Mathematik IV: Statistik

D-UWIS & D-ERDW & D-AGRL

Frühlingssemester 2013

Peter Bühlmann und Hansruedi Künsch
Seminar für Statistik
ETH Zürich

Inhaltsverzeichnis

Vorbemerkungen

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 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.

Für weitere Erläuterungen verweisen wir jeweils auf das folgende Buch:
Werner A. Stahel, Statistische Datenanalyse. Eine Einführung für Naturwissenschaftler. 4. Aufl. (2002), Vieweg, Braunschweig.

Manchmal erklären wir die grundlegenden Begriffe auch an Hand von Glücksspielen, obwohl wir wissen, dass Sie nicht speziell an solchen interessiert sind. Es gibt dort einfach weniger Verständnis- und Interpretationsprobleme als bei andern Beispielen. Wir hoffen auf Ihr Verständnis.

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 Wahrscheinlichkeitsmodelle (Stahel, Kap. 4.1, 4.2)

Wir betrachten **Zufallsexperimente**, bei denen der Ausgang nicht exakt vorhersagbar ist. Ein **Wahrscheinlichkeitsmodell** beschreibt, welche Ergebnisse in einem solchen Experiment möglich sind und welche Chancen die verschiedenen Ergebnisse haben. Ein Wahrscheinlichkeitsmodell erlaubt mittels Simulation mögliche Ergebnisse zu erzeugen und so eine Vorstellung der zufälligen Variabilität zu gewinnen.

Ein Wahrscheinlichkeitsmodell hat die folgenden Komponenten:
- **Grundraum** Ω, bestehend aus den **Elementarereignissen** ω,
- **Ereignissen** $A, B, C, \ldots$,
- **Wahrscheinlichkeit** P.

Elementarereignisse sind einfach mögliche Ergebnisse oder Ausgänge des Experiments, die zusammen den Grundraum bilden:

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

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:

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

"Ein Ereignis A tritt ein" bedeutet, dass das Ergebnis ω des Experiments zu A gehört.

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.

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

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.

$$\begin{aligned} &P(A^c) = 1 - P(A), \\ &P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad \text{(Additionssatz)}. \end{aligned}$$

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**)

Wir behandeln in diesem Kapitel diskrete Wahrscheinlichkeitsmodelle, bei denen der Grundraum endlich oder “abzählbar” ist (d.h. man kann die Elementarereignisse durchnumerieren). Zum Beispiel ist $\Omega = \{0, 1, \ldots, 10\}$ endlich und deshalb diskret; $\Omega = \mathbb{N}_0 = \{0, 1, 2, \ldots\}$ ist zwar unendlich, aber noch abzählbar und daher trotzdem diskret; $\Omega = \mathbb{R}$ ist nicht abzählbar.

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

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

Beispiel (Forts.) Beim Wurf zweier Münzen ist es plausibel, dass alle 4 Elemente von Ω gleich wahrscheinlich sind. Wegen $P(\Omega) = 1$ müssen sich die Wahrscheinlichkeiten zu Eins addieren:

$$P(KK) = P(KZ) = P(ZK) = P(ZZ) = \frac{1}{4}.$$

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

2.2 Der Begriff der Unabhängigkeit (Stahel, Kap. 4.6)

Wenn man die Wahrscheinlichkeiten $P(A)$ und $P(B)$ kennt, lässt sich im Allgemeinen daraus nicht $P(A \cap B)$ berechnen: Es sind alle Werte zwischen 0 und dem Minimum von $P(A)$ und $P(B)$ möglich. Ein wichtiger Spezialfall liegt vor, wenn folgende Produktformel gilt

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

Dann heissen A und B **stochastisch unabhängig**.

Beispiel (Forts.): Es sei $A = K$ im 1. Wurf und $B = K$ im 2. Wurf. Dann gilt $P(A) = P(KK) + P(KZ) = \frac{1}{2}$ und analog $P(B) = \frac{1}{2}$. Wegen $P(A \cap B) = P(KK) = \frac{1}{4}$, sind also A und B unabhängig.

Viel wichtiger ist jedoch der umgekehrte Schluss. Wenn zwischen den Ereignissen A und B kein kausaler Zusammenhang besteht (d.h. es gibt keine gemeinsamen Ursachen oder Ausschliessungen), dann **postuliert** man stochastische Unabhängigkeit und nimmt damit an, dass obige Produktformel gilt. In diesem Fall kann also $P(A \cap B)$ aus $P(A)$ und $P(B)$ berechnet werden (Dieses Vorgehen wurde auch schon als "Unabhängigkeitserklärung" bezeichnet).

Beispiel (Forts.): Beim Wurf von zwei Münzen können wir die Wahrscheinlichkeiten auch mit folgender Überlegung finden. Es sei wieder $A =$ "K im 1. Wurf" und $B =$ "K im 2. Wurf". Wir postulieren, dass $P(A) = P(B) = \frac{1}{2}$ und dass A und B unabhängig sind (es gibt keinen Kausalzusammenhang zwischen den beiden Würfen !). Dann folgt $P(KK) = \frac{1}{4}$, und mit dem Additionssatz folgt ferner $P(KZ) = P(ZK) = P(ZZ) = \frac{1}{4}$. Wir erhalten also das Gleiche wie zuvor.

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

$$P(A_1 \cap A_2) = P(A_1)P(A_2),$$
$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3).$$

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

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

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 ungewohnt für die Bezeichung einer Funktion, ist aber üblich in der Wahrscheinlichkeitsrechnung. Sie haben hoffentlich in der Analysis gesehen, dass man Funktionen wie gewöhnliche Variablen x addieren oder multiplizieren kann (man addiert oder multipliziert einfach die Funktionswerte).

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.

Wenn der Grundraum Ω diskret ist, dann muss auch der Wertebereich $W = W_X$ (Menge der möglichen Werte von X) diskret sein, d.h. endlich oder abzählbar. Wir werden in diesem Kapitel bloss diskrete Zufallsvariablen genauer diskutieren. Insbesondere sind Anzahlen stets diskret, während Messungen meist eher den Wertebereich $\mathbb{R}$ haben und damit nicht diskret, sondern kontinuierlich sind (praktisch kann man aber nur mit endlicher Genauigkeit messen).

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

x	0	2	3	4	10	11
$P(X = x)$	4/9	1/9	1/9	1/9	1/9	1/9

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:

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

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

Diese Verteilung heisst **Bernoulli(π)-Verteilung**. Es ist eine triviale Mathematisierung, um das Eintreffen oder Nicht-Eintreffen eines Ereignis zu beschreiben.

Etwas interessanter ist das n-malige Werfen einer Münze. Der Grundraum Ω besteht dann aus allen "Wörtern" der Länge n, welche man mit den Buchstaben K und Z schreiben kann. Ω hat also 2^n Elemente. Wir betrachten die Zufallsvariablen

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

Um die Verteilung von X bestimmen zu können, müssen wir eine Wahrscheinlichkeit auf Ω festlegen. Wir postulieren, dass die Ereignisse $X_i = 1$ (also "K im i-ten Wurf") alle die Wahrscheinlichkeit π haben und unabhängig sind. Dann 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 in obigem Beispiel motiviert, ist die Binomialverteilung angebracht für die Zufallsvariable "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. Vor allem wenn n gross sind, sind die meisten Wahrscheinlichkeiten $P(X = x)$ verschwindend klein.

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 **Standardabweichung** $\sigma(X)$.

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

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

Die Standardabweichung beschreibt die Streuung der Verteilung. Rechnerisch ist das Quadrat der Standardabweichung, die sogenannte **Varianz** bequemer:

$$\begin{aligned} \text{Var}(X) &= \sum_{x \in W_X} (x - \mathcal{E}(X))^2 P(X = x) \\ \sigma(X) &= \sqrt{\text{Var}(X)}. \end{aligned}$$

Die Standardabweichung hat dieselbe Einheit wie X, während die Einheit der Varianz deren Quadrat ist: Wird z.B. X in Metern (m) gemessen, so besitzt

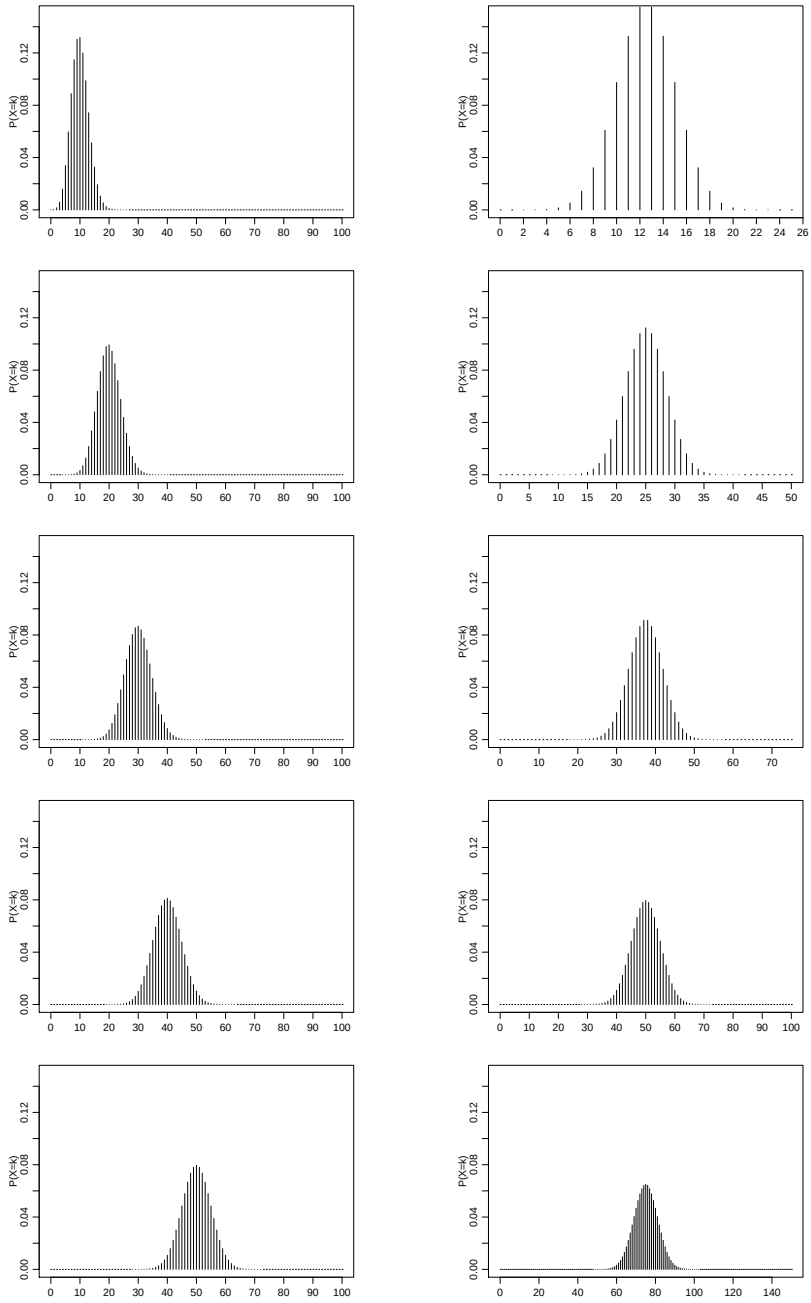

Abbildung 2.1: Die Binomialwahrscheinlichkeiten $P(X = x)$ als Funktion von x für verschiedene n's und π's. Links ist $n = 100$ und $\pi = 0.1, 0.2, 0.3, 0.4, 0.5$ und rechts ist $\pi = 0.5$ und $n = 25, 50, 75, 100, 150$.

$\mathrm{Var}(X)$ die Dimension Quadratmeter (m^2) und $\sigma(X)$ wiederum die Dimension 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)):

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

Die Kennzahlen fassen also sehr gut zusammen, was wir in der Abbildung 2.1 gesehen haben: Die Verteilung ist um den Erwartungswert konzentriert, die Streuung wächst mit n, aber langsamer als n, und ist für festes n maximal bei $\pi = 1/2$.

2.5.1 Kumulative Verteilungsfunktion

Manchmal ist es für Rechnungen nützlicher, statt der "Liste" $P(X = x)$ (für alle x) die sukzessiven Summen

$$\sum_{y \in W_X; y \le x} P(X = y) = P(X \le x)$$

anzugeben. Dabei läuft x ebenfalls über den Wertebereich W_X von X. Man kann in dieser Definition aber auch beliebige reelle Werte x betrachten und erhält dann eine Funktion

$$F(x) = P(X \le x) = \sum_{y \in W_X; y \le x} P(X = y),$$

die sogenannte **kumulative Verteilungsfunktion**. Diese springt an den Stellen, die zum Wertebereich gehören, und ist dazwischen konstant. 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.

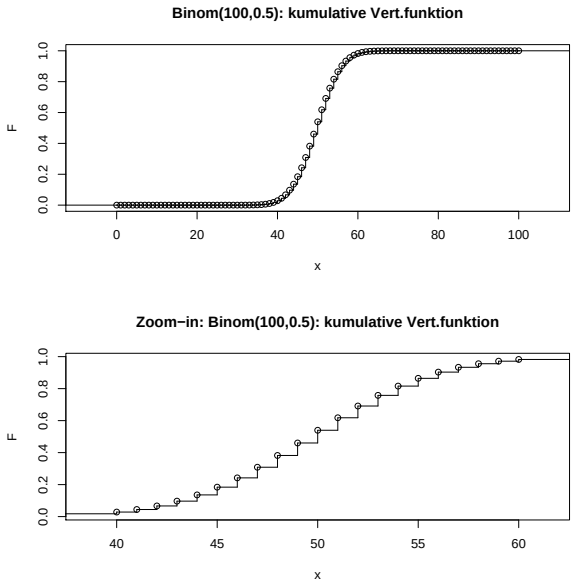

Abbildung 2.2: Kumulative Verteilungsfunktion $F(\cdot)$ für $X \sim$ Binomial(100,0.5). Unten: zoom-in für die Werte $x \in [40, 60]$. Die Kreise zeigen an, dass an den Sprungstellen der obere Wert gilt.

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 als Modell gebraucht werden:
Anzahl Schadenmeldungen eines Versicherten pro Jahr,
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).

2.6.2 Summen von Poisson-verteilten Zufallsvariablen

Die Poisson-Verteilung hat die folgende Additionseigenschaft: Wenn $X \sim$ Poisson(λ_X) und $Y \sim$ Poisson(λ_Y) unabhängig sind, dann ist $X + Y \sim$ Poisson$(\lambda_X + \lambda_Y)$. Wenn also zum Beispiel die Anzahlen spontaner Ereignisse in einer Nervenzelle in zwei disjunkten Zeitintervallen Poisson-verteilt und unabhängig sind, dann ist auch das Total wieder Poisson-verteilt. Wir erhalten also eine Poisson-Verteilung für alle Intervalle. Weil sich bei der Addition der Zufallsvariablen die Parameter der Poisson-Verteilung addieren, ist üblicherweise λ proportional zur Länge des betrachteten Zeitintervalls.

Kapitel 3

Statistik für Zähldaten

3.1 Drei Grundfragestellungen der Statistik (Stahel Kap. 7.1)

Die Statistik fasst Beobachtungen als **Realisierung** von Zufallsvariablen auf. Sie untersucht dann, was für Schlüsse man aus den Beobachtungen über die zu Grunde liegende Verteilung ziehen kann. Meist zieht man nur Verteilungen in Betracht, die bis auf einen (oder wenige) Parameter bestimmt sind und möchte dann Fragen über diesen Parameter beantworten. Man unterscheidet die folgenden drei Grundfragestellungen:

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.): Sei $x = 11$ die effektive Anzahl weiblicher gezüchteter Kuhkälber beim Spermasexing (vgl. Kapitel 2.4). Wir fassen $x = 11$ auf als **Realisierung** einer Zufallsvariablen X auf, und nehmen an, dass X Binom$(12, \pi)$-verteilt ist. Wir möchten jetzt Schlüsse ziehen von der Beobachtung $x = 11$ auf den unbekannten Parameter π. Die Grundfragen lauten hier:

1. Welches ist der plausibelste Wert π (zu der Beobachtung $x = 11$)?
2. Ist die Beobachtung $x = 11$ kompatibel mit $\pi = 0.5$ (reiner Zufall) oder mit $\pi = 0.7$ (ökonomisch zu wenig interessant) ?

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 ersetzt den Erwartungswert durch die Beobachtung. Somit ergibt sich die relative Häufigkeit

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

als Schätzung der Erfolgswahrscheinlichkeit. Im Beispiel ist also die geschätzte Erfolgswahrscheinlichkeit für das Verfahren gleich $\frac{11}{12} = 0.917$. Es leuchtet unmittelbar ein, dass bei regelmässiger Anwendung des Verfahrens langfristig nicht genau 11 von 12 Kälbern das gewünschte Geschlecht haben werden. Mit andern Worten: Die wahre Erfolgswahrscheinlichkeit π ist nicht das Gleiche wie die geschätzte Erfolgswahrscheinlichkeit $\hat{\pi}$.

3.2.2 Statistischer Test

Beispiel: Wir haben eine Münze, bei der wir uns fragen, ob sie fair ist oder systematisch eher Kopf ergibt. Um dies zu beantworten, wird die Münze 100-mal geworfen, und wir erhalten 58 mal Kopf.
Betrachte $X =$ Anzahl Kopf (K) bei 100 Würfen. Es ist vernünftig, das Modell zu benutzen: $X \sim$ Binomial$(100, \pi)$. Beobachtet (realisiert) wurde $x = 58$, d.h. die geschätzte Wahrscheinlichkeit für Kopf ist $\hat{\pi} = 0.58$. Wir fragen uns, ob $\pi = 1/2$, oder 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 Ereignisse von der Form $\{X \geq c\}$ für "grosse" Werte c, weil grosse

Werte eher für $\pi > 1/2$ sprechen. Wir wollen damit herausfinden, wie plausibel die beobachtete Anzahl $x = 58$ ist bei einer fairen Münze. 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 $x = 58$ noch als plausibel bei einer fairen Münze eingestuft werden kann. Hätte man aber 59-mal Kopf beobachtet, so würde man dies als nicht mehr genügend 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:

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

Am häufigsten ist die Nullhypothese $H_0 : \ \pi = 1/2$ (d.h. $\pi_0 = 1/2$), also "reiner Zufall" oder "kein Effekt". Meist führt man einen Test durch, weil man glaubt, dass die Alternative richtig ist und man auch Skeptiker davon überzeugen möchte.

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:

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

Quantitativ wird K so berechnet, dass

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

Dabei bedeutet $\overset{\approx}{\leq}$, dass die linke Seite kleiner oder gleich der rechten Seite sein soll, aber so nahe wie möglich.

Beispiel (Forts.): Bei 100-maligem Münzwurf.
$H_0: \ \pi = 1/2$ 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 (H_0 ist dann statistisch widerlegt, die Abweichung von der Nullhypothese ist "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: Hypothesen können nur falsifiziert und nicht verifiziert werden.

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

Beispiel (Forts.): Beim Spermasexing (vgl. Kapitel 2.4) wurden $x = 11$ Kuhkälber geboren von insgesamt 12 Kälbern. Es scheint ziemlich klar zu sein, dass dies nicht reiner Zufall sein kann. Wir wollen trotzdem noch sehen, was der Test sagt:
Modell: $X \sim$ Binomial$(12, \pi)$,
$H_0: \ \pi = \pi_0 = 0.5$,
$H_A: \ \pi > \pi_0 = 0.5$.
Unter der Nullhypothese gilt

	$c = 8$	$c = 9$	$c = 10$	$c = 11$	$c = 12$
$P(X \geq c)$	0.194	0.073	0.019	0.003	0.0002

Für das, Signifikanzniveau $\alpha = 0.05$ ist also der Verwerfungsbereich $K = \{10, 11, 12\}$, und für das Niveau $\alpha = 0.01$ ist $K = \{11, 12\}$. Für beide Niveaus wird die Nullhypothese also verworfen, d.h. der Effekt der Methode "Spermasexing" ist statistisch signifikant, sowohl auf dem 5%- als auch auf dem 1%-Niveau.

Wenn jemand nur an einer Methode interessiert ist, deren Erfolgswahrscheinlichkeit grösser als 70% ist, dann wird er wie folgt vorgehen:
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: $P_{\pi=0.7}(X \in K) \overset{\approx}{\leq} 0.05 \rightsquigarrow \ K = \{12\}$

Entscheid: H_0 wird belassen, d.h. eine Erfolgswahrscheinlichkeit von über 70% ist nicht signifikant nachgewiesen.

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(Fehler 2. Art) wird grösser falls α 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, wenn $\pi = 0.8$ ökonomisch interessant wäre. Bei der kleinen Anzahl von 12 Versuchen, kann man einfach nur sehr schlecht zwischen $\pi = 0.7$ und $\pi = 0.8$ entscheiden. Beachte, dass die Wahrscheinlichkeit für einen Fehler 1. Art $\overset{\approx}{\leq} 0.05$, also klein 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 qualitativ Folgendes gilt:

Verwerfungsbereich $K = K(\alpha)$ wird kleiner mit kleiner werdendem α,

denn α ist ja die Wahrscheinlichkeit für einen Fehler 1. Art, und diese wird natürlich dann klein, wenn wir weniger oft die Nullhypothese H_0 verwerfen.

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.

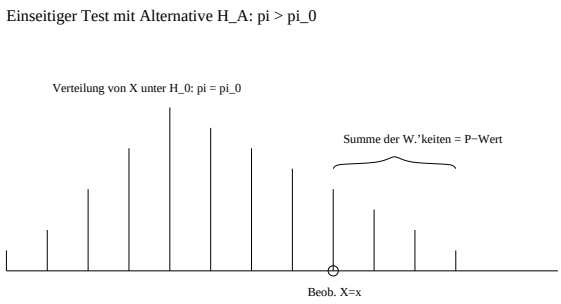

Abbildung 3.1: Schematische Darstellung des P-Werts bei einer einseitigen Alternative $H_A: \ \pi > \pi_0$.

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:

verwerfe H_0 falls P-Wert $\leq \alpha$
belasse H_0 falls P-Wert $> \alpha$.

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:

P-Wert ≈ 0.05 : schwach signifikant
P-Wert ≈ 0.01 : signifikant

$$\text{P-Wert } \approx 0.001 : \text{ stark signifikant}$$
$$\text{P-Wert } \leq 10^{-4} : \text{ äusserst signifikant}$$

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 Vertrauensintervallen 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:

$$P_\pi(\pi \in I(X)) \overset{\approx}{>} 1 - \alpha \quad \text{für jedes } \pi.$$

Dies kann so interpretiert werden, dass das Konfidenzintervall I den wahren Parameter π mit Wahrscheinlichkeit $1 - \alpha$ mit einschliesst.

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), erhält man - unter Verwendung der pragmatischen Schätzung von $E(X)$ mittels der Beobachtung x - 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:

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

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 $1 - \alpha = 0.95$:

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

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 $I = I(x) \approx (507.9, 600.1)$.

3.3.4 Mehrere Beobachtungen

Oft hat man nicht nur eine Anzahl x beobachtet, sondern n verschiedene Anzahlen $x_1, x_2, \ldots, x_n$, z.B. die Anzahl spontaner Ereignisse in einer Nervenzelle während n verschiedenen, nicht-überlappenden Zeitintervallen von einer Sekunde. Weil die Summe von unabhängigen Poisson-verteilten Zufallsvariablen wieder Poisson-verteilt ist (siehe Abschnitt 2.6.2), bilden wir einfach die Summe $x = x_1 + x_2 + \cdots + x_n$ der Beobachtungen und wenden die obigen Verfahren an für den Parameter $\lambda' = n\lambda$. Im Unterschied zu den Tests in Kapitel 3.2.2

basiert in diesem Fall der Test auf *mehreren* Beobachtungen, die mittels der Summe zu einem einzigen Wert zusammengefasst werden.

Wenn wir bei der Punktschätzung oder beim Vertrauensintervall an λ und nicht an λ' interessiert sind, dividieren wir das Ergebnis am Schluss einfach noch durch n:

$$\hat{\lambda} = \frac{x_1 + \cdots + x_n}{n},$$

$$I(x_1, \ldots, x_n) \approx \frac{x_1 + \cdots + x_n}{n} \pm \frac{1.96}{\sqrt{n}} \frac{\sqrt{x_1 + \cdots + x_n}}{\sqrt{n}} = \hat{\lambda} \pm \frac{1.96}{\sqrt{n}} \sqrt{\hat{\lambda}}.$$

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 Indi-

viduen (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 *arithmetische 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.

Quantile

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

$$\frac{1}{2}(x_{(\alpha n)} + x_{(\alpha n+1)}) \quad \text{falls } \alpha \cdot n \text{ eine ganze Zahl ist,}$$
$$x_{(k)} \text{ wobei } k = \alpha n + \frac{1}{2} \text{ gerundet auf eine ganze Zahl;} \quad \text{falls } \alpha \cdot n \text{ keine ganze Zahl ist.}$$

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
Das arithmetische Mittel der $n = 13$ Messungen ist $\overline{x} = 80.02$ und die Standardabweichung ist $s_x = 0.024$. Ferner ist für $n = 13$ $0.25n = 3.25$, $0.5n = 6.5$ 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.

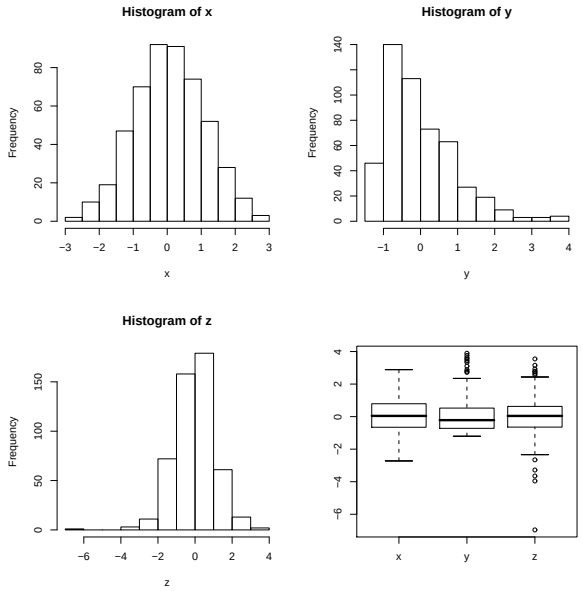

Abbildung 4.1: Boxplots für die zwei Methoden zur Bestimmung der Schmelzwärme von Eis.

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.

Mehrere Variablen

Wenn wir bei jeder Versuchseinheit 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

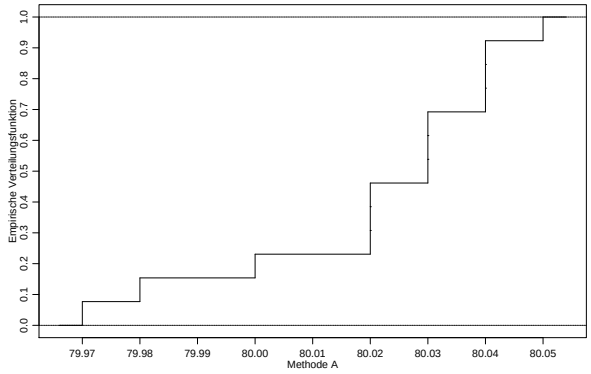

Abbildung 4.2: Empirische kumulative Verteilungsfunktion der Messungen der Schmelzwärme von Eis mit Methode A.

"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.

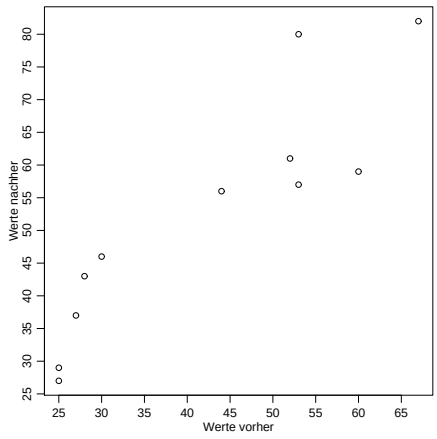

Abbildung 4.3: Streudiagramm der Blutplättchen-Aggregation vor und nach dem Rauchen einer Zigarette.

Für die numerische Zusammenfassung der Abhängigkeit ist die **empirische Korrelation** r (oder auch mit $\hat{\rho}$ bezeichnet) 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 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.

Weitere Ausführungen sind in Kapitel 5.1 zu finden.

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:

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

Dazu genügt es, die kumulative Verteilungsfunktion $F(x) = P(X \leq x)$ anzugeben, denn es gilt

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

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

Weil für stetige Zufallsvariablen $P(X = a) = P(X = b) = 0$, spielt es keine Rolle, ob wir $<$ oder $\leq$ schreiben. Obige Formeln sind jedoch auch richtig für diskrete Zufallsvariable.

4.3.1 (Wahrscheinlichkeits-)Dichte

Für stetige Zufallsvariablen können wir einen analogen Begriff zur "Punkt"-Wahrscheinlichkeit $P(X = x)$ für diskrete Variablen mit Hilfe der Ableitung gewinnen.

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

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

Damit erhalten wir folgende Interpretation:

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

Die Begründung dafür ist:

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

wobei die letzte Approximation aus der Definition einer Ableitung folgt.

Aus der Dichte kann man die kumulative Verteilungsfunktion zurückgewinnen:

$$F(x) = \int_{-\infty}^{x} f(y)dy$$

(weil F eine Stammfunktion von f ist und $F(-\infty) = 0$). Ausserdem gelten die folgenden Eigenschaften:

1. $f(x) \geq 0$ für alle x (da $F(\cdot)$ monoton wachsend ist)
2. $P(a < X \leq b) = F(b) - F(a) = \int_a^b f(x)dx$
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: Der Erwartungswert beschreibt die mittlere Lage der Verteilung und die Standardabweichung deren Streuung. Die Formeln ergeben sich, indem wir beim diskreten Fall $P(X = x)$ durch $f(x)dx$ und die Summe durch ein Integral ersetzen:

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

Die **Quantile** $q(\alpha)$ $(0 < \alpha < 1)$ einer Zufallsvariablen X, bzw. deren Verteilung, 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**.

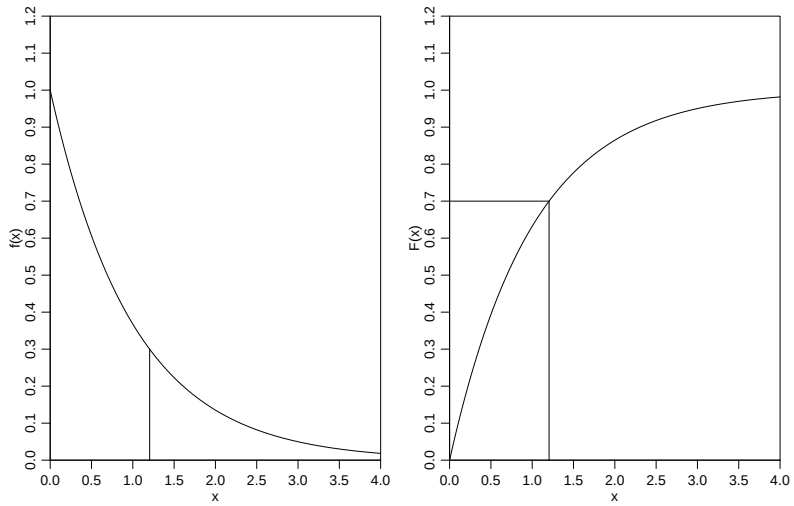

Abbildung 4.4: Illustration des 70%-Quantils $q(0.7)$. Links ist die Dichte gezeichnet, mit der Fläche von $-\infty$ (oder hier auch 0) bis $q(0.7)$ gleich 0.7. Rechts die kumulative Verteilungsfunktion und $q(0.7)$ als Wert der Umkehrfunktion an der Stelle 0.7.

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

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

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

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

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

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

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 Exponentialverteilt mit Parameter $\lambda \in \mathbb{R}^+$ ($X \sim \mathrm{Exp}(\lambda)$) falls

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

Die zugehörige kumulative Verteilungsfunktion ist

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

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:

$$\mathcal{E}(X) = \frac{1}{\lambda},$$
$$\mathrm{Var}(X) = \frac{1}{\lambda^2}, \quad \sigma_X = \frac{1}{\lambda}.$$

Es besteht folgender Zusammenhang zwischen der Exponential- und Poisson-Verteilung: 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 $\sigma^2 \in \mathbb{R}^+$ (in Formeln $X \sim \mathcal{N}(\mu, \sigma^2)$) falls

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

Die zugehörige kumulative Verteilungsfunktion $F(x) = \int_{-\infty}^{x} f(y)dy$ ist nicht explizit darstellbar mit Standardfunktionen wie $\exp(x)$, $\log(x)$ etc.

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

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

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.

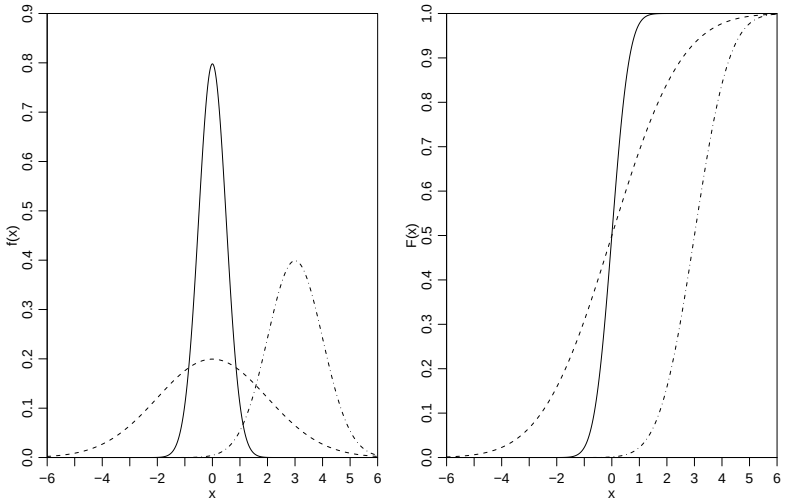

Abbildung 4.5: Dichten (links) und kumulative Verteilungsfunktionen (rechts) der Normalverteilungen mit $\mu = 0, \sigma = 0.5$ (ausgezogen), $\mu = 0, \sigma = 2$ (gestrichelt) und $\mu = 3, \sigma = 1$ (Strich-Punkte).

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:

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

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

4.4.4 Transformationen

Wenn $g : \mathbb{R} \to \mathbb{R}$ eine Funktion von $\mathbb{R}$ nach $\mathbb{R}$ ist, dann ist die Zusammensetzung

$$Y = g(X),$$

eine neue Zufallsvariable. Die Zusammensetzung bedeutet einfach, dass zu jeder Realisierung x von X die Realisierung $y = g(x)$ von Y gehört. Solche Transformationen treten häufig auf.

Die kumulative Verteilungsfunktion und die Dichte von Y sind im Prinzip durch die die Verteilungsfunktion und die Dichte von X bestimmt. Die Berechnung kann aber je nach Eigenschaften von g kompliziert sein. Für den Erwartungswert gilt jedoch stets die folgende Formel

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

Wenn wir speziell $g(x) = (x - \mathcal{E}(X))^2$ wählen, so erhalten wir

$$\mathcal{E}\left((X - \mathcal{E}(X))^2\right) = \int_{-\infty}^{\infty} (x - \mathcal{E}(X))^2 f(x) dx = \mathrm{Var}(X).$$

Das heisst, die Varianz ist der Erwartungswert der Zufallsvariable "Quadrierte Abweichung vom Erwartungswert".

Lineare Transformationen

Wir betrachten hier zuerst den Fall einer linearen Transformation

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

Für $Y = a + bX$ gilt dann

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

Die erste Gleichung folgt aus

$$\int_{-\infty}^{\infty} (a + bx) f_X(x) dx = a \int_{-\infty}^{\infty} f_X(x) dx + b \int_{-\infty}^{\infty} x f_X(x) dx,$$

und die zweite aus der ersten mit

$$\mathrm{Var}(a + bX) = \mathcal{E}\left((a + bX - (a + b\,\mathcal{E}(X))^2\right) = \mathcal{E}\left(b^2 (X - \mathcal{E}(X))^2\right).$$

Überdies gelten für $b > 0$:

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

Beispiel: Wenn $X \sim \mathcal{N}(\mu, \sigma)$, dann ist $Y \sim \mathcal{N}(a + b\mu, b^2\sigma^2)$, denn nach der Regel (4.2) gilt

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma b} \exp\left(-\frac{(y-a-b\mu)^2}{2\sigma^2 b^2}\right).$$

Eine linear transformierte Normalverteilung ist also 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.

Standardisieren einer Zufallsvariablen

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 $X \sim \mathcal{N}(\mu, \sigma^2)$, so ist die standardisierte Zufallsvariable wieder normalverteilt, wie wir im obigen Beispiel gesehen haben:

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

Damit lassen sich Wahrscheinlichkeiten für beliebige Normalverteilungen mit Hilfe der Standard-Normalverteilung berechnen.

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 $b = 1/\sigma = 1/2$):

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

Die Lognormal-Verteilung

Für positive Zufallsvariablen X wird oft die Logarithmus-Transformation verwendet. Führt diese Transformation auf eine Normalverteilung, dann heisst X **Lognormal**-verteilt. In andern Worten, wenn $Y \sim \mathcal{N}(\mu, \sigma^2)$, dann heisst $X = \exp(Y)$ Lognormal-verteilt mit Parametern $\mu \in \mathbb{R}$ und $\sigma^2 \in \mathbb{R}^+$. Die Lognormal-Verteilung ist nicht mehr symmetrisch und es gilt: $\mathcal{E}(X) = \exp(\mu + \sigma^2/2) > \exp(\mathcal{E}(Y))$.

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.

Es ist jedoch wichtig, die Unterschiede zwischen den empirischen und theoretischen Grössen zu verstehen: Mit Hilfe der Statistik werden wir quantifizieren, wie gross die Unsicherheit ist, wenn wir die theoretischen Grössen mit Hilfe der empirischen Grössen schätzen.

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, bei dem 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 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

(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).

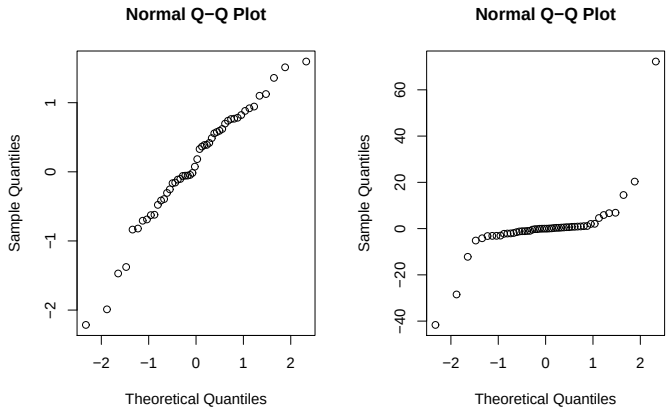

Abbildung 4.6: Links: Normal-Plot für 50 Realisierungen von $\mathcal{N}(0,1)$. Rechts: Normal-Plot für 50 Realisierungen von Cauchy-Verteilung (sehr langschwänzig).

Die Messungen $x_1, x_2, \ldots, x_n$ fassen wir als Realisierungen der Zufallsvariablen $X_1, \ldots, X_n$ auf. Diese Auffassung ist oft bequemer als die Interpretation, dass die Messungen n unabhängige Realisierungen einer Zufallsvariablen X sind. Oft sind wir an Funktionen der Messwerte $x_1, x_2, \ldots, x_n$ interessiert: $y = g(x_1, \ldots, x_n)$ wobei $g : \mathbb{R}^n \to \mathbb{R}$. Wir betrachten hier vor allem die Funktion "arithmetisches Mittel"

$$g(x_1, \ldots, x_n) = \overline{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Wenn $x_1, x_2, \ldots, x_n$ Realisierungen der Zufallsvariablen $X_1, \ldots, X_n$ sind, dann ist y eine Realisierung der Zufallsvariable

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

Das arithmetische Mittel der Daten $\overline{x}_n$ ist also eine Realisierung der Zufallsvariablen $\overline{X}_n$.

Wir sind an der Verteilung der Zufallsvariablen $\overline{X}_n$ interessiert: Die Kenntnis dieser Verteilung wird uns erlauben, Statistik aufgrund von arithmetischen Mitteln von Daten zu machen.

4.5.1 Unabhängigkeit und i.i.d. Annahme

Oft treffen wir die Annahme, dass die Zufallsvariablen $X_1, \ldots, X_n$ **unabhängig** voneinander sind. Anschaulich heisst das, es gibt keine gemeinsamen Faktoren, die den Ausgang der verschiedenen Messungen beeinflussen, und keine "carry over" Phänomene von einer Messung zur nächsten. Die mathematische Definition lautet: $X_1, \ldots, X_n$ sind unabhängig, falls die Ereignisse $\{X_1 \leq b_1\}, \ldots, \{X_n \leq b_n\}$ unabhängig sind für beliebige $b_1, \ldots, b_n \in \mathbb{R}$.

Wenn die Zufallsvariablen $X_1, \ldots, X_n$ unabhängig sind und alle **dieselbe** Verteilung haben, dann schreiben wir das kurz als

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

Die Abkürzung i.i.d. steht für: **i**ndependent, **i**dentically **d**istributed. Wir werden meistens mit dieser i.i.d. Annahme arbeiten. Welche Verteilung die X_i's haben, lassen wir offen. Es kann, aber muss nicht eine Normalverteilung sein.

Die Unabhängigkeit spielt eine Rolle bei den **Regeln für Erwartungswerte und Varianzen** von Summen.

$$\mathcal{E}(X_1 + X_2) = \mathcal{E}(X_1) + \mathcal{E}(X_2)$$

gilt immer,

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2)$$

jedoch nur, wenn X_1 und X_2 unabhängig sind.

4.5.2 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: $\mathcal{E}(X_i) = \mu,\ \mathrm{Var}(X_i) = \sigma_X^2$.

Die Kennzahlen von $\overline{X}_n$ folgen dann aus den allgemeinen Regeln für Erwartungwert und Varianz:

$$\mathcal{E}(\overline{X}_n) = \mu,$$
$$\mathrm{Var}(\overline{X}_n) = \frac{\sigma_X^2}{n}, \quad \sigma(\overline{X}_n) = \frac{\sigma_X}{\sqrt{n}}.$$

Die Standardabweichung von $\overline{X}_n$ heisst auch der **Standard-Fehler** des arithmetischen Mittels.

Der Erwartungswert von $\overline{X}_n$ ist also gleich demjenigen einer einzelnen Zufallsvariablen X_i, die *Streuung nimmt jedoch ab mit wachsendem n.*. Für $n \to \infty$ geht die Streuung gegen null, das heisst es gilt das **Gesetz der grossen Zahlen:** Falls $X_1, \ldots, X_n$ i.i.d. , dann

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

Die Streuung des arithmetischen Mittels ist jedoch nicht proportional zu $1/n$, sondern nur zu $1/\sqrt{n}$. Das bedeutet, dass man für eine doppelte Genauigkeit nicht doppelt so viele, sondern vier Mal so viele Messungen braucht.

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

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

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

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

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.

Beispiel:. Wenn $X_1, \ldots, X_{10}$ i.i.d. Uniform(0,100)-verteilt sind, dann ist $\mu = 50$ und $\sigma_X^2 = 100^2/12 = 2500/3$. Die Wahrscheinlichkeit, dass $\overline{X}_{10}$ um mehr als $1.96\sqrt{250/3} = 17.9$ von 50 abweicht, ist also ungefähr 95%.

Für eine exakte Formulierung des Zentralen Grenzwertsatzes betrachtet man die standardisierte Zufallsvariable

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

Diese ist ungefähr $\mathcal{N}(0,1)$ verteilt, was bedeutet, dass für alle x gilt

$$\lim_{n\to\infty} P(Z_n \leq x) = \Phi(x).$$

.

Der Zentrale Grenzwertsatz gilt auch für diskrete Zufallsvariablen X_i. Insbesondere können wir den Spezialfall von n unabhängigen Wiederholungen eines binären Experiments mit den beiden möglichen Ergenissen "Erfolg", bzw. "Misserfolg" betrachten. Wenn wir setzen $X_i = 1$ falls die i-te Wiederholung ein Erfolg ist, und $X_i = 0$ falls die i-te Wiederholung ein Misserfolg ist, dann sind die X_i i.i.d. Bernoulli(π)-verteilt, wobei π die Wahrscheinlichkeit für einen Erfolg bei einer Durchführung ist (siehe Abschnitt 2.4). Das arithmetische Mittel $\overline{X}_n$ ist dann nichts anderes als die relative Häufigkeit der Erfolge. Gemäss dem Zentralen Grenzwertsatz gilt

$$P(|\overline{X}_n - \pi| \leq \varepsilon) = P(\pi - \varepsilon \leq \overline{X}_n \leq \pi + \varepsilon) \approx \Phi(\frac{\varepsilon}{\sqrt{\pi(1-\pi)/n}}) - \Phi(\frac{-\varepsilon}{\sqrt{\pi(1-\pi)/n}}),$$

denn $E(X_i) = \pi$, $\text{Var}(X_i) = \pi(1-\pi)$. Wenn wir statt des arithmetischen Mittels die Summe $S_n = X_1 + \ldots + X_n$ betrachten, welche eine Binomial(n, π)-Verteilung hat, dann ist diese genähert $\mathcal{N}(n\pi, n\pi(1-\pi))$-verteilt. Das heisst

$$P(|S_n - n\pi| \leq c) \approx \Phi(\frac{c}{\sqrt{n\pi(1-\pi)}}) - \Phi(\frac{-c}{\sqrt{n\pi(1-\pi)}}).$$

Dass bei den beiden rechten Seiten n einmal im Zähler und einmal im Nenner steht, ist kein Druckfehler: Damit die beiden linken Seiten gleich sind, muss

$c = n\varepsilon$ sein. Daraus erhalten wir die folgende Näherung für den zweiseitigen Binomialtest: Verwirf $H_0 : \pi = \pi_0$, falls

$$|\overline{X}_n - \pi_0| > \frac{\sqrt{\pi_0(1-\pi_0)}}{\sqrt{n}}\Phi^{-1}(1-\alpha/2),$$

bzw. äquivalent dazu

$$|S_n - n\pi_0| > \sqrt{n\pi_0(1-\pi_0)}\Phi^{-1}(1-\alpha/2).$$

4.5.3 Verletzung der Unabhängigkeit

Was ändert sich, wenn die Zufallsvariablen $X_1, \ldots, X_n$ nicht mehr unabhängig sind ? Die Annahme der identischen Verteilung behalten wir bei, weil diese Zufallsvariablen die Wiederholung von Messungen unter identischen Bedingungen modellieren sollen. Dann gilt

$$\begin{aligned}
&\mathcal{E}(\overline{X}_n) = \mu, \\
&\mathrm{Var}(\overline{X}_n) = \frac{\sigma_X^2}{n}\left(1 + \frac{2}{n}\sum_{1\le i<j\le n}\rho(X_i, X_j)\right).
\end{aligned}$$

Dabei ist $\rho(X_i, X_j)$ die Korrelation zwischen den Zufallsvariablen X_i und X_j, das theoretische Gegenstück zur empirischen Korrelation in Abschnitt 4.2.2, siehe auch Abschnitt 5.1:

$$\rho(X_i, X_j) = \frac{\mathrm{Cov}(X_i, X_j)}{\sigma_{X_i}\sigma_{X_j}}, \quad \mathrm{Cov}(X_i, X_j) = \mathcal{E}((X_i - \mathcal{E}(X_i))(X_j - \mathcal{E}(X_j))).$$

Dies bedeutet, dass Abhängigkeit zwischen den Messungen die Genauigkeit des arithmetischen Mittels beeinflusst. Bei positiver Korrelation nimmt die Genauigkeit ab, bei negativer Korrelation kann sie auch zunehmen. Es ist daher wichtig, die Messungen so anzulegen, dass sie als unabhängig angenommen werden können, wenn man die in den folgenden Kapiteln beschriebenen statistischen Methoden anwenden will.

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:

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

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:

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

Wir verwenden einen einseitigen Test, weil es bei der Planung der Studie bereits klar war, dass man nur an einer Erhöhung der Aggregation durch Nikotin interessiert ist.

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**

$$Z = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{\sigma_X} = \frac{\text{beobachtet} - \text{erwartet}}{\text{Standardfehler}}.$$

Unter der Nullhypothese gilt (siehe Abschnitt 4.5):

$$Z \sim \ \mathcal{N}(0,1).$$

Der Verwerfungsbereich für die Teststatistik Z für die 2-seitige Alternative H_A: $\mu \neq \mu_0$ ist dann

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

4. Verwerfe H_0, falls der realisierte Wert z in K liegt, oder äquivalent dazu, falls das arithmetische Mittel der Daten $\overline{x}_n$ um mehr als $\frac{\sigma_X}{\sqrt{n}}\Phi^{-1}(1-\frac{\alpha}{2})$ von μ_0 abweicht (ansonsten belasse H_0).

Bei einseitigen Alternativen ist der Verwerfungsbereich ebenfalls einseitig, und man benutzt das $(1-\alpha)$-Quantil. Zusammenfassend geht der z-Test also wie folgt:

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

Im Unterschied zu den Tests in Kapitel 3.2.2 basiert der z-Test auf *mehreren* Beobachtungen. Diese werden aber mittels einer realisierten Teststatistik z zusammengefasst (eine Funktion der Daten). Ansonsten sind die Konzepte genau gleich wie in Kapitel 3.2.2.

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 stattdessen 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

$$t = \frac{\sqrt{n}(\overline{x}_n - \mu_0)}{\hat{\sigma}_X} = \frac{\text{beobachtet} - \text{erwartet}}{\text{geschätzter Standardfehler}}.$$

Deren Verteilung unter der Nullhypothese $H_0: \ \mu = \mu_0$ ist eine sogenannte t-Verteilung mit $n-1$ Freiheitsgraden, die wir mit t_{n-1} bezeichnen.

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:

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

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.

Den Verwerfungsbereich beim Test erhalten wir, indem wir einen Bereich der Wahrscheinlichkeit α bei der t_{n-1}-Verteilung abschneiden (je nach Alternative auf einer Seite, oder je die Hälfte auf beiden Seiten). Dazu brauchen wir die Quantile $t_{\nu,\alpha}$, welche tabelliert sind (siehe z. B. Stahel, Tabelle 8.5.g, p. 187) oder mittels Computer berechnet werden können.

Zusammenfassend ist der t-Test wie folgt:

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

Da das Quantil der t-Verteilung grösser ist als das Quantil der Normalverteilung, erhält man einen etwas kleineren Verwerfungsbereich als beim z-Test. 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:

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

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

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 $\mathcal{N}(\mu, \sigma_X^2)$.

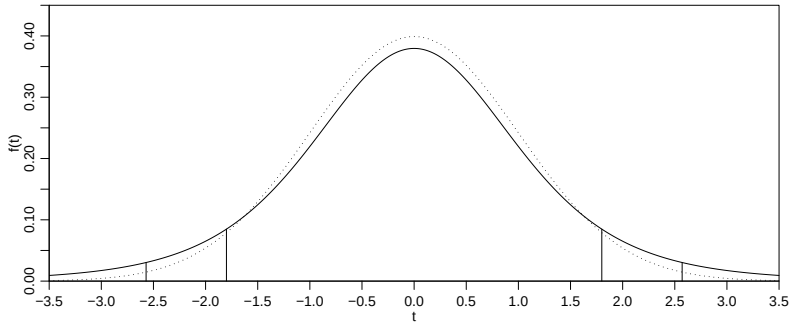

Abbildung 4.7: Dichte der t-Verteilung mit 5 Freiheitsgraden (ausgezogen) und der Normalverteilung (gestrichelt). Die Flächen ausserhalb der vertikalen Linien sind gleich 2.5% bzw. gleich dem halben P-Wert für einen hypothetischen Datensatz mit $\sqrt{6}|\overline{x}_6 - \mu_0|/\hat{\sigma}_X = 1.8$.

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 natürlich doppelt so gross: $P_{H_0}[|T| > 4.27] = 2P_{H_0}[T > 4.27] = 0.0016$.

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 $1 - \alpha$:

$$\overline{x}_n \pm \Phi^{-1}(1 - \alpha/2) \ \frac{\sigma_X}{\sqrt{n}} \quad \text{falls } \sigma_X \text{ bekannt,}$$

$$\overline{x}_n \pm t_{n-1,1-\alpha/2} \ \frac{\hat{\sigma}_X}{\sqrt{n}} \quad \text{falls } \sigma_X \text{ unbekannt.}$$

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. μ = 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:

Fehler 1. Art = fälschliches Verwerfen von H_0, obschon H_0 richtig ist,

und

Fehler 2. Art(μ) = (fälschliches) Beibehalten von H_0 falls $\mu (\in H_A)$ 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:

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

Der Vorzeichen-Test

Wir betrachten die Situation, wo die Daten Realisierungen von (4.4) sind, wobei die einzelnen X_i nicht normalverteilt sein müssen. Der Vorzeichentest testet Hypothesen über den Median der Verteilung von X_i, den wir hier mit μ bezeichnen; im Falle einer symmetrischen Verteilung ist $\mu = \mathcal{E}(X_i)$.

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

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

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

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 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.

Vom Standpunkt der Macht gibt es keine eindeutige Antwort, ob der Vorzeichen- oder der t-Test besser ist. Wenn die Verteilung der X_i langschwänzig ist, kann der Vorzeichentest grössere Macht haben. Weil der Vorzeichentest die Information nicht ausnützt, um wieviel die X_i von dem Wert μ_0 abweichen (siehe die Definition der Teststatistik V oben), kann die Macht aber auch wesentlich schlechter sein als beim t-Test.

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 bezüglich $\mu = \mathcal{E}(X_i)$. Wir verzichten auf die Formel für die Teststatistik und die Berechnung

der Verteilung der Teststatistik unter der Nullhypothese $\mu = \mu_0$, da der P-Wert mit statistischer Software berechnet werden kann.

Beispiel (Forts.): Blutplättchen-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.

Der Wilcoxon-Test ist in den allermeisten Fällen vorzuziehen: er hat in vielen Situationen oftmals wesentlich grössere Macht als der t- und als der Vorzeichen-Test, und selbst in den ungünstigsten Fällen ist er nie viel schlechter.

Wenn man trotzdem den t-Test verwendet, dann sollte man die Daten auch grafisch ansehen, damit wenigstens grobe Abweichungen von der Normalverteilung entdeckt werden. Insbesondere sollte der Normal-Plot (siehe Kap. 4.4.6) angeschaut werden.

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

Wir besprechen hier Methoden, um einen Vergleich zweier Methoden (Gruppen, Versuchsbedingungen, Behandlungen) hinsichtlich der Lage der Verteilung machen.

4.7.1 Gepaarte und ungepaarte Stichproben

Bei einem solchen Vergleich ist es naheliegend, dass man jede Versuchseinheiten einer der beiden Versuchsbedingung zuordnet. Um systematische Effekte zu vermeiden, soll diese Zuordnung durch das Los erfolgen (vgl. Abschnitt 4.8 unten). Man hat dann Beobachtungen (realisierte Zufallsvariablen)

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

Bei einer solchen **zufälligen Zuordnung von Versuchseinheiten zu einer von zwei verschiedenen Versuchsbedingungen** spricht man von einer **ungepaarten Stichprobe**. Im Allgemeinen ist in einer ungepaarten Stichprobe $m \neq n$, aber nicht notwendigerweise. Entscheidend ist, dass x_i und y_i zu verschiedenen Versuchseinheiten gehören und als unabhängig angenommen werden können.

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.

Meist sind ungepaarte Vergleiche aber nicht ideal, wenn die Versuchseinheiten stark verschieden sind. Wenn möglich sollte man eine Versuchseinheit beiden Versuchbedingungen unterwerfen: Es liegt eine **gepaarte Stichprobe** vor, wenn **beide Versuchsbedingungen an derselben Versuchseinheit eingesetzt** werden. Die Daten sind dann von der folgenden Struktur:

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

Notwendigerweise ist dann die Stichprobengrösse n für beide Versuchsbedingungen dieselbe. Zudem sind x_i und y_i abhängig, weil die Werte von der gleichen Versuchseinheit kommen.

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,

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

welche wir als Realisierungen von i.i.d. Zufallsvariablen $U_1, \ldots, U_n$ auffassen. Kein Unterschied zwischen den beiden Versuchsbedingungen heisst dann einfach $E[U_i] = 0$ (oder auch 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 $x_1, \ldots, x_n$ und $y_1, \ldots, y_m$ (siehe Kapitel 4.7.1), welche wir als Realisierungen der folgenden Zufallsvariablen auffassen:

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

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:

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

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 den geschätzten Standardfehler von $\overline{X}_n - \overline{Y}_m$ dividiert und dies als Teststatistik benutzt:

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

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 angegebenen 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. In den meisten Fällen erhält man ähnliche P-Werte wie unter der Annahme von gleichen Varianzen.

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:

$$\begin{aligned} &X_1, \ldots, X_n \text{ i.i.d.} \sim F_X, \\ &Y_1, \ldots, Y_m \text{ i.i.d.} \sim F_Y, \\ &F_X \text{ beliebige stetige Verteilungsfunktion, } F_Y(x) = F_X(x - \delta). \end{aligned}$$

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

5.1 Korrelation und empirische Korrelation

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

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

wobei $\sigma_X = \sqrt{\mathrm{Var}(X)}$, und analog für σ_Y.

Die Korrelation ρ_{XY} ist eine dimensionslose, normierte Zahl
mit Werten $\rho_{XY} \in [-1, 1]$.

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

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

Überdies gilt:

$$X \text{ und } Y \text{ unabhängig} \quad \Longrightarrow \quad \mathrm{Corr}(X,Y) = 0. \tag{5.1}$$

Die Umkehrung gilt i.A. nicht.

5.1.1 Die empirische Korrelation

In Kapitel 4.2.2 und Abbildung 4.3 haben wir ein Beispiel gesehen, wo die Daten $(x_1, y_1), \ldots, (x_n, y_n)$ als Realisierungen von i.i.d. Zufalls-vektoren $(X_1, Y_1), \ldots, (X_n, Y_n)$ aufgefasst werden können.

Die empirischen Korrelation ist dann

$$\widehat{\mathrm{Corr}}(X,Y) = \hat{\rho}_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}.$$

Analog zur Korrelation gelten die folgenden Eigenschaften:

$\hat{\rho}_{XY} \in [-1, 1]$,
$\hat{\rho}_{XY} = +1 \;\Leftrightarrow\; y_i = a + bx_i$ für alle $i = 1, \ldots, n$, und für ein $a \in \mathbb{R}$ und ein $b > 0$,
$\hat{\rho}_{XY} = -1 \;\Leftrightarrow\; y_i = a + bx_i$ für alle $i = 1, \ldots, n$, und für ein $a \in \mathbb{R}$ und ein $b < 0$.

5.2 Einfache lineare Regression

Wir betrachten das folgende Beispiel aus der Chemie. Die Dimerisation von 1,3-Butadien läuft nach einem Reaktionsmodell zweiter 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.

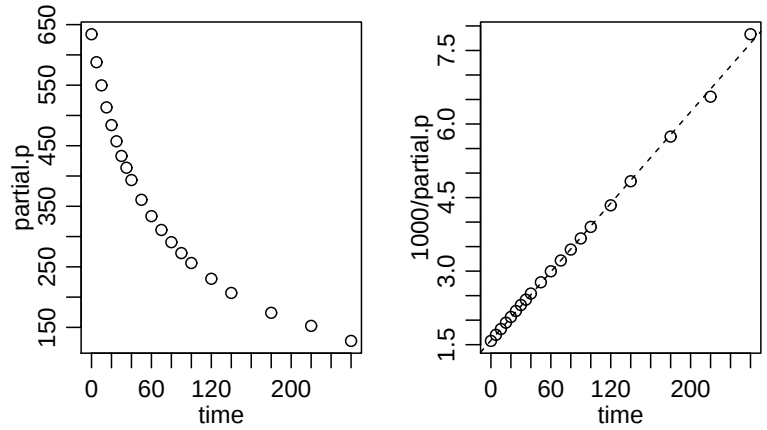

Abbildung 5.1: Partialdruck von Butadien (links) und Kehrwert 1000 × 1/Partialdruck (rechts), gegen die Zeit aufgetragen

5.2.1 Das Modell der einfachen linearen Regression

Im obigen Beispiel haben wir Daten

$$(x_1, y_1), \ldots, (x_n, y_n),$$

wobei x_i der Zeitpunkt der i-ten Messung und y_i der Kehrwert des Partialdrucks der i-ten Messung bezeichnen. Diese fassen wir auf als Realisierungen des folgenden Modells:

$$\begin{aligned} &Y_i = h(x_i) + E_i \ (i = 1, \ldots, n), \\ &E_1, \ldots, E_n \text{ i.i.d. }, \ \mathcal{E}(E_i) = 0, \ \mathrm{Var}(\mathcal{E}_i) = \sigma^2. \end{aligned}$$

Die Y-Variable ist die **Zielvariable** (engl: response variable) und die x-Variable ist die **erklärende Variable** oder **Co-Variable** (engl: explanatory variable; predictor variable; covariate). Die Zufallsvariablen E_i werden öfters als Fehler-Variablen oder Rausch-Terme bezeichnet. Sie besagen, dass der Zusammenhang zwischen der erklärenden und der Ziel-Variablen nicht exakt ist. Die erklärenden Variablen x_i $(i = 1, \ldots, n)$ sind deterministisch, hingegen sind die Ziel-Variablen Y_i Zufallsvariablen (wegen den Zufalls-Variablen E_i).

Die Modelle für die Funktion $h(\cdot)$ sind:

$$\begin{aligned} h(x) &= \beta_0 + \beta_1 x \quad &&: \text{einfache lineare Regression,} \\ h(x) &= \beta_1 x &&: \text{einfache lineare Regression durch Nullpunkt.} \end{aligned}$$

Meistens betrachten wir das allgemeinere Modell mit einem Achsenabschnitt β_0. Das Modell ist illustriert in Abbildung 5.2, wo für die Fehler-Variable eine $\mathcal{N}(0, 0.1^2)$-Verteilung spezifiziert wurde.

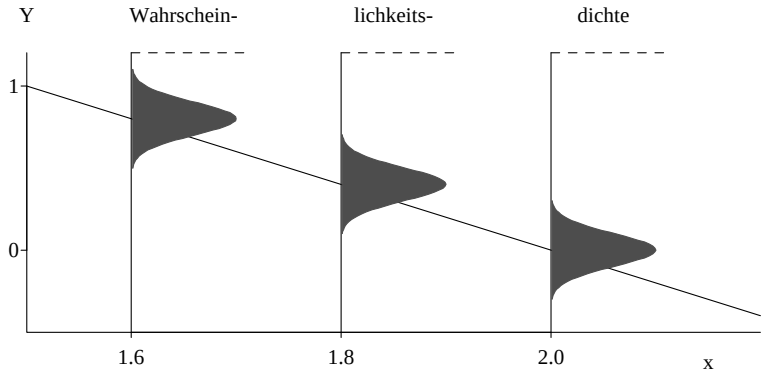

Abbildung 5.2: Veranschaulichung des Regressionsmodells $Y_i = 4 - 2x_i + E_i$ mit $E_i \sim \mathcal{N}(0, 0.1^2)$ für drei Beobachtungen.

5.2.2 Parameterschätzungen

Die unbekannten Modell-Parameter in der einfachen linearen Regression sind β_0, β_1 und auch die Fehlervarianz σ^2. Die Methode der Kleinsten-Quadrate

liefert die folgenden Schätzungen:

$$\hat{\beta}_0, \hat{\beta}_1 \text{ sind Minimierer von } \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 x_i))^2.$$

Die Lösung dieses Optimierungsproblem ergibt:

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

Das Prinzip der Kleinsten-Quadrate liefert sogenannt erwartungstreue Schätzungen:

$$\mathcal{E}(\hat{\beta}_0) = \beta_0, \quad \mathcal{E}(\hat{\beta}_1) = \beta_1.$$

Das heisst, dass die Schätzungen keinen systematischen Fehler haben, sondern um die wahren Parameter herum streuen (wäre z.B. $\mathcal{E}(\hat{\beta}_1) > \beta_1$, dann wäre $\hat{\beta}_1$ systematisch zu gross). Man kann auch die Standardabweichungen der Schätzungen, die sogenannten Standardfehler berechnen. Von Interesse ist insbesondere das Resultat für $\hat{\beta}_1$:

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

Für die Schätzung von σ^2 benützen wir das Konzept der Residuen. Falls wir Realisationen der Fehler-Terme E_i beobachten könnten, so könnten wir die empirische Varianzschätzung für σ^2 verwenden. Hier approximieren wir zuerst die unbeobachteten Fehler-Variablen E_i durch die **Residuen**:

$$R_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \ (i = 1, \ldots, n).$$

Da $E_i = Y_i - (\beta_0 + \beta_1 x_i)$ scheint die Approximation $R_i \approx E_i$ vernünftig. Als Varianzschätzung benutzt man dann:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} R_i^2. \tag{5.2}$$

Dabei ist zu beachten, dass (bei einfacher linearer Regression mit einem Achsenabschnitt β_0) gilt: $\sum_{i=1}^{n} R_i = 0$. Das heisst, die Varianzschätzung in (5.2) ist wie die empirische Varianz bei einer Stichprobe (siehe Kapitel 4.6.1), ausser dass wir den Faktor $1/(n-2)$ anstelle von $1/(n-1)$ nehmen. Dieser Faktor entspricht der folgenden Faustregel: $1/(n - \text{Anzahl Parameter})$, wobei die Anzahl Parameter ohne den zu schätzenden Varianz-Parameter zu zählen ist (in unserem Falle sind dies die Parameter $\beta_0,\ \beta_1$).

Bei einem Datensatz mir realisierten y_i $(i = 1, \ldots, n)$ werden die Schätzungen mit den Werten y_i anstelle von Y_i gerechnet. Zum Beispiel sind dann die realisierten Residuen von der Form $r_i = y_i - (\hat{\beta}_0 - \hat{\beta}_1 x_i)$.

5.2.3 Tests und Konfidenzintervalle

Wir diskutieren hier die 2. und 3. Grundfragestellung (siehe Kapitel 3.1) im Kontext der einfachen linearen Regression. Dabei werden wir entscheidend mehr Schlussfolgerungen ziehen, als bloss eine best passende Regressionsgerade zu finden.

Der t-Test in Regression

Wir betrachten hier als Beispiel den folgenden Datensatz. Es wurden $n = 111$ Messungen gemacht von mittlerer täglicher Temperatur (x-Variable) und mittlerem täglichem Ozongehalt (Y-Variable). Die Daten und die Regressionsgerade $\hat{\beta}_0 + \hat{\beta}_1 x$ sind in Abbildung 5.2.3 ersichtlich. Die interessierend Frage in der Praxis lautet: Hat die Temperatur einen Einfluss auf den Ozongehalt? Diese Frage

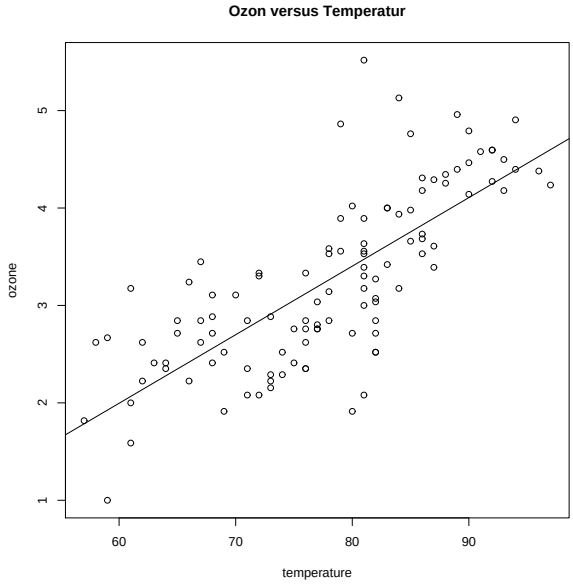

Abbildung 5.3: Streudiagramm und angepasste Regressionsgerade für den Ozon-Temperatur Datensatz.

kann man in ein Test-Problem übersetzen:

$$H_0 : \ \beta_1 = 0,$$
$$H_A : \ \beta_1 \neq 0.$$

Es wird hier "per default" ein zwei-seitiger Test durchgeführt, nämlich der t-Test für die Steigung in der einfachen linearen Regression.

Wir machen hier die Annahme, dass

$$E_1, \ldots, E_n \text{ i.i.d. } \mathcal{N}(0, \sigma^2). \tag{5.3}$$

Die Teststatistik ist

$$T = \frac{\text{beobachtet} - \text{erwartet}}{\text{geschätzter Standardfehler}} = \frac{\hat{\beta}_1 - 0}{\widehat{\text{s.e.}}(\hat{\beta}_1)}.$$

Dabei ist der geschätzte Standardfehler

$$\widehat{\text{s.e.}}(\hat{\beta}_1) = \sqrt{\widehat{\text{Var}}(\hat{\beta}_1)} = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}}.$$

Unter der Nullhypothese und der Annahme von normalverteilten Fehlern in (5.3) gilt:

$$T \sim t_{n-2} \text{ unter } H_0 : \ \beta_1 = 0.$$

Der P-Wert dieses zweiseitigen t-Test kann dann analog wie in Kapitel 4.6.2 berechnet werden (mit $n-2$ anstelle von $n-1$ Freiheitsgraden), und er wird auch von statistischer Software geliefert.

Völlig analog erhält man auch einen Test für $H_0 : \ \beta_0 = 0$ bei zweiseitiger Alternative $H_A : \ \beta_0 \neq 0$. Der entsprechende P-Wert, unter Annahme der Normalverteilung in (5.3), wird von statistischer Software geliefert.

Der Computer-Output vom `R` bei dem Anpassen einer einfachen linearen Regression für den Datensatz von Ozon als Funktion von Temperatur sieht wie folgt aus:

```
Call:
  lm(formula = ozone ~ temperature)

Residuals:
       Min        1Q    Median        3Q       Max
-1.49016 -0.42579  0.02521  0.36362  2.04439

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.225984   0.461408  -4.824 4.59e-06 ***
temperature   0.070363   0.005888  11.951  < 2e-16 ***
---

Residual standard error: 0.5885 on 109 degrees of freedom
Multiple R-Squared: 0.5672,Adjusted R-squared: 0.5632
F-statistic: 142.8 on 1 and 109 DF,  p-value: < 2.2e-16
```

Die zweite Kolonne bei "Coefficients" beschreibt die Punktschätzer $\hat{\beta}_i$ $(i = 0, 1)$; die dritte Kolonne die geschätzten Standardfehler s.e.$(\hat{\beta}_i)$ $(i = 0, 1)$; die vierte Kolonne die Teststatistik $\hat{\beta}_i/\text{s.e.}(\hat{\beta}_i)$ $(i = 0, 1)$, welche sich aus der zweiten dividiert durch die dritte Kolonne ergibt; die fünfte Kolonne bezeichnet den P-Wert für $H_0 : \ \beta_i = 0$ und $H_A : \ \beta_i \neq 0$ $(i = 0, 1)$. Überdies ist die geschätzte Standardabweichung für den Fehler $\hat{\sigma}$ ersichtlich unter "Residual standard error"; die "degrees of freedom" sind gleich $n - 2$.

Konfidenzintervalle

Basierend auf der Normalverteilungsannahme erhält man die folgenden zweiseitigen Konfidenzintervalle für β_i $(i = 0, 1)$ zum Niveau $1 - \alpha$:

$$\hat{\beta}_0 \pm \widehat{\text{s.e.}}(\hat{\beta}_0) t_{n-2;1-\alpha/2} \quad \text{für } \beta_0,$$
$$\hat{\beta}_1 \pm \widehat{\text{s.e.}}(\hat{\beta}_1) t_{n-2;1-\alpha/2} \quad \text{für } \beta_1.$$

5.2.4 Das Bestimmtheitsmass R^2

Die Güte eines Regressionsmodells kann mit dem sogenannten Bestimmtheitsmass R^2 quantifiziert werden. Dazu betrachten wir eine Beziehungen zwischen verschiedenen Variations-Quellen. Wenn wir mit $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ den Wert der angepassten Geraden beim Wert x_i bezeichnen, dann gilt

$$\underbrace{\sum_{i=1}^{n} (y_i - \overline{y})^2}_{SS_Y} = \underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}_{SS_E} + \underbrace{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}_{SS_R}. \tag{5.4}$$

(Das Besondere an dieser Gleichung ist, dass das Doppelprodukt $2\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \overline{y})$ gleich Null ist.) Dabei beschreint SS_Y die totale Variation der Zielvariablen (ohne Einfluss der erklärenden Variablen x), SS_E die Variation des Fehlers (Residuen-Quadratsumme), und SS_R die Variation, welche durch die Regression erklärt wird (Einfluss der erklärenden Variablen x). Das Bestimmtheitsmass ist dann wie folgt definiert:

$$R^2 = \frac{SS_R}{SS_Y}, \tag{5.5}$$

und beschreibt den Anteil der totalen Variation, welche durch die Regression erklärt wird. Wegen 5.4 gilt, dass $0 \leq R^2 \leq 1$: falls R^2 nahe bei 1 ist, so erklärt das Regressionsmodell viel der totalen Variation und ist somit gut; falls $R^2 \approx 0$ taugt das Regressionsmodell nicht besonders viel. Die Realisation von R^2 ist im Computer-Ouput zu finden unter "Multiple R-squared".

Im Falle der einfachen linearen Regression gilt auch:

$$R^2 = \hat{\rho}_{XY}^2,$$

d.h. R^2 ist gleich der quadrierten empirischen Korrelation.

5.2.5 Allgemeines Vorgehen bei einfacher linearer Regression

Grob zusammengefasst kann bei einfacher linearer Regression folgendemassen vorgegangen werden.

1. Anpassen der Regressionsgeraden; d.h. Berechnung der Punktschätzer $\hat{\beta}_0,\ \hat{\beta}_1$.

2. Testen ob erklärende Variable x einen Einfluss auf die Zielvariable Y hat mittels t-Test für $H_0: \beta_1 = 0$ und $H_a: \beta_1 \neq 0$. Falls dieser Test nicht-signifikantes Ergebnis liefert, so ist das Problem "in der vorliegenden Form uninteressant".

3. Testen ob Regression durch Nullpunkt geht mittels t-Test für $H_0: \beta_0 = 0$ und $H_A: \beta_0 \neq 0$. Falls dieser Test nicht-signifikantes Ergebnis liefert, so benützt man das kleinere Modell mit Regression durch Nullpunkt.

4. Bei Interesse Angabe von Konfidenzintervallen für β_0 und β_1.

5. Angabe des Bestimmtheitsmass R^2. Dies ist in gewissem Sinne eine informellere (und zusätzliche) Quantifizierung als der statistische Test in Punkt 2.

6. Überprüfen der Modell-Voraussetzungen mittels Residuenanalyse. Dieser wichtige Schritt wird ausführlicher in Kapitel 5.2.6 beschrieben.

5.2.6 Residuenanalyse

Wir werden hier graphische Methoden beschreiben, basierend auf realisierten Residuen $r_i (i = 1, \ldots, n)$, welche zur Überprüfung der Modell-Voraussetzungen für die einfache lineare Regression eingesetzt werden können. Die Modell-Voraussetzungen sind, in prioritärer Reihenfolge, die folgenden.

1. $\mathcal{E}(E_i) = 0$.
   Somit gilt $\mathcal{E}(Y_i) = \beta_0 + \beta_1 x_i$, das heisst: es gibt keinen systematischen Fehler im Modell.
   Abweichungen von dieser Annahme könnten zum Beispiel durch einen nicht-linearen Zusammenhang zwischen x und Y verursacht sein.

2. $E_1, \ldots, E_n$ i.i.d.
   Identische Verteilung heisst insbesondere, dass die Varianz aller Fehler gleich ist. Abweichungen von dieser Annahme könnten also durch verschiedene Genauigkeiten der Beobachtungen oder durch Abhängigkeiten verursacht sein.

3. $E_1, \ldots, E_n$ i.i.d. $\mathcal{N}(0, \sigma^2)$.
   Abweichungen könnte durch eine lang-schwänzige Fehlerverteilung verursacht sein.

Der Tukey-Anscombe Plot

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: gleichmässige Streuung der Punkte um Null.
Abweichungen:

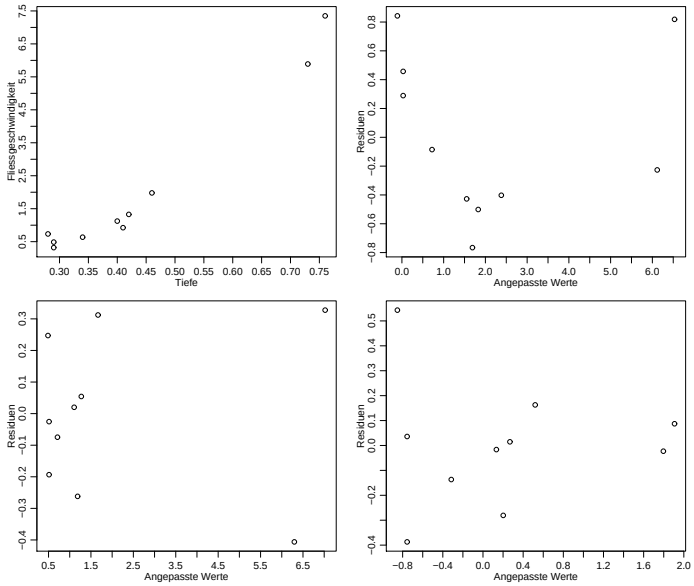

Abbildung 5.4: Streudiagramm von Tiefe und Fliessgeschwindigkeit (oben links), Tukey-Anscombe Plots für einfache lineare Regression (oben rechts), für quadratische Regression (siehe Kapitel 5.3.1) (unten links) und für einfache lineare Regression mit logarithmierten Variablen $\log(Y)$ und $\log(x)$ (unten rechts).

- kegelförmiges Anwachsen der Streuung mit $\hat{y}_i$
  evtl. kann man die Zielvariable logarithmieren (falls Y_i's positiv sind), d.h. man benutzt das neue Modell

$$\log(Y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

- Ausreisserpunkte
  evtl. können robuste Regressions-Verfahren verwendet werden (siehe Literatur)
- unregelmässige Struktur
  Indikation für nichtlinearen Zusammenhang
  evtl. Ziel und.oder erklärende Variablen transformieren (siehe auch das Beispiel in Abbildung 5.1.

Für den Ozon-Datensatz ist der Tukey-Anscombe Plot in Abbildung 5.5 gezeigt.

Nichtlineare Zusammenhänge können in der Praxis natürlich vorkommen: sie zeigen an, dass die Regressionsfunktion nicht korrekt ist. Abhilfe schaffen die Aufnahme zusätzlicher erklärender Variablen (z.B. quadratische Terme, siehe Kapitel 5.3.1) oder - wie bereits oben angedeutet - Transformationen der erklärenden und/oder der Ziel-Variablen. Ein einfaches Beispiel ist in Abbildung 5.4 gezeigt, bei dem es 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 (siehe Kapitel 5.3.1) oder wenn man bei-

de Variablen logarithmiert (d.h. einen Potenzzusammenhang anpasst mit dem Modell

$$\log(Y_i) = \beta_0 + \beta_1 \log(x_i) + \varepsilon_i \ (i = 1, \ldots, n).$$

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-Anscombe Plot besser zu sehen.

Plot bezüglich serieller Korrelation

Um die Unabhängigkeitsannahme der $E_1, \ldots, E_n$ zu überprüfen, kann der folgende Plot gemacht werden: plotte r_i gegen die Beobachtungsnummer i.

Im Idealfall: gleichmässige Streuung der Punkte um Null.
Abweichungen:
- langfristiges Zonen mit durchwegs positiven oder negativen Residuen
  die Punktschätzungen sind immer noch OK, aber die Tests und Konfidenzintervalle stimmen nicht mehr evtl. Regression mit korrelierten Fehlern verwenden (siehe Literatur)

Für den Ozon-Datensatz ist der serielle Korrelations-Plot in Abbildung 5.5 gezeigt.

Der Normalplot

Mit dem Normalplot (siehe Kapitel 4.4.6) können wir die Normalverteilungsannahme in (5.3) überprüfen.

Im Idealfall: approximativ eine Gerade
Abweichungen:
- Abweichung von einer Geraden Evtl. robuste Regression benutzten (siehe Literatur)

Für den Ozon-Datensatz ist der Normalplot in Abbildung 5.5 gezeigt.

Das Auffinden eines guten Modells

Oftmals werden mehrere Modelle in einer Art "workflow-feedback" Prozeß betrachtet und angepasst. Man beginnt mit einem ersten Modell; dann, aufgrund von Residuenanalyse wird das Modell modifiziert. Das modifizierte Modell (immer noch als linear angenommen in evtl. transformierten Variablen) wird wiederum mit linearer Regression angepasst, und mit Residuenanalyse wird das neue Modell beurteilt. Dieses Vorgehen wird iteriert bis man ein "zufriedenstellendes" Modell gefunden und angepasst hat.

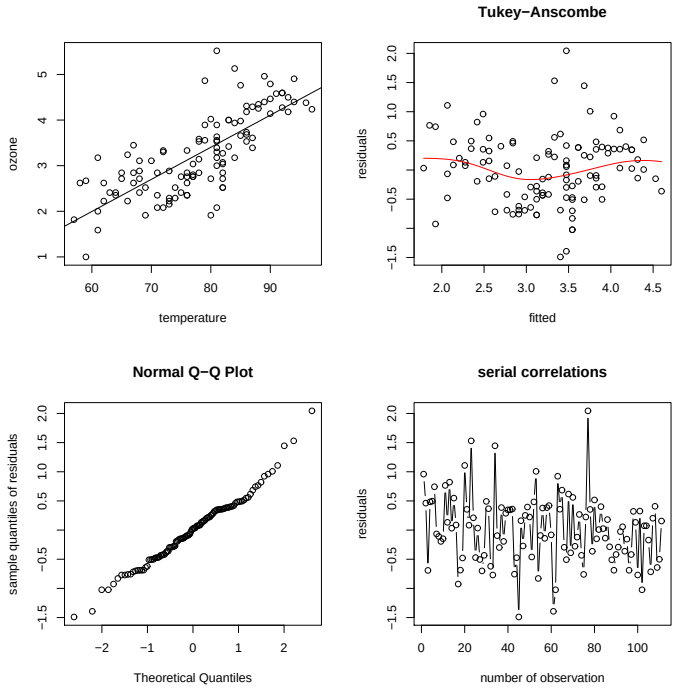

Abbildung 5.5: Ozon-Datensatz: Streudiagramm mit angepasster Regressiongerade (oben links); Tukey-Anscombe Plot (oben rechts); Normalplot (unten links); serieller Korrelations-Plot (unten rechts).

5.3 Multiple lineare Regression

Oftmals hat man mehrere erklärende Variablen $x_{i,1}, \ldots, x_{i,p-1}$ $(p > 2)$.

5.3.1 Das Modell der multiplen linearen Regression

Das Modell ist wie folgt:

$$Y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{i,j} + E_i,$$
$$E_1, \ldots, E_n \text{ i.i.d. }, \ \mathcal{E}(E_i) = 0, \ \mathrm{Var}(\mathcal{E}_i) = \sigma^2.$$

Wie bei der einfachen linearen Regression nehmen wir an, dass die erklärenden Variablen deterministisch sind. Es ist oftmals nützlich, das obige Modell in Matrix-Schreibweise darzustellen:

$$\underset{n\times 1}{Y} = \underset{n\times p}{X} \times \underset{p\times 1}{\beta} + \underset{n\times 1}{E} \tag{5.6}$$

wobei X eine $(n\times p)$-Matrix ist mit Kolonnenvektoren $(1, 1, \ldots, 1)^T$, $(x_{1,1}, x_{2,1}, \ldots, x_{n,1})^T$ und letztendlich $(x_{1,p-1}, x_{2,p-1}, \ldots, x_{n,p-1})^T$.

Beispiele von multipler linearer Regression sind unter anderen:

Einfache lineare Regression: $Y_i = \beta_0 + \beta_1 x_i + E_i$ $(i = 1, \dots n)$.

$$p = 2 \qquad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}.$$

Quadratische Regression: $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + E_i$ $(i = 1, \dots n)$.

$$p = 3, \qquad X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

Zu beachten ist, dass die Funktion quadratisch ist in den x_i's, aber *linear* in den Koeffizienten β_j und deshalb ein Spezialfall des multiplen linearen Regressions Modells.

Regression mit transformierten erklärenden Variablen:
$Y_i = \beta_0 + \beta_1 \log(x_{i2}) + \beta_2 \sin(\pi x_{i3}) + E_i$ $(i = 1, \dots n)$.

$$p = 3, \qquad X = \begin{pmatrix} 1 & \log(x_{12}) & \sin(\pi x_{13}) \\ 1 & \log(x_{22}) & \sin(\pi x_{23}) \\ \vdots & \vdots & \vdots \\ 1 & \log(x_{n2}) & \sin(\pi x_{n3}) \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

Wiederum, das Modell ist *linear* in den Koeffizienten β_j, aber nichtlinear in den x_{ij}'s.

5.3.2 Parameterschätzungen und t-Tests

Analog zur einfachen linearen Regression wird meist die Methode der Kleinsten Quadrate benutzt:

$\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_{p-1}$ sind Minimierer von $\sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 x_{i,1} + \dots + \beta_{p-1} x_{i,p-1}))^2$.

Die eindeutige Lösung dieser Optimierung ist explizit darstellbar falls $p < n$:

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

wobei $\hat{\beta}$ den $p \times 1$ Vektor $(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_{p-1})^T$ bezeichnet, und X, Y wie in (5.6).

Die Schätzung der Fehlervarianz ist

$$\frac{1}{n-p} \sum_{i=1}^{n} R_i^2, \quad R_i = Y_i - (\hat{\beta}_0 + \sum_{j=1}^{p-1} \hat{\beta}_j x_{i,j}).$$

Unter der Annahme, dass die Fehler normalverteilt sind wie in (5.3), können auch ähnlich zur einfachen Regression t-Tests für die folgenden Hypothesen gemacht werden:

$$H_{0,j}: \ \beta_j = 0; \quad H_{A,j}: \ \beta_j \neq 0 \ (j = 0, \ldots, p-1).$$

Der wesentliche Unterschied besteht aber in der Interpretation der Parameter:

β_j misst den linearen Effekt
der j-ten erklärenden Variablen auf die Zielvariable Y
nach Elimination der linearen Effekte
aller anderen Variablen auf Y $(j = 1, \ldots, p-1)$

Insbesondere impliziert dies, dass man die Koeffizienten β_j nicht einfach durch einzelne, individuelle simple lineare Regressionen von Y auf die j-te erklärende erhalten kann.

Beispiel: Wir betrachten $p = 3$ und 2 erklärende Variablen. Wir nehmen an, dass die beiden erklärenden Variablen empirisch stark korreliert sind. Es kann dann durchaus geschehen, dass:
sowohl $H_{0,1}: \ \beta_1 = 0$ als auch $H_{0,2}: \ \beta_2 = 0$ werden nicht verworfen,
obschon mindestens einer der Koeffizienten β_1 oder β_2 ungleich Null ist.
Um den Trugschluss zu vermeiden, dass es keine Effekt der erklärenden Variable auf die Ziel-Variable gibt, muss man den sogenannten F-Test betrachten.

5.3.3 Der F-Test

Der (globale) F-Test quantifiziert die Frage, ob es mindestens eine erklärende Variable gibt, welche einen relevanten Effekt auf die Zielvariable (im Sinne der linear Regression). Die folgende Nullhypothese wird beim (globalen) F-Test betrachtet:

$$H_0: \ \beta_1 = \ldots = \beta_{p-1} = 0$$
$$H_A: \ \text{mindestens ein } \beta_j \neq 0 \ (j = 1, \ldots, p-1).$$

Der P-Wert des (globalen) F-Tests ist im Computer-Output gegeben unter "F-statistic".

5.3.4 Das Bestimmtheitsmass R^2

Das Bestimmtheitsmass R^2 ist in der multiplen linearen Regression über die Formel (5.5) definiert (mit Hilfe der Zerlegung in (5.4). Die Interpretation im Sinne einer quadrierten Stichproben-Korrelation zwischen der Ziel-Variablen und den erklärenden Variablen ist komplizierter als im Fall der einfachen Regression.

5.3.5 Residuenanalyse

Die Residuenanalyse geht völlig analog zu Kapitel 5.2.6. Das allgemeine Vorgehen bei multipler linearer Regression ist wie in Kapitel 5.2.5, unter Einbezug des F-Tests nach dem Schritt 1.

5.3.6 Strategie der Datenanalyse: ein abschliessendes Beispiel

Wir betrachten ein Beispiel wo die Asphalt-Qualität als Funktion von 6 erklärenden Variablen analysiert wird.

```
  y = RUT  : "rate of rutting" = change of rut depth in inches per million
             wheel passes
             ["rut":= 'Wagenspur", ausgefahrenes Geleise]
x1 = VISC : viscosity of asphalt
x2 = ASPH : percentage of asphalt in surface course
x3 = BASE : percentage of asphalt in  base   course
x4 = RUN  : '0/1' indicator for two sets of runs.
x5 = FINES: 10* percentage of fines in surface course
x6 = VOIDS: percentage of voids in surface course
```

Die Daten sind in Abbildung 5.6 dargestellt. Die Zusammenhänge werden linea-

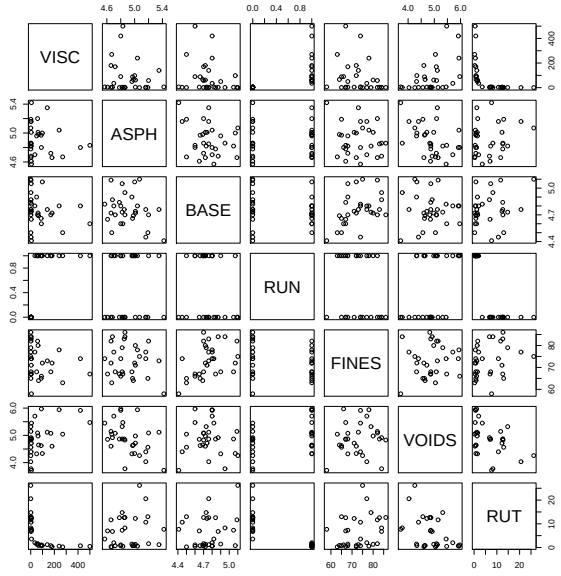

Abbildung 5.6: Paarweise Streudiagramme für den Asphalt-Datensatz. Die Zielvariable ist "RUT".

rer, wenn man die Zielvariable "RUT" logarithmiert und ebenfalls die erklärende Variable "VISC".

```
 y = LOGRUT  : log("rate of rutting") = log(change of rut depth in inches
               per million wheel passes)
               ["rut":= 'Wagenspur", ausgefahrenes Geleise]
x1 = LOGVISC : log(viscosity of asphalt)
x2 = ASPH : percentage of asphalt in surface course
x3 = BASE : percentage of asphalt in  base   course
x4 = RUN  : '0/1' indicator for two sets of runs.
x5 = FINES: 10* percentage of fines in surface course
x6 = VOIDS: percentage of voids in surface course
```

Die transformierten Daten sind in Abbildung 5.7 dargestellt.

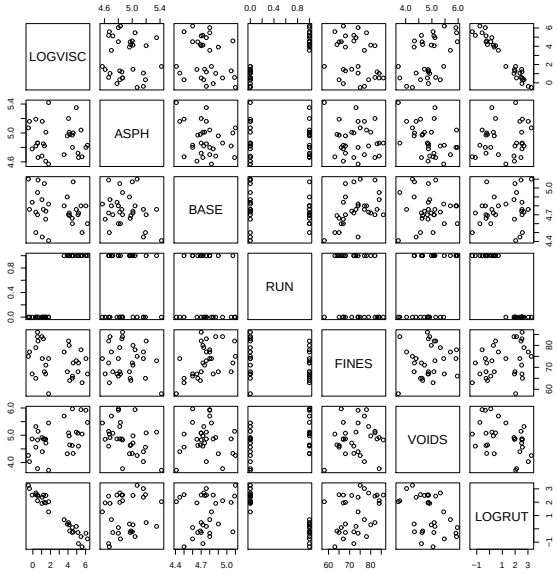

Abbildung 5.7: Paarweise Streudiagramme für den transformierten Asphalt-Datensatz. Die Zielvariable ist "LOGRUT", die log-transformierte ursprüngliche Variable "RUT". Die erklärende Variable "LOGVISC" ist ebenfalls die log-transformierte ursprüngliche Variable "VISC".

Mittels R wird ein multiples lineares Modell angepasst. Der Output sieht wie folgt aus:

```
Call:
lm(formula = LOGRUT ~ ., data = asphalt1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.48348 -0.14374 -0.01198  0.15523  0.39652

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -5.781239   2.459179  -2.351 0.027280 *
LOGVISC     -0.513325   0.073056  -7.027 2.90e-07 ***
ASPH         1.146898   0.265572   4.319 0.000235 ***
BASE         0.232809   0.326528   0.713 0.482731
RUN         -0.618893   0.294384  -2.102 0.046199 *
FINES        0.004343   0.007881   0.551 0.586700
VOIDS        0.316648   0.110329   2.870 0.008433 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2604 on 24 degrees of freedom
Multiple R-Squared: 0.9722,     Adjusted R-squared: 0.9653
F-statistic: 140.1 on 6 and 24 DF,  p-value: < 2.2e-16
```

Wir sehen, dass die Variablen “LOGVISC”, “ASPH” und “VOID” signifikant oder sogar hoch-signifikant sind; die Variable “RUN” ist bloss schwach signifikant. Der F-Test ist hoch-signifikant, das Bestimmtheitsmass R^2 sehr nahe bei 1. Die degrees of freedom sind hier $n - p = 24$ mit $p = 7$, d.h. $n = 31$. Die Residuenanalyse ist mittels Tukey-Anscombe und Normalplot in Abbildung 5.8 zusammengefasst: die Normalverteilungsannahme für die Fehler ist eine vernünftige Approximation. Der Tukey-Anscombe Plot zeigt etwas systematische Variation was durch Nichtlinearität induziert sein könnte; das das R^2 aber bereits sehr nahe bei 1 liegt, so kann man trotzdem sagen, dass die multiple linear Regression sehr viel der totalen Variation erklären kann.

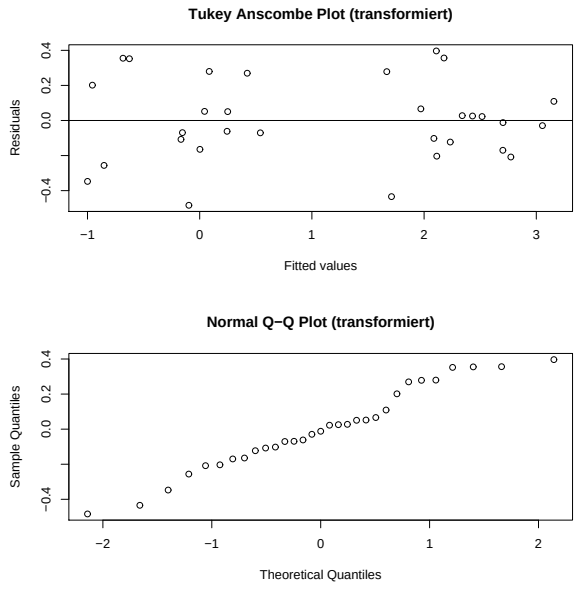

Abbildung 5.8: Tukey-Anscombe Plot (oben) und Normalplot (unten) beim Asphalt-Datensatz mit den transformierten Variablen “LOGRUT” und “LOGVISC”.

Ohne log-Transformationen, d.h. das untransformierte Modell wie in Abbildung 5.6, ist das Bestimmtheitsmass $R^2 = 0.7278$, also wesentlich schlechter als im transformierten Modell.