Vorlesungskurzskript ,,Statistik”

Christoph Kühn

Wintersemester 2010/11

Skript zur Zeit im Aufbau, Hinweise auf Fehler oder Verbesserungsvorschläge sind natürlich sehr willkommen, letzte Aktualisierung: 18. Februar 2011

Inhaltsverzeichnis

Das Kurzskript soll bei der Nachbereitung der Vorlesung und der Klausurvorbereitung helfen. Den aufmerksamen Vorlesungsbesuch kann es aber natürlich nicht ersetzen. Auf Zeichnungen und graphische Veranschaulichungen aus der Vorlesung wird hier verzichtet. Das Kurzskript wird fortlaufend aktualisiert und soll der Vorlesung immer etwas voraus sein.

Was ist Statistik ?

In der Statistik unterscheidet man zwischen der **deskriptiven** (beschreibenden) Statistik und der **induktiven** (schließenden) Statistik. Die deskriptive Statistik besteht eigentlich nur in der Komprimierung und ansprechenden Darstellung von Daten. Es werden gewisse **statistische Kenngrößen** ausgerechnet, die als aussagekräftig für einen (meist großen) Datensatz angesehen werden. Dagegen versucht die induktive Statistik, ausgehend von beobachteten Daten Rückschlüsse auf (noch) nicht beobachtete Daten zu ziehen. Dafür werden die beobachteten Daten als Realisation eines (nicht genau bekannten) Zufallsexperimentes interpretiert und es wird versucht, aus einer Klasse möglicher zugrundeliegende stochastischer Modelle auf ein bestimmtes Modell zu schließen, das unter den Beobachtungen besonders plausibel erscheint, oder es wird versucht die Plausibilität eines vorgegebenen stochastischer Modells zu testen. Mit Informationen über das zugrundeliegende stochastische Modell können dann Rückschlüsse auf (noch) nicht beobachtete Daten (insbesondere Daten, die sich auf die Zukunft beziehen) gezogen werden. Die deskriptive Statistik fusst dagegen nicht notwendigerweise auf dem Glauben an wahrscheinlichkeitstheoretische Modelle. Ihre Möglichkeiten sind dafür entsprechend bescheiden.

1 Deskriptive Statistik

Seien $(x_k)_{k=1,2,\dots,n}$ reellwertige Beobachtungen. $x = (x_k)_{k=1,2,\dots,n}$ bezeichnen wir im Folgenden auch als **Datensatz**.

- Mittelwert (arithmetisches Mittel)

$$\bar{x} = \frac{1}{n} \sum_{k=1}^{n} x_k$$

- Geometrisches Mittel

$$\bar{x}_{\text{geom}} = \left(\prod_{k=1}^{n} x_k \right)^{1/n}, \qquad \text{für } x_k \geq 0 \ \forall k.$$

  Das geometrische Mittel ist eine adäquate Mittelung von *Wachstumsfaktoren*. Nehme an, ein Bestand vermehrt sich im k-ten Jahr *um* den Faktor $x_k - 1$, d.h. der Wachstumsfaktor im k-ten Jahr beträgt x_k. Nach n Jahren ist der Bestand dann genauso stark gewachsen als wenn er sich jedes Jahr um den Faktor $\bar{x}_{\text{geom}} - 1$ gewachsen wäre, d.h. der jährliche Wachstumsfaktor jedes Jahr $\bar{x}_{\text{geom}}$ betragen hätte.

- Harmonisches Mittel

$$\bar{x}_{\text{harm}} = \frac{1}{\frac{1}{n}\sum_{k=1}^{n}\frac{1}{x_k}}, \qquad \text{für } x_k > 0 \ \forall k.$$

Das harmonische Mittel ist z.B. bei Geschwindigkeiten ein sinnvoller Durschnittswert. Nehme etwa an, ein Auto hat n Streckeneinheiten zu überwinden und fährt im k-ten Streckenstück mit der Geschwindigkeit x_k. Die Durchschnittsbeschwindigkeit der Fahrt beträgt dann $\bar{x}_{\text{harm}}$ (für das k-te Streckenstück braucht das Auto $1/x_k$ Zeiteinheiten und damit $\sum_{k=1}^{n} 1/x_k$ Zeiteinheiten für die Strecke der Länge n).

Es gilt $\bar{x}_{\text{harm}} \leq \bar{x}_{\text{geom}} \leq \bar{x}$.

- Ordnungsstatistiken

$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ bezeichnet die der Größe nach geordneten Daten $x_1, x_2, \ldots, x_n$. Formal lassen sich $x_{(k)}$ wie folgt definieren

$$x_{(k)} := \min\{x_i \mid \#\{j \mid x_j \leq x_i\} \geq k\}, \quad k = 1, \ldots, n \tag{1.1}$$

- Empirische Verteilungsfunktion

$$F_x(y) := \frac{1}{n} \cdot \#\{k \mid x_k \leq y\}, \qquad y \in \mathbb{R}$$

- Quantile

$$q_x(p) := \inf\{y \mid F_x(y) \geq p\}, \qquad p \in (0, 1] \tag{1.2}$$

bezeichnet das p-Quantil des Datensatzen $x = (x_1, \ldots, x_n)$. $q_x(p)$ ist also der kleinste Wert, so dass mindestens $p \times 100$ Prozent der Daten kleiner oder gleich $q_x(p)$ sind. Die Funktion $p \mapsto q_x(p)$ ist linksstetig, d.h. es gilt $\lim_{h \to 0,\ h<0} q_x(p+h) = q_x(p)$ für alle $p \in (0, 1]$. Es gilt zudem $q_x(k/n) = x_{(k)}$ für $k = 1, \ldots, n$. Für $p = 0$ würde sich in obiger Definition $q_x(0) = -\infty$ ergeben, was auch Sinn macht, aber gewöhnlich nicht betrachtet wird.

- Median

$$x_{\text{med}} := \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{für } n \text{ ungerade} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & \text{für } n \text{ gerade} \end{cases}$$

Für ungerade n gilt offenbar $x_{\text{med}} = q_x(0.5)$, d.h. der Median ist in diesem Fall das 50%-Qunantil der empirischen Verteilung. Für gerade n gilt $x_{\text{med}} \geq x_{(n/2)} = q_x(0.5)$, aber die Differenz ist natürlich für große Datensätze zumeist vernachlässigbar.

Bei Verteilungsfunktionen von Zufallsvariablen wird der Median auch oft als 50%-Qunantil definiert. x_{med} wie oben definiert angewandt auf die *Realisationen* eines Zufallsexperiments nennt man dann auch *Stichprobenmedian*, was zumeist ein guter Schätzer für den Median der theoretischen Verteilungsfunktion ist (aber dazu mehr in der induktiven Statistik).

Der Median hat gegenüber dem Mittelwert eines Datensatzes den Vorteil, dass er bei einer montonen Transformationen des Datensatzes (im wesentlichen) einfach mittransformiert wird. Sei n ungerade und y ein weiterer Datensatz mit $y_k = f(x_k)$, $k = 1, \ldots, n$ für eine nichtfallende Funktion f. Dann gilt $y_{\text{med}} = f(x_{\text{med}})$. Dies ist eine wünschenswerte Eigenschaft, wenn es keine kanonische Skalierung gibt (wenn man sich einen Datensatz genauso gut auf einer logarithmischen Skala anschauen könnte, also statt $(x_1, \ldots, x_n)$ lieber $(\ln(x_1), \ldots, \ln(x_n))$ betrachtet).

Der Median x_{med} (und im Falle eines geraden n alle Werte im Intervall $[x_{(n/2)}, x_{(n/2+1)}]$) minimiert die Summe der absoluten Abweichungen der Daten zu einem festen anderen Wert $\widetilde{x}$, d.h.

$$\sum_{k=1}^{n} |x_k - x_{\text{med}}| = \min_{\widetilde{x} \in \mathbb{R}} \left(\sum_{k=1}^{n} |x_k - \widetilde{x}| \right)$$

(der Mittelwert $\bar{x}$ minimiert ja bekanntlich die Summe der quadratischen Abweichungen).

- Empirische Varianz

$$\operatorname{var}(x) := \frac{1}{n} \sum_{k=1}^{n} (x_k - \bar{x})^2$$

- Empirische Standardabweichung (radizierte empirische Varianz)

$$\sigma(x) := \sqrt{\operatorname{var}(x)} = \sqrt{\frac{1}{n} \sum_{k=1}^{n} (x_k - \bar{x})^2}$$

  Die Standardabweichung besitzt gegenüber der Varianz den Vorteil, dass ihre Dimension mit der von x_k übereinstimmt.

- Lorenzkurve

  Die Lorenzkurve wurde Anfang des 20. Jahrhunderts von Max O. Lorenz entwickelt und zur Charakterisierung der Disparität (Ungleichheit) der Haushaltseinkommen einer Volkswirtschaft benutzt. Allgemein kann man mit ihr die Konzentration von Verteilungen nichtnegativer Größen untersuchen. Formal lässt sich die Lorenzkurve

eines Datensatzes $x = (x_1, x_2 \dots, x_n)$, für den gilt $x_k \geq 0$ für alle $k = 1, 2, \dots, n$ und $x_k > 0$ für mindestens ein k, wie folgt definieren

$$L := \left\{ (u, v) \in [0,1]^2 \mid v = \frac{\int_0^u q_x(p)\, dp}{\int_0^1 q_x(p)\, dp} \right\},$$

d.h. L ist der Graph der Funktion l mit

$$u \mapsto l(u) := \frac{\int_0^u q_x(p)\, dp}{\int_0^1 q_x(p)\, dp}. \tag{1.3}$$

Es gilt $l(0) = 0$ und

$$l(k/n) = \frac{\sum_{i=1}^k x_{(i)}}{\sum_{i=1}^n x_{(i)}}, \quad \forall k = 1, \dots, n.$$

L lässt sich wie folgt konstruieren: zum geordneten Datensatz $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ verbinde man die Punkte

$$(0,0), (u_1, v_1), (u_2, v_2), \dots, (u_n, v_n) = (1,1),$$

wobei

$$u_k := \frac{k}{n}$$

und

$$v_k := l(k/n) = \frac{\sum_{j=1}^k x_{(j)}}{\sum_{j=1}^n x_{(j)}}.$$

Man sieht, dass die Funktion l monoton nicht-fallend und konvex ist (da die Steigung ist nicht-fallend) und, dass $l(u) \leq u$ für alle $u \in [0,1]$ (da $p \mapsto q_x(p)$ nicht-fallend), d.h. l verläuft unterhalb der Diagonalen.

- Gini-Koeffizient

  Die Ungleichheit der Verteilung kommt durch den Abstand der Lorenzkurve von der Diagonalen zum Ausdruck. Ein naheliegendes Maß für die Disparität ist daher der Flächeninhalt zwischen Diagonale und Lorenzkurve.

$$\begin{aligned} G &= \frac{\text{Fläche zwischen Diagonale und Lorenzkurve}}{\text{Fläche zwischen Diagonale und } u\text{-Achse}} \\ &= 1 - \frac{\text{Fläche zwischen Lorenzkurve und } u\text{-Achse}}{\text{Fläche zwischen Diagonale und } u\text{-Achse}} \\ &= 1 - 2 \cdot \text{Fläche zwischen Lorenzkurve und } u\text{-Achse} \end{aligned}$$

  Formal sind Flächen natürlich über Integrale definierbar, also „Fläche zwischen Lorenzkurve und u-Achse" $:= \int_0^1 l(u)\, du$ bzw.

$$G := 1 - 2 \int_0^1 l(u)\, du$$

Proposition 1.1. *Es gilt*

$$G = \frac{2\sum_{k=1}^{n} k x_{(k)}}{n\sum_{k=1}^{n} x_k} - \frac{n+1}{n}. \tag{1.4}$$

Proof. Es gilt

$$\begin{aligned}
\int_0^1 l(u)\,du &= \frac{1}{2n}\sum_{k=1}^{n}(v_k + v_{k-1}) \\
&= \frac{1}{2n}\left(2\sum_{k=1}^{n} v_k - v_n + v_0\right) \\
&= \frac{1}{2n}\left(\frac{2\sum_{k=1}^{n}\sum_{j=1}^{k} x_{(j)}}{\sum_{j=1}^{n} x_{(j)}} - 1\right) \\
&= \frac{1}{n}\left(\frac{\sum_{j=1}^{n} x_{(j)}\sum_{k=j}^{n} 1}{\sum_{j=1}^{n} x_{(j)}} - \frac{1}{2}\right) \\
&= \frac{1}{n}\left(\frac{(n+1)\sum_{j=1}^{n} x_{(j)} - \sum_{j=1}^{n} j x_{(j)}}{\sum_{j=1}^{n} x_{(j)}} - \frac{1}{2}\right) \\
&= 1 + \frac{1}{2n} - \frac{1}{n}\frac{\sum_{j=1}^{n} j x_{(j)}}{\sum_{j=1}^{n} x_{(j)}}
\end{aligned}$$

Daraus folgt

$$G = 1 - 2\int_0^1 l(u)\,du = -1 - \frac{1}{n} + \frac{2\sum_{k=1}^{n} k x_{(k)}}{n\sum_{k=1}^{n} x_k}.$$

□

Die Werte, die der Gini-Koeffizient annehmen kann, liegen im Intervall $[0, 1 - 1/n]$. Die beiden Extremsituation treten genau dann auf, wenn

$$x_1 = x_2 = \ldots = x_n \quad \rightsquigarrow G = 0$$

und

$x_k = 0$ für alle k außer einem, bei dem der Wert größer als 0 ist $\quad \rightsquigarrow G = 1 - 1/n$.

An der Darstellung (1.4) sieht man sofort, dass die minimalen/maximalen Werte der Koeffizienten genau in obigen Fällen angenommen werden (man beachte dabei die Nebenbedingung, dass $x_k \geq 0$ für alle k und $x_k > 0$ für mindestens ein k).

Bemerkung 1.2. *Der maximale Wert hängt offenbar von n ab, auch wenn dies für große n nicht sonderlich ins Gewicht fällt. Um diese Abhängigkeit von n zu vermeiden, wird oft der norminierte Gini-Koeffizient (Lorenz-Münzner-Koeffizient)*

$$G^* = \frac{n}{n-1}G$$

betrachtet.

Nach einer Erhebung aus dem *United Nations Human Development Report* im Jahre 2004 liegt der Gini-Koeffizient der Verteilung der verfügbaren Einkommen der Haushalte in Deutschland bei 0,274 (2003), in Frankreich bei 0,327 (1995), in Großbritannien bei 0,360 (1999), in Japan bei 0,249 (1993) und in den USA bei 0,408 (2000). Natürlich ist zwischen Einkommen und Vermögen und vielen anderen Dingen wie verschiedenen Einkommensarten, von der öffentlichen Hand kostenfrei bereitgestellte Leistungen etc. genau zu unterscheiden, weswegen unterschiedliche Erhebungen natürlich zu abweichenden Ergebnissen führen. Eine Auseinandersetzung mit den vielfältigen Erhebungen von Einkommensverteilungen ist aber nicht Gegenstand dieser Vorlesung.

2 Induktive Statistik

Aufgabe der **Wahrscheinlichkeitstheorie** ist die mathematische Modellierung von Zufallsexperimenten. Die formale Beschreibung erfolgt meist über einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, P)$.

Ω: ,,beliebige" Menge, $\omega \in \Omega$ nennen wir ein **Ergebnis**

$\mathcal{F}$: Mengensystem bestehend aus Teilmengen von Ω, d.h. $\mathcal{F} \subset 2^\Omega$. $A \in \mathcal{F}$ nennen wir ein **Ereignis**. $\mathcal{F}$ soll zudem eine σ-Algebra sein, d.h. folgende Eigenschaften besitzen

(i) $\Omega \in \mathcal{F}$

(ii) $A \in \mathcal{F} \implies A^c := \Omega \setminus A \in \mathcal{F}$

(iii) $A_1, A_2, \ldots \in \mathcal{F} \implies \bigcup_{n=1}^\infty A_n \in \mathcal{F}$

(iv) Wahrscheinlichkeitsmaß P: $\mathcal{F} \to [0,1]$, $P(\Omega) = 1$, σ-additiv, d.h. für jede Folge von disjunkten Ereignissen $(A_n)_{n\in\mathbb{N}} \subset \mathcal{F}$ gilt $P(\bigcup_{n=1}^\infty A_n) = \sum_{n=1}^\infty P(A_n)$.

Interpretation: $\mathcal{F} = \{A \subset \Omega \mid A \text{ ist beobachtbares Ereignis}\}$.

Das Experiment besteht in dem ,,Ziehen" eines $\omega \in \Omega$. Meistens ist man nicht an dem Ergebnis ω selber interessiert, sondern an einer Größe, die von ω abhängt. Diese Abhängigkeit modelliert man durch eine Abbildung

$$X : \Omega \to \mathbb{R}^n.$$

Von der Abbildung X wird gefordert, dass sie $\mathcal{F} - \mathcal{B}(\mathbb{R}^n)$-messbar ist, d.h. $\{X \in B\} = \{\omega \in \Omega \mid X(\omega) \in B\} \in \mathcal{F}$ für alle $B \in \mathcal{B}(\mathbb{R}^n)$. Dabei bezeichnet $\mathcal{B}(\mathbb{R}^n)$ die Borelsche-σ-Algebra von $\mathbb{R}^n$. Es ist eine Menge von ,,geeigneten" Teilmengen des $\mathbb{R}^n$, denen man Wahrscheinlichkeiten zuordnen kann. Ein Erzeuger von $\mathcal{B}(\mathbb{R}^n)$ sind die kartesischen Produkte von offenen Intervallen, d.h. Mengen der Form $(a_1, b_1) \times (a_2, b_2) \times \ldots \times (a_n, b_n)$. Details über die Beschaffenheit von $\mathcal{B}(\mathbb{R}^n)$ müssen uns in dieser Vorlesung nicht interessieren. X bezeichnet man als **Zufallsvariable**. Bisweilen ist auch die Zufallsvariable

X die Basisgröße des Modells bei Hintanstellung des Grundraumes Ω. Man arbeitet mit $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), P_X)$, wobei

$$P_X(B) := P(X \in B) := P(\{\omega \in \Omega \mid X(\omega) \in B\}), \quad B \in \mathcal{B}(\mathbb{R}^n)$$

das **Bildmaß** von P unter der Abbildung X bezeichnet. Aus Beobachtungen einer Realisation von X können nämlich nur Rückschlüsse auf die Verteilung P_X von X gezogen werden und keine weitergehenden Rückschlüsse auf die genaue Beschaffenheit des Grundraumes Ω, auf dem die Zufallsvariable X definiert ist.

Beispiel 2.1 (n-faches Würfeln)**.**

$$\Omega = \{1,2,3,4,5,6\}^n = \{\omega = (\omega_1, \omega_2, \dots, \omega_n) \mid \omega_k \in \{1,2,3,4,5,6\}\},$$

$\mathcal{F} = 2^\Omega$ *und* $P(A) = \frac{\#A}{6^n}$*. Beispiel für eine Zufallsvariable: die Augensumme*

$$X(\omega) = X(\omega_1, \omega_2, \dots, \omega_n) = \sum_{k=1}^{n} \omega_k$$

Anzahl der Fünfen

$$Y(\omega) = Y(\omega_1, \omega_2, \dots, \omega_n) = \sum_{k=1}^{n} 1_{\{\omega_k = 5\}}$$

Anders als beim Würfeln ist in der **Statistik** das zugrundeliegende Wahrscheinlichkeitsmodell nicht vollständigt bekannt und man versucht aus Beobachtungen, die man als Realisation $X(\omega)$ des Zufallsexperimentes interpretiert, auf das zugrundeliegende stochastische Modell zu schließen, d.h. auf die **Verteilung** P_X der Zufallsvariablen X. Diese Vorgehensweise hat sich in verschiedenartigsten Anwendungen in den Sozial- bis zu den Naturwissenschaften als extrem erfolgreich erwiesen.

Formal beschreiben wir das Zufallsexperiment durch einen Wahrscheinlichkeitsraum, auf dem eine **Familie von Wahrscheinlichkeitsmaßen** operiert, d.h.

$$(\Omega, \mathcal{F}, (P_\vartheta)_{\vartheta \in \Theta}).$$

Interpretation: Es gibt einen ,,wahren" Parameter $\vartheta \in \Theta$ und das Zufallsexperiment wird gemäß des Wahrscheinlichkeitsmaßes P_ϑ durchgeführt. Dem Beobachter ist der wahre Parameter jedoch nicht bekannt. Er weiß nur, dass dieser in der Menge Θ liegt.

Ausgangspunkt der praktischen Modellierung sind zumeist Zufallsvariablen und deren stochastische Verteilung(en). Da sich eine vorgegebene Verteilung P^X einer Zufallsvariablen X auf den verschiedensten Wahrscheinlichkeisräumen modellieren lässt, ist es zweckmässig direkt zu den **Bildräumen** überzugehen. Für das statistische Modell mit mehreren Wahrscheinlichkeitsmaßen sieht dies wie folgt aus:

Definition 2.2. *Wir nennen*

$$\mathcal{M} = (\mathcal{X}, \mathcal{B}, (P_\vartheta^X)_{\vartheta\in\Theta})$$

ein **statistisches Modell**. *Dabei ist die Menge $\mathcal{X}$ der Zustandsraum einer i.A. mehrdimensionalen Zufallsvariablen X und $\mathcal{B}$ ist eine σ-Algebra auf der Menge $\mathcal{X}$. P_ϑ^X ist die Verteilung der Zufallsvariablen X mit Zustandsraum $\mathcal{X}$ unter dem Parameter ϑ. Θ ist eine mindestens zweielementige Indexmenge.*

Sei $\mathcal{T} \neq \emptyset$ eine Menge und $\mathcal{D}$ eine σ-Algebra auf $\mathcal{T}$. Eine $\mathcal{B}-\mathcal{D}$-messbare Abbildung $T : \mathcal{X} \to \mathcal{T}$ bezeichnen wir als eine **Stichprobenfunktion** *oder* **Statistik** *(T bewirkt i.A. eine Reduktion des Datenmaterials).*

Meistens ist die Beobachtung endlich-dimensional, d.h. $\mathcal{X} = \mathbb{R}^n$ und $\mathcal{B} = \mathcal{B}(\mathbb{R}^n)$. $\mathcal{X}$ kann jedoch auch der Raum der reellwerigen Folgen sein, d.h. $\mathcal{X} = \mathbb{R}^\infty$.

Definition 2.3. *Ein statistisches Modell $\mathcal{M} = (\mathcal{X}, \mathcal{B}, (P_\vartheta^X)_{\vartheta\in\Theta})$ heißt ein* **parametrisches Modell**, *wenn $\Theta \subset \mathbb{R}^d$ für ein $d \in \mathbb{N}$. Ist $d = 1$, so heißt $\mathcal{M}$ ein* einparametriges *Modell.*

Man spricht auch von *parametrischer Statistik*, wenn die Verteilungen $(P_\vartheta^X)_{\vartheta\in\Theta}$ alle aus einer bestimmen Klasse kommen. Also z.B. der zweiparametrischen Verteilungsfamilie der eindimensionalen Normalverteilung mit unbekanntem Erwartungswert und Varianz, d.h. $(P_\vartheta^X)_{\vartheta\in\Theta} = ((\mathcal{N}_{\mu,\sigma})^{\otimes n})_{\mu\in\mathbb{R},\ \sigma\in\mathbb{R}_+}$, wobei für ein Maß Q der Ausdruck $Q^{\otimes n}$ für das n-fache Produktmaß steht. Θ kann jedoch, wie im folgenden Beispiel, auch alle (Rand-)Verteilungen indizieren.

Beispiel 2.4. *Sei $X = (X_1, X_2, \dots, X_n)$, X_k i.i.d., $P_F(X_k \leq x) = F(x)$, $\forall x \in \mathbb{R}$, $k = 1, \dots, n$, wobei $F(x)$ eine eindimensionale Verteilungsfunktion ist. Dies bedeutet:*
$\mathcal{X} = \mathbb{R}^n$, $\mathcal{B} = \mathcal{B}(\mathbb{R}^n)$, $F_{X_1,\dots,X_n} = F^{\otimes n}$, $\vartheta = F \in \{F : \mathbb{R} \to [0,1] \mid F$ ist Verteilungsfunktion$\}$.

Ein Beispiel für ein einparametrisches Modell, mit einer spezielleren Verteilungsannahme, ist das folgende

Beispiel 2.5. *[Binomialmodell] Sei $X = (X_1, \dots, X_n)$. Es sei bekannt, dass X_k, $k = 1, \dots, n$ i.i.d. sind und dass X_1 binomialverteilt ist und nur den Wert 0 und 1 annehmen kann. $X_k = 1$ bedeutet etwa, dass das k-te Experiment erfolgreich war und $X_k = 0$ bedeutet, dass das k-te Experiment fehlschlug.*

Nur der Parameter p der Binomialverteilung, also die Wahrscheinlichkeit, mit der X_1 den Wert 1 annimmt, sei unbekannt. Formal lautet das statistische Modell hierzu

$$(\mathcal{X}, \mathcal{B}, (P_\vartheta^X)_{\vartheta\in\Theta}) = (\{0,1\}^n, \mathcal{P}(\{0,1\}^n), ((B(1,p))^{\otimes n})_{p\in[0,1]})$$

(wobei $\mathcal{P}(\mathcal{X})$ die Potenzmenge der Menge $\mathcal{X}$ bezeichnet). Es gilt

$$(B(1,p))^{\otimes n}(\{(x_1, \dots, x_n)\}) = p^{\#\{i|x_i=1\}}(1-p)^{\#\{i|x_i=0\}}, \quad x_i \in \{0,1\}.$$

In einem guten statistischen Modell darf die Anzahl der unbekannten Parameter, *über die man etwas erfahren will*, nicht zu groß im Vergleich zur Anzahl der Beoachtungen (hier gleich n) sein, weil es ansonsten unmöglich wird, aus den Beobachtungen einigermaßen verlässliche Rückschlüsse auf die Parameter zu ziehen.

Ein solches Problem würde auftreten, wenn im Modell aus Beispiel 2.5 nicht a priori Unabhängigkeit von $X_1, \dots, X_n$ angenommmen würde und man stattdessen aus den Daten etwas über die Abhängigkeiten erfahren möchte. Dann müssten aus den Daten Hinweise auf die Wahrscheinlichkeiten aller Ereignisse $\{X_1 = x_1, \dots, X_n = x_n\}, \quad x_k \in \{0,1\}$ für $k = 1, \dots, n$ gewonnen werden. Den n Beoachtungen stünden 2^n Parameter mit $2^n - 1 - (n-1) = 2^n - n$ Freiheitsgraden gegenüber (wenn Randverteilungen für alle X_k gleich sein sollen), während es in Beispiel 2.5 lediglich einen Freiheitgrad gibt.

2.1 Testen von Hypothesen

Beim Testen beginnt man mit einer **Hypothese** (auch Nullhypothese genannt), etwa

$$H : \vartheta \in \Theta_0.$$

D.h. der ,,wahre" Parameter der Verteilung liege in der Teilmenge $\Theta_0 \subset \Theta$. Die **Gegenhypothese** hierzu ist

$$K : \vartheta \in \Theta \setminus \Theta_0.$$

Θ_0 könnte z.B. einelementig sein, also $\Theta_0 = \{\vartheta_0\}$. Man kann das Experiment wie folgt interpretieren. Im Prinzip glaubt man an die Hypothese H. Nur wenn die Daten, also die Realisierung $X(\omega)$ nur sehr schwerlich mit der Hypothese vereinbar erscheint, wird die Hypothese verworfen. Ein Test ist nun eine Regel, die besagt, bei welchem Datensatz, die Hypothese zu verwerfen ist und bei welchem nicht. Da dabei i.A. keine sicheren Schlüsse gezogen werden können, muss ein Zielkonflikt gelöst werden. Einerseits: wenn die Hypothese stimmt, soll die Wahrscheinlichkeit, dass sie fälschlicherweise verworfen wird, möglichst klein sein. Andererseits: wenn die Hypothese falsch ist, soll die Wahrscheinlichkeit, dass sie fälschlicherweise nicht verworfen wird, möglichst klein sein. Man kann nun die Wahrscheinlichkeit, die Hyothese zu verwerfen, obwohl sie richtig ist, durch eine Zahl $\alpha \in [0,1]$ beschränken und versuchen dazu einen Test zu entwickeln, bei dem die Wahrscheinlichkeit nicht zu verwerfen, obwohl die Hypothese falsch ist, möglichst klein wird.

Wir definieren dies nun formal

Definition 2.6. *Unter einem* **Test** *zum Niveau $\alpha \in (0,1)$ für eine Hypothese $H : \vartheta \in \Theta_0$ gegen $K : \vartheta \in \Theta \setminus \Theta_0$ versteht man eine ($\mathcal{B} - \mathcal{B}([0,1])$-messbare) Abbildung*

$$\varphi : \mathcal{X} \to [0,1]$$

$$\textit{mit} \qquad E_\vartheta(\varphi(X)) \leq \alpha, \quad \forall \vartheta \in \Theta_0 \quad \text{,, } \textbf{Fehler 1. Art} \leq \alpha\text{''}.$$

φ entscheidet, bei welchen Kombination von $(X_1, \dots, X_n)$ die Hypothese abgelehnt wird. $\varphi = \mathbf{1}$ **bedeutet, dass die Hypothese H verworfen wird (Entscheidung für K).** $\varphi = \mathbf{0}$ **bedeutet, dass die Hypothese H nicht verworfen wird. Bei** $\varphi = \delta \in (\mathbf{0}, \mathbf{1})$ **wird mit Wahrscheinlichkeit** δ **verworfen**.

Als **Fehler 2. Art** bezeichnet man die Erwartungswerte

$$E_\vartheta(1 - \varphi(X)), \quad \vartheta \notin \Theta_0$$

also im Wesentlichen die Wahrscheinlichkeiten $P_\vartheta(\varphi(X) = 0)$, $\vartheta \notin \Theta_0$, also die Wahrscheinlichkeiten, die Hypothese nicht zu verwerfen, obwohl sie falsch ist. $E_\vartheta(\varphi(X))$ für $\vartheta \notin \Theta_0$ wird als **Macht** des Tests bezeichnet.

Die Funktion $G_\varphi : \Theta \to [0, 1]$ mit $G_\varphi(\vartheta) = E_\vartheta(\varphi(X))$ wird **Gütefunktion** des Tests φ genannt.

Einen Test φ, der nur Werte in $\{0, 1\}$ annimmt, nennt man nichtrandomisiert. In diesem Fall sagt man, dass $\{x \in \mathcal{X} \mid \varphi(x) = 1\}$ der **Ablehungsbereich** oder der **Verwerfungsbereich** des Test und $\{x \in \mathcal{X} \mid \varphi(x) = 0\}$ der **Annahmebereich** ist.

Definition 2.7. *Ein Test φ der Nullhypothese Θ_0 gegen $\Theta \setminus \Theta_0$ heißt ein (gleichmäßig) bester Test zum Niveau α, wenn er vom Niveau α ist und für jeden anderen Test ψ zum Niveau α gilt, dass*

$$G_\varphi(\vartheta) \geq G_\psi(\vartheta) \quad \forall \vartheta \in \Theta \setminus \Theta_0.$$

In der englischsprachigen Literatur verwendet man den Begriff "UMP test" für "uniformly most powerful test".

Definition 2.8. *Ein Test φ der Nullhypothese Θ_0 gegen $\Theta \setminus \Theta_0$ heißt* **unverfälscht** *zum Niveau α, wenn*

$$G_\varphi(\vartheta_0) \leq \alpha \leq G_\varphi(\vartheta_1) \quad \forall \vartheta_0 \in \Theta_0,\ \vartheta_1 \in \Theta \setminus \Theta_0.,$$

d.h. wenn man sich mit größerer Wahrscheinlichkeit für eine Alternative entscheidet, wenn sie richtig ist als wenn sie falsch ist.

Bemerkung 2.9. *Die Möglichkeit, die Entscheidung für oder gegen die Hypothese zu* **randomisieren**, *d.h. von einem stochastisch unabhängigen Zufallsexperiment abhängig zu machen, ist von Interesse, wenn die Statistk $T(X)$, auf deren Grundlage entschieden werden soll, nicht stetig verteilt ist. Durch eine Radomisierung kann erreicht werden, dass das Testniveau α exakt eingehalten wird. Bei einer Unterschreitung würde man zu selten ablehnen und damit die Fehlerwahrscheinlichkeit 2. Art unnötig groß machen.*

Man beachte, dass der Parameter ϑ zwar in die Analyse der Fehlerwahrscheinlichkeit eingeht, nicht aber in die Abbildung φ. Die Entscheidung wird also ausschließlich aufgrund der Daten gefällt. Der wahre Parameter ist ja nicht bekannt. α ist klein, typische Werte sind 0.01 oder 0.05.

Beispiel 2.10 (**Binomialtest**). *Diplom-Mathematiker Reiner D. hat ein Verfahren entwickelt, mit dem man zukünftige Aktienkurse besser als der Markt prognostizieren kann. Damit bewirbt er sich um die Stelle eines Abteilungsleiters einer großen Fondgesellschaft. Da sich Herr D. im Vorstellungsgespräch mit Details über sein Prognoseverfahren sehr bedeckt hält, möchte der Vorstand die Qualität durch einen möglichst einfachen statistischen Test überprüfen. Herr D. soll für ausgwählte $n = 20$ Aktien sagen, ob sie innerhalb eines bestimmten Zeitintervalls steigen oder fallen werden (der Fall, dass der Preis genau auf demselben Niveau bleibt, soll der Einfachheit halber ausgeschlossen werden). Die Aktien und Zeiträume sind so gewählt, dass man nährungsweise davon ausgehen kann, dass das Auf-oder-ab der verschiedenen Aktien stochastisch unabhängig ist und dass die Wahrscheinlichkeit aus Sicht eines Laien, dass eine Aktie steigt, $\frac{1}{2}$ beträgt. Z.B. soll die Drift der Aktie nicht stören, was realistisch ist, wenn die Zeiträume sehr kurz sind und damit die zufälligen Schwankungen dominieren. Ein geeignetes einfaches statistisches Modell scheint dann das Binomialmodell aus Beispiel 2.5 zu sein.*

$X_k = 1$ bedeutet, dass Herr D. bei der k-ten Aktie mit seiner Prognose richtig liegt. Bei $X_k = 0$ liegt er falsch. Das statistische Modell, mit dem der Vorstand arbeitet, lautet

$$(\{0,1\}^n, \mathcal{P}(\{0,1\}^n), ((B(1,p))^{\otimes n})_{p\in[1/2,1]})$$

wobei p die unbekannte Trefferwahrscheinlichkeit von Herrn D. bezeichnet. Wenn das p von Herrn D. echt größer als $1/2$ ist, dann ist er der richtige Kandidat für die Stelle. Der Vorstand ist erstmal skeptisch. Man möchte nicht Gefahr laufen die wichtige Stelle mit einem Scharlatan zu besetzen. Die Wahrscheinlichkeit, dass Herr D. die Stelle bekommt, obwohl seine Ergebnisse durch reines Raten zustande kommen, soll sehr klein sein. Man einigt sich auf 5%.

Getestet wird die Nullhypothese, dass das Verfahren von Herrn D. nur Rauschen produziert, also

$$H : \Theta_0 = \{1/2\}$$

gegen die Hypothese

$$K : \Theta \setminus \Theta_0 = (1/2, 1]$$

dass es gute Prognosen liefert. Der Test soll auf der Statistik

$$T(X_1, \dots, X_n) = \sum_{k=1}^{n} X_k$$

basieren und die Nullhypothese soll abgelehnt werden, wenn T groß ist. Die Ablehnung der Nullhypothese bedeutet, dass Herr D. die Stelle bekommt. T nennt man die **Teststatistik** *des Tests. Wie findet man nun die Anzahl der Erfolge, ab der Herr D. die Stelle bekommt ? In einer Binomialtabelle steht*

$$P_{0.5}(\sum_{k=1}^{20} X_k \geq 14) \approx 0.0577 > 0.05$$

$$P_{0.5}(\sum_{k=1}^{20} X_k \geq 15) \approx 0.0207 < 0.05$$

wobei $P_{0.5}$ das Maß der Nullhypothese ist. Bei einer Ablehnung der Nullhypothese, genau dann wenn die Anzahl der Erfolge ≥ 14 ist, wäre der Fehler 1. Art also etwas zu groß. Wenn man nur ablehnt, wenn $\sum_{k=1}^{n} X_k \geq 15$ wäre das Testniveau deutlich zu klein. Also wird bei $\sum_{k=1}^{n} X_k = 14$ randomisiert. Im Fall $\sum_{k=1}^{n} X_k = 14$ wird mit Wahrscheinlichkeit δ verworfen. Damit der Test das Niveau α hat, muss gelten

$$\delta \cdot P_{0.5}(\sum_{k=1}^{20} X_k = 14) + P_{0.5}(\sum_{k=1}^{20} X_k \geq 15) = 0.05.$$

Dies bedeutet

$$\delta = \frac{0.05 - P_{0.5}(\sum_{k=1}^{20} X_k \geq 15)}{P_{0.5}(\sum_{k=1}^{20} X_k = 14)} \approx 0.792.$$

Der Test lautet also

$$\varphi(x_1, \ldots, x_{20}) = 1_{\{\sum_{k=1}^{20} x_k \in \{15,\ldots,20\}\}}(x_1, \ldots, x_{20}) + 0.792 \; 1_{\{\sum_{k=1}^{20} x_k = 14\}}(x_1, \ldots, x_{20}) \quad (2.1)$$

Tatsächlich hat das Prognoseverfahren von Herrn D. von den 20 Aktienkursen nur 5 richtig vorhergesagt. Damit ist die Nullhypothese in dem Modell nicht abgelehnt und Herr D. bekommt die Stelle nicht.

Einem Vorstandsmitglied kommen aber leichte Zweifel an der Entscheidung. Die Prognosen sind, auch wenn Herr D. nur geraten hätte, ungewöhnlich schlecht. Sie scheinen systematische Fehldiagnosen zu sein. Hätte man den gespiegelten Test $H : \Theta_0 = \{1/2\}$ gegen $K : \Theta \setminus \Theta_0 = [0, 1/2)$ durchgeführt, wäre die Nullhypothese abgelehnt worden. Daher könnte Herr D. für die Fondgesellschaft durchaus nützlich werden. Man müsste ihn nach seiner Einschätzung fragen und dann immer das genaue Gegenteil machen. Daran hatte bei der Versuchsplanung niemand gedacht. Im Sinne der statistischen Testtheorie wäre es aber unzulässig Herrn D. einzustellen (auch wenn wir voraussetzen, dass die Fondgesellschaft ihn im Falle $p < 1/2$ gerne einstellen würde). Der Grund ist, dass bei einem Test immer erst der Ablehnungsbereich definiert werden muss und dann das Experiment durchgeführt wird. Nach Bekanntwerden des Ergebnisses kann der Ablehnungsbereich nicht mehr verändert werden.

Nehme an Herr D. hat doch nur geraten und ist zufällig auf das ungewöhnlich schlechte Ergebnis gekommen. Durch die Konstruktion des Tests hat der Vorstand Herrn D. eine Chance von 5% gegeben, die Stelle durch reines Raten zu bekommen. Würde er die Stelle jetzt trotzdem bekommen, wäre diese Wahrscheinlichkeit überschritten.

Wäre $p < 1/2$ genauso gut wie $p > 1/2$ hätte der Vorstand von vorne herein einen **zweiseitigen Test** *konstruieren müssen. Dieser hätte dann die Gestalt*

$$\varphi(x_1, \ldots, x_{20}) = 1_{\{\sum_{k=1}^{20} x_k \in \{0,\ldots c\} \cup \{20-c,\ldots,20\}\}}(x_1, \ldots, x_{20}) + \delta 1_{\{\sum_{k=1}^{20} x_k \in \{c+1,19-c\}\}}(x_1, \ldots, x_{20})$$

gehabt.

Unter der Voraussetzung, dass man Herrn D. im Fall $p < 1/2$, an den der Vorstand erst gar nicht gedacht hat, auch nicht haben will, hätte man den gleichen Test (2.1) durchgeführt. Formal wäre nun $H : \Theta_0 = [0, 1/2]$ gegen $K : \Theta \setminus \Theta_0 = (1/2, 1]$ getestet worden. Da jedoch für $p < 1/2$ die Wahrscheinlichkeit, in den Ablehnungsbereich zu kommen, sogar noch kleiner ist als für $p = 1/2$, wäre der Fehler 1. Art auch für diese Elemente der Nullhypothese eingehalten worden.

Bemerkung 2.11. *Nehme an, ein Test soll auf der Teststatistik $T : \mathbb{R}^n \to \mathbb{R}$ basieren und die Gestalt haben, dass bei $T(X_1, \dots, X_n) > c_\alpha$ die Nullhypothese abgelehnt wird und bei $T(X_1, \dots, X_n) \leq c_\alpha$ nicht, also*

$$\varphi(X_1, \dots, X_n) = 1_{\{T(X_1,\dots,X_n)>c_\alpha\}}.$$

Der Einfachheit halber soll $T(X_1, \dots, X_n)$ zumindest unter der Nullhypothese stetig verteilt sein (also anders als in Beispiel 2.10). c_α wird so gewählt, dass der Test das Niveau α bekommt. α wird aber zunächst nicht fixiert, sondern man schaut erstmal auf die Daten. Wenn eine Hypothese mit $\alpha = 0.05$ nicht abgelehnt werden kann, sagt man auch die Daten sind nicht signifikant. *Bei einer Ablehnung mit $\alpha = 0.05$ spricht man üblicherweise von* statistischer Signifikanz. *Bei einer Ablehnung mit $\alpha = 0.01$ wären die Daten sogar* hochsignifikant. *Der p-Wert einer Beobachtung $T(X_1, \dots, X_n)$ wird als das größte α definiert, bei dem die Hypothese noch nicht abgelehnt wird.*

Bemerkung 2.12. *Eine häufige* **Fehlinterpretation** *obigen Tests besteht darin zu meinen, dass das Testergebnis eine Wahrscheinlichkeit angibt, mit der θ in Θ_0 bzw. in $\Theta \setminus \Theta_0$ fällt. Dies kann der Test nicht leisten, da die Auswahl von ϑ aus der Menge Θ nicht über ein stochastisches Experiment modelliert wird, also keine a priori Wahrscheinlichkeiten über den wahren Parameter (vor Beobachtung der Daten) festgelegt sind.*

Bemerkung 2.13. *Um einen Test methodisch sauber zu konstruieren, muss man zunächst die Funktion φ festlegen und darf erst dann einen Blick in die Daten werfen. An einem hochdimensionalen Datensatz $(x_1, x_2, \dots, x_n)$ findet man nämlich sehr oft irgendetwas scheinbar untypisches, also etwas, was etwa der Hypothese ,,X_k sind alle nach der Verteilungsfunktion F verteilt" scheinbar widerspricht. Wenn man dann getrieben von dieser Beobachtung einen Test entwirft, der darauf abzielt, dass die Beobachtung unter der Verteilungsfunktion F unwahrscheinlich erscheint, wird man die Hypothese ,,X_k sind alle nach der Verteilungsfunktion F verteilt"' relativ leicht verwerfen.*

Ob dies in der Praxis tatsächlich so erfolgt, erscheint fraglich. Allerding haben die Funktionen oft eine kanonische Gestalt, was das methodische Problem verkleinert.

2.1.1 Der Rangsummentest von Wilcoxon

Beispiel 2.14 (Einkommensunterschiede Frauen/Männer)**.** *Aus den Daten der Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften 2006 wurden zufällig 20 Personen gezogen und ihr Nettoeinkommen ermittelt:*

Rang	*Nettoeinkommen*	*Geschlecht*
1	0	*M*
2	400	*W*
3	500	*M*
4	550	*W*
5	600	*M*
6	650	*W*
7	750	*M*
8	800	*M*
9	900	*W*
10	950	*W*
11	1000	*M*
12	1100	*M*
13	1200	*W*
14	1500	*M*
15	1600	*W*
16	1800	*M*
17	1900	*M*
18	2000	*M*
19	2200	*M*
20	3500	*M*

Auf Grundlage dieser Stichprobe, soll geprüft werden, ob das Einkommen der Männer und Frauen in der Bevölkerung gleich ist (zweiseitiger Test) oder ob das Einkommen der Frauen geringer ist (einseitiger Test).

Für die stochastische Modellierung erweist es sich hier als zweckmässig, nur den Rang des Einkommens in der Gruppe und nicht die absolute Höhe zu betrachten. Der Grund hierfür ist, dass für den Rang die spezielle (theoretische) Verteilung der Einkommen nicht relevant ist, und die zu entwickelnde Teststatistik damit ,,ziemlich universell" einsetzbar ist.

Um dies zu verdeutlichen, modellieren wird zunächst die Einkommenshöhen und leiten daraus die Ränge ab. Sei also

$$X = (X_1, \dots, X_m, X_{m+1}, \dots, X_{m+n}),$$

wobei X_i das zufällige Einkommen der i-ten Person bezeichnet. Die ersten m Personen sind weiblich (bzw. gehören Gruppe 1 an) und die letzten n Personen sind männlich (diese Anzahlen werden nicht als Ausgang eines Zufallsexperiments interpretiert !). Formal sind die Bildmaße von X auf $\mathcal{B}(\mathbb{R}^{m+n})$ gegeben durch

$$Q^{\otimes m} \otimes \widetilde{Q}^{\otimes n}$$

wobei Q und $\widetilde{Q}$ alle Maße auf $\mathbb{R}_+$ durchlaufen, die **keine Punktmasse** haben. Dies bedeutet, alle $m+n$ Beobachtungen sind stochastisch unabhängig und die ersten m soweis die letzten n Beobachtungen sind untereinander identisch verteilt. Da wir ausschließen, dass $P(X_i = x) > 0$ unter einem möglichem Maß und zudem stochastische Unabhängigkeit gilt, folgt mit der Übungsaufgabe 2, Teil b, auf Blatt 1, dass $P(X_i \neq X_j \ \forall i, j) = 1$. Damit gilt für die Ordnungstatistiken (siehe die Definition in (1.1))

$$X_{(1)} < X_{(2)} < \ldots < X_{(m+n)}$$

Wenn viele Daten übereinstimmen, ist die Idealisierung, dass $P(X_i \neq X_j \ \forall i, j) = 1$ offenbar unbrauchbar. Wenn aber im Datensatz z.B. nur zwei Daten übereinstimmen, wäre ein pragmatischer Ansatz, für jede der beiden möglichen Reihungen den Test durchführen und zu sehen, ob sich die Ergebnisse stark unterscheiden. Streng genommen wäre das statistische Modell aber schon falsifiziert, da wir die Annahme, dass sich alle Werte unterscheiden, auch für die Maße aus der Gegenhypothese gemacht haben. Bei obigem Datensatz tritt das Problem aber nicht auf.

Die Positionen der weiblichen Personen $1, 2, \ldots, m$ sind nun m Zufallsvariablen $U_1, \ldots, U_m$, die rein zufällig aus der Menge $\{1, 2, \ldots, m+n\}$ **ohne Zurücklegen** gezogen werden. $U_k = j$ mit $k \in \{1, \ldots, m\}$ und $j \in \{1, 2, \ldots, m+n\}$ würde bedeutet, dass beim k-ten Zug der Rang j gezogen wird.

Unter der Nullhypothese $Q = \widetilde{Q}$ besitzt $(U_1, \ldots, U_m)$ die entsprechende Verteilung (unabhängig von Q solange Q wie gefordert keine Punktmasse hat). Als Teststatistik betrachten wir die Summe

$$S_{m,n} := \sum_{k=1}^{m} U_k$$

Nun ist die Verteilung von $S_{m,n}$ unter der Nullhypothese zu untersuchen. Der Erwartungswert von $S_{m,n}$ lässt sich einfach bestimmen. Es gilt

$$E(U_1) = \sum_{k=1}^{m+n} k \frac{1}{m+n} = \frac{(m+n)(m+n+1)}{2} \frac{1}{m+n} = \frac{m+n+1}{2}$$

Des weiteren gilt

$$E(S_{m,n}) = mE(U_1)$$

(wieso ?) und damit

$$E(S_{m,n}) = \frac{m(m+n+1)}{2}$$

Wie kann man die Varianz von $S_{m,n}$ bestimmen ? Man ziehe dazu, wiederum rein zufällig und ohne Zurücklegen, n-mal aus der verbleibenden Menge $\{1, 2, \ldots, m+n\} \backslash \{U_1, \ldots, U_m\}$. Die so konstruierten Zufallsvariablen nennen wir $V_1, \ldots, V_n$. Es gilt nun

$$U_k = \sum_{i=1}^{m} 1_{\{U_i \leq U_k\}} + \sum_{i=1}^{n} 1_{\{V_i \leq U_k\}}$$

Es gilt

$$S_{m,n} = \sum_{k=1}^{m} U_k = \frac{m(m+1)}{2} + \sum_{k=1}^{m}\sum_{i=1}^{n} 1_{\{V_i \le U_k\}}$$

und damit

$$\begin{aligned} \mathrm{Var}(S_{m,n}) &= \mathrm{Var}(\sum_{k=1}^{m}\sum_{i=1}^{n} 1_{\{V_i \le U_k\}}) \\ &= \sum_{k=1}^{m}\sum_{i=1}^{n} \mathrm{Var}(1_{\{V_i \le U_k\}}) + \sum_{(k_1,i_1)\neq(k_2,i_2)} \mathrm{Cov}(1_{\{V_{i_1} \le U_{k_1}\}} 1_{\{V_{i_2} \le U_{k_2}\}}) \end{aligned}$$

Es gilt

$$\mathrm{Var}(1_{\{V_i \le U_k\}}) = E(1_{\{V_i \le U_k\}}) - (E(1_{\{V_i \le U_k\}}))^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}.$$

Für $i_1 \neq i_2$ *und* $k_1 \neq k_2$ sind die Zufallsvariablen $1_{\{V_{i_1} \le U_{k_1}\}}$ und $1_{\{V_{i_2} \le U_{k_2}\}}$ stochastisch unabhängig und die Kovarianz verschwindet. Für $i_1 = i_2$ und $k_1 \neq k_2$ gilt

$$\mathrm{Cov}(1_{\{V_{i_1} \le U_{k_1}\}} 1_{\{V_{i_1} \le U_{k_2}\}}) = E(1_{\{V_{i_1} \le U_{k_1}\}} 1_{\{V_{i_1} \le U_{k_2}\}}) - E(1_{\{V_{i_1} \le U_{k_1}\}}) E(1_{\{V_{i_1} \le U_{k_2}\}}) = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

Analoges gilt für $i_1 \neq i_2$ und $k_1 = k_2$. Insgesamt gilt

$$\mathrm{Var}(S_{m,n}) = mn\frac{1}{4} + mn(n-1)\frac{1}{12} + m(m-1)n\frac{1}{12} = \frac{mn(m+n+1)}{12}.$$

Die sogenannte **Mann-Whitney-Statistik** ist definiert als

$$W := \sum_{i=1}^{n}\sum_{k=1}^{m} 1_{\{V_i \le U_k\}} = \sum_{k=1}^{m} U_k - \frac{m(m+1)}{2}$$

Kommen wir nun auf Beispiel 2.14 zurück. Die Einkommen der $m = 7$ Frauen werden zufällig aus $m + n = 20$ Werten gezogen. Wir könnten nun prüfen, ob das Einkommen der Männer und Frauen gleich ist (zweiseitiger Test) oder das Einkommen der Frauen geringer (einseitiger Test).

```
W <- scan ()
400 550 650 900 950 1200 1600

M <- scan ()
0 500 600 750 800 1000 1100 1500 1800 1900 2000 2200 3500

boxplot(W, M)
```

```
wilcox.test(W,M)
Wilcoxon rank sum test
data: W and M W = 31, p-value = 0.2749 alternative hypothesis: true location
shift is not equal to 0
```

Bei den üblichen Irrtumswahrscheinlichkeiten (z.B. $p = 0.05$ oder $p = 0.01$) wird man die Nullhypothese noch nicht verwerfen. Dies liegt jedoch an dem sehr kleinen Stichprobenumfang $n + m = 20$.

Der in R implementierte Test beruht bei einer Parametereingabe von $m + n \leq 50$ auf der exakten Verteilung der Statistik W unter der Nullhypothese, die durch kombinatorische Überlegungen leicht ermittelt werden kann. Der Rechenaufwand wächst für große m und n jedoch sehr schnell. Ein gute Approximation liefert die Normalverteilung. Da wir $E(S_{m,n})$ und $\mathrm{Var}(S_{m,n})$ bereits ausgerechnet haben, ist das Konfidenzintervall für die Approximation bekannt. Für $m + n > 50$ wird in R die Normalapproximation verwendet, es sei denn man spezifiziert

```
wilcox.test(W,M,exact=TRUE)
```

Satz 2.15 (Hoeffding 1949)**.** *Für $m, n \to \infty$ gilt*

$$\frac{S_{m,n} - E(S_{m,n})}{\sqrt{\mathrm{Var}(S_{m,n})}} \to \mathcal{N}(0,1) \quad \textit{in Verteilung.}$$

Wir werden den Satz hier nicht beweisen. Die Aussage erscheint jedoch plausibel. Betrachtet man die n^2m^2 Paare, die aus den nm Summanden gebildet werden können, so sind $m(m-1)n(n-1)$ stochastisch unabhängig. Dieser Anteil dominiert für $m, n \to \infty$. Ein Beweis findet sich z.B. im Lehrbuch von Herrn Georgii (Satz 11.28).

Achtung: Der Wilcoxon-Test testet die Nullhyothese, dass beide Verteilungen gleich sind, gegen die Alternative, dass die eine gegenüber der anderen verschoben ist. Er ist nicht sensitiv gegenüber anderen Alternativen, wie etwa „*die Varianz der einen Verteilung ist größer, aber die Erwartungswerte sind gleich*" (also etwa die Alternative, dass Männer ganz viel oder ganz wenig verdienen und Frauen eher durchschnittlich). Bei nur einer Alternative existiert aber ein bester Test

2.2 Alternativtests und das Lemma von Neyman-Pearson

Zunächst definieren wir, was man üblicherweise unter einem Standardmodell versteht.

Definition 2.16. *Ein statistisches Modell*

$$\mathcal{M} = (\mathcal{X}, \mathcal{B}, (P_\vartheta^X)_{\vartheta \in \Theta})$$

heißt **Standardmodell***, wenn es entweder diskret ist, d.h. der Zustandsraum $\mathcal{X}$ ist abzählbar oder wenn es stetig bzgl. des Lebesgue-Maßes auf dem $\mathbb{R}^n$ ist, d.h. $\mathcal{X} = \mathbb{R}^n$, $\mathcal{B} = \mathcal{B}(\mathbb{R}^n)$ und jedes Maß P_ϑ^X besitzt eine Dichtefunktion ρ_ϑ bzgl. des Lebesgue-Maßes λ^n auf $\mathbb{R}^n$.*

Letzteres bedeutet, dass für alle $B \in \mathcal{B}(\mathbb{R}^n)$ gilt

$$\begin{aligned} P_\vartheta(X \in B) &= P_\vartheta^X(B) = \int_B \rho_\vartheta(x)\, \lambda^n(dx) \\ &= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} 1_B(x_1, \dots, x_n) \rho_\vartheta(x_1, \dots, x_n)\, dx_1 \dots dx_n. \end{aligned}$$

Bei mehreren Alternativen muss ein gleichmäßig bester Test im Sinne von Definition 2.7 nicht existieren. Testet man etwa für $\Theta = \mathbb{R}$ die Nullhypothese $H : \vartheta = \vartheta_0$ gegen $K : \vartheta \neq \vartheta_0$ so werden höhere Verwerfungswahrscheinlichkeiten für $\vartheta < \vartheta_0$ typischerweise höheren Verwerfungswahrscheinlichkeiten für $\vartheta > \vartheta_0$ entgegenstehen.

Definition 2.17. *Sei $\mathcal{M}$ ein zweielementiges statistisches Standardmodell. Ein Test φ von P_0 gegen P_1 heißt ein Neyman-Pearson-Test, wenn es ein $c \in \mathbb{R}_+$ gibt mit*

$$\varphi(x) = \begin{cases} 1 & \text{falls } \rho_1(x) > c\rho_0(x) \\ 0 & \text{falls } \rho_1(x) < c\rho_0(x), \end{cases}$$

wobei im diskreten Fall $\rho_i(x) = P_i(\{x\})$ und andernfalls ρ_i die Dichte von P_i bzgl. des Lebesgue-Maßes ist (Der Fall $\rho_1 = c\rho_0$ ist in der Bedingung bewusst ausgespart).

Satz 2.18 (Lemma von Neyman-Pearson, 1932)**.** *Sei $\mathcal{M}$ ein zweielementiges Standardmodell mit Wahrscheinlichkeitsmaßen P_0 und P_1. Wir testen die Nullhypothese P_0 gegen die Alternative P_1 zu vorgegebenem Niveau $\alpha \in (0, 1)$. Dann gilt*

(a) Es existiert ein Neyman-Pearson-Test φ mit $E_{P_0}(\varphi) = \alpha$ (also ein Test, der das Niveau α voll ausschöpft).

(b) Jeder Neyman-Pearson-Test φ mit $E_{P_0}(\varphi) = \alpha$ ist ein bester Test zum Niveau α.

Proof. Zunächst müssen wir zu gegebenem $\alpha \in (0, 1)$ einen Neyman-Pearson-Test finden, der das Niveau α voll ausschöpft. Wähle

$$c_\alpha := \inf\{c \in \mathbb{R}_+ \mid P_0(\{x \mid \rho_1(x) > c\rho_0(x)\}) \leq \alpha$$

Wir schreiben den Beweis für den stetigen Fall auf, also z.B.

$$P_0(\{x \mid \rho_1(x) > c\rho_0(x)\}) = \int 1_{\{x \mid \rho_1(x) > c\rho_0(x)\}}(x) \rho_0(x)\, dx.$$

Beim diskreten Fall müssen einfach Integrale durch Summen ersetzt werden, also

$$P_0(\{x \mid \rho_1(x) > c\rho_0(x)\}) = \sum_{x \in \mathcal{X} \text{ mit } \rho_1(x) > c\rho_0(x)} P_0(\{x\}).$$

Der Ausdruck $\int 1_{\{x \mid \rho_1(x) = c_\alpha \rho_0(x)\}}(x) \rho_0(x)\, dx$ ist selbst im stetigen Fall (d.h. mit Dichten bzgl. des Lebesgue-Maßes) i.A. ungleich Null (nehme das trivale Beispiel $P_0 = P_1$, oder die Dichten sind in einem Bereich gleich). D.h. das Infimum ist i.A. kein Minimum und

ein nichtrandomisierender Test mit der Schranke c_α würde das Niveau α nicht einhalten.

1. Fall: $P_0(\{x \mid \rho_1(x) = c_\alpha \rho_0(x)\}) = 0$. Hier ist keine Randomisierung nötig und

$$\varphi(x) = \begin{cases} 1 & \text{falls } \rho_1(x) > c_\alpha \rho_0(x) \\ 0 & \text{falls } \rho_1(x) \leq c_\alpha \rho_0(x), \end{cases}$$

ist ein Neyman-Pearson-Test mit $E_{P_0}(\varphi) = \alpha$.

Fall 2: $P_0(\{x \mid \rho_1(x) = c_\alpha \rho_0(x)\}) > 0$. Setze

$$\delta := \frac{\alpha - P_0(\{x \mid \rho_1(x) > c_\alpha \rho_0(x)\})}{P_0(\{x \mid \rho_1(x) = c_\alpha \rho_0(x)\})}$$

und

$$\varphi(x) = \begin{cases} 1 & \text{falls } \rho_1(x) > c_\alpha \rho_0(x) \\ \delta & \text{falls } \rho_1(x) = c_\alpha \rho_0(x) \\ 0 & \text{falls } \rho_1(x) < c_\alpha \rho_0(x), \end{cases}$$

Der so definierte Test φ ist ein Neyman-Pearson-Test mit $E_{P_0}(\varphi) = P_0(\{x \mid \rho_1(x) > c_\alpha \rho_0(x)\}) + \delta P_0(\{x \mid \rho_1(x) = c_\alpha \rho_0(x)\}) = \alpha$. Damit ist Teil (a) bewiesen.

Sei nun φ ein Neyman-Pearson-Test mit Schwellenwert c und $E_{P_0}(\varphi(X)) = \alpha$ sowie ψ ein beliebiger anderer Test zum Niveau α, d.h. $E_{P_0}(\psi(X)) \leq \alpha$.

Ist $\varphi(x) > \psi(x)$, so ist $\varphi(x) > 0$ und dann $\rho_1(x) \geq c\rho_0(x)$. Ist andererseits $\varphi(x) < \psi(x)$, so ist $\varphi(x) < 1$ und dann $\rho_1(x) \leq c\rho_0(x)$. Also gilt

$$(\varphi(x) - \psi(x))\rho_1(x) \geq c(\varphi(x) - \psi(x))\rho_0(x), \quad \forall x \in \mathbb{R}^n.$$

Integriert man nun beide Seiten, dann folgt

$$\begin{aligned} E_{P_1}(\varphi(X)) - E_{P_1}(\psi(X)) &= \int_{\mathbb{R}^n} (\varphi(x) - \psi(x))\rho_1(x)\, dx \\ &\geq c \int_{\mathbb{R}^n} (\varphi(x) - \psi(x))\rho_0(x)\, dx \\ &= c \underbrace{\int_{\mathbb{R}^n} \varphi(x)\rho_0(x)\, dx}_{=\alpha} - c \underbrace{\int_{\mathbb{R}^n} \psi(x)\rho_0(x)\, dx}_{\leq \alpha} \\ &\geq 0. \end{aligned}$$

□

Wie gut ist der optimale Test ? Wir nehmen an, dass das Experiment n mal unabhängig wiederholt wird und dass die Verteilung in jedem Versuch entweder durchgehend P_0 oder durchgehend P_1 ist. Formal führt dies für festes $n \in \mathbb{N}$ auf das folgende Produktmodell

$$(E^n, \mathcal{E}^{\otimes n}, (P_i^{\otimes n})_{i=0,1}). \tag{2.2}$$

E ist der Wertebereich der Zufallsvariablen, die die Ergebnisse der einzelnen Experimente modellieren (auch diese können bereits mehrdimensional sein). Die Vermutung, dass man bei einer immer größer werdenden Anzahl von unabhängigen Wiederholungen beliebig genau zwischen den Alternativen unterscheiden kann, wird in diesem Abschnitt bestätigt werden.

Definition 2.19 (Relative Entropie)**.** *Für zwei Wahrscheinlichkeitsmaße P_0, P_1 mit Dichten ρ_0, ρ_1 sei die* **relative Entropie** *oder Kullback-Leibler-Information von P_0 bzgl. P_1 definiert als*

$$H(P_0; P_1) := \begin{cases} E_{P_0}\left(\ln\left(\frac{\rho_0}{\rho_1}\right)\right) & \text{falls } P_0(\rho_1 = 0) = 0 \\ \infty & \text{falls } P_0(\rho_1 = 0) > 0, \end{cases}$$

wobei beim Integrieren mit $0 \cdot \log(0) = 0$ gearbeitet werden soll

(Im stetigen Fall ist $E_{P_0}\left(\ln\left(\frac{\rho_0}{\rho_1}\right)\right) = \int_E \rho_0(x) \ln\left(\frac{\rho_0(x)}{\rho_1(x)}\right) dx$ und im diskreten Fall ist $E_{P_0}\left(\ln\left(\frac{\rho_0}{\rho_1}\right)\right) = \sum_{x\in E} P_0(\{x\}) \ln\left(\frac{P_0(\{x\})}{P_1(\{x\})}\right)$).

Es gilt $H(P_0; P_1) \geq 0$ und Gleichheit gilt genau dann, wenn $P_0(\rho_0 = \rho_1) = 1$. Der Beweis benutzt die Abschätzung $\ln(x) \leq x - 1$ und geht so:

$$\begin{aligned} \int_E \rho_0(x) \ln\left(\frac{\rho_0(x)}{\rho_1(x)}\right) dx &= -\int_E \rho_0(x) \ln\left(\frac{\rho_1(x)}{\rho_0(x)}\right) dx \\ &\geq \int_E \rho_0(x) \left(1 - \frac{\rho_1(x)}{\rho_0(x)}\right) dx \\ &= \int_E \rho_0(x)\, dx - \int_E \rho_0(x) \frac{\rho_1(x)}{\rho_0(x)}\, dx \\ &= \int_E \rho_0(x)\, dx - \int_E \rho_1(x)\, dx \\ &= 1 - 1 = 0. \end{aligned}$$

Da $\ln(x) = x - 1$ nur für $x = 1$ gilt, verschwindet die Entropie nur wenn

$$0 = \int_E \rho_0(x) 1_{\{\rho_0(x) \neq \rho_1(x)\}}(x)\, dx = P_0(\rho_0 \neq \rho_1).$$

Satz 2.20 (Lemma von C. Stein, 1952)**.** *Sei $\alpha \in (0, 1)$. Im statistischen Modell (2.2) betrachten wir zu jedem n einen Neyman-Perrson-Test φ_n mit $E_0(\varphi_n(X_1, \ldots, X_n)) = \alpha$ (d.h. der n-te Test hängt von den Beobachtungen $X_1, X_2, \ldots, X_n$ ab). Dann strebt die Macht $E_1(\varphi_n(X_1, \ldots, X_n))$ für $n \to \infty$ mit exponentieller Geschwindigkeit gegen 1. Genauer gilt*

$$\lim_{n\to\infty} \frac{1}{n} \ln[1 - E_1(\varphi_n(X_1, \ldots, X_n))] = -H(P_0; P_1),$$

d.h. $E_1(\varphi_n(X_1, \ldots, X_n)) \approx 1 - \exp(-nH(P_0; P_1))$ für großes n

(E_i bezeichnet den Erwartungswert bzgl. des Produktmaßes $(P_i)^{\otimes n}$)

Proof. Für $n \geq 1$ ist $(\rho_i)^{\otimes n}(X_1, \dots, X_n) = \prod_{k=1}^n \rho_i(X_k)$ die Dichte von $(P_i)^{\otimes n}$ (d.h. bei einem Produktmaß ist die Dichte das Produkt der Dichten der Maße). Ein Neyman-Pearson-Test φ_n erfüllt die Bedingung

$$\varphi_n(X_1, \dots, X_n) = \begin{cases} 1 & \text{falls } \prod_{k=1}^n \rho_1(X_k) > c_n \prod_{k=1}^n \rho_0(X_k) \\ 0 & \text{falls } \prod_{k=1}^n \rho_1(X_k) < c_n \prod_{k=1}^n \rho_0(X_k) \end{cases}$$

für geeignete Konstanten $c_n > 0$. Auf der Menge

$$A_n := \{\rho_0(X_k) > 0 \quad \forall k = 1, \dots, n\} \cup \{\rho_1(X_k) > 0 \quad \forall k = 1, \dots, n\}$$

gilt

$$\ln\left(\frac{\prod_{k=1}^n \rho_1(X_k)}{\prod_{k=1}^n \rho_0(X_k)}\right) = \sum_{k=1}^n \ln\left(\frac{\rho_1(X_k)}{\rho_0(X_k)}\right) \tag{2.3}$$

(mit den üblichen Konventionen $1/0 = \infty$ etc.) und damit

$$\varphi_n(X_1, \dots, X_n) = \begin{cases} 1 & \text{falls } \sum_{k=1}^n \ln\left(\frac{\rho_1(X_k)}{\rho_0(X_k)}\right) > \ln(c_n) \\ 0 & \text{falls } \sum_{k=1}^n \ln\left(\frac{\rho_1(X_k)}{\rho_0(X_k)}\right) < \ln(c_n) \end{cases}$$

Auf der Menge A_n ist also (2.3) wohldefiniert, da $\infty - \infty$ nicht vorkommen kann. Das Komplement A_n^c von A_n muss jedoch bei den folgenden Rechnungen nicht interessieren, da sowohl

$$(P_0)^{\otimes n}(A_n^c) \leq (P_0)^{\otimes n}(\cup_{k=1,\dots,n}\{\rho_0(X_k) = 0\}) \leq nP_0(\{\rho_0(X_1) = 0\}) = 0$$

als auch $(P_1)^{\otimes n}(A_n^c) = 0$. Es gilt

$$E_{P_0} \ln\left(\frac{\rho_1(X_1)}{\rho_0(X_1)}\right) = -E_{P_0} \ln\left(\frac{\rho_0(X_1)}{\rho_1(X_1)}\right) = -H(P_0; P_1). \tag{2.4}$$

Aus der Forderung, dass die Tests ein Niveau $\alpha \in (0, 1)$ einhalten und dem schwachen Gesetz der großen Zahlen folgt somit, dass

$$\ln(c_n) = -H(P_0; P_1) \cdot n + o(n), \quad n \to \infty.$$

Des weiteren gilt

$$\begin{aligned} 1 &\geq E_0(1 - \varphi_n(X_1, \dots, X_n)) \\ &= E_0\left((1 - \varphi_n(X_1, \dots, X_n)) 1_{\{\prod_{k=1}^n \frac{\rho_1(X_k)}{\rho_0(X_k)} \leq c_n\}}\right) \\ &\geq \frac{1}{c_n} E_0\left((1 - \varphi_n(X_1, \dots, X_n)) \frac{\prod_{k=1}^n \rho_1(X_k)}{\prod_{k=1}^n \rho_0(X_k)}\right) \\ &= \frac{1}{c_n} E_1\left(1 - \varphi_n(X_1, \dots, X_n)\right). \end{aligned}$$

Also gilt

$$0 \geq \frac{1}{n}\left[-\ln(c_n) + \ln\left(E_1\left(1-\varphi_n(X_1,\ldots,X_n)\right)\right)\right]$$

und damit

$$\limsup_{n\to\infty} \frac{1}{n} E_1\left(1-\varphi_n(X_1,\ldots,X_n)\right) \leq \lim_{n\to\infty} \frac{\ln(c_n)}{n} = -H(P_0;P_1).$$

Nun wollen wir $E_1\left(1-\varphi_n(X_1,\ldots,X_k)\right)$ nach unten abschätzen. Es gilt für alle $a > H(P_0;P_1)$ und n hinreichend groß

$$\begin{aligned}
E_1\left(1-\varphi_n(X_1,\ldots,X_n)\right) &= E_0\left(\left(1-\varphi_n(X_1,\ldots,X_n)\right)\frac{\prod_{k=1}^n \rho_1(X_k)}{\prod_{k=1}^n \rho_0(X_k)}\right) \\
&\geq E_0\left(\exp(-an)(1-\varphi_n(X_1,\ldots,X_n))1_{\{\prod_{k=1}^n \frac{\rho_1(X_k)}{\rho_0(X_k)} \geq \exp(-an)\}}\right) \\
&\geq \exp(-an) E_0\left(1_{\{\frac{1}{n}\sum_{k=1}^n \ln\left(\frac{\rho_1(X_k)}{\rho_0(X_k)}\right) \geq -a\}} - \varphi_n(X_1,\ldots,X_n)\right) \\
&\geq \exp(-an)\left(\frac{1+\alpha}{2} - \alpha\right) = \exp(-an)\frac{1-\alpha}{2}.
\end{aligned}$$

Hierbei geht ein, dass wegen (2.4) $E_{P_0} \ln\left(\frac{\rho_1(X_1)}{\rho_0(X_1)}\right) = -H(P_0;P_1) > -a$ und damit mit dem schwachen Gesetz der großen Zahlen $P_0(\frac{1}{n}\sum_{k=1}^n \ln\left(\frac{\rho_1(X_k)}{\rho_0(X_k)}\right) \geq -a) \to 1$ für $n \to \infty$. Es folgt

$$\liminf_{n\to\infty} \frac{1}{n} E_1\left(1-\varphi_n(X_1,\ldots,X_n)\right) \geq \lim_{n\to\infty}\left(-a + \frac{1}{n}\ln\left(\frac{1-\alpha}{2}\right)\right) = -a.$$

Da diese Aussage für beliebige a mit $a > H(P_0;P_1)$ gilt, folgt

$$\liminf_{n\to\infty} \frac{1}{n} E_1\left(1-\varphi_n(X_1,\ldots,X_n)\right) \geq -H(P_0;P_1)$$

und insgesamt

$$\lim_{n\to\infty} \frac{1}{n} E_1\left(1-\varphi_n(X_1,\ldots,X_n)\right) = -H(P_0;P_1).$$

□

2.3 Schätzen

Definition 2.21 (Punktschätzung)**.** *Sei $\gamma(\vartheta)$ ein sogenannter abgeleiteter Parameter. Eine (messbare) Abbildung*

$$d : \mathcal{X} \to \gamma(\Theta)$$

heißt Punktschätzung (messbar bzgl. der σ-Algebra $\mathcal{B}$ auf dem Urbildraum $\mathcal{X}$ und einer geeigneten σ-Algebra auf dem Bildraum Θ). Bei beobachtetem x heißt $d(x)$ Schätzwert für $\gamma(\vartheta)$.

Definition 2.22 (Erwartungstreue). *Sei $\gamma(\Theta) \subset \mathbb{R}$. Ein Punktschätzer d mit $E_\vartheta(|d(X)|) < \infty$ für alle $\vartheta \in \Theta$ heißt erwartungstreu, wenn*

$$E_\vartheta(d(X)) = \gamma(\vartheta) \quad \forall \vartheta \in \Theta$$

Beispiel 2.23 (Schätzung des Erwartungswertes). *Sei $X = (X_1, \dots, X_n)$, wobei $X_1, \dots, X_n$ (unter allen P_ϑ) identisch verteilt sind und endlichen Erwartungswert besitzen. Dann ist der Punktschätzer $d(X) := \bar{X} := \frac{1}{n}\sum_{k=1}^n X_k$ offenbar erwartungstreu für $\gamma(\vartheta) := E_\vartheta(X_1)$*
.

Beispiel 2.24 (Schätzung der Varianz). *Sei $X = (X_1, \dots, X_n)$, wobei $X_1, \dots, X_n$ unter allen P_ϑ i.i.d. sind und endliche zweite Momente besitzen. Dann ist der Punktschätzer*

$$d(X_1, \dots, X_n) := \frac{1}{n-1}\sum_{k=1}^n (X_k - \bar{X})^2$$

erwartungstreu für $\gamma(\vartheta) := \mathrm{Var}_\vartheta(X_1)$. Dies ergibt sich aus der folgenden Rechnung. Sei $\mu_\vartheta := E_\vartheta(X_1)$. Es gilt

$$\begin{aligned}
E_\vartheta\left(\frac{1}{n-1}\sum_{k=1}^n (X_k - \bar{X})^2\right) &= \frac{n}{n-1} E_\vartheta\left(X_1 - \bar{X}\right)^2 \\
&= \frac{n}{n-1} E_\vartheta\left(\frac{n-1}{n}(X_1 - \mu_\vartheta) - \frac{1}{n}\sum_{k=2}^n (X_k - \mu_\vartheta)\right)^2 \\
&= \frac{n}{n-1}\left(\frac{(n-1)^2}{n^2}\mathrm{Var}_\vartheta(X_1) + \frac{1}{n^2}\sum_{k=2}^n \mathrm{Var}_\vartheta(X_k)\right) \\
&= \mathrm{Var}_\vartheta(X_1), \qquad (2.5)
\end{aligned}$$

wobei für die dritte Gleichung benutzt wird, dass die Varianz einer Summe unabhängiger Zufallsvariablen die Summe der Varianzen ist.

$\frac{1}{n-1}\sum_{k=1}^n (X_k - \bar{X})^2$ wird **korrigierte Stichprobenvarianz** *(oder oft auch einfach nur Stichprobenvarianz) genannt.*

Beispiel 2.25 (Erraten des Bereichs von Zufallszahlen). *In einer Fernsehshow führt ein Moderator einen Zufallsgenerator vor, der gleichverteilte Zufallsvariablen erzeugt. Die Zufallsvariablen $X_1, X_2, \dots, X_n$ sind stochastisch unabhängig und liegen im Bereich $[0, \vartheta]$, wobei die vom Moderator gewählte obere Grenze $\vartheta > 0$ nicht bekanntgegeben wird. Die Spieler sollen nun aufgrund der Beobachtungen ϑ möglichst gut erraten.*

Ein ähnliches Modell könnte man anwenden, wenn aus dem Alter gewisser Objekte, die man zu einem festen Zeitpunkt beobachtet, auf deren maximale Lebensdauer zu schließen ist. *Die Gleichverteilungsannahme ist hier natürlich leicht angreifbar, weshalb wird uns auf das künstliche Beispiel des Zufallsgenerators zurückziehen.*

Das formale statistische Modell bei n unabhängigen Beoachtungen ist gegeben durch

$$(\mathcal{X}, \mathcal{B}, (P_\vartheta^X)_{\vartheta\in\Theta}) = ([0,\infty)^n, \mathcal{B}([0,\infty)^n), ((\mathcal{U}_{[0,\vartheta]})^{\otimes n})_{\vartheta>0})$$

wobei $\mathcal{U}_{[0,\vartheta]}$ *die Gleichverteilung auf* $[0,\vartheta]$ *bezeichnet.*

Spieler A *bedenkt, dass* $E_\vartheta(X_k) = \vartheta/2$ *und erinnert sich an das Gesetz der großen Zahlen. Sein Schätzer für* ϑ *ist daher*

$$T_n^A := \frac{2}{n}\sum_{k=1}^n X_k$$

In der Tat gilt $P_\vartheta(|T_n^A - \vartheta| > \varepsilon) \to 0$ *für* $n \to \infty$*, für alle* $\varepsilon > 0$ *und alle* $\vartheta > 0$*.*

Spieler B *bedenkt, dass das Beobachtungsmaximum* $\max\{X_1, \dots, X_n\}$ *zwar stets kleiner als* ϑ *ist, aber für große n zumeist nahe bei* ϑ *liegt. Da* $\max\{X_1, \dots, X_n\}$ *den Parameter* ϑ *systematisch unterschätzen würde, soll er um einen geeigneten Faktor korrigiert werden. Es gilt* $P_\vartheta(\max\{X_1, \dots, X_n\} \le y) = (y/\vartheta)^n$*. Damit hat* $\max\{X_1, \dots, X_n\}$ *die Dichte* $ny^{n-1}\vartheta^{-n}$ *und es gilt*

$$E_\vartheta(\max\{X_1, \dots, X_n\}) = \int_0^\vartheta y n y^{n-1}\vartheta^{-n}\, dy = n\vartheta^{-n}\int_0^\vartheta y^n\, dy = \frac{n}{n+1}\vartheta, \quad \forall\vartheta.$$

Ein naheliegender erwartungstreuer Schätzer für Spieler B ist daher

$$T_n^B := \frac{n+1}{n}\max\{X_1, \dots, X_n\}.$$

Auch T_n^B *konvergiert* P_ϑ*-stochastisch gegen* ϑ *für* $n \to \infty$*.*

Welcher Schätzer ist nun aber besser ? Ein Kriterium ist die Varianz. Welcher Schätzer streut weniger ? Es gilt

$$\mathrm{Var}(T_n^A) = \frac{4}{n^2} n \mathrm{Var}(X_1) = \frac{4}{n}\int_0^\vartheta \frac{1}{\vartheta}\left(x - \frac{\vartheta}{2}\right)^2 dx = \frac{4}{n}\cdot\frac{1}{3}\frac{1}{\vartheta}\left(x-\frac{\vartheta}{2}\right)^3\Big|_0^\vartheta = \frac{\vartheta^2}{3n}.$$

Für die Varianz des Schätzers von Spieler B gilt

$$\begin{aligned}
\mathrm{Var}(T_n^B) &= E_\vartheta((T_n^B)^2) - (E_\vartheta(T_n^B))^2 \\
&= \frac{(n+1)^2}{n^2}\int_0^\vartheta y^2 n y^{n-1}\vartheta^{-n}\, dy - \vartheta^2 \\
&= \left(\frac{(n+1)^2}{n^2}\frac{n}{n+2} - 1\right)\vartheta^2 = \frac{\vartheta^2}{n(n+2)}.
\end{aligned}$$

Offenbar ist der Schätzer von Spieler B für alle $\vartheta \in \mathbb{R}$ *und alle* $n \ge 2$ *besser (für* $n = 1$ *stimmen die beiden Schätzer ja überein) und ist für* $n \to \infty$ *von besserer Ordnung.*

2.3.1 Schätzen von Quantilen

Ähnlich wie Quantile von Datensätzen, vgl. (1.2), definieren wir Quantile einer Verteilung einer reellwertigen Zufallsvariablen.

Definition 2.26 (Quantile)**.** *Sei X eine reellwertige Zufallsvariable und $\alpha \in (0,1)$. Eine reelle Zahl q ist ein α-***Quantil** *der Verteilung von X unter einem Maß P, wenn gilt*

$$P(X \leq q) \geq \alpha \quad \textit{und} \quad P(X < q) \leq \alpha.$$

Ein **Median** *ist ein $\frac{1}{2}$-Quantil.*

Proposition 2.27. *Die Menge der α-Quantile ist das Intervall $[q_\alpha^-, q_\alpha^+]$ mit*

$$q_\alpha^- := \inf\{y \mid P(X \leq y) \geq \alpha\}$$

und

$$q_\alpha^+ := \inf\{y \mid P(X \leq y) > \alpha\} = \sup\{y \mid P(X < y) \leq \alpha\}$$

Proof. Es existiert eine Folge $(y_n)_{n\in\mathbb{N}} \subset \mathbb{R}$ mit $y_n \geq q_\alpha^-$ für alle n und $y_n \to q_\alpha^-$ für $n \to \infty$ mit

$$P(X \leq y_n) \geq \alpha. \tag{2.6}$$

Es gilt $\{X \leq q_\alpha^-\} = \cap_{n\in\mathbb{N}}\{X \leq y_n\}$ und damit $P(X \leq q_\alpha^-) = \lim_{n\to\infty} P(X \leq y_n)$. Also folgt aus (2.6), dass $P(X \leq q_\alpha^-) \geq \alpha$. Die zweite Quantilseigenschaft von q_α^- zeigen wir durch Widerspruch.

(Widerspruchs-)Annahme: $P(X < q_\alpha^-) > \alpha$. Wegen

$$\{X < q_\alpha^-\} = \cup_{n\in\mathbb{N}}\{X \leq q_\alpha^- - 1/n\}$$

existiert dann auch ein $n \in \mathbb{N}$ mit $P(X \leq q_\alpha^- - 1/n) > \alpha$. Dies ist aber ein Widerspruch zur Minimalität von q_α^-. Also ist q_α^- ein α-Quantil. Der Beweis, dass auch q_α^+ ein α-Quantil ist, funktioniert analog. Zudem gilt offensichtlich $q_\alpha^- \leq q_\alpha^+$ und jeder Wert zwischen q_α^- und q_α^+ ist auch ein α-Quantil. Andererseits kann $q < q_\alpha^-$ kein α-Quantil sein, da dies ein Widerspruch zur Minimalität in der Definition von q_α^- wäre. Analoges gilt für $q > q_\alpha^+$ mit Benutzung der Darstellung als Supremum. □

q_α^- ist also der kleinste Wert, so dass mindestens $\alpha \times 100$ Prozent der Wahrscheinlichkeitsmasse kleiner oder gleich q_α^- ist. q_α^+ ist der größte Wert, so dass höchstens $\alpha \times 100$ Prozent der Wahrscheinlichkeitsmasse strikt kleiner als q_α^+ ist.

Die Funktion $\alpha \mapsto q_\alpha^-$ ist linksstetig, d.h. es gilt $\lim_{h\to 0,\ h<0} q_{\alpha+h}^- = q_\alpha^-$ für alle $\alpha \in (0,1]$. Die Funktion $\alpha \mapsto q_\alpha^+$ ist rechtsstetig, d.h. es gilt $\lim_{h\to 0,\ h>0} q_{\alpha+h}^+ = q_\alpha^+$ für alle $\alpha \in (0,1]$. Für $\alpha = 0$ würde sich in obiger Definition $q_0^- = -\infty$ ergeben, was auch Sinn macht, aber gewöhnlich nicht betrachtet wird.

Bemerkung 2.28. *Es sind zwei Feinheiten zu beachten. Zum einen kann es vorkommen, dass das Quantil nicht eindeutig ist, also* $q_\alpha^- < q_\alpha^+$. *Dies tritt genau dann auf* $P(X \leq q_\alpha^-) = \alpha$ *und* $P(q_\alpha^- < X < q_\alpha^+) = 0$. *Zum anderen kann es bei Eindeutigkeit (!) des Quantils passieren, dass* $P(X \leq q_\alpha^-) = P(X \leq q_\alpha^+) > \alpha$. *Ist die Verteilungsfunktion dagegen stetig, dann gilt immer* $P(X \leq q_\alpha^-) = P(X \leq q_\alpha^+) = \alpha$. *Besonders bequem ist also der Fall einer stetigen und strikt monoton steigenden Verteilungsfunktion. Dann gilt* $q_\alpha^- = q_\alpha^+$ *und* $P(X \leq q_\alpha^-) = \alpha$.

Beispiel 2.29. *Sei* X *eine zufällige* **finanzielle Position** *am Ende eines Handelstages und* $\alpha \in (0,1)$. X *kann der Gewinn (bzw. im negativen Fall der Verlust) eines Wertpapierhändlers während eines Tages sein. Der* **Value at Risk** *zum Niveau* α *ist üblicherweise definiert als*

$$V@R(X) := -q_\alpha^+(X) = q_{1-\alpha}^-(-X) = \inf\{y \in \mathbb{R} \mid P(X + y < 0) \leq \alpha\}$$

(für die letzte Gleichheit beachte man die Äquivalenz $P(-X \leq y) \geq 1 - \alpha \Leftrightarrow P(-X > y) \leq \alpha$*). $V@R(X)$ ist der minimale Geldbetrag, den man zuschießen muss, damit die Position am Ende des Tages höchstens mit Wahrscheinlichkeit* α *negativ wird. Ein typischer Wert für* α *ist* 0.05. *Der Value at Risk dient als ein aussagekräftiges Maß für das Risiko eines Investments. Zudem basieren viele Eigenkapitalvorschriften der Bankenregulierung auf dem Value at Risk. $V@R(X)$ ist dann das Eigenkapital, das ein Finanzinstitut "reservieren" muss, um das Investment X zu tätigen. Der Value at Risk hat jedoch viele Nachteile. Ein offensichlicher Nachteil ist, dass Verluste, die mit einer kleineren Wahrscheinlichkeit als α eintreten, nicht in den Value at Risk eingehen.*

Sei $X = (X_1, \ldots, X_n)$. Der **Stichprobenmedian** ist definiert als

$$M_n := \begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{für } n \text{ ungerade} \\ \frac{1}{2}\left(X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}\right) & \text{für } n \text{ gerade} \end{cases} \tag{2.7}$$

Satz 2.30. *Seien* $X_1, \ldots, X_n$ *unter einem Maß* P *i.i.d. und die Verteilung von* X_1 *besitze einen Median* m, *in dem die Verteilungsfunktion* $y \mapsto P(X_1 \leq y) =: F_{X_1}(y)$ *eine strikt positive Ableitung hat, d.h.* $F'_{X_1}(m) > 0$ *(in diesem Fall ist der Median eindeutig und es gilt zudem* $F_{X_1}(m) = 1/2$*). Dann gilt für alle* $a \in \mathbb{R}$

$$P(\sqrt{n}(M_n - m) \leq a) \to P(Z \leq 2aF'_{X_1}(m)), \quad n \to \infty,$$

wobei Z eine standardnormalverteilte Zufallsvariable ist. D.h.

$$\sqrt{n}(M_n - m) \to \mathcal{N}(0, (2F'_{X_1}(m))^{-2}) \quad n \to \infty \quad \textit{in Verteilung.}$$

Der Stichprobenmedian ist also asymptotisch normalverteilt und die Standardabweichung ist von der Ordnung $\frac{1}{2\sqrt{n}F'_{X_1}(m)}$.

Eine Anwendung: Betrachtet man Verteilungsklassen, die um einen Punkt θ symmetrische Dichten f_θ besitzen, etwa

$$f_\theta(x) = \frac{1}{2\theta} 1_{[0,2\theta]}(x) \quad \text{(Gleichverteilung)},$$

oder

$$f_\theta(x) = \frac{\alpha}{2} \exp\left(-\alpha|x-\theta|\right) \quad \text{(zweiseitige Exponentialverteilung)},$$

oder

$$f_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\theta)^2}{2\sigma^2}\right) \quad \text{(Normalverteilung)}$$

dann ist sowohl das arithmetische Mittel als auch der Stichprobenmedian ein erwartungstreuer Schätzer für θ. Es stellt sich die Frage, welcher Schätzer die kleinere Varianz hat. Eine asymptotische Antwort auf diese Frage liefert Satz 2.30. Dabei kann je nach Verteilung sowohl das arithmetische Mittel als auch der Stichprobenmedian weniger schwanken.

Für die Gleichverteilung auf $[0, 2\theta]$ gilt $\mathrm{Var}_\theta(X_1) = \frac{\theta^2}{3}$ (vgl. die Rechnung in Beispiel 2.25) und $\frac{1}{(2F'_{X_1}(m))^2} = \theta^2$. Also ist bei der Gleichverteilung die Varianz des arithmetischen Mittels asymptotisch kleiner als die Varianz des Stichprobenmedians.

Beweis von Satz 2.30. Sei $a \in \mathbb{R}$. Definiere für alle $n \in \mathbb{N}$ und $i \leq n$ die Bernoulliverteilten Zufallsvariablen

$$Y_{i,n} := \begin{cases} 1 & \text{falls } X_i \leq m + \frac{a}{\sqrt{n}} \\ 0 & \text{sonst} \end{cases}$$

Es gilt

$$P(Y_{i,n} = 1) = F_{X_1}\left(m + \frac{a}{\sqrt{n}}\right) =: p_n$$

Für ungerade n gilt

$$\begin{aligned} P(\sqrt{n}(M_n - m) \leq a) &= P\left(\sum_{i=1}^n Y_{i,n} \geq \frac{n+1}{2}\right) \\ &= P\left(\frac{\sum_{i=1}^n Y_{i,n} - np_n}{\sqrt{np_n(1-p_n)}} \geq \frac{(n+1)/2 - np_n}{\sqrt{np_n(1-p_n)}}\right) \end{aligned} \tag{2.8}$$

Die normierte Zufallsvariable $\frac{\sum_{i=1}^n Y_{i,n} - np_n}{\sqrt{np_n(1-p_n)}}$ konvergiert für $n \to \infty$ in Verteilung gegen die Normalverteilung, d.h. für alle $b \in \mathbb{R}$ gilt

$$P\left(\frac{\sum_{i=1}^n Y_{i,n} - np_n}{\sqrt{np_n(1-p_n)}} \geq b\right) \to P(Z \geq b), \quad n \to \infty, \tag{2.9}$$

wobei Z eine standardnormalverteilte Zufallsvariable ist. Für festes n sind nämlich $Y_{1,n}, Y_{2,n}, \ldots, Y_{n,n}$ i.i.d. Wir können den Zentralen Grenzwertsatz von Lindeberg-Feller anwenden (siehe etwa Satz 15.43 in Klenke, Wahrscheinlichkeitstheorie, 1. Auflage), der

den Zentralen Grenzwertsatz dahingehend verallgemeinert, dass bei einer Summe bis n die Verteilung der Summanden auch von n abhängen darf. Die Lindeberg-Bedingung ist in unserem Beispiel erfüllt, d.h. kein einzelner Summand dominiert alle anderen. Man beachte, dass die Verteilung von $Y_{i,n}$ nur recht schwach mit n variiert, so dass wir nicht so weit vom gewöhnliche Zentralen Grenzwertsatz entfernt sind, bei dem die Verteilung der Summanden nicht von n abhängt.

Die Schranke

$$\frac{(n+1)/2 - np_n}{\sqrt{np_n(1-p_n)}}$$

hängt natürlich auch noch von n ab, konvergiert aber für $n \to \infty$ gegen $-2aF'(m)$. Es gilt nämlich mit Benutzung von $p_n \to 1/2$

$$\begin{aligned}
\frac{(n+1)/2 - np_n}{\sqrt{np_n(1-p_n)}} &= \frac{-n(p_n - 1/2)}{\sqrt{np_n(1-p_n)}} + \frac{1/2}{\sqrt{np_n(1-p_n)}} \\
&= -\frac{1}{\sqrt{p_n(1-p_n)}} \frac{p_n - 1/2}{1/\sqrt{n}} + \frac{1/2}{\sqrt{np_n(1-p_n)}} \\
&= -\frac{1}{\sqrt{p_n(1-p_n)}} \frac{F_{X_1}(m + a/\sqrt{n}) - F_{X_1}(m))}{1/\sqrt{n}} + \frac{1/2}{\sqrt{np_n(1-p_n)}} \\
&\to -2aF'_{X_1}(m), \quad n \to \infty.
\end{aligned}$$

Aus (2.9) folgt somit

$$P\left(\frac{\sum_{i=1}^n Y_{i,n} - np_n}{\sqrt{np_n(1-p_n)}} \geq \frac{(n+1)/2 - np_n}{\sqrt{np_n(1-p_n)}} \right) \to P(Z \geq -2aF'(m)) = P(Z \leq 2aF'(m)).$$

(Zu einem vorgegenen Fehler $\varepsilon > 0$ ist der Abstand zwischen der Schranke und $-2aF'(m)$ für n groß genug kleiner als ε. Nun schätzt man die Schranke gegen $-2aF'(m) \pm \varepsilon$ ab benutzt (2.9) für feste Schranken. Dann lässt man ε gegen 0 laufen)

Zusammen mit (2.8) folgt die Behauptung (erstmal gilt Konvergenz nur entlang von ungeraden n, für gerade n muss man die Wahrscheinlichkeiten in (2.8) abschätzen, was zwar etwas unangenehmer ist, aber auf dem gleichen Weg zum Ziel führt). □

2.3.2 Schätzung des Medians mit Konfidenz

Neben der Punktschätzung für einen abgeleiteten Parameter ist es wünschenswert einen Bereich anzugeben, in dem der abgeleitete Parameter mit ,,großer Wahrscheinlichkeit" liegt.

Definition 2.31 (Bereichsschätzung)**.** *Sei $\gamma(\vartheta)$ ein sogenannter abgeleiteter Parameter. Eine (messbare) Abbildung*

$$C : \mathcal{X} \to 2^{\gamma(\Theta)}$$

heißt Bereichsschätzung (Konfidenzschätzung) für $\gamma(\vartheta)$ zum (Konfidenz-)Niveau $1-\alpha$, falls

$$P_\vartheta(C(X) \ni \gamma(\vartheta)) \geq 1-\alpha \quad \forall \vartheta \in \Theta.$$

Bei beobachtetem x heißt $C(x)$ **Konfidenzbereich** *für $\gamma(\vartheta)$ zum Niveau $1-\alpha$.*

Bei eindimensionalem $\gamma(\vartheta)$ ist man i.d.R. an Intervallen $C(x)$ interessiert und im mehrdimensionalen Fall an konvexen Menge. Hier ist $C(x)$ jedoch erstmal eine beliebige Teilmenge des Parameterraums.

Sei $X = (X_1, \ldots, X_n)$ mit X_k i.i.d. unter allen P_ϑ und X_1 habe unter allen $\vartheta \in \Theta$ eine stetige und strikt monoton steigende Verteilungsfunktion. Damit ist der Median $m(\vartheta)$ eindeutig und erfüllt die Bedingung $P_\vartheta(X_k \leq m(\vartheta)) = 1/2$. Der Stichprobenmedian M_n aus (2.7) konvergiert für $n \to \infty$ fast sicher gegen $m(\vartheta)$ (Übungsaufgabe).

Gesucht: Ein Intervall $I(X)$ mit

$$P_\vartheta(m(\vartheta) \in I(X)) \geq 1-\alpha$$

($I(X) = I(X_1, \ldots, X_n)$ ist ein reellwertiges Intervall, dessen Endpunkte von der Realisation $X_1(\omega), \ldots, X_n(\omega)$ abhängen können)

Beispiel: $n = 10$, $\alpha = 0.05$.

1. Vorschlag: $I(X) := [X_{(1)}, X_{(10)}]$. Es gilt

$$\begin{aligned} P_\vartheta(I(X) \not\ni m(\vartheta)) &= P_\vartheta(X_k > m(\vartheta)\ \forall k = 1, \ldots, n) + P_\vartheta(X_k < m(\vartheta)\ \forall k = 1, \ldots, n) \\ &= 2 \cdot (1/2)^{10} = 2^{-9} = \frac{1}{512} \approx 0.00195 \end{aligned}$$

2. Vorschlag: $I(X) := [X_{(2)}, X_{(9)}]$. Es gilt

$$\begin{aligned} P_\vartheta(I(X) \not\ni m(\vartheta)) &= P_\vartheta(\text{höchstens eines der } X_k \text{ ist } \leq m(\vartheta)) \\ &\quad + P_\vartheta(\text{höchstens eines der } X_k \text{ ist } \geq m(\vartheta)) \\ &= 2\left(\binom{10}{0} + \binom{10}{1}\right) 2^{-10} = 11 \cdot 2^{-9} \approx 0.0215 \end{aligned}$$

3. Vorschlag: $I(X) := [X_{(3)}, X_{(8)}]$. Es gilt

$$\begin{aligned} P_\vartheta(I(X) \not\ni m(\vartheta)) &= P_\vartheta(\text{höchstens zwei der } X_k \text{ sind } \leq m(\vartheta)) \\ &\quad + P_\vartheta(\text{höchstens zwei der } X_k \text{ sind } \geq m(\vartheta)) \\ &= 2\left(\binom{10}{0} + \binom{10}{1} + \binom{10}{2}\right) 2^{-10} = 56 \cdot 2^{-9} \approx 0.1094 \end{aligned}$$

Damit ist $[X_{(2)}, X_{(9)}]$ das kleinste symmetrische Intervall $I(X)$, dessen Endpunkte Beobachtungswerte sind und das in den Beobachtungswerten symmetrisch ist, mit

$$P_\vartheta(m(\vartheta) \in I(X)) \geq 0.95.$$

Würde man auch Intervalle zulassen, dessen Endwerte, zwar von den Beobachtungswerten bestimmt werden, aber selber keine Beobachtungswerte sind, würden die Wahrscheinlichkeiten von der Randverteilung abhängen.

Allgemein: Zu n und $1-\alpha$ bestimme man l möglichst groß, so dass die Verteilungsfunktion der Binomialverteilung $B(n,1/2)$ an der Stelle l kleiner oder gleich $\alpha/2$ ist und definiere zu diesem l

$$I(X) := [X_{(l)}, \dots, X_{(n-l+1)}]$$

Man beachte, dass die Wahrscheinlichkeiten völlig unabhängig von der Randverteilung der X_k sind (außer dass wir die Stetigkeit der Randverteilung benutzt haben, um auszuschließen, dass der Median und Realisationen zusammenfallen). Damit haben wir einen Konfidenzschätzer für den Median bestimmt, der für alle P_ϑ etwa das Konfidenzniveau $1-\alpha$ einhält. Dies ist natürlich ein großer Glücksfall. Die folgende Überlegung zeigt, dass es für den Erwartungswert **kein** endliches stochastisches Intervall $I(X)$ geben kann mit $\alpha < 1$ und $P_\vartheta(E_\vartheta(X_1) \in I(X)) \geq 1-\alpha$ für alle $\vartheta \in \Theta$, wenn Θ die Menge aller Randverteilungen mit endlichem Erwartungswert indiziert.

Um dies zu sehen, betrachten wir eine unangenehme Teilmenge von Randverteilungen. Seien wie gehabt $X = (X_1, \dots, X_n)$ mit X_k i.i.d. unter allen $(P_\vartheta)_{\vartheta\in(0,1]}$ mit

$$P_\vartheta(X_k = 0) = 1-\vartheta \quad \text{und} \quad P_\vartheta\left(X_k = \frac{1}{\vartheta^2}\right) = \vartheta. \tag{2.10}$$

Nun schaue man sich einen beliebigen Bereichsschätzer C an, dessen Werte *endliche* Intervalle sind und der auf $(X_1, \dots, X_n)$ angewandt wird (entscheidend ist die Endlichkeit und nicht dass $C(x)$ ein Intervall ist). Es gilt $E_\vartheta(X_1) = \frac{1}{\vartheta} \to \infty$ für $\vartheta \to 0$. Daraus folgt für ein beliebiges zufälliges endliches Intervall C, dass $E_\vartheta(X_1) \notin C(0, \dots, 0)$ für ϑ klein genug. Damit gilt für ϑ klein genug

$$P_\vartheta(E_\vartheta(X_1) \in C(X_1, \dots, X_n)) \leq P_\vartheta((X_1, \dots, X_n) \neq (0, \dots, 0)) \tag{2.11}$$

Andererseits gilt

$$P_\vartheta((X_1, \dots, X_n) \neq (0, \dots, 0)) = 1 - (1-\vartheta)^n \to 0, \quad \vartheta \to 0. \tag{2.12}$$

Zusammen ergeben (2.11) und (2.12) die Behauptung.

Für große n gibt es jedoch einen asymptotischen Ersatz. Dieser beruht darauf, dass für jede Verteilung von X_1 mit endlichem zweiten Moment gilt, dass

$$\frac{\sqrt{n}(\bar{X} - E_\vartheta(X_1))}{\sqrt{\frac{1}{n-1}\sum_{k=1}^n (X_k - \bar{X}_n)^2}} \to \mathcal{N}(0,1) \quad \text{in Verteilung für } n \to \infty. \tag{2.13}$$

Dies folgt daraus, dass zum einen mit dem Zentralen Grenzwertsatz

$$\frac{\sqrt{n}(\bar{X} - E_\vartheta(X_1))}{\sqrt{\mathrm{Var}_\vartheta(X_1)}} \to \mathcal{N}(0,1) \quad \text{in Verteilung für } n \to \infty.$$

Zum anderen konvergiert die Stichprobenvarianz gegen die theoretische Varianz. Es gilt nämlich

$$\begin{aligned}
&\frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X})^2 \\
&= \frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - E_\vartheta(X_1) + E_\vartheta(X_1) - \bar{X}\right)^2 \\
&= \frac{1}{n-1}\sum_{k=1}^{n}(X_k - E_\vartheta(X_1))^2 + 2(E_\vartheta(X_1) - \bar{X})\frac{1}{n-1}\sum_{k=1}^{n}(X_k - E_\vartheta(X_1)) \\
&\quad + \frac{n}{n-1}(E_\vartheta(X_1) - \bar{X})^2
\end{aligned}$$

Mit dem starken Gesetz der großen Zahlen konvergiert der erste Summand fast sicher gegen $\mathrm{Var}_\vartheta(X_1)$ und der zweiten und dritte Summand gegen Null.

Vermöge (2.13) kann nun ein asymptotisches Konfidenzintervall konstruiert werden. Bezeichne q_α das α-Quantil von $\mathcal{N}(0,1)$. Für große n gilt

$$\begin{aligned}
& P_\vartheta\left(\frac{\sqrt{n}(\bar{X} - E_\vartheta(X_1))}{\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X}_n)^2}} \in [q_{\alpha/2}, q_{1-\alpha/2}]\right) \approx 1-\alpha \\
\Leftrightarrow\quad & P_\vartheta\left(\bar{X} - E_\vartheta(X_1) \in \left[\frac{\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X}_n)^2}}{\sqrt{n}}q_{\alpha/2}, \frac{\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X}_n)^2}}{\sqrt{n}}q_{1-\alpha/2}\right]\right) \\
& \approx 1-\alpha \\
\Leftrightarrow\quad & P_\vartheta\left(E_\vartheta(X_1) \in \left[\bar{X} - \frac{\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X}_n)^2}}{\sqrt{n}}q_{1-\alpha/2}, \bar{X} - \frac{\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X}_n)^2}}{\sqrt{n}}q_{\alpha/2}\right]\right) \\
& \approx 1-\alpha
\end{aligned} \tag{2.14}$$

In diesem Sinne ist

$$I(X_1, \ldots, X_n) := \left[\bar{X} - \frac{\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X}_n)^2}}{\sqrt{n}}q_{1-\alpha/2}, \bar{X} - \frac{\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X}_n)^2}}{\sqrt{n}}q_{\alpha/2}\right]$$

ein asymptotisches Konfidenzintervall zum Niveau $1-\alpha$. Dies ist kein Widerspruch zur Nicht-Existenz eines endlichen Konfidenzintervalls im Beispiel (2.10). Der Grund hierfür ist, dass das n ab dem die Normalverteilung eine gute Approximation darstellt, von der Verteilung von X_1 abhängt. Die Approximation 2.14 gilt also zwar für alle ϑ, aber nicht gleichmäßig in ϑ.

2.4 Student, Fisher und die schöne Welt der Normalverteilung

Wir erinnern an folgende Transformationsformel von Dichten

Proposition 2.32. *Sei X eine $\mathbb{R}^n$-wertige Zufallsvariable, die bzgl. des Wahrscheinlichkeitsmaßes P eine Dichte ρ besitzt, d.h.*

$$P(X \in B) = \int 1_B(x)\rho(x)\,dx, \quad \forall B \in \mathcal{B}(\mathbb{R}^n),$$

deren gesamte Masse in der offenen Menge $\mathcal{X} \subset \mathbb{R}^n$ liegt, d.h. $\int 1_{\mathcal{X}}(x)\rho(x)\,dx = 1$. Sei ferner $\mathcal{Y} \subset \mathbb{R}^n$ eine offene Menge und $h : \mathcal{X} \to \mathcal{Y}$ ein Diffeomorphismus, d.h. eine stetig differenzierbare Bijektion mit Jacobi-Determinante $\det Dh(x) \neq 0$ *für alle $x \in \mathcal{X}$. Dann hat die Zufallsvariable $Y := h(X)$ auf $\mathcal{Y}$ die Dichte*

$$\widetilde{\rho}(y) = \rho(h^{-1}(y))|\det\, Dh^{-1}(y)|$$

Satz 2.33 (Multivariate Normalverteilung)**.** *Seien $X_1, \ldots, X_n$ reellwertige, unabhängige und normalverteilte Zufallsvariablen mit Erwartungswert $= 0$ und Varianz $= 1$. Sei $X = (X_1, \ldots, X_n)^\top$ der dazugehörige zufällige Spaltenvektor, $B \in \mathbb{R}^{n\times n}$ eine reguläre Matrix und $m \in \mathbb{R}^n$ ein Spaltenvektor. Dann hat $Y := BX + m$ die Verteilungsdichte*

$$\rho_{m,C}(y) = (2\pi)^{-n/2}|\det\, C|^{-1/2}\exp(-1/2(y-m)^\top C^{-1}(y-m)), \quad \forall y \in \mathbb{R}^n,$$

wobei $C = BB^\top$. Für die Koordinaten des Zufallsvektors Y gilt $E_P(Y_i) = m_i$ und $\mathrm{Cov}_p(Y_i, Y_j) = C_{ij}$.

Proof. Da $X_1, \ldots, X_n$ unabhängig sind, besitzt der Vektor $(X_1, \ldots, X_n)$ die Produktdichte

$$(2\pi)^{-n/2}\exp(-\frac{1}{2}x^\top x), \quad \forall x \in \mathbb{R}^n,$$

Mit Proposition 2.32 besitzt Y die Dichte

$$\begin{aligned}&(2\pi)^{-n/2}\exp(-\frac{1}{2}(B^{-1}(y-m))^\top(B^{-1}(y-m)))|\det\, DB^{-1}(y)|\\ = (2\pi)^{-n/2}&\exp(-\frac{1}{2}(y-m))^\top C^{-1}(y-m))|\det\, DC(y)|^{-1/2}, \quad \forall x \in \mathbb{R}^n. \qquad (2.15)\end{aligned}$$

Des weiteren gilt

$$E_P(Y_i) = E_P(\sum_{j=1}^n B_{ij}X_j + m_i) = \sum_{j=1}^n B_{ij}E_P(X_j) + m_i = m_i$$

und

$$\begin{aligned}\mathrm{Cov}_P(Y_iY_j) &= \mathrm{Cov}_P(\sum_{k=1}^n B_{ik}X_k, \sum_{l=1}^n B_{jl}X_l)\\ &= \sum_{k=1}^n\sum_{l=1}^n B_{ik}B_{jl}\mathrm{Cov}_P(X_k, X_l) = \sum_{k=1}^n B_{ik}B_{jk} = C_{ij}.\end{aligned}$$

□

Definition 2.34. *Zufallsvariablen $(Y_1, \ldots, Y_n)$ mit gemeinsamer Dichte (2.15) nennt man* **multivariat normalverteilt** *(oder auch gemeinsam normalverteilt) mit Erwartungswertvektor m und Varianz-Kovarianz-Matrix C.*

Für multivariat normalverteilte Zufallsvariablen (Y_1, Y_2) folgt aus der Unkorreliertheit bereits die Unabhängigkeit (da bei Unkorreliertheit nach Satz 2.33 $C_{12} = C_{21} = 0$ gilt und damit die Produktdichte faktorisierbar ist).

Man beachte, dass es natürlich zwei Zufallsvariablen geben kann, die beide normalverteilt sind, die aber als Paar nicht multivariat (bzw. bivariat) normalverteilt sind. Beispiel ?

Definition 2.35 (Gamma-Verteilung)**.** *Für $\alpha, r > 0$ heißt die Verteilung $\Gamma(\alpha, r)$ auf $(0, \infty)$ mit Dichte*

$$\gamma_{\alpha,r} : \quad x \mapsto \frac{\alpha^r}{\Gamma(r)} x^{r-1} \exp(-\alpha x)$$

Gamma-Verteilung *mit Größenparameter α und Formparameter r. Dabei bezeichnet $\Gamma(x) := \int_0^\infty t^{x-1} \exp(-t)\, dt$ die Gamma-Funktion (für $n \in \mathbb{N}$ gilt $\Gamma(n) = (n-1)!$).*

Definition 2.36 (Beta-Verteilung)**.** *Für $r, s > 0$ heißt die Verteilung $\beta(r, s)$ auf $(0, 1)$ mit Dichte*

$$x \mapsto \frac{x^{r-1}(1-x)^{s-1}}{B(r,s)}, \quad \text{wobei } B(r,s) := \int_0^1 y^{r-1}(1-y)^{s-1}\, dy,$$

Beta-Verteilung *mit Parametern r und s.*

Aus dem Beweis von Proposition 2.38 wird sich nebenbei ergeben, dass für den Normierungsfaktor gilt

$$B(r,s) = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)}.$$

Lemma 2.37. *Ist X eine $\mathcal{N}(0,1)$-verteilte Zufallsvariable, dann hat X^2 eine Gamma-Verteilung $\Gamma(1/2, 1/2)$.*

Proof. Sei X $\mathcal{N}(0,1)$-verteilt. Aus Symmetriegründen besitzt $|X|$ auf $(0,\infty)$ die doppelte Dichte wie X (das Ereignis $X = 0$, das nur mit Wahrscheinlichkeit 0 eintritt, kann bei dieser Überlegung ignoriert werden). Es gilt nämlich für alle $0 < a < b$

$$\begin{aligned} P(|X| \in (a,b)) &= P(X \in (a,b) \text{ oder } X \in (-b,-a)) \\ &= P(X \in (a,b)) + P(X \in (-b,-a)) \\ &= 2P(X \in (a,b)). \end{aligned} \tag{2.16}$$

Da die Abbildung: $h : (0,\infty) \to (0,\infty)$ $x \mapsto x^2$ ein Diffeomorphismus ist mit Umkehrabbildung $h^{-1}(y) = \sqrt{y}$ können wir mit Proposition 2.32 schließen, dass $X^2 = h(|X|)$ auf $(0,\infty)$ die Dichtefunktion

$$\widetilde{\rho}(y) = 2\rho(\sqrt{y})\frac{1}{2}y^{-1/2} = \frac{1}{\sqrt{2\pi}} \exp(-y/2) y^{-1/2} = \frac{\Gamma(1/2)}{\sqrt{\pi}} \gamma_{1/2,1/2}(y)$$

besitzt, wobei $\rho(x) = (2\pi)^{-1/2}\exp(-x^2/2)$ die Dichtefunktion der Standard-Normalverteilung bezeichne. Da sich Dichten zu 1 integrieren folgt $\Gamma(1/2) = \sqrt{\pi}$ und damit die Behauptung. □

Proposition 2.38 (Zusammenhang zwischen Gamma- und Beta-Verteilungen)**.** *Seien $\alpha, r, s > 0$ und X, Y unabhängige Zufallsvariablen mit Gamma-Verteilung $\Gamma(\alpha, r)$ bzw. $\Gamma(\alpha, s)$. Dann sind die Zufallsvariablen $X + Y$ und $X/(X + Y)$ unabhängig voneinander mit Gamma-Verteilung $\Gamma(\alpha, r + s)$ bzw. Beta-Verteilung $\beta(r, s)$.*

Bemerkung 2.39. *Für $r = 1$ stimmt die Gammaverteilung $\Gamma(\alpha, r)$ mit der Exponentialverteilung mit Parameter α überein. Die Proposition besagt also auch, dass sie die Gammaverteilung als Summe unabhängiger Exponentialverteilungen ergibt.*

Proof. Die gemeinsame Verteilung von (X, Y) besitzt auf $\mathcal{X} = (0, \infty)^2$ die Dichte

$$\frac{\alpha^{r+s}}{\Gamma(r)\Gamma(s)} x^{r-1} y^{s-1} \exp(-\alpha(x + y)), \quad x, y > 0.$$

Wir betrachten den Diffeomorphismus

$$h(x, y) := \left(x + y, \frac{x}{x + y} \right)$$

von $\mathcal{X}$ nach $\mathcal{Y} := (0, \infty) \times (0, 1)$. h hat die Umkehrabbildung

$$h^{-1}(u, v) = (uv, u(1 - v))$$

mit Jacobi-Matrix

$$Dh^{-1}(u, v) = \begin{pmatrix} v & u \\ 1 - v & -u \end{pmatrix}$$

Es folgt $|\det Dh^{-1}(u, v)| = u$. Mit Proposition 2.32 besitzt also der Zufallsvektor

$$h(X, Y) = \left(X + Y, \frac{X}{X + Y} \right)$$

die Dichte

$$\begin{aligned} &\frac{\alpha^{r+s}}{\Gamma(r)\Gamma(s)} (uv)^{r-1} (u(1 - v))^{s-1} \exp(-\alpha(uv + u(1 - v)))u \\ &= \frac{\alpha^{r+s}}{\Gamma(r)\Gamma(s)} u^{r+s-1} \exp(-\alpha u) v^{r-1} (1 - v)^{s-1} \\ &= \frac{1}{\Gamma(r)\Gamma(s)} \Gamma(r + s) \gamma_{\alpha, r+s}(u) B(r, s) \beta_{r,s}(v), \quad u > 0, 0 < v < 1. \end{aligned}$$

Da sich sowohl die zweidimensionale Dichte als auch die eindimensionalen Dichten $\gamma_{\alpha,r+s}$ und $\beta_{r,s}$ zu 1 integrieren, muss der Vorfaktor gleich 1 sein, also

$$B(r, s) = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r + s)} \tag{2.17}$$

und es folgt die Behauptung. (2.17) haben wir damit nebenbei gezeigt. □

Satz 2.40. *Seien $X_1, \dots, X_n$ unabhängige $\mathcal{N}(0,1)$-verteilte Zufallsvariablen. Dann hat die Zufallsvariable $\sum_{i=1}^n X_i^2$ die Gamma-Verteilung $\Gamma(1/2, n/2)$.*

Proof. Lemma 2.37 und Proposition 2.38 □

Definition 2.41. *Für jedes $n \geq 1$ heißt die Gamma-Verteilung $\chi^2(n) := \Gamma(1/2, n/2)$ zu den Parametern $1/2, n/2$ mit Dichtefunktion*

$$\chi_n^2(x) := \gamma_{1/2,n/2}(x) = \frac{x^{n/2-1}}{\Gamma(n/2)2^{n/2}} \exp(-x/2)$$

auch die Chiquadrat-Verteilung mit n **Freiheitsgraden** *bzw. kurz die $\chi^2(n)$-Verteilung.*

Satz 2.42 (Die Fisher-Verteilungen)**.** *Seien $X_1, \dots, X_m, Y_1, \dots, Y_n$ unabhängige und $\mathcal{N}(0,1)$-verteilte Zufallsvariablen. Dann hat der Quotient*

$$F_{m,n} := \frac{\frac{1}{m}\sum_{i=1}^m X_i^2}{\frac{1}{n}\sum_{j=1}^n Y_j^2}$$

die Verteilungsdichte

$$f_{m,n}(x) = \frac{m^{m/2}n^{n/2}}{B(m/2, n/2)} \frac{x^{m/2-1}}{(n+mx)^{(m+n)/2}}, \quad x > 0. \tag{2.18}$$

Proof. Nach Satz 2.40 hat $X := \sum_{i=1}^m X_i^2$ die Gamma-Verteilung $\Gamma(1/2, m/2)$ und $Y := \sum_{j=1}^n Y_j^2$ die Gamma-Verteilung $\Gamma(1/2, n/2)$. Da X und Y zudem unabhängig sind, hat nach Proposition 2.38 die Zufallsvariable $Z := X/(X+Y)$ die Beta-Verteilung $\beta(m/2, n/2)$. Nun gilt aber

$$F_{m,n} = \frac{n}{m}\frac{X}{Y} = \frac{n}{m}\frac{Z}{1-Z} = h(Z),$$

für den Diffeomorphismus $h : (0,1) \to (0,\infty),\ z \mapsto \frac{n}{m}\frac{z}{1-z}$. Die Umkehrabbildung ist gegeben durch $h^{-1}(u) = \frac{mu}{n+mu}$ und mit Proposition 2.32 hat $F_{m,n}$ die Verteilungsdichte

$$\beta_{m/2,n/2}\left(\frac{mu}{n+mu}\right)\frac{mn}{(n+mu)^2} = f_{m,n}(u), \quad u \in (0,\infty).$$

□

Definition 2.43 (Fisher-Verteilung)**.** *Die Verteilung $\mathcal{F}(m,n)$ auf $(0,\infty)$ mit Dichtefunktion $f_{m,n}$ gemäß (2.18) für $m, n \in \mathbb{N}$ heißt nach R.A. Fisher* **Fischer-Verteilung** *mit m und n Freiheitsgraden..*

Bemerkung 2.44 (Beziehung zwischen Fisher- und Beta-Verteilung)**.** *Im Beweis von Satz 2.42 wurde folgende Beziehung zwischen Fisher- und Beta-Verteilung herausgearbeitet:*

Sei Z eine $\beta_{m/2,n/2}$-verteilte Zufallsvariable, dann ist $R := T(Z)$ mit $T(x) = \frac{n}{m}\frac{x}{1-x}$ eine $F_{m,n}$-verteilte Zufallsvariable. Daraus folgt

$$P(R \leq y) = P(Z \leq T^{-1}(y)) = P\left(Z \leq \frac{my}{n+my}\right), \quad \forall y \in \mathbb{R}_+.$$

Somit kann man die Quantile der Beta-Verteilung und die Quantile der Fisher-Verteilung leicht ineinander umrechnen (für halbzahlige Parameter der Beta-Verteilung).

Corollary 2.45 (Student-Verteilung)**.** *Seien $X, Y_1, \ldots, Y_n$ unabhängige und $\mathcal{N}(0,1)$-verteilte Zufallsvariablen. Dann hat der Quotient*

$$T := \frac{X}{\sqrt{\frac{1}{n}\sum_{i=1}^n Y_i^2}}$$

die Verteilungsdichte

$$\tau_n(x) = \frac{1}{B(1/2, n/2)\sqrt{n}} \frac{1}{\left(1+\frac{x^2}{n}\right)^{\frac{n+1}{2}}}, \quad x \in \mathbb{R}. \tag{2.19}$$

Proof. Gemäß Satz 2.42 besitzt T^2 die Dichte $f_{1,n}$, die auf $(0,\infty)$ konzentriert ist. Mit Proposition 2.32 angewandt auf T^2 und den Diffeomorphismus $h : (0,\infty) \to (0,\infty),\ x \mapsto \sqrt{x}$ mit Umkehrabbildung $h^{-1}(y) = y^2$ hat $|T| = \sqrt{T^2}$ die Dichte $f_{1,n}(y^2)2y$. Da nun aber wegen der Symmetrie der Normalverteilung T und $-T$ die gleiche Verteilung haben, folgt mit der gleichen Überlegung wie in (2.16), dass die Dichte von T an der Stelle $y \in \mathbb{R}$ halb so groß wie die Dichte von $|T|$ an der Stelle $|y|$ ist, also

$$\begin{aligned}
\tau_n(y) &= f_{1,n}(y^2)2|y|\frac{1}{2} \\
&= f_{1,n}(y^2)|y| \\
&= \frac{n^{n/2}}{B(1/2,n/2)} \frac{|y|^{-1}}{(n+y^2)^{(1+n)/2}}|y| \\
&= \frac{1}{B(1/2,n/2)\sqrt{n}\left(1+\frac{y^2}{n}\right)^{(1+n)/2}}, \quad \forall y \in \mathbb{R}.
\end{aligned}$$

□

Definition 2.46 (Student'sche t-Verteilung)**.** *Die Verteilung t_n auf $\mathbb{R}$ mit Dichtefunktion τ_n gemäß (2.19) für $n \in \mathbb{N}$ heißt* **Student'sche t-Verteilung** *mit n Freiheitsgraden.*

Satz 2.47. *[Student 1908] Sei $\bar{X} := \frac{1}{n}\sum_{k=1}^n X_k$ und $S := \sqrt{\frac{1}{n-1}\sum_{k=1}^n (X_k - \bar{X})^2}$. Im Gauß'schen Produktmodell $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), (\mathcal{N}(\mu,\sigma^2))_{\mu\in\mathbb{R},\sigma^2>0}^{\otimes n})$ gelten für alle (μ,σ^2) bzgl. P_{μ,σ^2} die folgenden Aussagen*

(i) $\bar{X}$ und S^2 sind unabhängig.

(ii) $\bar{X}$ *hat die Verteilung* $\mathcal{N}(\mu, \sigma^2/n)$ *und* $\frac{n-1}{\sigma^2}S^2$ *die Verteilung* χ^2_{n-1}

(iii) Die Statistik

$$T := \frac{\sqrt{n}(\bar{X} - \mu)}{S}$$

hat die Verteilung t_{n-1}

Bemerkung 2.48. *Unter der Annahme, dass* $X_1, \dots, X_n$ *i.i.d. und zudem normalverteilt sind, können also analog zu dem asymptotischen Konfidenzintervall (wie in (2.14)) für den Erwartungswert (bei unbekannter Varianz) auch exakte Konfidenzintervalle für endliche Stichprobenumfänge angegeben werden. Zu* $\alpha \in (0,1)$ *seien* $t_{n-1,\alpha/2}$ *und* $t_{n-1,1-\alpha/2}$ *das* $\alpha/2$*-Quantil und das* $1-\alpha/2$*-Quantil der* t_{n-1}*-Verteilung. Ein Konfidenzintervall zum Niveau* α *ist gegeben durch*

$$\left[\bar{X} - \frac{\sqrt{\frac{1}{n-1}\sum_{k=1}^n (X_k - \bar{X}_n)^2}}{\sqrt{n}} t_{n-1,1-\alpha/2}, \bar{X} - \frac{\sqrt{\frac{1}{n-1}\sum_{k=1}^n (X_k - \bar{X}_n)^2}}{\sqrt{n}} t_{n-1,\alpha/2}\right]$$

Mit Satz 2.47 gilt nämlich für alle $\mu \in \mathbb{R}$, $\sigma > 0$

$$\begin{aligned}
&P_{\mu,\sigma^2}\left(\mu \in \left[\bar{X} - \frac{\sqrt{\frac{1}{n-1}\sum_{k=1}^n (X_k - \bar{X}_n)^2}}{\sqrt{n}} t_{n-1,1-\alpha/2}, \bar{X} - \frac{\sqrt{\frac{1}{n-1}\sum_{k=1}^n (X_k - \bar{X}_n)^2}}{\sqrt{n}} t_{n-1,\alpha/2}\right]\right) \\
&= P_{\mu,\sigma^2}\left(\frac{\sqrt{n}(\bar{X} - \mu)}{\sqrt{\frac{1}{n-1}\sum_{k=1}^n (X_k - \bar{X}_n)^2}} \in [t_{n-1,\alpha/2}, t_{n-1,1-\alpha/2}]\right) \\
&= 1 - \alpha.
\end{aligned}$$

Proof. Schritt 1: Man mache sich klar, dass es reicht die Aussage für $\mu = 0$ und $\sigma^2 = 1$ zu zeigen, da sich der allgemeine Fall durch entsprechendes Kürzen darauf zurückführen lässt.

Schritt 2: Sei $X := (X_1, \dots, X_n)^\top$ (also ein Spaltenvektor). Man betrachte den Zufallsvektor $Y := OX$ für eine geeignete orthogonale $n \times n$-Matrix O, d.h. $OO^\top = I$ (I Einheitsmatrix). Genauer soll für die erste Zeile von O gelten

$$O_{1j} = \frac{1}{\sqrt{n}}$$

(damit gilt $\sum_{j=1}^n O_{1j} O_{j1}^\top = 1$). Für die erste Koordinate von Y gilt also $Y_1 = \frac{1}{\sqrt{n}}\sum_{k=1}^n X_k = \sqrt{n}\bar{X}$. Die weiteren Zeilen werden beliebig orthogonal ergänzt. Mit Satz 2.33 und wegen $OO^\top = I$ gilt, dass Y die Dichte

$$\rho_{0,I}(y) = (2\pi)^{-n/2} \exp(-1/2 y^\top y), \quad \forall y \in \mathbb{R}^n,$$

besitzt, also wie X standardnormalverteilt ist. Dies bedeutet aber, dass $Y_1, Y_2, \ldots, Y_n$ stochastisch unabhängig sind. Zudem gilt

$$\sum_{k=1}^{n} Y_k^2 = Y^\top Y = X^\top O^\top O X = X^\top X = \sum_{k=1}^{n} X_k^2$$

und damit

$$\begin{aligned}(n-1)S^2 = \sum_{k=1}^{n}(X_k - \bar{X})^2 &= \sum_{k=1}^{n} X_k^2 - 2\bar{X}\sum_{k=1}^{n} X_k + n\bar{X}^2 \\ &= \sum_{k=1}^{n} X_k^2 - n\bar{X}^2 \\ &= \sum_{k=1}^{n} Y_k^2 - Y_1^2 = \sum_{k=2}^{n} Y_k^2.\end{aligned}$$

Folglich ist S^2 unabhängig von $\bar{X}$ (da $\sum_{k=2}^{n} Y_k^2$ unabhängig von Y_1 ist) und $(n-1)S^2$ ist χ^2_{n-1}-verteilt. □

2.5 Der Maximum-Likelihood-Schätzer

Definition 2.49. *Sei $(\mathcal{X}, \mathcal{B}, (P_\vartheta^X)_{\vartheta\in\Theta})$ ein statistisches Standardmodell mit Dichte- bzw. Gewichtsfunktionen $\rho_\vartheta(x) = \rho(\vartheta, x)$. Ein Schätzer $T : \mathcal{X} \to \Theta$ heißt ein* **Maximum-Likelihood-Schätzer**, *wenn gilt*

$$\rho(T(x), x) = \max_{\vartheta\in\Theta} \rho(\vartheta, x), \quad \forall x \in \mathcal{X}.$$

*In der englischsprachigen Literatur ist die Abkürzung MLE für „***maximum likelihood estimator***" gebräuchlich.*

Beispiel 2.50 (Normalverteilung)**.** *Wir betrachten das Gauß'sche Produktmodell*

$$\left(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), (\mathcal{N}(\mu, \sigma^2))_{\mu\in\mathbb{R},\sigma^2>0}^{\otimes n}\right)$$

mit $n \geq 2$. Wir wollen den Maximum-Likelihood-Schätzer für das unbekannte Paar (μ, σ^2) bestimmen. Zunächst ist nicht klar, ob es einen „MLE nur für μ" gibt, solange σ^2 unbekannt ist, d.h. ob es ein $\widehat{\mu}$ gibt, das für alle σ die Wahrscheinlichkeit maximiert (Man beachte, dass der MLE als Schätzer für den Parameter ϑ definiert wurde und nicht allgemeiner für einen abgeleiteten Parameter $\gamma(\vartheta)$). Für die Produktdichte gilt

$$\rho(\mu, \sigma^2, x_1, \ldots, x_n) = (2\pi\sigma^2)^{-n/2} \exp\left(-\sum_{k=1}^{n} \frac{(x_k - \mu)^2}{2\sigma^2}\right). \tag{2.20}$$

Um (2.20) zu maximieren, muss **für jedes** σ

$$\widehat{\mu} = \bar{x} := \frac{1}{n}\sum_{k=1}^{n} x_k$$

gewählt werden. Dies folgt aus der schon öfters benutzen Verschiebungsformel

$$\sum_{k=1}^{n}(x_k-\mu)^2=\sum_{k=1}^{n}(x_k-\bar{x})^2+n(\bar{x}-\mu)^2.$$

Nun muss also

$$(2\pi\sigma^2)^{-n/2}\exp\left(-\sum_{k=1}^{n}\frac{(x_k-\bar{x})^2}{2\sigma^2}\right)$$

über alle σ^2 maximiert werden. Differentiation des Logarithmuses des Ausdruckes nach σ ergibt

$$-\frac{n}{\sigma}+\frac{\sum_{k=1}^{n}(x_k-\bar{x})^2}{\sigma^3} \tag{2.21}$$

Und damit

$$\widehat{\sigma}^2:=\frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{x})^2$$

Offenbar ist die First-Order-Condition auch hinreichend für Maximalität, da die Ableitung (2.21) genau dann >0, wenn $\sigma^2<\frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{x})^2$. Man beachte, dass die Argumentation nur für $\sum_{k=1}^{n}(x_k-\bar{x})^2>0$ funktioniert. Dies gilt jedoch mit Wahrscheinlichkeit 1, da wir $n\geq 2$ vorausgesetzt haben. Im Fall $n=1$ offenbart sich eine gewisse Schwäche des Maximum-Likelihood-Ansatzes: die maximale Plausibilität der Beobachtung würde näherungsweise erreicht, indem die Varianz gegen 0 geht und würde schließlich in der Einpunkt-Verteilung δ_{x_1} angenommen.

Beispiel 2.51. *[Schätzen von Anteilen] In einer Population gebe es l sich gegenseitig ausschließende Merkmale $i=1,\dots,l$. Das Merkmal i trete mit der Wahrscheinlichkeit p_i auf. Es gilt*

$$\sum_{i=1}^{l}p_i=1. \tag{2.22}$$

Nun wird n mal mit Zurücklegen gezogen. Die Wahrscheinlichkeit, dass der entstehende Zufallsvektor $(X_1,\dots,X_n)$ den Wert $(x_1,\dots,x_n)\in\{1,\dots l\}^n$ annimmt, ist durch

$$\rho(p_1,\dots,p_l,x_1,\dots,x_n)=p_1^{k_1}\cdot\ldots\cdot p_l^{k_l}$$

gegeben, wobei $k_i:=\#\{k\mid x_k=i\}$. Die Maximierung von $\ln(\rho(p_1,\dots,p_l,x_1,\dots,x_n)$ über $p_1,\dots,p_l$ unter der Nebenbedingung (2.22) und bei gegebenen $x_1,\dots,x_n$ führt mit der Lagrange-Methode auf die Gleichungen

$$\frac{\partial}{\partial p_i}\ln(\rho(p_1,\dots,p_l,x_1,\dots,x_n))=\frac{k_i}{p_i}$$

$$\frac{\partial}{\partial p_j}(\sum_{i=1}^{l} p_i - 1) = 1$$

Also

$$\frac{k_i}{p_i} = \lambda,$$

wobei λ der Lagrange-Multiplikator ist und damit, wegen (2.22) und $\sum_{i=1}^{l} k_i = n$, $\lambda = n$ und für den ML-Schätzer folgt

$$\widehat{p_i} = \frac{k_i}{n}, \quad i = 1, \dots, l$$

2.6 Varianzminimierende Schätzer und die Cramér-Rao-Ungleichung

In diesem Abschnitt wollen wir die Optimalität von Schätzern etwas systematischer diskutieren. Unter allen erwartungstreuen Schätzern, wollen wir den Schätzer mit der kleinsten Varianz bestimmen.

Definition 2.52. *Sei $(\mathcal{X}, \mathcal{B}, (P_\vartheta^X)_{\vartheta\in\Theta})$ ein statistisches Modell. Ein erwartungstreuer Schätzer T für eine reellwertigen abgeleiteten Parameter $\gamma(\vartheta)$ heißt varianzminimierend bzw. gleichmäßig bester Schätzer, wenn für jeden weiteren erwartungstreuen Schätzer gilt*

$$\mathrm{Var}_\vartheta(T(X)) \leq \mathrm{Var}_\vartheta(S(X)), \quad \forall \vartheta \in \Theta.$$

I.A. muss ein gleichmäßig bester Schätzer nicht existieren. Ein erwartungstreuer Schätzer könnte für ein ϑ_1 eine besonders geringe Varianz liefern, ein anderer erwartungstreuer Schätzer für ein ϑ_2. Beispiel ?

Proposition 2.53. *Es kann höchstens einen varianzminimierenden Schätzer geben.*

Proof. Seinen T und S varianzminimierende Schätzer für $\gamma(\cdot)$. Insbesondere gilt $\mathrm{Var}_\vartheta(T) = \mathrm{Var}_\vartheta(S)$ für alle $\vartheta \in \Theta$.

Man betrachte das arithmetische Mittel der Schätzer. Dieser ist offenbar ebenso ein erwartungstreuer Schätzer. Nun schätze man unter einem Parameter ϑ die Varianz des gemittelten Schätzers ab

$$\begin{aligned}
&\mathrm{Var}_\vartheta\left(\frac{1}{2}(T(X) + S(X))\right) \\
&= \frac{1}{4}\mathrm{Var}_\vartheta(T(X)) + \frac{1}{2}\mathrm{Cov}_\vartheta(T(X), S(X)) + \frac{1}{4}\mathrm{Var}_\vartheta(S(X)) \\
&= \frac{1}{4}\mathrm{Var}_\vartheta(T(X)) + \frac{1}{2}\frac{1}{2}\left[\mathrm{Var}_\vartheta(T(X)) + \mathrm{Var}_\vartheta(S(X)) - \mathrm{Var}_\vartheta(T(X) - S(X))\right] + \frac{1}{4}\mathrm{Var}_\vartheta(S(X)) \\
&= \frac{1}{2}\mathrm{Var}_\vartheta(T(X)) + \frac{1}{2}\mathrm{Var}_\vartheta(S(X)) - \frac{1}{4}\mathrm{Var}_\vartheta(T(X) - S(X)) \\
&= \mathrm{Var}_\vartheta(T(X)) - \frac{1}{4}\mathrm{Var}_\vartheta(T(X) - S(X)).
\end{aligned}$$

Da der gemittelte Schätzer keine strikt kleinere Varianz als T und S haben kann, folgt $\mathrm{Var}_\vartheta\,(T(X) - S(X)) = 0$. Da beide Schätzer erwartungstreu sind, folgt $P_\vartheta(T(X) = S(X)) = 1$. □

Definition 2.54. *Ein einparametriges Standardmodell $(\mathcal{X}, \mathcal{B}, (P_\vartheta^X)_{\vartheta\in\Theta})$ heißt regulär, falls*

- Θ *ein offenes Intervall in* $\mathbb{R}$ *ist.*
- *Die Likelihood Funktion $\rho : \Theta \times \mathcal{X} \to \mathbb{R}_+$ ist strikt positiv und nach ϑ stetig differenzierbar. Insbesondere besitzt die Funktion $\ln(\rho(\vartheta, x))$ eine Ableitung nach ϑ mit*

$$U(\vartheta, x) := \frac{\partial}{\partial\vartheta} \ln(\rho(\vartheta, x)) = \frac{\frac{\partial}{\partial\vartheta}\rho(\vartheta, x)}{\rho(\vartheta, x)}.$$

- *Für jedes $\vartheta \in \Theta$ existiert die Varianz*

$$I(\vartheta) := \mathrm{Var}_\vartheta(U(\vartheta, X)) \tag{2.23}$$

uns ist > 0 und es gilt die Vertauschungsrelation

$$\int \frac{\partial}{\partial\vartheta}\rho(\vartheta, x)\,dx = \frac{\partial}{\partial\vartheta} \underbrace{\int \rho(\vartheta, x)\,dx}_{=1 \quad \forall\vartheta} \quad (= 0). \tag{2.24}$$

Die Funktion I aus (2.23) heißt **Fisher-Information** *des Modells (nach dem britischen Statistiker Sir Ronald A. Fisher, 1880-1962).*

Als Konsequenz aus der Vertauschbarkeit (2.24) ergibt sich für alle $\vartheta \in \Theta$

$$E_\vartheta(U(\vartheta, X)) = \int \left(\frac{\partial}{\partial\vartheta} \ln(\rho(\vartheta, x))\right) \rho(\vartheta, x)\,dx = \int \frac{\partial}{\partial\vartheta}\rho(\vartheta, x)\,dx = 0 \tag{2.25}$$

und damit

$$I(\vartheta) = E_\vartheta(U(\vartheta, X)^2) - (E_\vartheta(U(\vartheta, X)))^2 = E_\vartheta(U(\vartheta, X)^2).$$

Der Begriff Information wird verwendet, da $I(\vartheta_0)$ eine Aussage darüber macht, wie gut durch Beoachtung des Ausgangs X des Zufallsexperiments zwischen dem Parameter ϑ_0 und einem Parameter ϑ mit $|\vartheta - \vartheta_0|$,,klein" unterschieden werden kann. Ein Extremfall, der hier allerdings formal ausgeschlossen ist, liegt vor, wenn ein $\varepsilon > 0$ existiert, so dass für alle $x \in \mathcal{X}$ und alle $\vartheta \in [\vartheta_0 - \varepsilon, \vartheta_0 + \varepsilon]$ $U(\vartheta, x) = 0$ gilt. In diesem Extremfall wären die Dichten für alle $\vartheta \in [\vartheta_0 - \varepsilon, \vartheta_0 + \varepsilon]$ identisch und die Information $I(\vartheta_0) = 0$. Es bestünde keine Möglichkeit zwischen den Parametern im Intervall $[\vartheta_0 - \varepsilon, \vartheta_0 + \varepsilon]$ zu unterscheiden. Trotzdem kann es natürlich möglich sein, dass Paramter außerhalb des Intervalls sehr gut unterschieden werden können.

Proposition 2.55 (Additivität der Fisher-Information)**.** *Ist $\mathcal{M} = (\mathcal{X}, \mathcal{B}, (P_\vartheta^X)_{\vartheta\in\Theta})$ ein reguläres Modell mit Fisher-Information I, so hat das Produktmodell $\mathcal{M}^n = (\mathcal{X}^n, \mathcal{B}^{\otimes n}, (P_\vartheta^{(X_1,\dots,X_n)})_{\vartheta\in\Theta})$ die Fisher-Information $I^{\otimes n} = nI$.*

Proof. Für die Produktdichte gilt

$$\ln(\rho(\vartheta, x_1, \dots, x_n)) = \ln\left(\prod_{k=1}^n \rho(\vartheta, x_k)\right) = \sum_{k=1}^n \ln\left(\rho(\vartheta, x_k)\right)$$

und damit

$$U(\vartheta, X_1, \dots, X_n) = \frac{\partial}{\partial\vartheta}\ln(\rho(\vartheta, X_1, \dots, X_n)) = \sum_{k=1}^n \frac{\partial}{\partial\vartheta}\ln\left(\rho(\vartheta, X_k)\right) = \sum_{k=1}^n U(\vartheta, X_k).$$

Des weiteren sind die Zufallsvariablen $U(\vartheta, X_k)$, $k = 1, \dots, n$ unter P_ϑ i.i.d., da selbiges für X_k, $k = 1, \dots, n$ gilt und es gilt wegen (2.24) $E_\vartheta(U(\vartheta, X_k)) = 0$. Damit folgt

$$E_\vartheta(U(\vartheta, X_1, \dots, X_n)) = nE_\vartheta(U(\vartheta, X_1)).$$

□

Noch eine kleine unvermeidbare Definition

Definition 2.56. *Ein Schätzer T heißt regulär, wenn*

$$\int T(x)\frac{\partial}{\partial\vartheta}\rho(\vartheta, x)\,dx = \frac{\partial}{\partial\vartheta}\int T(x)\rho(\vartheta, x)\,dx.$$

Satz 2.57 (Cramér-Rao-Ungleichung)**.** *Sei $\mathcal{M} = (\mathcal{X}, \mathcal{B}, (P_\vartheta^X)_{\vartheta\in\Theta})$ ein reguläres statistisches Modell, $\gamma : \Theta \to \mathbb{R}$ eine stetig differenzierbare Funktion mit $\gamma'(\vartheta) \neq 0$ für alle ϑ und T ein regulärer, erwartungstreuer Schätzer für $\gamma(\vartheta)$. Dann gilt*

$$\mathrm{Var}_\vartheta(T(X)) \geq \frac{(\gamma'(\vartheta))^2}{I(\vartheta)} \quad \forall\vartheta \in \Theta. \tag{2.26}$$

Gleichheit für alle $\vartheta \in \Theta$ gilt genau dann, wenn

$$T(X) - \gamma(\vartheta) = \frac{\gamma'(\vartheta)U(\vartheta, X)}{I(\vartheta)} \quad \forall\vartheta \in \Theta, \tag{2.27}$$

d.h. wenn das Modell die Likelihood-Funktion

$$\rho(\vartheta, x) = \exp\left(a(\vartheta)T(x) - b(\vartheta)\right)h(x), \quad \forall\vartheta \in \Theta, \tag{2.28}$$

besitzt, wobei $a : \Theta \to \mathbb{R}$ eine Stammfunktion von I/γ' ist, $h : \mathcal{X} \to (0, \infty)$ eine beliebige messbare Funktion ist und

$$b(x) = \ln\left(\int \exp(a(\vartheta)T(x))h(x)\,dx\right)$$

sich eine aus der Normiertheit von Dichten ergebene Funktion ist.

Definition 2.58 (Exponentialfamilie)**.** $\mathcal{M} = (\mathcal{X}, \mathcal{B}, (P_\vartheta^X)_{\vartheta\in\Theta})$ *heißt ein exponentielles Modell und* $\{P_\vartheta \mid \vartheta \in \Theta\}$ *eine Exponentialfamilie* **bezüglich einer Statistik** $T : \mathcal{X} \to \mathbb{R}$*, wenn die Likelihood-Funktion die Gestalt (2.28) besitzt mit einer stetig differenzierbaren Funktion* $a : \Theta \to \mathbb{R}$ *mit* $a' \neq 0$ *und einer messbaren Funktion* $h : \mathcal{X} \to (0, \infty)$*.*

Proof. Wegen der in (2.25) gezeigten Zentriertheit von $U(\vartheta, X)$ und der Erwartungstreue des Schätzers T gilt

$$\begin{aligned}
\mathrm{Cov}_\vartheta(T(X), U(\vartheta, X)) &= E_\vartheta(T(X)U(\vartheta, X)) \\
&= \int T(x)\frac{\partial}{\partial\vartheta}\rho(\vartheta, x)\,dx \\
&= \frac{\partial}{\partial\vartheta}\int T(x)\rho(\vartheta, x)\,dx \\
&= \frac{\partial}{\partial\vartheta}E_\vartheta(T(X)) \\
&= \gamma'(\vartheta), \quad \forall\vartheta \in \Theta. \qquad (2.29)
\end{aligned}$$

Mit der Cauchy-Schwarz'schen Ungleichung folgt bereits

$$\mathrm{Var}_\vartheta(T(X))I(\vartheta) = \mathrm{Var}_\vartheta(T(X))\mathrm{Var}_\vartheta(U(\vartheta, X)) \geq \left(\mathrm{Cov}_\vartheta(T(X), U(\vartheta, X))\right)^2 = \left(\gamma'(\vartheta)\right)^2,$$

also (2.26). Wann gilt Gleichheit und Optimalität ? Da, wie in (2.29) gesehen, die Kovarianz von $T(X)$ und $U(\vartheta, X)$ durch die Erwartungstreue von $T(X)$ vorgegeben ist, wird die Varianz von $T(X)$ minimiert, wenn $T(X)$ und $U(\vartheta, X)$ perfekt korreliert sind. Das Problem ist natürlich, dass perfekte Korrelation gleichzeitig für alle ϑ erfüllt sein muss, was i.A. nicht erreichbar ist. Mit der Konstanten $c(\vartheta) := \gamma'(\vartheta)/I(\vartheta)$ und unter Benutzung von (2.29) errechnet man

$$\begin{aligned}
0 &\leq \mathrm{Var}_\vartheta(T(X) - c(\vartheta)U(\vartheta, X)) \\
&= \mathrm{Var}_\vartheta(T(X)) - 2c(\vartheta)\mathrm{Cov}_\vartheta(T(X), U(\vartheta, X)) + (c(\vartheta))^2\mathrm{Var}_\vartheta(U(\vartheta, X)) \\
&= \mathrm{Var}_\vartheta(T(X)) - 2\frac{(\gamma'(\vartheta))^2}{I(\vartheta)} + \frac{(\gamma'(\vartheta))^2}{I^2(\vartheta)}I(\vartheta) \\
&= \mathrm{Var}_\vartheta(T(X)) - \frac{(\gamma'(\vartheta))^2}{I(\vartheta)}.
\end{aligned}$$

Für festes ϑ gilt Gleichheit genau dann, wenn die Zufallsvariable $T(X) - c(\vartheta)U(\vartheta, X)$ konstant ist, also mit ihrem Erwartungswert übereinstimmt, d.h.

$$T(X) - c(\vartheta)U(\vartheta, X) = \gamma(\vartheta), \qquad (2.30)$$

also (2.27). Eine Umformung ergibt

$$\frac{I(\vartheta)}{\gamma'(\vartheta)}T(X) - \frac{\gamma(\vartheta)I(\vartheta)}{\gamma'(\vartheta)} = \frac{\frac{\partial}{\partial\vartheta}\rho(\vartheta, X)}{\rho(\vartheta, X)}. \qquad (2.31)$$

Wenn ρ die Form (2.28) besitzt, dann folgt durch Differentiation nach ϑ, dass

$$\frac{\frac{\partial}{\partial\vartheta}\rho(\vartheta, X)}{\rho(\vartheta, X)} = a'(\vartheta)T(X) - b'(\vartheta). \tag{2.32}$$

Mit $a'(\vartheta) = I(\vartheta)/\gamma'(\vartheta)$ und Erwartungswertbildung auf beiden Seiten von (2.32) folgt $b'(\vartheta) = \frac{\gamma(\vartheta)I(\vartheta)}{\gamma'(\vartheta)}$ und damit (2.31).

Setze umgekehrt (2.31) für alle $\vartheta \in \Theta$ voraus. Dann gilt

$$\int_{\vartheta_0}^{\cdot} \left[\frac{I(\vartheta)}{\gamma'(\vartheta)}T(X) - \frac{\gamma(\vartheta)I(\vartheta)}{\gamma'(\vartheta)}\right] d\vartheta = \int_{\vartheta_0}^{\cdot} \frac{\frac{\partial}{\partial\vartheta}\rho(\vartheta, X)}{\rho(\vartheta, X)}\, d\vartheta = \ln(\rho(\cdot, X)) + C(X)$$

für ein beliebige $\vartheta_0 \in \Theta$ (mit der üblichen Konvention, dass $\int_b^a \ldots := -\int_a^b \ldots$ für $a < b$) und damit

$$\rho(\cdot, X) = \exp\left(T(X)\int_{\vartheta_0}^{\cdot} \frac{I(\vartheta)}{\gamma'(\vartheta)}\, d\vartheta - \int_{\vartheta_0}^{\cdot} \frac{\gamma(\vartheta)I(\vartheta)}{\gamma'(\vartheta)}\, d\vartheta - C(X)\right).$$

Man beachte, dass wegen der geforderten Stetigkeit von γ' und der Forderung $\gamma'(\vartheta) \neq 0$ für alle ϑ, γ' auf Kompakta gleichmäßig von der Null entfernt ist. □

Bemerkung 2.59. *Aus Satz 2.57 in Verbindung mit Proposition 2.55 folgt, dass in regulären Modellen bei n-facher unabhängiger Wiederholung die Varianz eines Schätzers bestenfalls mit der Ordnung $1/n$ fallen kann. In Beispiel 2.25, beim Erraten des Bereichs einer Zufallsvaraiblen, hatten wir beim Schätzen des Parameters ϑ der Gleichverteilung auf $[0, \vartheta]$ den Schätzer*

$$T_n^B := \frac{n+1}{n}\max\{X_1, \ldots, X_n\}.$$

betrachtet mit

$$\mathrm{Var}(T_n^B) \;=\; \frac{\vartheta^2}{n(n+2)},$$

d.h. die Varianz geht schneller als mit der Ordnung $1/n$ gegen 0. Dies ist jedoch kein Widerspruch, da die Gleichverteilungen wegen der fehlenden Differenzierbarkeit der Dichten an den Rändern kein reguläres statistisches Modell ist.

Beispiel 2.60 (Binomialmodell)**.** *Wir betrachten erneut Beispiel 2.5, also das Binomialmodell $(B(1,p))_{p\in[0,1]}^{\otimes n}$. Dies ist ein Spezialfall von Beispiel 2.51 für $l = 2$ Merkmale. Damit ist der unbekannte Parameter eindimensional (und Satz 2.57 ist anwendbar). Es wurde gezeigt, dass*

$$T(x_1, \ldots, x_n) := \frac{\#\{i \mid x_i = 1\}}{n}$$

der ML-Schätzer für p ist. Offenbar bildet das Binomialmodell zusammen mit dem ML-Schätzer eine Exponentialfamilie. Es gilt nämlich

$$\begin{aligned}\rho(p, x_1, \dots, x_n) &= p^{\#\{i \mid x_i=1\}}(1-p)^{n-\#\{i \mid x_i=1\}} \\ &= \exp\left(\#\{i \mid x_i = 1\}\ln(p) + (n - \#\{i \mid x_i = 1\})\ln(1-p)\right) \\ &= \exp\left(\#\{i \mid x_i = 1\}\ln\left(\frac{p}{1-p}\right) + n\ln(1-p)\right).\end{aligned}$$

Wähle also $a(p) = n\ln\left(\frac{p}{1-p}\right)$, $h(x_1, \dots, x_n) = 1$ und $b(p) = -n\ln(1-p)$. Mit Satz 2.57 folgt, dass der ML-Schätzer im Binomialmodell varianzminimierend ist.

2.7 Lineare Regression

Bei der linearen Regression geht man davon aus, dass Beobachtungwerte (affin-)linear von einer Kontrollvariablen abhängen, diese Abhängigkeit aber von einem zufälligen Rauschen überlagert wird. Die lineare Abhängigkeit soll nun möglichst gut aus den Beobachtungen herausgefiltert werden. Seien die Kontrollvariablen die bekannten und nicht als Zufallsvariablen modellierten Punkte $t_1, t_2, \dots, t_n$ und $X_1, X_2, \dots, X_n$ die zufälligen Beobachtungen. Die t_k dürfen zusammenfallen, aber nicht alle gleich sein. Betrachte das Modell

$$X_k = \mu_0 + \mu_1 t_k + \sigma\xi_k, \quad k = 1, \dots, n, \tag{2.33}$$

wobei $\mu_0, \mu_1 \in \mathbb{R}$ und $\sigma \in \mathbb{R}_+$ unbekannte Parameter sind. ξ_k sind standardisierte Zufallsvariablen, d.h. $E(\xi_k) = 0$ und $\mathrm{Var}(\xi_k) = 1$ (erst durch diese Normierung bekommen die Paramter $\mu_0, \mu_1 \in \mathbb{R}$ und $\sigma \in \mathbb{R}_+$ ihre beabsichtigte Bedeutung). An die ξ_k werden später bei Bedarf weitere Voraussetzungen gestellt (sie müssen zunächst nicht i.i.d. sein).

Die Variable mit den Werten $t_1, t_2, \dots, t_n$ wird **Regressionsvariable** und die Variable mit Werten $X_1, X_2, \dots, X_n$ wird **Zielvariable** genannt.

Bemerkung 2.61. *Unser gewohntes statistischen Modell würden wir erhalten, indem wir bei fester (und bekannter) Verteilung des Zufallvektors $(\xi_1, \xi_2, \dots, \xi_n)$ mit P_{μ_0,μ_1,σ^2} die Verteilung des Zufallsvektors $(X_1, \dots, X_n) = (\mu_0 + \mu_1 t_1 + \sigma\xi_1, \dots, \mu_0 + \mu_1 t_n + \sigma\xi_n)$ bezeichnen. Aber Vorsicht, $(X_1, \dots, X_n)$ ist dann* **nicht** *der Zufallsvektor im entstehenden Modell*

$$(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), (P_{\mu_0,\mu_1,\sigma^2})_{\mu_0,\mu_1\in\mathbb{R},\sigma^2\in\mathbb{R}_+}).$$

Vielmehr haben wir X_k benutzt, um das Maß P_{μ_0,μ_1,σ^2} zu spezifizieren. X_k, als Abbildung von Ω nach $\mathbb{R}$, hängt ja von den Parametern μ_0, μ_1, σ^2 und damit vom Maß P_{μ_0,μ_1,σ^2} ab, was der Zufallsvektor in unserem gewohnten Modell nicht darf. Man könnte nun aber einen Zufallsvektor $(\widetilde{X}_1, \dots, \widetilde{X}_n)$ angeben (unabhängig von den Parametern μ_0, μ_1, σ^2), der unter P_{μ_0,μ_1,σ^2} wie $(X_1, \dots, X_n)$ verteilt ist.

Beim **Prinzip der kleinsten Quadrate** wird bei gegebenen Realisationen von $X_1, \ldots, X_n$ die Summe der quadratischen Abweichungen zwischen X_k und $\mu_0 + \mu_1 t_k$ minimiert, also der Ausdruck

$$Q(\mu_0, \mu_1) := \sum_{k=1}^{n} \left(X_k - (\mu_0 + \mu_1 t_k)\right)^2 .$$

Die eindeutige Minimalstelle $\widehat{\mu} := (\widehat{\mu}_0, \widehat{\mu}_1)$ von Q ist also eine Zufallsvariable und scheint ein vielversprechender Schätzer für $\mu := (\mu_0, \mu_1)$ zu sein. Die Minimalität des quadratischen Abstandes *definiert* hier erst den Schätzer, sie ist nicht zu verwechseln mit der minimalen Varianz eines Schätzers aus dem vorherigen Abschnitt !

Die First-Order-Conditions lauten

$$\frac{\partial}{\partial \mu_0} Q(\mu_0, \mu_1) = -2 \sum_{k=1} (X_k - \mu_0 - \mu_1 t_k) \stackrel{!}{=} 0$$

und

$$\frac{\partial}{\partial \mu_1} Q(\mu_0, \mu_1) = -2 \sum_{k=1} t_k (X_k - \mu_0 - \mu_1 t_k) \stackrel{!}{=} 0$$

Dies führt zu den Gleichungen

$$\mu_0 + \mu_1 \bar{t} = \bar{X} \tag{2.34}$$

und

$$\mu_0 \bar{t} + \mu_1 \frac{1}{n} \sum_{k=1}^{n} t_k^2 = \frac{1}{n} \sum_{k=1}^{n} t_k X_k, \tag{2.35}$$

wobei $\bar{t} := \frac{1}{n} \sum_{k=1}^{n} t_k$ und $\bar{X} := \frac{1}{n} \sum_{k=1}^{n} X_k$. Wir bezeichnen wie in Kapitel 1 die **empirische** Varianz des Datensatzes $t := (t_1, \ldots, t_n)$ und der Beobachtungen $X := (X_1, \ldots, X_n)$ mit

$$\mathrm{var}(t) = \frac{1}{n} \sum_{k=1}^{n} (t_k - \bar{t})^2 = \frac{1}{n} \sum_{k=1}^{n} t_k^2 - \bar{t}^2$$

bzw.

$$\mathrm{var}(X) = \frac{1}{n} \sum_{k=1}^{n} (X_k - \bar{X})^2 = \frac{1}{n} \sum_{k=1}^{n} X_k^2 - \bar{X}^2$$

und deren empirische Kovarianz mit

$$\mathrm{cov}(t, X) = \frac{1}{n} \sum_{k=1}^{n} (t_k - \bar{t})(X_k - \bar{X}) = \frac{1}{n} \sum_{k=1}^{n} t_k X_k - \bar{t}\bar{X}.$$

Damit ist (2.35) äquivalent zu

$$\mu_0\bar{t} + \mu_1\left[\mathrm{var}(t) + \bar{t}^2\right] = \mathrm{cov}(t, X) + \bar{t}\bar{X},$$

Mit (2.34) ist die rechte Seite $\mathrm{cov}(t, X) + \bar{t}(\mu_0 + \mu_1\bar{t})$ und damit

$$\mu_1\mathrm{var}(t) = \mathrm{cov}(t, X).$$

Da nach Voraussetzung nicht alle t_k übereinstimmen, gilt $\mathrm{var}(t) > 0$ und

$$\mu_1 = \frac{\mathrm{cov}(t, X)}{\mathrm{var}(t)}.$$

Für μ_0 ergbt dies

$$\mu_0 = \bar{X} - \frac{\mathrm{cov}(t, X)}{\mathrm{var}(t)}\bar{t}.$$

Satz 2.62 (Regressionsgerade)**.** *Die Statistiken*

$$\widehat{\mu}_0(X) := \bar{X} - \frac{\mathrm{cov}(t, X)}{\mathrm{var}(t)}\bar{t}.$$

und

$$\widehat{\mu}_1(X) := \frac{\mathrm{cov}(t, X)}{\mathrm{var}(t)}$$

sind die eindeutigen Kleinste-Quadrate-Schätzer für die Paramter μ_0 und μ_1. Beide Schätzer sind erwartungstreu.

Proof. Es bleibt die Erwartungstreue zu zeigen. Aus $E(X_k) = \mu_0 + \mu_1 t_k$ folgt

$$\begin{aligned}
E(\widehat{\mu}_1(X)) &= E\left(\frac{\mathrm{cov}(t, X)}{\mathrm{var}(t)}\right) \\
&= \frac{1}{\mathrm{var}(t)}\left(\frac{1}{n}\sum_{k=1}^{n} t_k E(X_k) - \bar{t}E(\bar{X})\right) \\
&= \frac{1}{\mathrm{var}(t)}\left(\frac{1}{n}\sum_{k=1}^{n} t_k(\mu_0 + \mu_1 t_k) - \bar{t}(\mu_0 + \mu_1\bar{t})\right) \\
&= \frac{1}{\mathrm{var}(t)}\mu_1\left(\frac{1}{n}\sum_{k=1}^{n} t_k^2 - \bar{t}^2\right) \\
&= \mu_1.
\end{aligned}$$

Weiter folgt

$$E(\widehat{\mu}_0(X)) = E(\bar{X}) - \bar{t}E(\widehat{\mu}_1(X)) = \mu_0 + \frac{1}{n}\sum_{k=1}^{n}\mu_1 t_k - \bar{t}\mu_1 = \mu_0.$$

□

Man beachte, dass für dieses Resultat noch keine Unabhängigkeit oder Unkorreliertheit der Störterme ξ_i gebraucht wird.

Bemerkung 2.63 (Zufällige Regressionsvariable)**.** *Es würde sich nicht viel ändern, wenn die Regressionvariable t auch zufällig gemacht würde. Zum Beispiel könnte man Zufallsvariablen $T_1, \dots, T_n$ erzeugen, die stochastisch unabhängig von den Störtermen $\xi_1, \dots, \xi_n$ sind und $X_1, \dots, X_n$ analog zu (2.33) definieren.*

Definition 2.64 (Linearer Schätzer)**.** *Ein linearer Schätzer ist ein Schätzer der Form $T(X) = \sum_{k=1}^n b_k X_k$ mit $b_k \in \mathbb{R}$.*

Satz 2.65 (Spezialfall vom Satz von Gauß)**.** *Seien die $(\xi_k)_{k=1,\dots,n}$ in (2.33) unkorreliert. Dann haben $\widehat{\mu}_0(X)$ und $\widehat{\mu}_1(X)$ unter allen erwartungstreuen linearen Schätzern für μ_0 bzw. μ_1 jeweils die kleinste Varianz.*

Schätzer mit der Eigenschaft aus Satz 2.65 werden **beste lineare Schätzer** genannt, mit der englischen Abkürzung **BLUE** für **"best linear unbiased estimator"**.

Beweis von Satz 2.65. Zeigen wir zunächst die Aussage für $\widehat{\mu}_1(X)$. Offenbar ist der Schätzer selber linear. Wir werden zunächst zeigen, dass

$$\mathrm{Cov}(\widehat{\mu}_1(X), \sum_{k=1}^n b_k X_k) = 0 \tag{2.36}$$

für alle $(b_1, \dots, b_n)$ mit

$$\sum_{k=1}^n b_k = 0 \tag{2.37}$$

und

$$\sum_{k=1}^n b_k t_k = 0. \tag{2.38}$$

Es gilt

$$\begin{aligned}
\mathrm{Cov}(\mathrm{var}(t)\widehat{\mu}_1(X), \sum_{k=1}^n b_k X_k) &= \mathrm{Cov}(\frac{1}{n}\sum_{k=1}^n t_k X_k - \bar{t}\bar{X}, \sum_{k=1}^n b_k X_k) \\
&= \frac{1}{n}\mathrm{Cov}(\sum_{k=1}^n t_k \sigma\xi_k, \sum_{k=1}^n b_k \sigma\xi_k) - \mathrm{Cov}(\bar{t}\sigma\bar{\xi}, \sum_{k=1}^n b_k \sigma\xi_k) \\
&= \frac{\sigma^2}{n}\sum_{k=1}^n t_k b_k - \frac{\sigma^2}{n}\bar{t}\sum_{k=1}^n b_k \\
&= 0.
\end{aligned}$$

Damit folgt (2.36). Sei nun $T(X)$ ein beliebiger linearer Schätzer für μ_1. T lässt sich schreiben als $T(X) = \widehat{\mu}_1(X) + \sum_{k=1}^n b_k X_k$, wobei die $(b_1, \dots, b_n)$ wegen der Erwartungstreue von T und $\widehat{\mu}_1$ die Bedingungen (2.37) und (2.37) erfüllt. Es folgt

$$\begin{aligned}
\mathrm{Var}(T(X)) &= \mathrm{Var}(\widehat{\mu}_1(X)) + 2\mathrm{Cov}(\widehat{\mu}_1(X), \sum_{k=1}^n b_k X_k) + \mathrm{Var}(\sum_{k=1}^n b_k X_k) \\
&\overset{(2.36)}{=} \mathrm{Var}(\widehat{\mu}_1(X)) + \mathrm{Var}(\sum_{k=1}^n b_k X_k) \\
&\geq \mathrm{Var}(\widehat{\mu}_1(X)).
\end{aligned}$$

Wegen

$$\mathrm{Cov}(\bar{X}, \sum_{k=1}^n b_k X_k) = 0$$

für alle $(b_1, \dots, b_n)$ mit (2.37), gilt nun wegen (2.36) auch

$$\mathrm{Cov}(\widehat{\mu}_0(X), \sum_{k=1}^n b_k X_k) = 0$$

für alle $(b_1, \dots, b_n)$, die (2.37) und (2.38) erfüllen. Damit folgt die Minimalität der Varianz von $\widehat{\mu}_0(X)$ unter allen erwartungstreuen, linearen Schätzern analog zu oben. □

Satz 2.66. *Seien die $(\xi_k)_{k=1,\dots,n}$ in (2.33) unkorreliert. Für $n \geq 3$ ist der Schätzer*

$$\widehat{\sigma}^2(X) := \frac{1}{n-2} \sum_{k=1}^n (X_k - \widehat{\mu}_0(X) - \widehat{\mu}_1(X) t_k)^2 = \frac{1}{n-2} \sum_{k=1}^n \left(X_k - \bar{X} - \frac{\mathrm{cov}(t, X)}{\mathrm{var}(t)} (t_k - \bar{t}) \right)^2$$

erwartungstreu für σ^2.

Proof. Sei $\xi := (\xi_1, \dots, \xi_n)$ und $\bar{\xi} := \frac{1}{n} \sum_{k=1}^n \xi_k$. Für jedes $k = 1, \dots, n$ gilt

$$\begin{aligned}
& X_k - \bar{X} - \frac{\mathrm{cov}(t, X)}{\mathrm{var}(t)} (t_k - \bar{t}) \\
&= \mu_0 + \mu_1 t_k + \sigma \xi_k - \mu_0 - \mu_1 \bar{t} - \sigma \bar{\xi} - \mu_1 \frac{\mathrm{cov}(t, t)}{\mathrm{var}(t)} (t_k - \bar{t}) - \sigma \frac{\mathrm{cov}(t, \xi)}{\mathrm{var}(t)} (t_k - \bar{t}) \\
&= \sigma \left(\xi_k - \bar{\xi} - \frac{\mathrm{cov}(t, \xi)}{\mathrm{var}(t)} (t_k - \bar{t}) \right). \qquad (2.39)
\end{aligned}$$

Des weiteren gilt

$$\begin{aligned}
&\sum_{k=1}^{n}\left(\xi_k-\bar{\xi}-\frac{\operatorname{cov}(t,\xi)}{\operatorname{var}(t)}(t_k-\bar{t})\right)^2\\
&=\sum_{k=1}^{n}\xi_k^2-2\sum_{k=1}^{n}\xi_k\bar{\xi}-2\sum_{k=1}^{n}\xi_k\frac{\operatorname{cov}(t,\xi)}{\operatorname{var}(t)}(t_k-\bar{t})+\sum_{k=1}^{n}\left(\bar{\xi}+\frac{\operatorname{cov}(t,\xi)}{\operatorname{var}(t)}(t_k-\bar{t})\right)^2\\
&=\sum_{k=1}^{n}\xi_k^2-2n\bar{\xi}^2-2\frac{\operatorname{cov}(t,\xi)}{\operatorname{var}(t)}\sum_{k=1}^{n}\xi_k(t_k-\bar{t})+n\bar{\xi}^2+\sum_{k=1}^{n}\left[\frac{\operatorname{cov}(t,\xi)}{\operatorname{var}(t)}(t_k-\bar{t})\right]^2\\
&=\sum_{k=1}^{n}\xi_k^2-2n\bar{\xi}^2-2\frac{\operatorname{cov}(t,\xi)}{\operatorname{var}(t)}\underbrace{\left[\sum_{k=1}^{n}\xi_k t_k-n\bar{\xi}\bar{t}\right]}_{=n\ \operatorname{cov}(t,\xi)}+n\bar{\xi}^2+\left[\frac{\operatorname{cov}(t,\xi)}{\operatorname{var}(t)}\right]^2\sum_{k=1}^{n}(t_k-\bar{t})^2\\
&=\sum_{k=1}^{n}\xi_k^2-n\bar{\xi}^2-\frac{n(\operatorname{cov}(t,\xi))^2}{\operatorname{var}(t)}. \qquad (2.40)
\end{aligned}$$

Die Erwartungswerte berechnen sich zu

$$E[n\bar{\xi}^2]=\frac{1}{n}E\left[\left(\sum_{k=1}^{n}\xi_k\right)^2\right]=\frac{1}{n}nE[\xi_1^2]=1$$

und

$$\begin{aligned}
E\left[\frac{n(\operatorname{cov}(t,\xi))^2}{\operatorname{var}(t)}\right] &= \frac{n}{\operatorname{var}(t)}E\left[(\operatorname{cov}(t,\xi))^2\right]\\
&= \frac{n}{\operatorname{var}(t)}E\left[\left(\frac{1}{n}\sum_{k=1}^{n}\xi_k(t_k-\bar{t})\right)^2\right]\\
&= \frac{n}{\operatorname{var}(t)}\frac{1}{n^2}\sum_{k=1}^{n}(t_k-\bar{t})^2\\
&= 1.
\end{aligned}$$

Es folgt

$$\begin{aligned}
E\left[\sum_{k=1}^{n}\left(X_k-\bar{X}-\frac{\operatorname{cov}(t,X)}{\operatorname{var}(t)}(t_k-\bar{t})\right)^2\right] &= \sigma^2E\left[\sum_{k=1}^{n}\left(\xi_k-\bar{\xi}-\frac{\operatorname{cov}(t,\xi)}{\operatorname{var}(t)}(t_k-\bar{t})\right)^2\right]\\
&= \sigma^2E\left[\sum_{k=1}^{n}\xi_k^2-n\bar{\xi}^2-\frac{n(\operatorname{cov}(t,\xi))^2}{\operatorname{var}(t)}\right]\\
&= \sigma^2(n-1-1)=\sigma^2(n-2)
\end{aligned}$$

und damit die Erwartungstreue des Schätzers $\widehat{\sigma}^2(X)$ für σ^2.

□

2.7.1 Das lineare Gauß-Modell

In diesem Abschnitt werden wir die zusätzliche Annahme machen, dass der Störvektor $\xi = (\xi_1, \dots, \xi_n)$ multivariat normalverteilt ist. Insbesondere sind die ξ_k nicht nur unkorreliert, sondern auch stochastisch unabhängig.

Satz 2.67. *Im Gauß'schen Produktmodell gelten für alle $\mu_0, \mu_1 \in \mathbb{R}$, $\sigma^2 > 0$ bzw. für alle entsprechenden P_{μ_0,μ_1,σ^2} die folgenden Aussagen*

(i) *$\bar{X}$, $\widehat{\mu}_1(X)$ und $\widehat{\sigma}^2(X)$ sind stochastisch unabhängig. $\widehat{\mu}_0(X)$ ist stochastisch unabhängig von $\widehat{\sigma}^2(X)$.*

(ii) *$\bar{X}$ hat die Verteilung $\mathcal{N}\left(\mu_0 + \mu_1\bar{t}, \frac{\sigma^2}{n}\right)$, $\widehat{\mu}_1(X)$ hat die Verteilung $\mathcal{N}\left(\mu_1, \frac{\sigma^2}{n\mathrm{var}(t)}\right)$, $\widehat{\mu}_0(X)$ hat die Verteilung $\mathcal{N}\left(\mu_0, \frac{\sigma^2}{n}\left(1 + \frac{\bar{t}^2}{\mathrm{var}(t)}\right)\right)$ und $\frac{n-2}{\sigma^2}\widehat{\sigma}^2(X)$ die Verteilung χ^2_{n-2}*

(iii) *Die Statistiken*

$$\frac{\sqrt{n\ \mathrm{var}(t)}(\widehat{\mu}_1(X) - \mu_1)}{\widehat{\sigma}(X)} \tag{2.41}$$

und

$$\frac{\sqrt{n}(\widehat{\mu}_0(X) - \mu_0)}{\sqrt{1 + \frac{\bar{t}^2}{\mathrm{var}(t)}}\widehat{\sigma}(X)} \tag{2.42}$$

haben beide die Verteilung t_{n-2}.

Satz 2.67 ähnelt dem Satz von Student (Satz 2.47). Der Unterschied besteht darin, dass das statistische Modell nun eine lineare Abhängigkeit der Daten von der Regressionsvariablen t erlaubt, die im Satz von Student nicht vorkommt. Würden wir alle t_k gleich 0 setzen, wären wir wieder in der Situation von Satz 2.47. Man beachte jedoch, dass dies formal ausgeschlossen ist, weil wir $\mathrm{var}(t) > 0$ vorausgesetzt haben. Bei $\mathrm{var}(t) = 0$ wäre μ_1 nicht schätzbar. $\frac{n-2}{\sigma^2}\widehat{\sigma}^2(X)$ hat nun $n-2$ statt $n-1$ Freiheitsgrade, da mehr Diversität durch das Modell erklärt und nicht auf den Zufall geschoben wird.

Proof. Schritt 1: Man mache sich klar, dass in den Statistiken (2.41) und (2.42) X_k durch ξ_k ersetzt werden kann. In den Zählern und Nennern von (2.41) und (2.42) heben sich μ_0 und μ_1 nämlich weg. In den Quotienten kürzt sich zudem σ weg (letzters folgt aus der Linearität von $\widehat{\mu}_0(X)$ und $\widehat{\mu}_1(X)$ sowie (2.39)). Also o.B.d.A. $\mu_0 = \mu_1 = 0$ und $\sigma^2 = 1$.

Schritt 2: $\bar{X}$, $\widehat{\mu}_1(X)$ und $\widehat{\mu}_0(X)$ sind als gewichtete Summen unabhängiger Normalverteilungen wieder normalverteilt. Es müssen nur noch die Varianzen von $\widehat{\mu}_1(X)$ und $\widehat{\mu}_0(X)$

ausgerechnet werden. Es gilt

$$\begin{aligned}
\mathrm{Var}(\widehat{\mu}_1(X)) &= \mathrm{Var}\left(\frac{\mathrm{cov}(t,X)}{\mathrm{var}(t)}\right) \\
&= \mathrm{Var}\left(\frac{\sum_{k=1}^n (t_k-\bar{t})X_k}{n\ \mathrm{var}(t)}\right) \\
&= \frac{\sum_{k=1}^n (t_k-\bar{t})^2}{n^2(\mathrm{var}(t))^2} \\
&= \frac{1}{n\ \mathrm{var}(t)}
\end{aligned}$$

und

$$\begin{aligned}
\mathrm{Var}(\widehat{\mu}_0(X)) &= \mathrm{Var}\left(\bar{X} - \frac{\mathrm{cov}(t,X)}{\mathrm{var}(t)}\bar{t}\right) \\
&= \mathrm{Var}\left(\sum_{k=1}^n X_k\left(\frac{1}{n} - \frac{(t_k-\bar{t})\bar{t}}{n\ \mathrm{var}(t)}\right)\right) \\
&= \sum_{k=1}^n \left(\frac{1}{n} - \frac{(t_k-\bar{t})\bar{t}}{n\ \mathrm{var}(t)}\right)^2 \\
&= \frac{1}{n} - \frac{2}{n}\sum_{k=1}^n \frac{(t_k-\bar{t})\bar{t}}{n\ \mathrm{var}(t)} + \sum_{k=1}^n \frac{(t_k-\bar{t})^2\bar{t}^2}{(n\mathrm{var}(t))^2} \\
&= \frac{1}{n}\left(1 + \frac{\bar{t}^2}{\mathrm{var}(t)}\right).
\end{aligned}$$

Schritt 3: Sei $X := (X_1, \ldots, X_n)^\top$ (also ein Spaltenvektor) multivariat standardnormalverteilt. Man betrachte den Zufallsvektor $Y := OX$ für eine geeignete orthogonale $n \times n$-Matrix O, d.h. $OO^\top = I$ (I Einheitsmatrix). Genauer definieren wir die erste Zeile von O durch

$$O_{1j} = \frac{1}{\sqrt{n}}$$

und die zweite Zeile durch

$$O_{2j} = \frac{1}{\sqrt{n\ \mathrm{var}(t)}}(t_j - \bar{t}).$$

Damit gilt $\sum_{j=1}^n O_{1j}O_{j1}^\top = 1$, $\sum_{j=1}^n O_{2j}O_{j2}^\top = 1$ und $\sum_{j=1}^n O_{1j}O_{j2}^\top = 0$. Für die erste Koordinate von Y gilt also $Y_1 = \frac{1}{\sqrt{n}}\sum_{k=1}^n X_k = \sqrt{n}\bar{X}$ und für die zweite Koordinate $Y_2 = \frac{1}{\sqrt{n\ \mathrm{var}(t)}}\sum_{j=1}^n (t_j-\bar{t})X_j = \sqrt{\frac{n}{\mathrm{var}(t)}}\mathrm{cov}(t,X)$.

Die weiteren Zeilen werden beliebig orthogonal ergänzt. Mit Satz 2.33 und wegen $OO^\top = I$ gilt, dass Y die Dichte

$$\rho_{0,I}(y) = (2\pi)^{-n/2}\exp(-1/2y^\top y), \quad \forall y \in \mathbb{R}^n,$$

besitzt, also wie X standardnormalverteilt ist. Dies bedeutet aber, dass $Y_1, Y_2, \ldots, Y_n$ stochastisch unabhängig sind. Insbesondere ist gezeigt, dass $\bar{X}$ und $\widehat{\mu}_1(X)$ stochastisch unabhängig sind. Zudem gilt

$$\sum_{k=1}^{n} Y_k^2 = Y^\top Y = X^\top O^\top O X = X^\top X = \sum_{k=1}^{n} X_k^2$$

und damit

$$\begin{aligned}
(n-2)\widehat{\sigma}^2(X) &= \sum_{k=1}^{n} \left(X_k - \bar{X} - \frac{\operatorname{cov}(t,X)}{\operatorname{var}(t)}(t_k - \bar{t}) \right)^2 \\
&\overset{(2.40)}{=} \sum_{k=1}^{n} X_k^2 - n\bar{X}^2 - \frac{n(\operatorname{cov}(t,X))^2}{\operatorname{var}(t)} \\
&= \sum_{k=1}^{n} Y_k^2 - Y_1^2 - Y_2^2 \\
&= \sum_{k=3}^{n} Y_k^2.
\end{aligned}$$

Folglich ist $\widehat{\sigma}^2(X)$ unabhängig von $(\bar{X}, \widehat{\mu}_1(X))$ (da $\sum_{k=3}^{n} Y_k^2$ unabhängig von (Y_1, Y_2) ist) und $(n-2)\widehat{\sigma}^2(X)$ ist χ^2_{n-2}-verteilt. Da $\widehat{\mu}_0(X) = \bar{X} - \widehat{\mu}_1(X)\bar{t}$ ist auch $\widehat{\mu}_0(X)$ stochastisch unabhängig von $\widehat{\sigma}^2(X)$. □