine Transformation ist, gilt:

$$\mathcal{E}(Y) = \mathcal{E}(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

Damit erhält man, dass

$$\mathrm{Var}(X) = \mathcal{E}\left((X - \mathcal{E}(X))^2\right).$$

Die obige Formel gilt auch für diskrete Zufallsvariablen. Die folgenden **Rechenregeln** (auch für diskrete Zufallsvariablen) erweisen sich oft als nützlich: für beliebige $a, b \in \mathbb{R}$,

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

Die **Quantile** (einer Verteilung von X) $q(\alpha)$ $(0 < \alpha < 1)$ sind wie folgt definiert:

$$P(X \leq q(\alpha)) = \alpha.$$

Das heisst:

$$F(q(\alpha)) = \alpha \Leftrightarrow q(\alpha) = F^{-1}(\alpha).$$

Dies kann auch so interpretiert werden, dass $q(\alpha)$ der Punkt ist, so dass die Fläche von $-\infty$ bis $q(\alpha)$ unter der Dichte $f(\cdot)$ gleich α ist. Siehe auch Abbildung 4.4. Das 50%-Quantil heisst der **Median**.

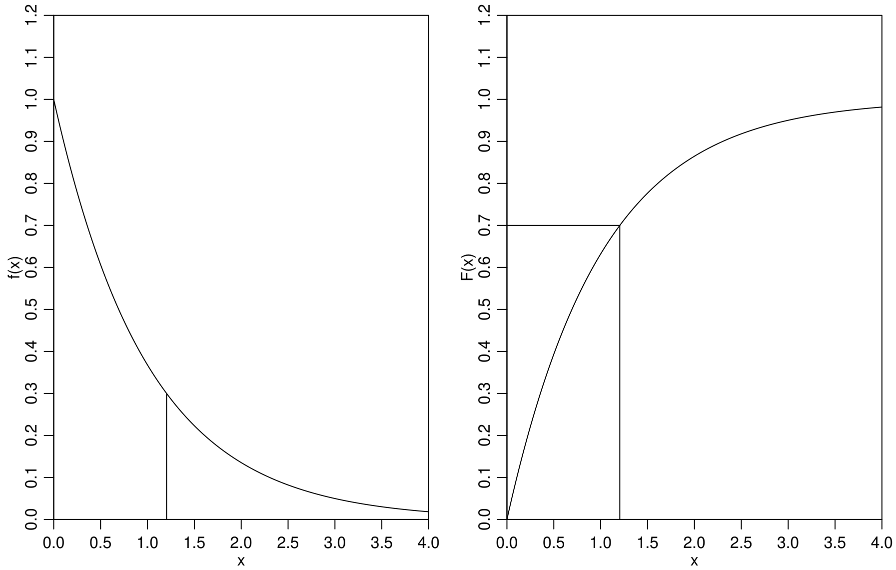

Abbildung 4.4: Illustration des 70%-Quantils $q(0.7)$. Links ist die Dichte gezeichnet, mit der Fläche von $-\infty$ (oder hier auch 0) bis $q(0.7)$ gleich 0.7. Rechts die kumulative Verteilungsfunktion und $q(0.7)$ als Wert der Umkehrfunktion an der Stelle 0.7.

4.4 Wichtige stetige Verteilungen (Stahel, Kap. 6.2, 6.4, 6.5, 11.2)

Wir haben in Kapitel 4.3 gesehen, dass wir die Wahrscheinlichkeitsverteilung einer stetigen Zufallsvariablen mit der kumulativen Verteilungsfunktion $F(\cdot)$ oder der Dichte $f(\cdot)$ charakterisieren können.

4.4.1 Uniforme Verteilung

Die Uniforme Verteilung tritt auf bei Rundungsfehlern und als Formalisierung der völligen "Ignoranz".

Eine Zufallsvariable X mit Wertebereich $W_X = [a, b]$ heisst Uniform($[a, b]$) verteilt, falls

$$f(x) = \begin{cases} 1/(b-a) & \text{falls } a \le x \le b \\ 0 & \text{sonst} \end{cases}$$

Die Dichte ist also konstant auf dem Intervall $[a, b]$. Das heisst, dass die gleiche Wahrscheinlichkeit vorliegt auf dem ganzen Wertebereich $W_X = [a, b]$, deshalb der Name "uniform".

Die zugehörige kumulative Verteilungsfunktion ist

$$F(x) = \begin{cases} 0 & \text{falls } x < a \\ (x-a)/(b-a) & \text{falls } a \le x \le b \\ 1 & \text{falls } x > b \end{cases}$$

Für $X \sim$ Uniform($[a, b]$) sind die Kennzahlen wie folgt:

$$\mathcal{E}(X) = (a+b)/2,$$
$$\text{Var}(X) = (b-a)^2/12, \quad \sigma_X = \sqrt{Var(X)}.$$

4.4.2 Exponential-Verteilung

Die Exponential-Verteilung ist das einfachste Modell für Wartezeiten auf Ausfälle.

Beispiel: Ionenkanäle
In Membranen von Muskel- und Nerven-Zellen gibt es viele Kanäle wo Ionen fliessen können, falls der Kanal offen ist. Simple kinetische Modelle motivieren, dass die Offenzeit eines Kanals mit der Exponential-Verteilung modelliert werden kann.

Eine Zufallsvariable X mit Wertebereich $W_X = \mathbb{R}^+ = [0, \infty)$ heisst Exponential-verteilt mit Parameter $\lambda \in \mathbb{R}^+$ (Exp(λ)) falls

$$f(x) = \begin{cases} \lambda \exp(-\lambda x), & \text{falls } x \ge 0 \\ 0 & \text{sonst} \end{cases}$$

Die zugehörige kumulative Verteilungsfunktion ist

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{falls } x \ge 0 \\ 0 & \text{falls } x < 0 \end{cases}$$

Die Dichte und kumulative Verteilungsfunktion für $\lambda = 1$ sind in Abbildung 4.4 zu sehen.

Für $X \sim$ Exp(λ) sind die Kennzahlen wie folgt:

$$\mathcal{E}(X) = 1/\lambda,$$
$$\text{Var}(X) = 1/\lambda^2, \quad \sigma_X = \sqrt{Var(X)}.$$

Einen Zusammenhang zwischen der Exponential- und Poisson-Verteilung ergibt sich wie folgt. 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.

4.4.3 Normal-Verteilung (Gauss-Verteilung)

Die Normal-Verteilung (manchmal auch Gauss-Verteilung genannt) ist die häufigste Verteilung für Messwerte.

Beispiel: Die Lichtmessungen eines "White Dwarf Sterns" können modelliert werden als Realisierungen von Normal-verteilten Zufallsvariablen.

Eine Zufallsvariable X mit Wertebereich $W_X = \mathbb{R}$ heisst Normal-verteilt mit Parametern $\mu \in \mathbb{R}$ und $\sigma^2 \in \mathbb{R}^+$ ($\mathcal{N}(\mu, \sigma^2)$) falls

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Die zugehörige kumulative Verteilungsfunktion $F(\cdot)$ ist nicht explizit darstellbar: allgemein gilt aber $F(x) = \int_{-\infty}^{x} f(y)dy$.

Für $X \sim \mathcal{N}(\mu, \sigma^2)$ sind die Kennzahlen wie folgt:

$$\begin{aligned} &\mathcal{E}(X) = \mu, \\ &\mathrm{Var}(X) = \sigma^2, \quad \sigma_X = \sqrt{Var(X)}. \end{aligned}$$

Das heisst, dass die Parameter μ und σ^2 eine natürliche Interpretation als Erwartungswert und Varianz der Verteilung haben. Drei Normalverteilungen mit verschiedenen Werten von μ und σ sind in Abbildung 4.5 dargestellt.

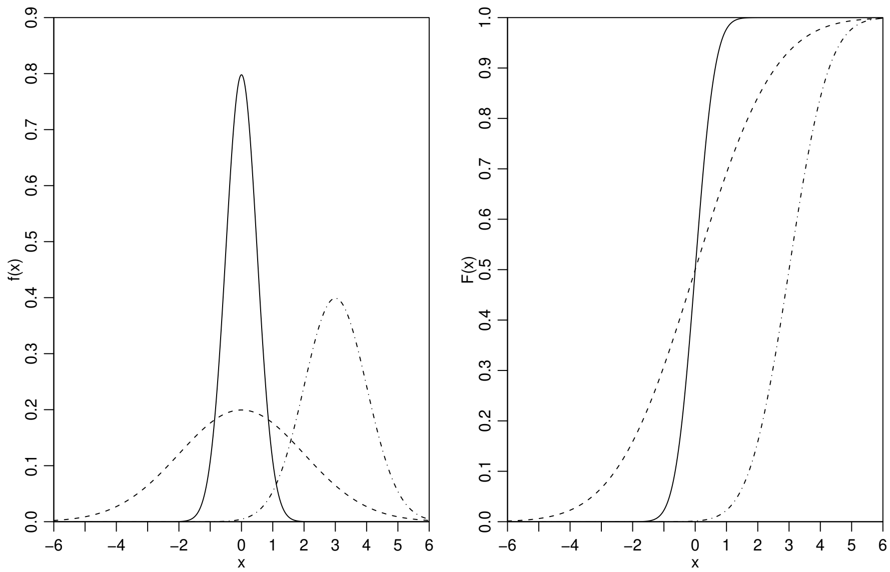

Abbildung 4.5: Dichten (links) und kumulative Verteilungsfunktionen (rechts) der Normalverteilungen mit $\mu = 0, \sigma = 0.5$ (ausgezogen), $\mu = 0, \sigma = 2$ (gestrichelt) und $\mu = 3, \sigma = 1$ (Strich-Punkte).

Die Standard-Normalverteilung

Die Normal-Verteilung mit $\mu = 0$ und $\sigma^2 = 1$ heisst Standard-Normalverteilung. Deren Dichte und kumulative Verteilungsfunktion werden wie folgt bezeichnet:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right),$$
$$\Phi(x) = \int_{-\infty}^{x} \varphi(y)dy.$$

Die Werte von $\Phi(\cdot)$ sind tabelliert. Wir werden unten sehen, dass eine Normal-Verteilung $|calN(\mu, \sigma^2)$ immer in eine Standard-Normalverteilung transformiert werden kann. Deshalb werden die Werte von $\Phi(\cdot)$ genügen um Wahrscheinlichkeiten und Quantile einer allgemeinen $\mathcal{N}(\mu, \sigma^2)$-Verteilung zu berechnen.

4.4.4 Transformationen

Es kann nützlich sein, eine stetige Zufallsvariable X zu transformieren:

$$Y = g(X),$$

wobei $g : \mathbb{R} \to \mathbb{R}$ eine Transformation ist.

Lineare Transformationen

Wir betrachten hier den Fall einer linearen Transformation

$$g(x) = a + bx \quad (a,\ b \in \mathbb{R}).$$

Für $Y = g(X)$ gilt dann (siehe auch Kapitel 4.3.1):

$$\begin{aligned} &\mathcal{E}(Y) = \mathcal{E}(a + bX) = a + b\,\mathcal{E}(X), \\ &\mathrm{Var}(Y) = \mathrm{Var}(a + bX) = b^2\,\mathrm{Var}(X), \quad \sigma_Y = b\sigma_X. \end{aligned} \tag{4.1}$$

Überdies gelten für $b > 0$:

$$\begin{aligned} &\alpha - \text{Quantil von } Y = q_Y(\alpha) = a + bq_X(\alpha), \\ &f_Y(y) = \frac{1}{b} f_X\left(\frac{y-a}{b}\right). \end{aligned} \tag{4.2}$$

Standardisieren einer Zufallsvariablen

Wir betrachten eine stetige Zufallsvariable X. Wir können X immer linear transformieren, so dass die transformierte Zufallsvariable Erwartungswert $= 0$ und Varianz $= 1$ hat. Dies geschieht wie folgt: betrachte die lineare Transformation

$$g(x) = -\frac{\mathcal{E}(X)}{\sigma_X} + \frac{1}{\sigma_X}x$$

und bilde die Transformierte

$$Z = g(X) = \frac{X - \mathcal{E}(X)}{\sigma_X}.$$

Mit Hilfe der Regeln in (4.1) gilt dann: $\mathcal{E}(Z) = 0,\ \mathrm{Var}(Z) = 1$.

Falls $X \sim \mathcal{N}(\mu, \sigma)$, so ist die standardisierte Zufallsvariable

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

Dies folgt aus der Regel in (4.2) für linear transformierte Dichten. Diese lineare Transformation überführt also eine Normal-Verteilung in eine Standard-Normalverteilung. (Allgemeiner gilt: eine linear transformierte Normalverteilung ist wiederum eine Normalverteilung. Diese Eigenschaft, dass man mit linearen Transformationen innerhalb der Verteilungs-Klasse bleibt ist eine spezielle Eigenschaft der Normalverteilung und im Allgemeinen nicht richtig).

Beispiel: Berechnung von Wahrscheinlichkeiten bei $\mathcal{N}(\mu, \sigma^2)$.
Wir betrachten $X \sim \mathcal{N}(2, 4)$ und möchten $P(X \leq 5)$ berechnen. Man geht dann wie folgt vor.

$$P(X \leq 5) = P(\frac{X - 2}{\sqrt{4}} \leq \frac{5 - 2}{\sqrt{4}}) = P(Z \leq 3/2),$$

wobei $Z \sim \mathcal{N}(0, 1)$. Somit hat man:

$$P(X \leq 5) = P(Z < 3/2) = 0.933,$$

wobei man den numerischen Wert mittels einer Tabelle oder Computer bestimmt.

Mit der Regel (4.2) findet man für das 95%-Quantil ((4.2) angewendet mit $a = -\mu/\sigma = 2/2 = 1$ und $b = 1/\sigma = 1/2$):

$$q_X(0.95) = (q_Z(0.95) + 1) \cdot 2 = (\Phi^{-1}(0.95) + 1) \cdot 2 = 5.290.$$

Nichtlineare Transformationen

Wenn $g : \mathbb{R} \to \mathbb{R}$ eine beliebige Transformation ist, so sind viele der oben diskutierten Eigenschaften komplizierter. Die folgende Formel kann aber nützlich sein. Für eine stetige Zufallsvariable X mit Dichte $f_X(\cdot)$ gilt:

$$\mathcal{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

Siehe auch Kapitel 4.3.1.

Eine oft gebrauchte Verteilung ist die **Lognormal-Verteilung**. Wenn $X \sim \mathcal{N}(\mu, \sigma^2)$, dann heisst $Y = \exp(X)$ Lognormal-verteilt mit Parametern $\mu \in \mathbb{R}$ und $\sigma^2 \in \mathbb{R}^+$. Die Lognormal-Verteilung ist nicht mehr symmetrisch und es gilt: $\mathcal{E}(Y) = \exp(\mu + \sigma^2/2)$.

4.4.5 Analogien zwischen Modellen und Daten

Zufallsvariablen und Verteilungen beschreiben die Population, d.h. was mit welcher Wahrscheinlichkeit passieren könnte. Daten $x_1, \ldots, x_n$ interpretieren wir als Realisierungen von Zufallsvariablen $X_1, \ldots, X_n$ (man könnte auch die n Daten als n Realisierungen von einer Zufallsvariablen X interpretieren; die Schreibweise mit mehreren Zufallsvariablen hat jedoch Vorteile, siehe Abschnitt 4.5).

Aus Daten können wir Rückschlüsse auf die zugrunde liegende Verteilung ziehen. Insbesondere haben alle Grössen, die für Zufallsvariablen definiert sind, ein Gegenstück für Datensätze gemäss folgender Tabelle. Die empirischen Grössen sind Schätzungen für die theoretischen Grössen. Diese werden mit wachsendem Stichprobenumfang n immer genauer.

Daten	Population (Modell)
Histogramm	Dichte
empirische kumulative Verteilungsfkt.	theoretische kumulative Verteilungsfkt.
empirische Quantile	theoretische Quantile
Arithmetisches Mittel	Erwartungswert
empirische Standardabweichung	theoretische Standardabweichung.

4.4.6 Überprüfen der Normalverteilungs-Annahme

Oft wollen wir überprüfen ob eine Verteilung ein brauchbares Modell für einen Datensatz darstellt. Das heisst, wir wollen überprüfen, ob ein Datensatz $x_1, \ldots, x_n$ als Realisierungen von einer Zufallsvariablen X mit einer Modell-Verteilung (z.B. mit einer kumulativen Verteilungsfunktion $F(\cdot)$) aufgefasst werden kann.

Im Prinzip kann man das Histogramm der empirischen Daten mit der Dichte der Modell-Verteilung vergleichen. Oft sieht man aber die Abweichungen oder Übereinstimmungen besser, wenn man stattdessen die Quantile benutzt.

Q-Q Plot

Die Idee des Q-Q-Plot (Quantil-Quantil Plot) ist die empirischen Quantile gegen die theoretischen Quantile der Modell-Verteilung zu plotten. Konkret: plotte für $\alpha = 0.5/n, 1.5/n, \ldots,$ $(n - 0.5)/n$ die theoretischen Quantile der Modell-Verteilung $q(\alpha)$ auf der x-Achse gegen die empirischen Quantile, welche den geordneten Beobachtungen $x_{[1]} < x_{[2]} < \ldots < x_{[n]}$ entsprechen, auf der y-Achse. Wenn die Beobachtungen gemäss der Modell-Verteilung erzeugt wurden, sollten diese Punkte ungefähr auf der Winkelhalbierenden $y = x$ liegen.

Normal-Plot

Meist will man nicht eine spezifische Verteilung, sondern eine ganze Klasse von Verteilungen prüfen, also zum Beispiel die Klasse der Normalverteilungen mit beliebigem μ und σ.

Ein Q-Q Plot wo die Modell-Verteilung gleich der Standard-Normalverteilung $\mathcal{N}(0,1)$ ist heisst Normal-Plot.

Falls die Daten Realisierungen von $X \sim \mathcal{N}(\mu, \sigma^2)$ sind, so gilt für die Quantile von X:

$$q(\alpha) = \mu + \sigma\Phi^{-1}(\alpha).$$

siehe (4.2). Wenn man also einen Normal-Plot macht, so sollten die Punkte im Normal-Plot ungefähr auf der Geraden $\mu + \sigma \cdot x$ liegen. Abbildung 4.6 zeigt zwei Normal-Plots: einmal

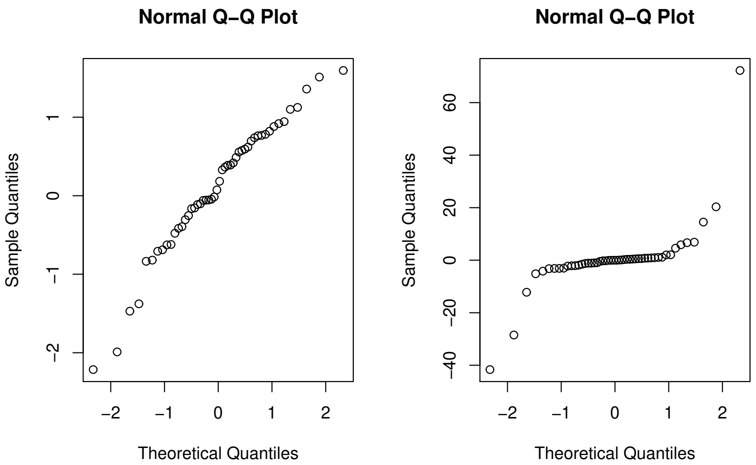

Abbildung 4.6: Links: Normal-Plot für 50 Realisierungen von $\mathcal{N}(0, 1)$. Rechts: Normal-Plot für 50 Realisierungen von Cauchy-Verteilung (sehr langschwänzig).

wo die Daten-generierende Verteilung eine Normal-Verteilung ist, und eine Situation wo die Daten von einer sehr langschwänzigen Verteilung erzeugt sind. Weitere Illustrationen sind in Abbildung 11.2. in Stahel ersichtlich.

4.5 Funktionen von Zufallsvariablen, Fehlerfortpflanzung (Stahel, Kap. 6.8 – 6.11)

In den meisten Anwendungen hat man es nicht mit einer, sondern mit mehreren Zufallsvariablen zu tun. Üblicherweise misst man die gleiche Grösse mehrmals (man hat mehrere Individuen, oder man wiederholt die Messungen).

Die Messungen $x_1, x_2, \ldots, x_n$ fassen wir als Realisierungen der Zufallsvariablen $X_1, \ldots, X_n$ auf. Diese Notation ist oft bequemer als die Interpretation, dass die Messungen als n unabhängige Realisierungen einer Zufallsvariablen X sind. Oft sind wir an Funktionen der Zufallsvariablen $X_1, \ldots, X_n$ interessiert:

$$Y = g(X_1, \ldots, X_n),$$

wobei $g : \mathbb{R}^n \to \mathbb{R}$, und Y wiederum eine Zufallsvariable ist. Wir betrachten hier vor allem die Funktion

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

Beachte die folgende Verbindung: wenn x_i Realisierungen der Zufallsvariablen X_i sind, so ist das arithmetische Mittel der Daten $\overline{x}_n = n^{-1} \sum_{i=1}^{n} x_i$ eine Realisierung der Zufallsvariablen $\overline{X}_n$.

Wir sind hier an der Verteilung der Zufallsvariablen $\overline{X}_n$ interessiert (die Kenntnis dieser Verteilung ist ein wichtiges Werkzeug um später Statistik aufgrund von arithmetischen Mitteln von Daten zu machen). Dazu machen wir üblicherweise die folgende Annahme.

Die i.i.d. Annahme

Oft treffen wir die Annahme, dass die Zufallsvariablen $X_1, \ldots, X_n$ **unabhängig** voneinander sind und dass alle **dieselbe** Verteilung haben. Dafür benutzt man die Notation:

$$X_1, \ldots, X_n \ \text{ i.i.d.}$$

Die Abkürzung i.i.d. steht für: **i**ndependent, **i**dentically **d**istributed.

Beispiel:

$$X_1, \ldots, X_n \ \text{ i.i.d.} \quad \sim \mathcal{N}(\mu, \sigma^2)$$

bedeutet, dass alle X_i's voneinander unabhängig sind und alle dieselbe Normal-Verteilung $\mathcal{N}(\mu, \sigma^2)$ haben.

Kennzahlen und Verteilung von $\overline{X}_n$

Wir nehmen in diesem Abschnitt an, dass

$$X_1, \ldots, X_n \text{ i.i.d.} \quad \sim \ \text{kumulative Verteilungsfkt. } F.$$

Wegen dem zweiten "i" in i.i.d. hat jedes X_i dieselbe Verteilung und dieselben Kennzahlen: $\mathcal{E}(X_i) = \mu$, $\mathrm{Var}(X_i) = \sigma_X^2$.

Zwei Kennzahlen von $\overline{X}_n$ sind dann:

$$\begin{aligned} \mathcal{E}(\overline{X}_n) &= \mu, \\ \mathrm{Var}(\overline{X}_n) &= \frac{\sigma_X^2}{n}. \end{aligned}$$

Der Erwartungswert von $\overline{X}_n$ ist also gleich demjenigen einer einzelnen Zufallsvariablen X_i, die *Varianz nimmt jedoch ab falls n gross ist.* Daraus folgt das folgende Gesetz:

Gesetz der Grossen Zahlen: falls $X_1, \ldots, X_n$ i.i.d. , dann

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

Beispiel: n-maliges Werfen eines Würfels
Betrachte $X_i =$ Augenanzahl im i-ten Wurf. Dann

$$\begin{aligned} \overline{X}_n \quad &= \quad \text{durschnittliche Augenanzahl bei } n \text{ Würfen} \\ &\longrightarrow \quad \mu = \mathcal{E}(X_i) = (1+2+3+4+5+6)/6 = 3.5 \ (n \to \infty). \end{aligned}$$

Das heisst, bei sehr vielen Würfen wird die durschnittliche Augenanzahl nahe bei 3.5 sein.

Die Verteilung von $\overline{X}_n$ ist im allgemeinen schwierig anzugeben. Ein Spezialfall ist der folgende:

$$\overline{X}_n \sim \mathcal{N}(\mu, \sigma_X^2/n) \text{ falls } X_1, \ldots, X_n \text{ i.i.d. } \sim \mathcal{N}(\mu, \sigma_X^2).$$

Falls die einzelnen X_i's nicht normal-verteilt sind, so gilt erstaunlicherweise die obige Verteilungs-Formel immer noch approximativ. Dies liefert der folgende berühmte Satz.

Zentraler Grenzwertsatz: falls $X_1, \ldots, X_n$ i.i.d. , dann

$$\overline{X}_n \approx \mathcal{N}(\mu, \sigma_X^2/n),$$

wobei die Approximation im Allgemeinen besser wird mit grösserem n. Überdies ist auch die Approximation besser, je näher die Verteilung von X_i bei der Normal-Verteilung $\mathcal{N}(\mu, \sigma_X^2)$ ist.

Ebenso ist die standardisierte Zufallsvariable

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma_X}$$

ungefähr $\mathcal{N}(0, 1)$ verteilt.

4.6 Statistik für eine Stichprobe (Stahel, Kap. 8.3 – 8.5, 9.3)

Wir betrachten Daten $x_1, \ldots, x_n$ welche als Realisierungen von $X_1, \ldots, X_n$ i.i.d. aufgefasst werden. Zwei Kennzahlen der Zufallsvariablen X_i sind: $\mathcal{E}(X_i) = \mu$ und $\mathrm{Var}(X_i) = \sigma_X^2$. Typischerweise sind diese (und andere) Kennzahlen unbekannt, und man möchte Rückschlüsse aus den Daten darüber machen.

Beispiel: Blutplättchen-Aggregation (siehe Abschnitt 4.1)
Die Blutplättchen-Aggregation ist ein Beispiel eines sogenannten *gepaarten Vergleichs*, wo man bei jeder Versuchseinheit eine Grösse unter zwei verschiedenen Bedingungen misst. Von Interesse ist, ob ein systematischer Unterschied bezüglich der Aggregation vor und nach dem Rauchen einer Zigarette besteht. Um dies zu untersuchen bilden wir die *Differenzen*: x_i = Aggregation "nachher" - Aggregation "vorher" ($i = 1, \ldots, 11$), und wir haben somit eine (uns interessierende) Stichprobe.

4.6.1 (Punkt-) Schätzungen

Die (Punkt-) Schätzungen für den Erwartungswert und die Varianz sind:

$$\hat{\mu} = n^{-1} \sum_{i=1}^{n} X_i,$$

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

Beachte dass die Schätzer hier als Funktionen der Zufallsvariablen $X_1, \ldots, X_n$ geschrieben sind: insbesondere sind $\hat{\mu}$ und $\hat{\sigma}_X^2$ selbst wieder Zufallsvariablen (die Verteilungs-Eigenschaften von $\hat{\mu}$ wurden in Abschnitt 4.5 diskutiert). Mit der Interpretation, dass die Daten x_i Realisierungen der Zufallsvariablen X_i sind, sind die realisierten Schätzer gleich dem arithmetischen Mittel und der empirischen Varianz der Daten.

4.6.2 Tests für μ

Beispiel: Blutplättchen-Aggregation (Forts.)
Wir wollen testen, ob ein systematischer Unterschied zwischen Aggregation "nachher" und Aggregation "vorher" besteht. Da x_i gerade die Differenz der Aggregationen zwischen "nachher" und "vorher" ist, betrachten wir das folgende Test-Problem:

$$H_0: \ \mu = 0, \quad H_A: \ \mu \neq 0.$$

Um auf den Parameter μ zu testen, machen wir vorerst einmal die Annahme, dass

$$X_1, \ldots, X_n \text{ i.i.d. } \mathcal{N}(\mu, \sigma_X^2). \tag{4.3}$$

Eine Abschwächung dieser Annahme wir später diskutiert.

Der z-Test

Wir nehmen an, dass die Daten $x_1, \ldots, x_n$ Realisierungen von (4.3) sind. Überdies machen wir die Annahme, dass σ_X^2 bekannt ist.

Der z-Test für den Parameter μ ist dann wie folgt.

1. Spezifiziere die Nullhypothese $H_0: \ \mu = \mu_0$
und die Alternative $H_A: \ \mu \neq \mu_0$ (oder "$<$" oder "$>$").

2. Lege das Signifikanzniveau α fest (z.B. $\alpha = 0.05$).

3. Betrachte die **Teststatistik** $\overline{X}_n$. Unter der Nullhypothese gilt (siehe Abschnitt 4.5):

$$\overline{X}_n \ \sim \ \mathcal{N}(\mu_0, \sigma_X^2/n).$$

Der Verwerfungsbereich für die Teststatistik $\overline{X}_n$ für die 2-seitige Alternative $H_A: \mu \neq \mu_0$ ist dann

$$K = (-\infty, \mu_0 - \Phi^{-1}(1-\alpha/2)\sigma_X/\sqrt{n}] \cup [\mu_0 + \Phi^{-1}(1-\alpha/2)\sigma_X/\sqrt{n}, \infty).$$

Somit folgt mit einfacher Umformung, dass

$$P_{H_0}[\overline{X}_n \in K] = P_{\mu_0}[|\overline{X}_n - \mu_0| > \frac{\sigma_X}{\sqrt{n}}\Phi^{-1}(1-\frac{\alpha}{2})] = \alpha,$$

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

4. Verwerfe H_0 falls das arithmetische Mittel der Daten $\overline{x}_n \in K$ (ansonsten belasse H_0).

Zusammenfassend ist der z-Test wie folgt:

$$\text{verwerfe } H_0, \text{ falls} \quad \begin{array}{ll} |\frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\sigma_X}| > \Phi^{-1}(1-\alpha/2) & \text{bei } H_A: \ \mu \neq \mu_0, \\ \frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\sigma_X} < -\Phi^{-1}(1-\alpha) & \text{bei } H_A: \ \mu < \mu_0, \\ \frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\sigma_X} > \Phi^{-1}(1-\alpha) & \text{bei } H_A: \ \mu > \mu_0. \end{array}$$

Im Unterschied zu den Tests in Kapitel 3.2.2 basiert der z-Test auf *mehreren* Beobachtungen. Diese werden aber mittels einer realisierten Teststatistik $\overline{x}_n$, oder auch in der standardisierten Form

$$z = \frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\sigma_X}$$

zusammengefasst (eine Funktion der Daten). Ansonsten sind die Konzepte genau gleich wie in Kapitel 3.2.2: insbesondere brauchen wir für die Bestimmung des Verwerfungsbereichs K die Verteilung der Zufallsvariablen unter der Nullhypothese $H_0: \ \mu = \mu_0$:

$$Z = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{\sigma_X} \sim \mathcal{N}(0,1).$$

Der t-Test

Wir vorhin nehmen wir an, dass die Daten Realisierungen von (4.3) sind. In der Praxis ist die Annahme, dass σ_X bekannt ist, oftmals unrealistisch. Wir können aber die Schätzung

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

benutzen. Dies führt aber zu einer zusätzlichen Unsicherheit, welche berücksichtigt werden muss.

Die Teststatistik beim t-Test ist

$$t = \frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\hat{\sigma}_X},$$

und deren Verteilung unter der Nullhypothese $H_0: \ \mu = \mu_0$ ist

$$T = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{\hat{\sigma}_X} \sim t_{n-1},$$

wobei t_{n-1} eine sogenannte t-Verteilung mit $n-1$ Freiheitsgraden ist.

Die t_ν-Verteilung ist eine symmetrsiche Verteilung um 0, welche langschwänziger ist als die Standard-Normalverteilung $\mathcal{N}(0,1)$. Für $T \sim t_\nu$ gilt:

$$\begin{aligned} \mathcal{E}(T) &= 0 \\ \mathrm{Var}(T) &= \frac{\nu}{\nu - 2}. \end{aligned}$$

Für grosse ν ist t_ν ähnlich zu $\mathcal{N}(0,1)$: insbesondere strebt die t_ν-Verteilung gegen die Standard-Normalverteilung $\mathcal{N}(0,1)$ falls $\nu \to \infty$. Abbildung 4.7 zeigt die Dichte einer t_5-Verteilung.

Zusammenfassend ist der t-Test wie folgt:

$$\begin{aligned} \text{verwerfe } H_0: \ \mu = \mu_0, \text{ falls} \qquad & |t| = |\frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\hat{\sigma}_X}| > t_{n-1,1-\alpha/2} \quad \text{bei } H_A: \ \mu \neq \mu_0, \\ & t = \frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\hat{\sigma}_X} < -t_{n-1,1-\alpha} \quad \text{bei } H_A: \ \mu < \mu_0, \\ & t = \frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\hat{\sigma}_X} > t_{n-1,1-\alpha} \quad \text{bei } H_A: \ \mu > \mu_0, \end{aligned}$$

wobei $t_{n-1;\alpha}$ das α-Quantile der t_{n-1}-Verteilung bezeichnet. Dieses Quantil ist tabelliert (siehe z. B. Stahel, Tabelle 8.5.g, p. 187) oder kann mittels Computer berechnet werden. Es ist etwas grösser als das Quantil der Normalverteilung und ergibt daher einen etwas kleineren Verwerfungsbereich. Für grosse n ist der Unterschied allerdings minim (da $t_{n-1} \approx \mathcal{N}(0,1)$ falls n gross).

Abbildung 4.7 illustriert den Verwerfungsbereich des t-Tests bei $n = 6$ Beobachtungen. Der P-Wert bei 2-seitiger Alternative $H_A: \ \mu \neq \mu_0$ kann wie folgt berechnet werden:

$$P - Wert = 2\left(1 - F_{t_{n-1}}\left(\frac{\sqrt{n}|\overline{x}_n - \mu_0|}{\hat{\sigma}_X}\right)\right),$$

wobei $F_{t_{n-1}}$ die kumulative Verteilungsfunktion der t-Verteilung mit $n-1$ Freiheitsgraden bezeichnet.

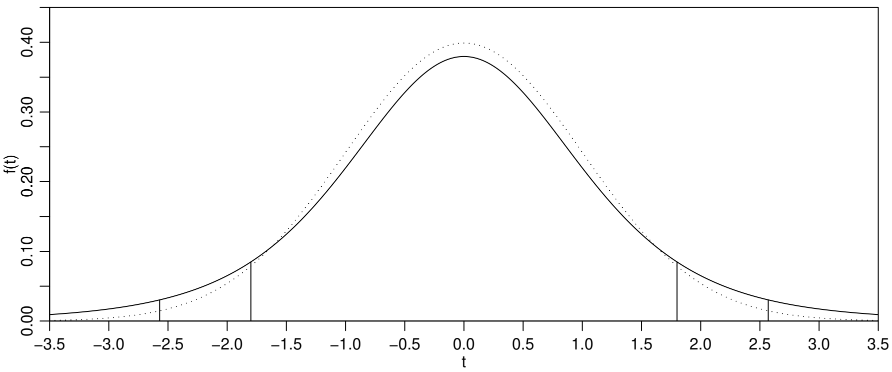

Abbildung 4.7: Dichte der t-Verteilung mit 5 Freiheitsgraden (ausgezogen) und der Normalverteilung (gestrichelt). Die Flächen ausserhalb der vertikalen Linien sind gleich 2.5% bzw. gleich dem halben P-Wert für einen hypothetischen Datensatz mit $\sqrt{6}|\overline{x}_6 - \mu_0|/\hat{\sigma}_X = 1.8$.

Beispiel (Forts.) Blutplättchen-Aggregation (siehe Abschnitt 4.1)
Wir betrachten die Differenzen x_i = Aggregation "nachher" - Aggregation "vorher" ($i = 1, \ldots, 11$) und fassen diese auf als i.i.d. Realisierungen von $\mathcal{N}(\mu, \sigma_X^2)$. Die interessierende Nullhypothese und Alternative ist $H_0: \ \mu = \mu_0 = 0$ und $H_A: \ \mu > \mu_0 = 0$. Die realisierte Teststatistik ist

$$\frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\hat{\sigma}_X} = 4.27$$

und das relevante Quantil für $\alpha = 0.05$ ist $t_{10;0.95} = 1.812$. Die Test-Entscheidung ist also: verwerfe H_0 auf dem 5% Signifikanz-Niveau. Der P-Wert ist

$$P_{H_0}[T > 4.27] = 1 - F_{10}(4.27) = 0.00082.$$

Dies bedeutet, dass der Einfluss von dem Rauchen einer Zigarette bezüglich der Blutplättchen-Aggregation hoch signifikant ist.
Wenn man anstelle der einseitigen die zweiseitige Alternative $H_A: \ \mu \neq \mu_0 = 0$ spezifiziert, so sind die Resultate wie folgt: das relevante Quantil für $\alpha = 0.05$ ist $t_{10;0.975} = 2.23$. Die Test-Entscheidung bleibt dieselbe: verwerfe H_0 auf dem 5% Signifikanz-Niveau. Der P-Wert ist

$$P_{H_0}[|T| > 4.27] = 2(1 - F_{10}(4.27)) = 0.0016,$$

4.6.3 Vertrauensintervall für μ

Analog wie bei Zähldaten in Kapitel 3.2.3 besteht das Vertrauensintervall aus denjenigen Werten μ, bei denen der entsprechende Test nicht verwirft.

Wir nehmen wiederum an, dass die Daten Realisierungen von (4.3) sind. Dies fuhrt dann auf die folgenden zweiseitigen Vertrauensintervalle (die dazugehörigen Tests sind zweiseitig mit Alterative $H_A: \mu \neq \mu_0$) zum Niveau $1 - \alpha$:

$$\begin{aligned} &\overline{x}_n \pm \Phi^{-1}(1 - \alpha/2)\ \frac{\sigma_X}{\sqrt{n}} \quad && \text{falls } \sigma_X \text{ bekannt,} \\ &\overline{x}_n \pm t_{n-1,1-\alpha/2}\ \frac{\hat{\sigma}_X}{\sqrt{n}} && \text{falls } \sigma_X \text{ unbekannt.} \end{aligned}$$

Beispiel (Forts.): Aggregation von Blutplättchen
Wir haben 10 Freiheitsgrade und $t_{10,0.975} = 2.23$. Das zweiseitige Konfidenzintervall für die Erhöhung der Bluttpplättchen-Aggregation nach dem Rauchen einer Zigarette ist somit (%-ige Zunahme)

$$I = 10.27 \pm 2.23 \cdot 7.9761/\sqrt{11} = [4.91, 15.63].$$

Insbesondere ist die Null nicht im Intervall I: das heisst, der Wert $\mu = 0$ ist nicht mit den Daten kompatibel (was wir bereits vom t-Test (siehe oben) wissen).

4.6.4 Tests für μ bei nicht-normalverteilten Daten

Der z- und t-Test sind optimal falls die Daten Realisierungen von normalverteilten Zufallsvariablen sind wie in (4.3). Optimalität bedeutet hier, dass dies die Tests sind, welche die beste Macht (-Kurve) haben, siehe unten.

Wir betrachten hier die allgemeinere Situation, wo die Daten Realisierungen sind von

$$X_1, \ldots, X_n \text{ i.i.d. }, \tag{4.4}$$

wobei X_i eine beliebige Verteilung hat. Wir bezeichen mit μ einen Lageparameter der Verteilung (z.B. $\mu =$ Median der Verteilung von X_i). Die Nullhypothese ist von der Form $H_0: \ \mu = \mu_0$.

Die Macht eines Tests

Wir haben in Kapitel 3.2.2 gesehen, dass es zwei Fehler für einen Test gibt:

$$\text{Fehler 1. Art} = \text{fälschliches Verwerfen von } H_0\text{, obschon } H_0 \text{ richtig ist,}$$

und

$$\text{Fehler 2. Art}(\mu) = \text{(fälschliches) Beibehalten von } H_0 \text{ falls } \mu (\in H_A) \text{ richtig ist.}$$

Die Wahrscheinlichkeit für einen Fehler 1. Art ist gerade gleich α; beim Fehler 2. Art (μ) betrachtet man oft die Macht:

$$\text{Macht}(\mu) = 1 - P(\text{Fehler 2. Art } (\mu)) = P(\text{Verwerfen von } H_0 \text{ falls } \mu \text{ stimmt}).$$

Für $\mu \in H_A$ kann man die Macht(μ) interpretieren als die Chance, dass man richtigerweise H_A entdeckt falls $\mu \in H_A$ stimmt. Für eine Teststatistik T und einen dazugehörigen Verwerfungsbereich K gilt dann:

$$\begin{aligned} &P_{\mu_0}(T \in K) = \alpha, \\ &\text{Macht}(\mu) = P_\mu(T \in K). \end{aligned}$$

Der Vorzeichen-Test

Wir betrachten die Situation wo die Daten Realisierungen von (4.4) sind, wobei die einzelnen X_i nicht-normalverteilt sind. Wir bezeichnen hier mit $\mu =$ Median der Verteilung von X_i; im Falle einer symmetrischen Verteilung ist $\mu = \mathcal{E}(X_i)$.

Der Vorzeichen-Test benutzt die folgende Teststatistik:

$$V = \text{ Anzahl } X_i\text{'s mit } (X_i > \mu_0).$$

Beachte dass $V =$ Anzahl positiver Vorzeichen von $(X_i - \mu_0)$ was den Namen des Tests erklärt.

Wir betrachten die Nullhypothese $H_0 : \ p = P(X_i > \mu_0) = 1/2$ und die Alternative $H_A : \ p \neq 1/2$ (oder einseitige Versionen). Somit ist unter der Nullhypothese H_0 die Teststatistik V folgendermassen verteilt:

$$V \sim \text{ Binomial}(n, 1/2),$$

und der Vorzeichentest wird somit zum Test für den Parameter p bei einer Binomialverteilung.

Der Vorzeichentest stimmt immer, falls die Daten Realisierungen von (4.4) sind: das heisst, die Wahrscheinlichkeit für einen Fehler 1. Art ist kontrolliert durch α bei beliebiger Verteilung der X_i's. Für den z- und t-Test stimmt dies nicht: wegen dem Zentralen Grenzwertsatz wird aber die Wahrscheinlichkeit für einen Fehler 1. Art approximativ kontrolliert durch α, zumindest falls n gross ist.

Die Macht des z- oder t-Tests wird aber im Allgemeinen schnell schlecht, wenn die X_i's in (4.4) nicht mehr normalverteilt sind. Deshalb hat bei nicht-normalverteilten Daten (nicht in guter Approximation normalverteilt) der Vorzeichentest oftmals eine höhere Macht als der z- oder t-Test. Ein Nachteil des Vorzeichentests ist, dass er die Information nicht ausnützt, um wieviel die X_i von dem Wert μ_0 abweichen (siehe die Definition der Teststatistik V oben).

Beispiel (Forts.): Bluttplättchen-Aggregation
Die Nullhypothese ist $H_0: \mu = \mu_0 = 0$. Die realisierte Teststatistik ist dann $v = 10$ und der P-Wert bei einseitiger Alternative $H_A: \mu > \mu_0 = 0$ ist 0.005 (beim t-Test war der P-Wert = 0.00082).

Der Wilcoxon-Test

Der Wilcoxon-Test ist ein Kompromiss, der keine Normalverteilung voraussetzt wie der t-Test und die Information der Daten besser ausnützt als der Vorzeichen-Test.

Die Voraussetzung für den Wilcoxon-Test ist: die Daten sind Realisierungen von (4.4) wobei die Verteilung der X_i's stetig und symmetrisch ist (symmetrische Dichte um $\mu = \mathcal{E}(X_i)$). Der P-Wert bei ein- oder zwei-seitiger Alternative kann mittels Computer berechnet werden.

Der Wilcoxon-Test ist in den allermeisten Fällen vorzuziehen: er hat in vielen Situationen oftmals bessere Macht als der t- und als der Vorzeichen-Test. 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. 4.4.6) grafisch überprüfen.

Beispiel (Forts.): Bluttplättchen-Aggregation
Die Nullhypothese ist $H_0: \mu = \mu_0 = 0$. Der P-Wert bei einseitiger Alternative $H_A: \mu > \mu_0 = 0$ ist 0.002528.

4.7 Tests bei zwei unabhängigen Stichproben (Stahel, Kap. 8.8)

Oft möchte man einen Vergleich zweier Methoden (Gruppen, Versuchsbedingungen, Behandlungen) hinsichtlich der Lage der Verteilung machen.

4.7.1 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.

Beispiel:
Datensatz zu latenter Schmelzwärme von Eis in Kapitel 4.1.

Andererseits liegt eine **gepaarte Stichprobe** vor, wenn **beide Versuchsbedingungen an derselben Versuchseinheit eingesetzt** werden. Die Daten sind dann von der folgenden Struktur:

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

Notwendigerweise ist dann die Stichprobengrösse n für beide Versuchsbedingungen dieselbe.

Beispiel:
Datensatz zu Blutplättchen-Aggregation, siehe Kapitel 4.1.

4.7.2 Gepaarte Tests

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

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

welche wir als Realisierungen von i.i.d. Zufallsvariablen $U_1, \dots, U_n$ auffassen. Kein Unterschied zwischen den beiden Versuchsbedingungen heisst dann einfach $E[U_i] = 0$ (oder auch $\text{Median}(U_i) = 0$). Tests dafür sind in Kapitel 4.6 beschrieben. Dabei ist zu beachten, dass die vorausgesetzte Symmetrie für die Verteilung von U_i beim Wilcoxon-Test immer gilt unter der Nullhypothese, dass X_i und Y_i dieselbe Verteilung haben.

4.7.3 Ungepaarte Tests

Bei ungepaarten Stichproben hat man Daten $x_1, \dots, x_n$ und $y_1, \dots, y_m$ (siehe Kapitel 4.7.1), welche wir als Realisierungen der folgenden Zufallsvariablen auffassen:

$$\begin{aligned} &X_1, \dots, X_n \ \text{ i.i.d. },\\ &Y_1, \dots, Y_m \ \text{ i.i.d. }, \end{aligned} \tag{4.5}$$

wobei auch alle X_i's von allen Y_j's unabhängig sind.

4.7.4 Zwei-Stichproben t-Test bei gleichen Varianzen

Das einfachste Problem lässt sich unter folgender Annahme an (4.5) lösen:

$$\begin{aligned} &X_1, \dots, X_n \text{ i.i.d. } \sim \mathcal{N}(\mu_X, \sigma^2),\\ &Y_1, \dots, Y_m \text{ i.i.d. } \sim \mathcal{N}(\mu_Y, \sigma^2). \end{aligned} \tag{4.6}$$

Die interessierende Null-Hypothese ist

$$H_0: \ \mu_X = \mu_Y.$$

Der Zwei-Stichproben t-Test (bei gleichen Varianzen) verwirft dann die Nullhypothese $H_0: \mu_X = \mu_Y$, falls

$$|T| = \frac{|\overline{X}_n - \overline{Y}_m|}{S_{pool}\sqrt{1/n + 1/m}} > t_{n+m-2,1-\alpha/2} \text{ bei Alternative } H_A: \mu_X \neq \mu_Y,$$
$$T = \frac{\overline{X}_n - \overline{Y}_m}{S_{pool}\sqrt{1/n + 1/m}} > t_{n+m-2,1-\alpha} \text{ bei Alternative } H_A: \mu_X > \mu_Y,$$
$$T = \frac{\overline{X}_n - \overline{Y}_m}{S_{pool}\sqrt{1/n + 1/m}} < t_{n+m-2,1-\alpha} \text{ bei Alternative } H_A: \mu_X < \mu_Y.$$

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 der Teststatistik T ergibt sich aus

$$\mathrm{Var}(\overline{X}_n - \overline{Y}_m) = \sigma^2(\frac{1}{n} + \frac{1}{m}). \tag{4.7}$$

Beweis von (4.7):
1. $\overline{X}_n$ und $\overline{Y}_m$ sind unabhängig, weil alle X_i's von allen Y_j's unabhängig sind.
2. Wegen der Unabhängigkeit von $\overline{X}_n$ und $\overline{Y}_m$ gilt:
$\mathrm{Var}(\overline{X}_n - \overline{Y}_m) = \mathrm{Var}(\overline{X}_n) + \mathrm{Var}(-\overline{Y}_m) = \mathrm{Var}(\overline{X}_n) + \mathrm{Var}(\overline{Y}_m)$.
3. $\mathrm{Var}(\overline{X}_n) = \sigma^2/n$ und $\mathrm{Var}(\overline{Y}_m) = \sigma^2/m$.
Somit ist mit Schritt 2: $\mathrm{Var}(\overline{X}_n - \overline{Y}_m) = \sigma^2(1/n + 1/m)$. □

Die Herleitung des Zwei-Stichproben t-Tests ist wie folgt. Man ersetzt die unbekannte Differenz $\mu_X - \mu_Y$ durch die Schätzung $\overline{X}_n - \overline{Y}_m$ und beurteilt, ob diese Schätung "nahe bei" 0 liegt ("weit weg von" 0 würde Evidenz für H_A bedeuten). Dies wird so quantifiziert, dass man durch die Wurzel der geschätzten Varianz dividiert, und dies als Teststatistik benutzt:

$$T = \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\widehat{\mathrm{Var}}(\overline{X}_n - \overline{Y}_m)}} = \frac{\overline{X}_n - \overline{Y}_m}{S_{pool}\sqrt{1/n + 1/m}}.$$

Unter der Annahme (4.6) und der Null-Hypothese $\mu_X = \mu_Y$ gilt dann:

$$T \ \sim \ t_{n+m-2}.$$

Somit kommt man zu der oben angegbenen Entscheidungsregel, analog zum t-Test für eine Stichprobe, siehe Kapitel 4.6.2.

Beispiel: Schmelzwärme von Eis, siehe Kapitel 4.1.
Die Null-Hypothese sei $H_0: \ \mu_X = \mu_Y$ und wir betrachten die Alternative $H_A: \ \mu_X \neq \mu_Y$. Die Kennzahlen des Datensatzes sind: $\overline{x}_{13} = 80.021$, $\overline{y}_8 = 79.979$, $s^2_{pool} = 7.2 \ 10^{-4}$. Damit hat die Testgrösse den Wert 3.47 was deutlich grösser ist als das 97.5% Quantil $t_{19,0.975} = 2.093$.

4.7.5 Weitere Zwei-Stichproben-Tests

Zwei-Stichproben t-Test bei ungleichen Varianzen

Anstelle der Annahme in (4.6) gelte:

$$\begin{aligned} X_1, \ldots, X_n \text{ i.i.d. } &\sim \mathcal{N}(\mu_X, \sigma_X^2), \\ Y_1, \ldots, Y_m \text{ i.i.d. } &\sim \mathcal{N}(\mu_Y, \sigma_Y^2). \end{aligned}$$

Die Verallgemeinerung des Zwei-Stichproben t-Tests für ungleiche Varianzen $\sigma_X^2 \neq \sigma_Y^2$ ist in der Literatur zu finden und in vielen statistischen Programmen implementiert.

Zwei-Stichproben Wilcoxon-Test (Mann-Whitney Test)

Die Voraussetzungen für den Zwei-Stichproben Wilcoxon-Test, manchmal auch Mann-Whitney Test genannt, bezüglich (4.5) sind wie folgt:

$$\begin{aligned} X_1, \ldots, X_n \text{ i.i.d. } &\sim \text{ beliebige kumulative Verteilungsfunktion } F(\cdot), \\ Y_1, \ldots, Y_m \text{ i.i.d. } &\sim F(\cdot - \delta). \end{aligned}$$

Dies bedeutet, dass die Verteilung von Y_j die um δ verschobene Verteilung von X_i ist, denn: $P(Y_j \leq x + \delta) = F_Y(x + \delta) = F_X(x + \delta - \delta) = F_X(x) = P(X_i \leq x)$.

Die Berechnung des P-Werts eines Zwei-Stichproben Wilcoxon-Tests kann mittels Computer erfolgen. Aus den gleichen Gründen wie im Fall einer Stichprobe (siehe Kapitel 4.6.4) ist der Wilcoxon-Test im Allgemeinen dem t-Test vorzuziehen.

4.8* Versuchsplanung (Stahel, Kap. 14.1 - 14.2)

Genauso wichtig wie die Auswertung der Daten sind Überlegungen, wie man die Daten gewinnen soll. Wir haben bisher vor allem Vergleiche zwischen zwei Behandlungen besprochen (gepaart oder ungepaart). Man soll dabei nie eine neue Behandlung vergleichen mit Resultaten für die Standardbehandlung aus früheren Studien. Es braucht immer eine **Kontrollgruppe** in der gleichen Studie, die sich möglichst wenig von der Gruppe mit der neuen Behandlung unterscheidet. Dann stellt sich natürlich die Frage, wie man die Zuordnung zu den beiden Gruppen durchführen soll. Im gepaarten Fall muss man analog entscheiden, in welcher Reihenfolge man die beiden Behandlungen durchgeführt werden. Systematische Unterschiede zwischen den Gruppen, bzw. systematische Effekte der Reihenfolge kann man am besten vermeiden, wenn man die Zuordnung zufällig macht (sogenannte **Randomisierung**). Zufällig heisst dabei nicht willkürlich, sondern mit Hilfe von Zufallszahlen.

Ein weiterer wichtiger Punkt ist, dass das Experiment wenn möglich **doppelblind** sein soll. Das heisst, dass weder die Person, welche die Behandlung durchführt oder beurteilt, noch die Versuchsperson die Gruppenzugehörigkeit kennen. Dies ist nötig, um Nebeneffekte auszuschalten (nicht die Behandlung wirkt, sondern der mit der Behandlung verbundene Aufwand).

Nicht immer ist ein randomisiertes, doppelblindes Experiment möglich (aus ethischen oder praktischen Gründen). Dies erschwert die Auswertung und Interpretation unter Umständen gewaltig, weil man Störeffekte praktisch nicht ausschliessen kann. Ein bekanntes Beispiel ist der Zusammenhang zwischen Rauchen und Lungenkrebs, der lange umstritten war, weil die genetische Veranlagung und der Effekt des Lebensstils nicht auszuschliessen waren.

Kapitel 5

Regression

Dieses Kapitel ist provisorisch und wird nochmals modifizert!

5.1 Von der Ausgleichsrechnung zum Modell der linearen Regression

Ähnliche Beispiele wie das folgende aus der Chemie treten in sehr vielen Anwendungen auf. Die Dimerisation von 1,3-Butadien läuft nach einem Reaktionsmodell 2. Ordnung ab und ist deshalb charakterisiert durch die Gleichung

$$\frac{d}{dt}C(t) = -\kappa C(t)^2 ,$$

wobei C hier den Partialdruck des Edukts und t die Zeit bedeutet. Diese Gleichung hat Lösungen der Form

$$\frac{1}{C(t)} = \frac{1}{C(0)} + \kappa t .$$

Messungen zu verschiedenen Zeitpunkten im Ablauf der Reaktion sind in Abbildung 5.1 dargestellt. Die letzte Gleichung zeigt, dass der Kehrwert des Partialdruckes linear von der Zeit abhängen sollte. Wegen zufälligen Messfehlern und kleinen systematischen Abweichungen vom einfachen Modell liegen die Punkte nicht genau auf einer Geraden.

Das allgemeine Problem, das wir hier behandeln, lässt sich wie folgt formulieren. Man möchte ausgehend von Beobachtungen angeben, wie eine **Zielgrösse** y (engl. target variable) von **Ausgangsgrössen** (auch Versuchsbedingungen oder **erklärende Variablen**, engl. explanatory variables, genannt) abhängt. Wir nehmen an, dass wir m solche Ausgangsgrössen haben, die wir mit $x^{(1)}$, $x^{(2)}$, $\ldots$, $x^{(m)}$ bezeichnen. Im Idealfall sollte also

$$y_i = h(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)})$$

für jede Beobachtung i gelten. Wir nehmen als einfachste funktionale Beziehung eine lineare, das heisst

$$h(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}) = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_m x_i^{(m)}.$$

Die Koeffizienten β_j $(j = 0, 1, \ldots, m)$ sind dabei in aller Regel unbekannt und müssen aus den Daten geschätzt werden. Im obigen Beispiel war die Zielgrösse der Kehrwert des Partialdrucks von Butadien, die erklärende Variable die Zeit, $\beta_0 = 1/C(0)$ und $\beta_1 = \kappa$.

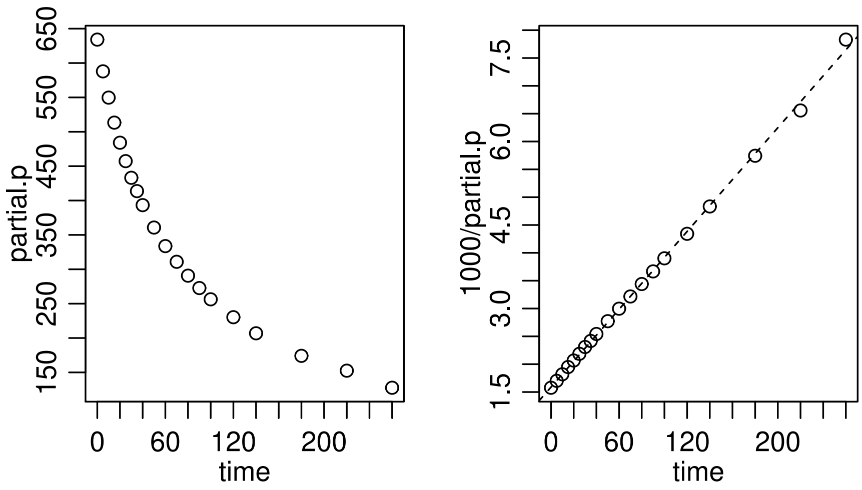

Abbildung 5.1: Partialdruck von Butadien (links) und Kehrwert davon (rechts), gegen die Zeit aufgetragen

Wenn die Anzahl n der Beobachtungen grösser ist als die Anzahl $p = m + 1$ unbekannter Koeffizienten, ist es nicht möglich, auf diese Weise einen perfekten Zusammenhang herzustellen (wir haben ein lineares Gleichungssystem mit mehr Gleichungen als Unbekannten). Das Beste, was wir tun können, ist daher, die Koeffizienten β_j so zu bestimmen, dass die Abweichungen r_i (auch **Residuen** oder Restfehler genannt),

$$r_i = y_i - (\beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_m x_i^{(m)}),$$

in einem vernünftigen Sinne minimal sind. Dies ist ein sogenanntes **Ausgleichsproblem**, vgl. Kap. 5.1 im Buch "Lineare Algebra" von Nipp und Stoffer. Die Ausgleichsrechnung nach Gauss misst die Grösse des Residuenvektors $r = (r_1, \ldots, r_n)^T$ mit der euklidischen Norm $\|r\| = \sqrt{\sum_{i=1}^n r_i^2}$. Für die Stelle, an der das Minimum angenommen wird, spielt es keine Rolle, ob wir $\|r\|$ oder $\|r\|^2$ anschauen. Das heisst wir bestimmen $\beta_0, \ldots, \beta_m$ so, dass

$$S(\beta) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_m x_i^{(m)}))^2$$

minimal wird. Den Vektor der optimalen Parameter, für die $S(\beta)$ minimal wird, bezeichnen wir mit $\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_m)^T$. Dieses Verfahren heisst **Methode der Kleinsten Quadrate**.

Die Bestimmung von $\hat{\beta}$ ist einfach. Der analytische Ansatz bildet die partiellen Ableitungen von $S(\beta)$ und setzt diese gleich Null. Der geometrische Ansatz betrachtet die n-Tupel $y = (y_1, \ldots, y_n)^T$, $x^{(0)} = (1, \ldots, 1)^T$ und $x^{(j)} = (x_1^{(j)}, \ldots, x_n^{(j)})^T$ als Vektoren im n-dimensionalen Raum. Das Minimierungsproblem besteht dann darin, denjenigen Vektor im Unterraum aufgespannt von $x^{(0)}, \ldots x^{(m)}$ zu finden, der am nächsten bei y ist. Dies ergibt den Vektor $\hat{y}$ der angepassten Werte $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i^{(1)} + \ldots + \hat{\beta}_m x_i^{(m)}$ als die orthogonale Projektion von y auf diesen Unterraum.

Bei beiden Ansätze bekommt man die optimalen Parameter $\hat{\beta}$ als Lösung eines linearen Gleichungssystems, der sogenannten Normalgleichungen, die kompakt wie folgt geschrieben werden können

$$(X^T X)\hat{\beta} = X^T y.$$

Dabei ist X eine $n \times (m+1)$-Matrix mit Spaltenvektoren $x^{(j)}$ $(j = 0, \ldots, m)$. Daher ist $X^T X$ eine symmetrische $(m+1) \times (m+1)$-Matrix, und man kann zeigen, dass sie invertierbar ist, wenn die Vektoren $x^{(j)}$ linear unabhängig sind. Man soll aber die optimalen Parameter $\hat{\beta}$ nicht durch Invertieren von $X^T X$ oder durch Lösen der Normalgleichungen bestimmen. Es gibt numerisch bessere Algorithmen, die auf der sogenannten QR-Zerlegung beruhen (in der Praxis wird man ohnehin Software benutzen, die hoffentlich die besten numerischen Verfahren implementiert hat).

Besonders einfach ist die Ausgleichsrechnung, wenn die Spaltenvektoren $x^{(j)}$ orthogonal sind, d.h. wenn für alle $k \neq j$ gilt

$$\sum_{i=1}^{n} x_i^{(j)} x_i^{(k)} = 0.$$

Dann ist nämlich $(X^T X)$ eine Diagonalmatrix, wie man sofort aus der Definition der Matrixmultiplikation sieht, und jedes $\hat{\beta}_j$ wird daher nur von $x^{(j)}$ und y bestimmt. Ohne die Orthogonalität ist das nicht der Fall: $\hat{\beta}_j$ hängt im Allgemeinen auch von $x^{(k)}$ ab für $k \neq j$, und wenn wir eine erklärende Variable weglassen oder hinzufügen, ändern sich *alle* geschätzten Koeffizienten (es kann sich sogar das Vorzeichen umkehren).

Explizit erwähnen wir die Lösung für $m = 0$ und $m = 1$. Für $m = 0$ passen wir eine Konstante an n Werte $y_1, \ldots, y_n$ an, und $\hat{\beta}_0$ ist einfach das arithmetische Mittel $\bar{y}_n$. Für $m = 1$ passen wir eine Gerade an n Punkte (x_i, y_i) in der Ebene an. Allerdings minimiert das Kriterium der Kleinsten Quadrate die Abstände Punkt-Gerade parallel zur y-Achse, und nicht orthogonal auf die Gerade, siehe Abbildung 5.3 unten. Der geschätzten Achsenabschnitt $\hat{\beta}_0$ und die geschätzte Steigung $\hat{\beta}_1$ sind

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n} (y_i - \bar{y}_n)(x_i - \bar{x}_n)}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2} \\ \hat{\beta}_0 &= \bar{y}_n - \hat{\beta}_1 \bar{x}_n. \end{aligned}$$

Die letzte Gleichung besagt, dass die Kleinste-Quadrate-Gerade durch den Punkt $(\bar{x}_n, \bar{y}_n)$ geht.

In der ganzen Diskussion bis hierher spielte die Statistik keine Rolle. Dies ändert sich, wenn wir die zufälligen Schwankungen der Daten (Messfehler, Zufallsauswahl der Versuchsobjekte) berücksichtigen wollen. Das heisst, wir fragen uns, wie stark verschieden die mit Hilfe der Ausgleichsrechnung bestimmten Parameter bei einem andern Experiment unter gleichwertigen Bedingungen wohl sein könnten, bzw. wie weit sich die Resultate einer Ausgleichsrechnung von den vorgegebenen Beobachtungen verallgemeinern lassen. Dazu brauchen wir ein stochastisches Modell, das wie folgt aussieht. Wir nehmen an, dass es einen wahren linearen Zusammenhang gibt und dass die Abweichungen von diesem Zusammenhang durch zufällige Fehler E_i zustande kommen:

$$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_m x_i^{(m)} + E_i.$$

In diesem Modell sind die Zielgrössen also Zufallsvariablen und werden daher mit Grossbuchstaben geschrieben. Die erklärenden Variablen sind hingegen fest, weil sie häufig von der Person, die das Experiment durchführt, festgelegt werden. Es gibt auch Fälle, wo auch die erklärenden Variablen zufällig sind, aber wir tun dann trotzdem so, als ob sie vorgegebene Grössen wären (Man kann dies so begründen, dass man auf die zufällig eingetretenen Werte bedingt).

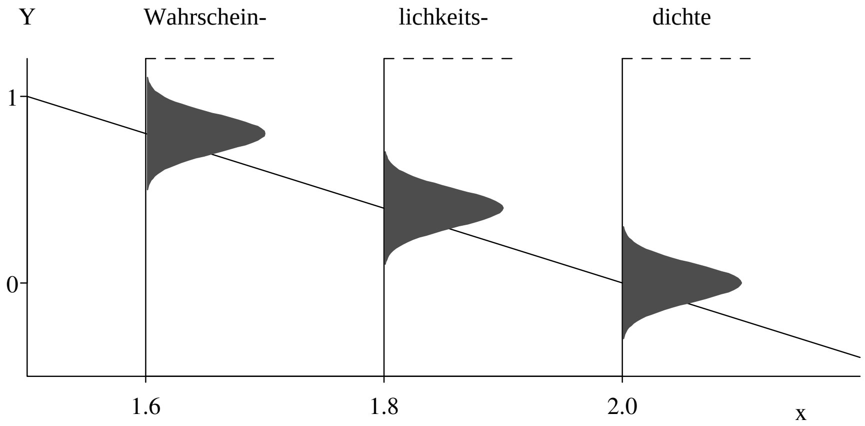

Abbildung 5.2: Veranschaulichung des Regressionsmodells $Y_i = 4 - 2x_i + E_i$ mit $E_i \sim \mathcal{N}(0, 0.1^2)$ für drei Beobachtungen.

Um das Modell vollständig zu spezifizieren, müssen wir noch die Verteilung der Fehler E_i (und damit der auch der Y_i) festlegen. Üblicherweise nimmt man an, dass die E_i Erwartungswert null haben (kein systematischer Fehler), unabhängig sind (keine gegenseitige Beeinflussung der Fehler), konstante Standardabweichung σ_E haben (überall die gleiche Genauigkeit) und normalverteilt sind. Das Modell ist illustriert in Abbildung 5.2.

Die Kleinste-Quadrate-Schätzungen sind Funktionen von $Y_1, \ldots, Y_n$ und somit ebenfalls Zufallsvariablen. Unter den obigen Annahmen sind die Y_i unabhängig und normalverteilt und die Normalengleichungen besagen, dass

$$\hat{\beta} = (X^T X)^{-1} X^T Y,$$

das heisst jedes $\hat{\beta}_j$ ist eine Linearkombination $Y_1, \ldots, Y_n$. Mit den Regeln aus Abschnitt 3.3 für den Erwartungswert und die Varianz einer Linearkombination von Zufallsvariablen folgt, dass

$$\mathcal{E}(\hat{\beta}_j) = \beta_j, \quad \sigma_{\hat{\beta}_j} = \sigma_E \sqrt{((X^T X)^{-1})_{jj}}.$$

Das heisst, die geschätzten Koeffizienten streuen um den unbekannten wahren Wert, und die Streuung ist proportional zur Streuung der Fehler, wobei die Proportionalitätskonstante von den erklärenden Variablen abhängt. Ausserdem ist $\hat{\beta}_j$ auch normalverteilt, wenn die Fehler E_i normalverteilt sind. Dies bildet die Basis für Tests und Vertrauensintervalle für die unbekannten Parameter, die wir im nächsten Abschnitt noch etwas ausführlicher behandeln.

Abbildung 5.3 zeigt 5 Punkte, die gemäss dem Modell $Y_i = 4 - 2x_i + E_i$ mit $E_i \sim \mathcal{N}(0, 0.1^2)$ simuliert wurden, zusammen mit der wahren und der best-passendsten Kleinste-Quadrate-Gerade. Man sieht auch den Unterschied zwischen dem Residuum (dem vertikalen Abstand eines Punktes von der Kleinsten-Quadrate-Gerade) und dem Zufallsfehler E_i (dem Abstand eines Punktes von der wahren Gerade). Andere simulierte Werte würden eine andere Kleinste-Quadrate-Gerade ergeben und damit andere Werte von $\hat{\beta}_0$ und $\hat{\beta}_1$.

Die oben gemachten Annahmen an E_i können auch (teilweise) überprüft werden. So können wir beurteilen, ob die angenommene lineare Beziehung zwischen der Ziel- und den erklärenden Variablen genügt, oder ob es systematische Abweichungen gibt, die einen komplizierteren Zusammenhang nahelegen. Dies ist das Thema der Residuenanalyse, auf die wir im übernächsten Abschnitt noch etwas näher eingehen.

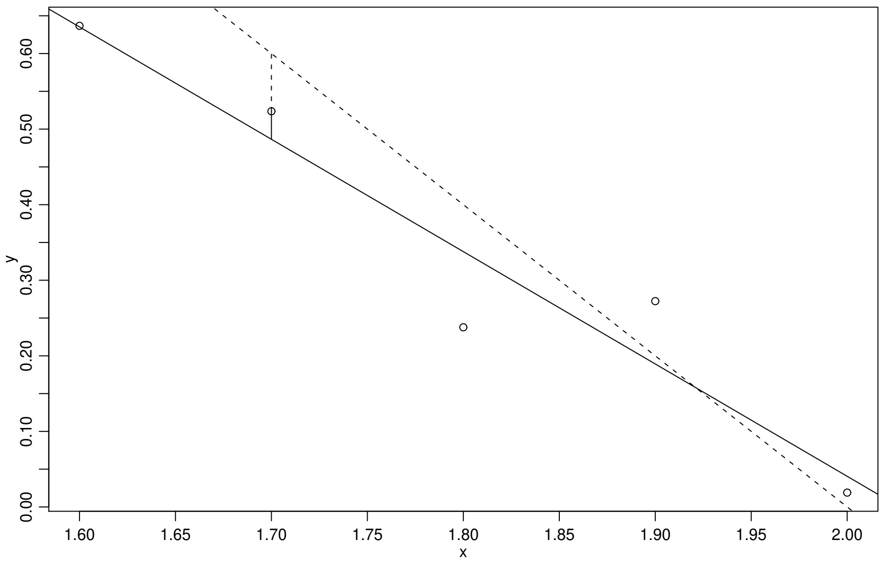

Abbildung 5.3: Fünf Punkte, die gemäss dem Modell $Y_i = 4 - 2x_i + E_i$ mit $E_i \sim \mathcal{N}(0, 0.1^2)$ simuliert wurden mit $x_1 = 1.6, \ldots x_5 = 2.0$. Gestrichelt sind die wahre Gerade und der Fehler E_2 gezeichnet, ausgezogen die Kleinste-Quadrate-Gerade und das Residuum r_2.

Die Annahme einer linearen Beziehung zwischen der Ziel- und den erklärenden Variablen ist nicht so einschränkend, wie es zunächst scheinen mag. Wie wir im Beispiel Butadien bereits gesehen haben, können wir die Zielvariable transformieren, und wir können auch neue erklärende Variablen durch Transformationen und Kombinationen kreieren. Es kann zum Beispiel $x_i^{(2)} = (x_i^{(1)})^2$ sein, das heisst, wir können ein Modell mit einer quadratischen Funktion betrachten

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + E_i$$

(und analog mit Polynomen höherer Ordnung). Entscheidend ist, dass die Koeffizienten β_j linear vorkommen ! Ein anderes wichtiges Beispiel ist ein Potenzzusammenhang $y = \alpha x^\gamma$, der durch Logarithmieren linear wird: $\log(y) = \beta_0 + \beta_1 \log(x)$ mit $\beta_0 = \log(\alpha)$ und $\beta_1 = \gamma$. Wenn man die Zielvariable ebenfalls transformiert, hat das aber Konsequenzen für die Annahmen über die Fehler: Wenn wir das Modell

$$\log(Y_i) = \beta_0 + \beta_1 \log(x_i) + E_i$$

zurücktransformieren, erhalten wir multiplikative statt additive Fehler:

$$Y_i = \exp(\beta_0)\ x_i^{\beta_1} \cdot \exp(E_i).$$

Wenn alle E_i die gleiche Verteilung haben, dann ist die Streuung der Y_i nicht mehr konstant, sondern proportional zum Erwartungswert. Glücklicherweise ist es oft so, dass nach einer Transformation, die zu einem linearen Zusammenhang führt, auch die Annahme additiver Fehler mit konstanter Streuung eher erfüllt zu sein scheint. Es gibt jedoch auch viele Anwendungen, bei denen es nicht möglich oder nicht sinnvoll ist zu linearisieren. Dann müssen die Methoden der nichtlinearen Regression verwendet werden, die wir hier nicht besprechen.

Die erklärenden Variablen können auch diskret sein und insbesondere Indikatoren für verschiedene Gruppen (Geschlecht, Behandlungsgruppe) darstellen. Auf diese Weise kann man zum Beispiel Geraden mit gleicher Steigung, aber unterschiedlichem Achsenabschnitt für die verschiedenen Gruppen anpassen. Ebenso lassen sich ungepaarte Vergleiche, die wir im vorangegangenen Kapitel diskutiert hatten, als Spezialfall des allgemeinen Regressionsmodells auffassen.

Die Euklid'sche Norm ist nicht die einzige sinnvolle Art, die Grösse der Abweichungen $(r_1, \ldots, r_n)$ zu messen. Von einigem Interesse ist insbesondere auch die L_1-Norm, $\|r\|_1 = \sum_{i=1}^{n} |r_i|$. Ausgleichsrechnung mit dieser Norm gibt Schätzungen $\hat{\beta}$ mit ziemlich anderem Verhalten. Dies sieht man am besten im Fall $m = 0$, d.h. Anpassen einer Konstanten, wo man anstelle des arithmetischen Mittels den Median von $(y_1, \ldots y_n)$ erhält. Bei der L_1-Ausgleichsrechnung gibt es zwar keine schönen Formeln, aber es gibt schnelle Algorithmen zur Berechnung. Vom statistischen Gesichtspunkt her ist das entscheidende Kriterium zur Wahl zwischen Kleinsten Quadraten und L_1 die Verteilung der Fehler E_i: Bei Normalverteilung sind Kleinste Quadrate vorzuziehen, bei längerschwänzigen Verteilungen kann aber L_1 deutlich besser sein.

5.1.1 Tests und Vertrauensintervalle

Wir haben schon erwähnt, dass unter den Modellannahmen jedes $\hat{\beta}_j$ normalverteilt ist mit Erwartungswert β_j, dem unbekannten wahren Wert des Parameters, und mit Standardabweichung

$$\sigma_{\hat{\beta}_j} = \sigma_E \sqrt{((X^T X)^{-1})_{jj}}.$$

Wenn wir die Standardabweichung σ_E kennen würden, könnten wir sofort die Nullhypothese testen, dass $\beta_j = 0$ ist, d.h. dass die j-te erklärende Variable im Modell weggelassen werden kann. Wir würden dies verwerfen, wenn

$$\left|\hat{\beta}_j\right| > z_{1-\alpha/2}\, \sigma_E \sqrt{((X^T X)^{-1})_{jj}}.$$

Wie im Ein- und Zweistichprobenfall ersetzen wir jetzt das unbekannte σ_E durch eine Schätzung und berücksichtigen die damit verbundene Unsicherheit, indem wir die Quantile der t- statt der Normalverteilung einsetzen. Wir verwenden die folgende Schätzung für σ_E (welche in Systat fälschlicherweise als "standard error of estimate" bezeichnet wird)

$$\hat{\sigma}_E = \sqrt{\frac{\sum_{i=1}^{n} r_i^2}{n-m-1}} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-m-1}},$$

wobei

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i^{(1)} + \hat{\beta}_2 x_i^{(2)} + \ldots + \hat{\beta}_m x_i^{(m)}$$

die angepassten Werte bezeichnet. Die richtige Anzahl Freiheitsgrade ist $n - m - 1$, d.h. wie üblich Anzahl Beobachtungen minus Anzahl geschätzter Parameter. Damit lautet der Verwerfungsbereich des **t-Tests**

$$\left|\hat{\beta}_j\right| > t_{n-m-1,1-\alpha/2}\, \hat{\sigma}_E \sqrt{((X^T X)^{-1})_{jj}}.$$

Statt auf den Wert Null können wir auf irgendeinen theoretisch vorgegebenen Wert für β_j testen. Dies ist insbesondere von Bedeutung für Vertrauensintervalle, die ja durch Umkehrung der entsprechenden Tests entstehen. In Formeln lautet das Vertrauensintervall für β_j also

$$\hat{\beta}_j \pm t_{n-m-1,1-\alpha/2}\, \hat{\sigma}_E \sqrt{((X^T X)^{-1})_{jj}}.$$

Im Fall der einfachen linearen Regression $Y_i = \beta_0 + \beta_1 x_i + E_i$ ergibt eine einfache Rechnung, dass

$$\sigma_{\hat{\beta}_1} = \frac{\sigma_E}{\sqrt{\sum_{i=1}^n (x_i - \overline{x}_n)^2}}.$$

Das bedeutet, dass die Steigung umso genauer bestimmt werden kann, je stärker die erklärende Variable streut. Dies sollte unmittelbar einleuchten.

Im Fall von mehreren erklärenden Variablen ist es wichtig zu verstehen, was die Nullhypothese $\beta_j = 0$, die mit dem t-Test überprüft wird, genau besagt. Es wird nämlich getestet, ob die j-te Variable weggelassen werden kann, unter der Voraussetzung, dass alle andern Variablen im Modell vorhanden sind. Häufig sind verschiedene erklärende Variablen untereinander verwandt, d.h. sie erfassen ähnliche Aspekte einer Situation. Dann kann es geschehen, dass eine Variable ohne grossen Verlust eine andere substituieren kann: Die Zusatzinformation, die man aus beiden Variablen gewinnt, bringt nicht wesentlich mehr, als was man schon aus einer Variablen erhält. In solchen Fällen könnte man also z.B. die erste Variable weglassen, so lange die zweite in der Regressionsgleichung bleibt, oder umgekehrt auch die zweite weglassen, so lange die erste in der Regressionsgleichung bleibt. In andern Worten, die Tests akzeptieren sowohl "$\beta_1 = 0$" als auch "$\beta_2 = 0$". Daraus folgt aber nicht, dass man beide Variablen weglassen kann. Sobald man eine Variable weggelassen hat, muss man die Regression neu berechnen.

Wegen dieser Komplikation will man oft zuerst die Frage beantworten, ob eventuell sogar kein Zusammenhang zwischen der Ziel- und der Gesamtheit aller erklärenden Variablen besteht. Die Nullhypothese lautet dann "Alle β_j sind $= 0$". Dafür verwendet man die Testgrösse

$$T = \frac{\sum_{i=1}^n (\hat{y}_i - \overline{y})^2/m}{\sum_{i=1}^n (y_i - \hat{y}_i)^2/(n-m-1)}.$$

Grosse Werte von T führen zur Ablehnung der Nullhypothese, und die kritische Grenze ist das Quantil der sogenannten F-Verteilung mit m und $n-m-1$ Freiheitsgraden, die wir bisher noch nicht angetroffen haben. Der Test heisst daher **F-Test**. Es gibt dafür Tabellen, und jedes statistische Programmpaket liefert automatisch den P-Wert.

Um besser zu verstehen, weshalb T eine vernünftige Testgrösse ist, ist folgende Identität nützlich

$$\sum_{i=1}^n (y_i - \overline{y})^2 = \sum_{i=1}^n (\hat{y}_i - \overline{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

Dies ist nichts anderes als der "Satz von Pythagoras" und die Tatsache, dass die Vektoren $(y_i - \hat{y}_i)$ und $(\hat{y}_i - \overline{y})$ senkrecht aufeinander stehen (weil $(\hat{y}_i)$ die Projektion von (y_i) auf den Unterraum aufgespannt von $x^{(0)}, \ldots x^{(m)}$ ist). Wenn alle wahren β_i null sind, sind alle angepassten Werte $\hat{y}_i$ bis auf zufällige Schwankungen gleich der best passendsten Konstanten $\overline{y}$, und man kann zeigen, dass dann der Zähler und der Nenner von T im Mittel gleich gross sind. Stimmt die Nullhypothese nicht, dann ist der Zähler wesentlich grösser als der Nenner.

Obige Zerlegung kann man in Worten beschreiben als "Varianz der Zielgrösse = Varianz der angepassten Werte plus Varianz der Abweichungen". (Man spricht daher auch von **Varianzanalyse**). Die angepassten Werte sind allein durch die erklärenden Variablen und die geschätzten Koeffizienten, d.h. durch das Modell, bestimmt. Man nennt daher den Koeffizienten

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^n (y_i - \overline{y})^2}$$

das **Bestimmtheitsmass** (coefficient of determination), oder den Anteil erklärter Varianz. Es liegt zwischen null und eins und misst, wie erfolgreich die erklärenden Variablen sind zur Vorhersage der Zielgrösse, und wird sehr oft in Berichten zitiert. Mit Umformungen, die wieder auf der Orthogonalität von $(y_i - \hat{y}_i)$ und $(\hat{y}_i - \overline{y})$ beruhen, erhält man eine zweite, äquivalente Form von R^2:

$$R^2 = \frac{\left(\sum_{i=1}^n (y_i - \overline{y})(\hat{y}_i - \overline{y})\right)^2}{\sum_{i=1}^n (\hat{y}_i - \overline{y})^2 \sum_{i=1}^n (y_i - \overline{y})^2}.$$

Dies ist offensichtlich das Quadrat einer Korrelation, und daher heisst R^2 auch das **Quadrat der multiplen Korrelation** (multiple R-squared). (Wir verzichten darauf, das Adjektiv "multipel" zu begründen.)

Häufig ist man weniger an den Parametern selber, sondern am Erwartungswert einer zukünftigen Beobachtung an einer vorgegebenen Stelle $(x_0^{(1)}, \ldots x_0^{(m)})$ der erklärenden Variablen interessiert, d.h. an einer Linearkombination $\beta_0 + \ldots + \beta_m x_0^{(m)}$. Diese Linearkombination wird natürlich als $\hat{\beta}_0 + \ldots + \hat{\beta}_m x_0^{(m)}$ geschätzt, und man kann dafür ebenfalls ein Vertrauensintervall angeben. Dies ist zu unterscheiden von einem **Vorhersageintervall** für eine zukünftige Beobachtung Y_0, das meist deutlicher grösser ist. Ich verweise dafür auf Kap. 13.4 im Buch von W. Stahel.

5.1.2 Residuenanalyse

Die obigen Schätz- und Testverfahren beruhen auf Modellannahmen über die Regressionsfunktion und die Verteilung der Fehler. Diese Annahmen zu überprüfen, ist meistens wesentlich. Es geht dabei nicht in erster Linie um eine Rechtfertigung, sondern um die Chance, aus allfälligen Abweichungen ein besseres Modell entwickeln zu können. Wir geben hier nur einen kurzen Überblick über die wichtigsten Methoden.

Die Fehler E_i kennen wir nicht, aber wir können stattdessen als Näherung die Residuen $r_i = y_i - \hat{y}_i$ verwenden. Selbst wenn das Modell genau stimmt, haben die Residuen – im Unterschied zu den Fehlern E_i – nicht alle genau die gleiche Streuung (und sie sind auch leicht abhängig). Man kann dies korrigieren, indem man "standardisierte" oder "studentisierte" Residuen betrachtet. Die Effekte sind jedoch klein, so dass es meist genügt, mit den gewöhnlichen Residuen zu arbeiten.

Zur Überprüfung der Normalverteilung dient ein Quantil-Quantil-Diagramm mit den Residuen. Ist die Verteilung schief, legt das eine Transformation der Zielgrösse nahe. Gibt es einzelne herausfallende Werte, soll man diese überprüfen und eventuell die Analyse ohne diese Werte wiederholen. Ist die Verteilung symmetrisch und langschwänzig, dann bedeutet dies, dass es genauere Schätzungen als die Kleinste Quadrate Schätzungen gibt, z.B. die L_1-Regression.

Der wichtigste Plot in der Residuenanalyse ist der Plot der Residuen r_i gegen die angepassten Werte $\hat{y}_i$, der sogenannte Tukey-Anscombe Plot. Im Idealfall zeigt sich ein Bild ohne jegliche Struktur. Aus den Normalengleichungen folgt, dass die Stichproben-Korrelation zwischen r_i und $\hat{y}_i$ null ist, das heisst es gibt keinen linearen Zusammenhang zwischen r_i und $\hat{y}_i$. Nichtlineare Zusammenhänge können jedoch vorkommen. Sie zeigen an, dass die Regressionsfunktion nicht korrekt ist. Abhilfe schaffen die Aufnahme zusätzlicher erklärender Variablen (z.B. quadratische Terme) oder Transformationen der erklärenden und/oder der Zielgrössen. Ein einfaches Beispiel ist in Abbildung 5.4 gezeigt, bei dem es

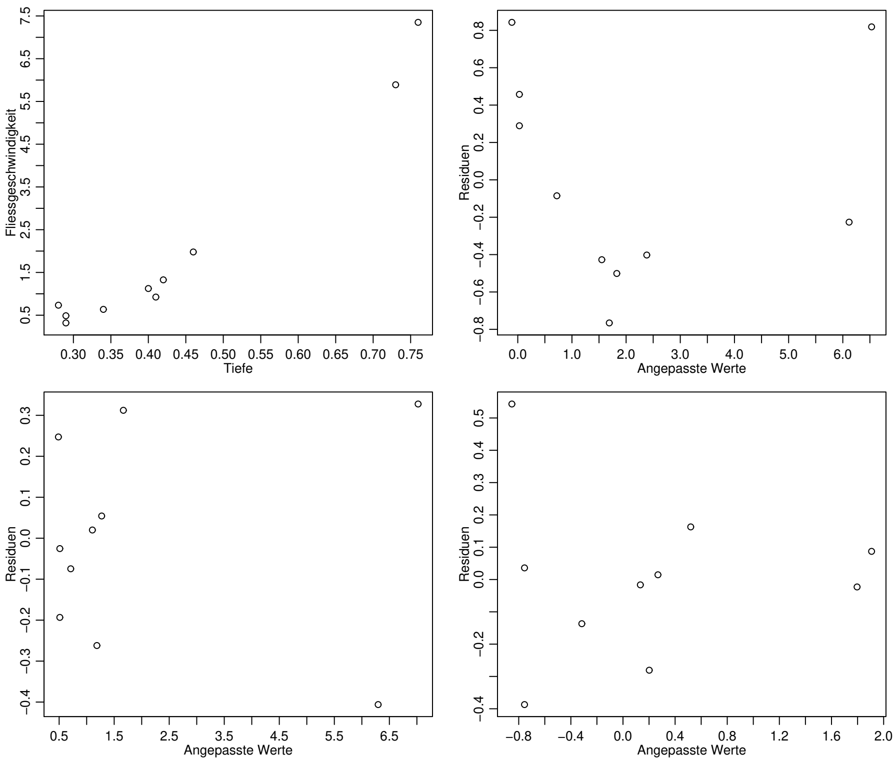

Abbildung 5.4: Streudiagramm von Tiefe und Fliessgeschwindigkeit (oben links), Tukey-Anscombe Plots für einfache lineare Regression (oben rechts), für quadratische Regression (unten links) und für einfache lineare Regression mit logarithmierten Variablen (unten rechts).

um den Zusammenhang zwischen Tiefe und Fliessgeschwindigkeit von Bächen geht. Bei einfacher Regression zeigt der Tukey-Anscombe Plot eine klare nichtlineare Struktur, die verschwindet, wenn man entweder einen quadratischen Term dazunimmt oder wenn man beide Variablen logarithmiert (d.h. einen Potenzzusammenhang anpasst). Mit so wenigen Daten kann man zwischen diesen beiden Modellen nicht unterscheiden. Die Nichtlinearität des Zusammenhangs ist natürlich auch im ursprünglichen Streudiagramm ersichtlich, wenn man genau hinschaut. Häufig sind aber Abweichungen von der Linearität im Tukey-Ansombe Plot besser zu sehen.

Im Tukey-Anscombe Plot sieht man auch, ob die Streuung der Residuen um die Nulllinie konstant ist oder mit wachsendem angepassten Wert ebenfalls zunimmt. Ist dies der Fall, dann hilft oft eine geeignete Transformation. Wie wir im Abschnitt über Fehlerfortpflanzung gesehen haben, ist

$$\mathrm{var}(g(Y_i)) \approx g'(\mathcal{E}(Y_i))^2 \,\mathrm{var}(Y_i),$$

und wir wollen die Transformation g so wählen, dass die Varianzen der transformierten Grössen $g(Y_i)$ konstant sind. Wenn also der Tukey-Anscombe Plot zeigt, dass die Streuung der Y_i proportional zu den vorausgesagten Werten $\hat{y}_i$ ist, dann soll $g'(x) = 1/x$ sein, also $g(x) = \log(x)$. Wenn die Streuung eher wie $\sqrt{\hat{y}_i}$ zunimmt, dann soll man $g'(x) = 1/\sqrt{x}$, d.h. $g(x) = \sqrt{x}$ wählen. Eine Transformation der Zielgrösse verändert natürlich das ganze Modell, aber wie bereits früher erwähnt, hat eine geeignete Transformation der Zielgrösse oft gleich drei positive Effekte: Der Zusammenhang mit den erklärenden Variablen wird linearer, die Streuung der Fehler ist eher konstant und die Fehlerverteilung ist näher bei einer Normalverteilung.

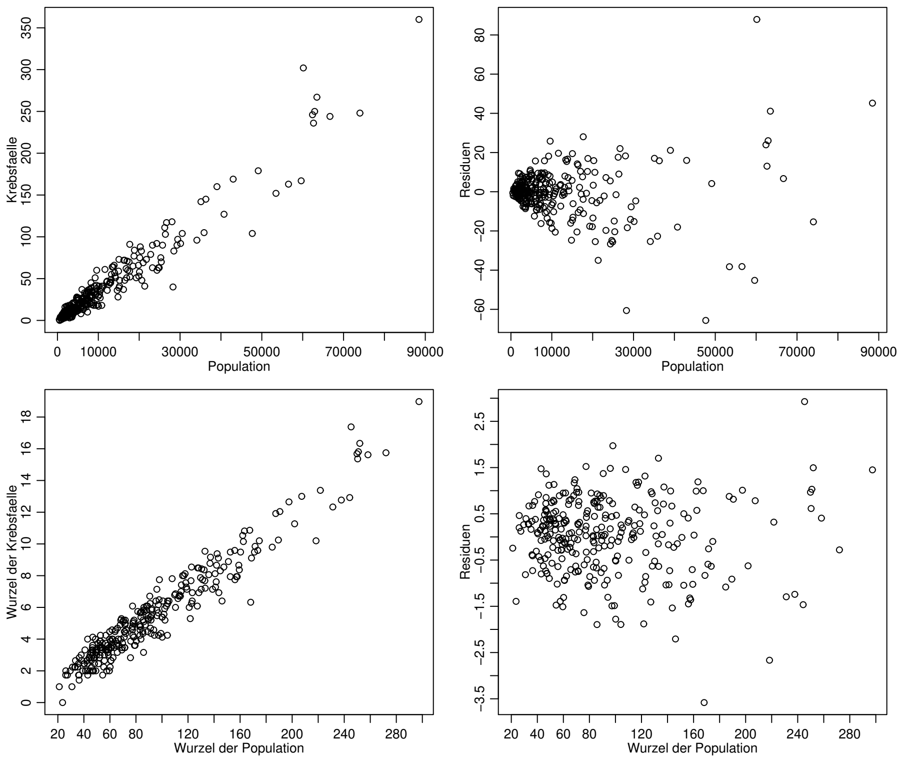

Abbildung 5.5: Fälle von Brustkrebs gegen die Grösse des Risikopopulation (oben links), Tukey-Anscombe Plot bei einfacher linearer Regression (oben rechts), sowie die gleichen Darstellungen nach Transformation mit der Wurzelfunktion (unten).

Abbildung 5.5 zeigt ein Beispiel. Die Zielvariable ist die Anzahl Fälle von Brustkrebs 1950-60 für 301 Bezirke ("counties") im Südosten der USA und die erklärende Variable ist Anzahl erwachsener Frauen, die in diesen Bezirken wohnen. Der Tukey-Anscombe Plot zeigt klar, dass die Streuung nicht konstant ist, sondern etwa wie die Wurzel der vorhergesagten Werte anwächst. Da wir in diesem Beispiel erwarten, dass y ungefähr proportional zu x ist, ist es daher naheliegend, das Modell $\sqrt{Y_i} = \beta_1 \sqrt{x_i} + E_i$ anzupassen. Wie die Abbildung zeigt, ist dann die Streuung der Fehler ungefähr konstant.

Man kann auch die Residuen r_i einzeln gegen die erklärenden Variablen $x_i^{(j)}$ für $j = 1, 2, \ldots m$ auftragen. Das kann hilfreich sein, um zu entscheiden, welche zusätzlichen Terme ins Modell aufgenommen werden sollen. Zeigt sich, dass die Streuung der Residuen von einigen wenigen erklärenden Variablen abhängt, kann man auch versuchen, diese Abhängigkeit durch eine geeignete Funktion zu beschreiben und dann sogenannte gewichtete Kleinste Quadrate zu verwenden (siehe dazu weiterführende Literatur).

Eine letzte wesentliche Voraussetzung ist die Unabhängigkeit der Fehler E_i. Ist diese verletzt, kann sich das Niveau der Tests und Vertrauensintervalle wesentlich ändern. Wenn die Residuen eine natürliche Reihenfolge haben (insbesondere zeitlich), dann soll man die Residuen gegen diese Reihenfolge auftragen. Verletzung der Unabhängigkeit äussert sich dann meist in der Form von langsam variierenden Trends. Programmpakete liefern einen Test dazu, den Durbin-Watson Test, den wir hier jedoch nicht besprechen. Methoden, die Abhängigkeiten der Fehler berücksichtigen, laufen unter der Bezeichnung verallgemeinerte Kleinste Quadrate.