Skript zur Vorlesung

Mathematische Statistik

ETHZ, D-Math, WS 2005/2006

H.R. Künsch
Seminar für Statistik

Oktober 2005

Inhaltsverzeichnis

Kapitel 1

Einführende Beispiele

Mathematische Statistik ist die Theorie der Analyse von Daten unter Verwendung von Wahrscheinlichkeitsmodellen. Ausgehend von einer Familie $\mathcal{P}$ von Wahrscheinlichkeitsverteilungen, einem sogenannten Modell, möchte man die folgenden Fragen beantworten:

a) Ist es plausibel, dass die Daten von einem $P \in \mathcal{P}$ erzeugt wurden?

b) Angenommen, dass die Verteilung der Daten zu $\mathcal{P}$ gehört, welches P oder welche Teilmenge von $\mathcal{P}$ erscheint plausibel?

Wie kommt man aber zu einem Modell $\mathcal{P}$? In erster Linie kommt es darauf an, aus welchem Gebiet die Daten stammen und wie sie gewonnen wurden. Daneben spielen aber auch Konvention und Erfahrung mit ähnlichen Situationen eine wichtige Rolle. Ausserdem ist statistische Analyse ein iterativer Prozess, in dessen Verlauf die Familie $\mathcal{P}$ im Licht bisheriger Erkenntnisse modifiziert wird. Diese Dinge sind schwierig zu formalisieren und zu unterrichten. Etwas enger betrachtet, werden wir uns vor allem mit den folgenden Fragen in einem fest vorgegebenen Modell $\mathcal{P}$ befassen:

c) Wie vergleicht man verschiedene Verfahren zur Beantwortung der Fragen a) und b)?

d) Wie quantifiziert man die Unsicherheit der Antworten?

Meist werden wir die Prinzipien anhand von einfachen Modellen erläutern. Die mehr praktischen Aspekte der Statistik, wo man auch mit komplexeren Modellen arbeitet, kommen in anderen Statistikvorlesungen zur Sprache. In diesem Kapitel stellen wir die Fragestellungen an zwei Beispielen etwas ausführlicher dar.

1.1 Bestimmung der Lichtgeschwindigkeit

Newcomb hat 1882 gemessen, wie lange das Licht braucht, um eine Strecke von 7'400 m zurückzulegen. Er erhielt die folgenden Werte (in 10^{-9}sec):
24'800 +

28	26	33	24	34	-44	27	16	40	-2	29	22	24	21
25	30	23	29	31	19	24	20	36	32	36	28	25	21
28	29	37	25	28	26	30	32	36	26	30	22	36	23
27	27	28	27	31	27	26	33	26	32	32	24	39	28
24	25	32	25	29	27	28	29	16	23				

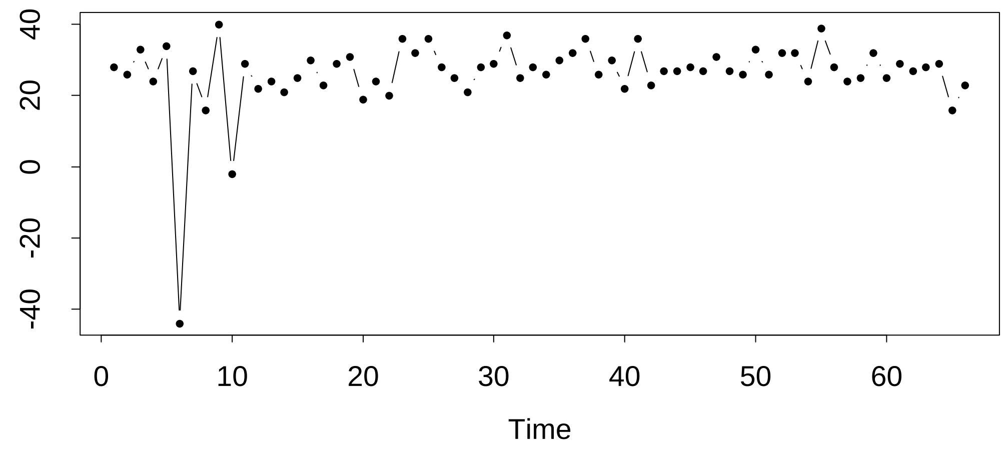
1.1.1 Modell

Das einfachste Modell ist

$$X_i = \mu + \varepsilon_i \quad (1 \le i \le n = 66)$$

wobei μ =wahre Zeit, die das Licht benötigt; und ε_i = Messfehler. Man nimmt an, dass die ε_i i.i.d. sind mit $\varepsilon_i \sim F$. Die Elemente von $\mathcal{P}$ sind also parametrisiert durch $\theta = (\mu, F) \in \mathbb{R} \times \mathcal{F}$. Für $\mathcal{F}$ gibt es verschiedene Möglichkeiten, z.B. $\mathcal{F} = \{F = \Phi(./\sigma); \sigma > 0\}$ oder $\mathcal{F} = \{F | F \text{ stetig}, F(-x) = 1 - F(x)\}$. In diesem Modell steckt die Annahme, dass es keine systematischen Fehler gibt. Von primärem Interesse ist der Parameter μ: Welches ist unsere beste Schätzung von μ bzw. in welchem Bereich etwa liegt μ? Die Verteilung F, bzw. die Streuung σ, interessieren uns erst in zweiter Linie. Es sind *Störparameter* (auf englisch nuisance parameter).

Zur Überprüfung des Modells sollte man mindestens die Daten gegen die Reihenfolge auftragen:



Man sieht, dass man zwar keine systematischen Effekte wie Trends, aber zwei deutliche Ausreisser hat. Die Annahme von identischer Verteilung und von Normalverteilung ist also sehr fragwürdig.

1.1.2 Punktschätzung

Ein Schätzer ist eine Abbildung $\hat{\mu} : \mathbb{R}^n \to \mathbb{R}$. Es gibt viele vernünftig aussehende Schätzer, z.B.

- Arithmetisches Mittel: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
- Median (mittlere Beobachtung):

$$\hat{\mu} = \begin{cases} X_{((n+1)/2)} & \text{falls } n \text{ ungerade} \\ \frac{1}{2}(X_{(n/2)} + X_{(n/2+1)}) & \text{falls } n \text{ gerade} \end{cases}$$

wobei $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ die geordnete Stichprobe bezeichnet.

- α - gestutztes Mittel ($0 < \alpha < \frac{1}{2}$):

$$\hat{\mu} = \frac{1}{n - 2[n\alpha]} \sum_{i=[n\alpha]+1}^{n-[n\alpha]} X_{(i)}$$

- Huber-Schätzer:

$$\hat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} \rho(X_i - \mu)$$

wobei

$$\rho(x) = \begin{cases} x^2/2 & \text{für } |x| \le c \\ c(|x| - c/2) & \text{für } |x| \ge c. \end{cases}$$

Die Schwelle c wird üblicherweise als Vielfaches einer Streuungsschätzung gewählt, z.B. $c = 1.5\mathrm{MAD}$, wobei $\mathrm{MAD} = \mathrm{median}(|X_i - \mathrm{median}(X_j)|)$ (MAD steht für "median absolute deviation", Median der absoluten Abweichungen vom Median).

Da $\arg\min_{\mu} \sum (X_i - \mu)^2$ das arithmetische Mittel ist und $\arg\min_{\mu} \sum |X_i - \mu|$ der Median (siehe Satz A.8), ergibt die Familie der Huber-Schätzer ebenfalls einen stetigen Übergang vom arithmetischen Mittel zum Median.

Da ρ konvex ist, kann der Huber-Schätzer auch durch

$$\sum_{i=1}^{n} \psi(X_i - \hat{\mu}) \stackrel{!}{=} 0$$

definiert werden mit $\psi = \rho'$ (stückweise, falls nötig)

Aufgrund welcher Kriterien soll man sich für einen Schätzer entscheiden ? Ein naheliegendes Kriterium ist eine möglichst kleine mittlere Abweichung, d.h. man minimiert

$$R(\hat{\mu}, \mu, F) := E_{\mu,F}[|\hat{\mu}(X_1, \ldots, X_n) - \mu|^p]$$

oder

$$R(\hat{\mu}, \mu, F) := P_{\mu,F}[|\hat{\mu}(X_1, \ldots, X_n) - \mu| > a].$$

Falls

$$\hat{\mu}(x_1 + c, \ldots, x_n + c) = \hat{\mu}(x_1, \ldots, x_n) + c \quad \forall x_1, \ldots, x_n, c$$

(sogenannte *Äquivarianz*, siehe Kapitel 4), dann ist $R(\hat{\mu}, \mu, F)$ unabhängig von μ. Die Abhängigkeit von F bleibt jedoch bestehen. Es zeigt sich auch, dass der optimale Schätzer wesentlich von F abhängt (siehe Kapitel 4), und F kennen wir ja gerade nicht. Dies ist ein Grundproblem der Statistik.

Ein weiteres Kriterium kann aber auch sein, wie empfindlich der Schätzer auf vereinzelte grob falsche Beobachtungen (Ausreisser) reagiert. Das kann erfasst werden mit der empirischen grobe-Fehler-Sensitivität

$$\gamma^* = \sup_x |\hat{\mu}(X_1, \ldots, X_n, x) - \hat{\mu}(X_1, \ldots, X_n)|$$

(maximaler Einfluss einer zusätzlichen Beobachtung) oder mit dem empirischen Bruchpunkt

$$\varepsilon^* = \frac{\min\{m; \sup\{|\hat{\mu}(y_1, \ldots, y_n)|\,; |\{i; y_i \neq x_i\}| = m\} = \infty\}}{n}.$$

Das ist der minimaler Anteil von Ausreissern, der zum Zusammenbruch des Schätzers führen kann.

Sowohl γ^* als auch ε^* hängen i.a. von den Beobachtungen ab. In Kapitel 6 werden wir eine einfache Näherung für γ^* kennenlernen.

1.1.3 Vertrauensintervalle

Definition 1.1. *Ein von den Daten abhängiges Intervall* $I = [\underline{\mu}(X_1, \dots, X_n), \overline{\mu}(X_1, \dots, X_n)]$ *heisst ein* Vertrauensintervall *(oder* Konfidenzintervall*) für* μ *zum Niveau* $1-\alpha$ *falls*

$$P_{\mu,F}[I \ni \mu] \geq 1 - \alpha \quad \forall \mu, \forall F \in \mathcal{F}.$$

Das Niveau bedeutet, dass bei N unabhängigen Wiederholungen des ganzen Experiments das zufällige Vertrauensintervall den unbekannten wahren Wert in etwa $(1-\alpha)N$ Fällen einfängt.

Wenn wir statt eines Intervalls eine zufällige Menge haben, sprechen wir von einem Vertrauensbereich. Die folgende Überlegung zeigt, dass Vertrauensbereiche äquivalent sind zu Tests: Betrachte für jedes μ_0 aus der Parametermenge einen Test der Nullhypothese $H_0 : \mu = \mu_0$ zum Niveau α, d.h. eine Familie von Abbildungen $\varphi(., \mu_0) : \mathbb{R}^n \to \{0, 1\}$ derart, dass

$$P_{\mu_0,F}[\varphi(X_1, \dots, X_n; \mu_0) = 1] \leq \alpha \quad \forall \mu_0, F.$$

Dann können wir den Bereich von Nullhypothesen betrachten, die nicht verworfen werden:

$$I(X_1, \dots, X_n) = \{\mu | \varphi(X_1, \dots, X_n; \mu) = 0\}.$$

Es gilt offensichtlich

$$P_{\mu,F}[I \ni \mu] = P_{\mu,F}[\varphi(X_1, \dots, X_n; \mu) = 0] \geq 1 - \alpha.$$

Ferner ist I ein Intervall, falls aus $\mu_1 < \mu_2 < \mu_3$ und $\varphi(X_1, \dots, X_n; \mu_1) = \varphi(X_1, \dots, X_n; \mu_3) = 0$ folgt, dass $\varphi(X_1, \dots, X_n; \mu_2) = 0$.

Wir formulieren diese Korrespondenz zwischen Vertrauensbereichen und Tests für einen abgeleiteten Parameter $\gamma = g(\theta)$ noch allgemein (oben war $\gamma = g(\mu, F) = \mu$).

Lemma 1.1. *Sei* $(P_\theta; \theta \in \Theta)$ *eine Familie von Wahrscheinlichkeitsverteilungen auf einem Raum* $\mathbb{X}$ *und* g *eine Abbildung* $\Theta \to \Gamma$.

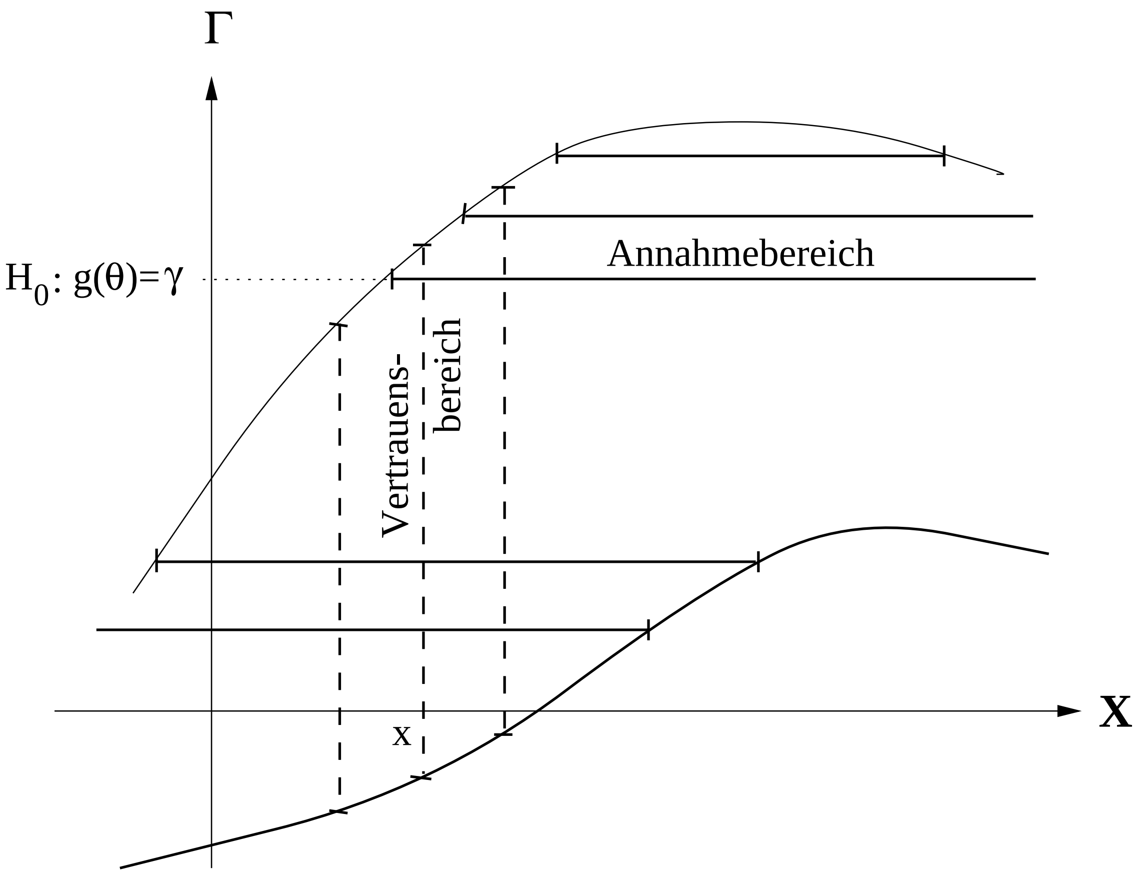
a) Wenn für jedes $\gamma \in \Gamma \quad \varphi(X; \gamma)$ *ein Test für* $H_0 : g(\theta) = \gamma$ *zum Niveau* α *ist, dann ist* $I(X) = \{\gamma | \varphi(X; \gamma) = 0\}$ *ein Vertrauensbereich für* $g(\theta)$ *zum Niveau* $1 - \alpha$.

b) Wenn $I(X) \subset \Gamma$ *ein Vertrauensbereich für* $g(\theta)$ *zum Niveau* $1 - \alpha$ *ist, dann ist* $\varphi(X; \gamma) = \mathbf{1}_{[\gamma \notin I(X)]}$ *ein Test für* $H_0 : g(\theta) = \gamma$ *zum Niveau* α.

Beweis. folgt aus den Definitionen, vgl. die nächste Figur. □

Tests und Vertrauensintervalle zu beliebigem Niveau erhalten wir bequem, wenn es gelingt, ein sogenanntes *Pivot* ("Tür-Angel") zu finden. Das ist eine Abbildung $T : \mathbb{X} \times \Gamma \to \mathbb{R}$ derart dass

$$P_\theta[T(X, g(\theta)) \leq t] = G(t) \quad \forall \theta.$$



Wir bestimmen dann ein Intervall $[t_1, t_2]$ derart, dass

$$P_\theta[T(X, g(\theta)) < t_1] \le \frac{\alpha}{2}, \quad P_\theta[T(X, g(\theta)) > t_2] \le \frac{\alpha}{2}$$

und t_1 möglichst gross, t_2 möglichst klein. Dies wird geleistet durch die beiden "Quantilfunktionen" von G

$$t_1 = q_+(\tfrac{\alpha}{2}) = \sup\{t|G(t) \le \tfrac{\alpha}{2}\}, \quad t_2 = q_-(1-\tfrac{\alpha}{2}) = \inf\{t|G(t) \ge 1-\tfrac{\alpha}{2}\}$$

(selber nachprüfen !). Also ist der Indikator der Menge

$$\{x \in \mathbb{X}|T(x,\gamma) \notin [q_+(\tfrac{\alpha}{2}), q_-(1-\tfrac{\alpha}{2})]\}$$

ein Test zum Niveau α, und die Menge

$$\{\gamma \in \Gamma|T(x,\gamma) \in [q_+(\tfrac{\alpha}{2}), q_-(1-\tfrac{\alpha}{2})]\}$$

ein Vertrauensbereich zum Niveau $1-\alpha$.

In unserem Modell für die Messung der Lichtgeschwindigkeit können wir Pivots finden. Sie hängen aber davon ab, welche Annahmen wir über $\mathcal{F}$ treffen.

- Wenn $\mathcal{F} = \{F\}$ und $\hat{\mu}$ äquivariant ist, dann ist

  $$T(X_1, \ldots, X_n; \mu) = \hat{\mu}(X_1, \ldots, X_n) - \mu$$

  ein Pivot.

- Wenn $\mathcal{F} = \{\Phi(./\sigma); \sigma > 0\}$, dann ist
$$T(X_1, \ldots, X_n; \mu) = \frac{n^{1/2}(\overline{X} - \mu)}{S_n}$$
ein Pivot, wobei $S_n^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \overline{X})^2$. Die Verteilung von T ist die sogenannte t-Verteilung mit $(n-1)$ Freiheitsgraden (siehe Satz A.5). Der zugehörige Test ist natürlich der bekannte t-Test.

- Wenn $\mathcal{F} = \{F | F(0) = \frac{1}{2}, F \text{ stetig an der Stelle } 0\}$, dann ist
$$T = \sum_{i=1}^n \mathbf{1}_{[X_i \geq \mu]}$$
ein Pivot, denn T ist Binomial $(n, \frac{1}{2})$ verteilt. Der zugehörige Test ist der Vorzeichentest. Das Vertrauensintervall hat die Form $[X_{(k_{\alpha/2})}, X_{(n-k_{\alpha/2}+1)}]$, wobei k_α das obere α-Quantil $q_+(\alpha)$ der Binomial$(n, \frac{1}{2})$-Verteilung ist.

Oft ist es aber nicht so einfach, Pivots zu konstruieren, und man gibt sich mit asymptotischen Pivots zufrieden. Darunter versteht man eine Folge (T_n) derart, dass
$$P_\theta[T_n(X_1, \ldots, X_n; g(\theta)) \leq t]$$
gegen einen Wert in $(0,1)$ konvergiert, der unabhängig von θ ist. Zum Beispiel konvergiert
$$P_{\mu,F}\left[\frac{n^{1/2}(\overline{X} - \mu)}{S_n} \leq t\right]$$
gegen $\Phi(t)$ für alle F mit $\int x F(dx) = 0, \int x^2 F(dx) < \infty$. Daher ist $n^{1/2}(\overline{X} - \mu)/S_n$ ein asymptotisches Pivot. Allgemein werden wir in Kapitel 6 sehen, wie man ausgehend von sehr vielen Schätzern $\hat{\mu}$ ein asymptotisches Pivot der Form
$$\frac{n^{1/2}(\hat{\mu} - \mu)}{\hat{\sigma}}$$
konstruieren kann.

Nach welchen Kriterien soll man verschiedene Vertrauensintervalle zum gleichen Niveau vergleichen? Naheliegend wäre z.B. die mittlere Länge, d.h. $E_{\mu,F}[|I(X_1, \ldots, X_n)|]$. Der Zusammenhang zwischen Vertrauensintervallen und Tests von vorhin legt auch das folgende Kriterium nahe:
$$P_{\mu,F}[I(X_1, \ldots, X_n) \ni \mu'] \stackrel{!}{=} \min$$
für ein festes $\mu' \neq \mu$. Dies ist die Wahrscheinlichkeit, dass das Vertrauensintervall einen falschen Parameter umfasst und entspricht dem Fehler 2. Art beim Testen.

1.1.4 Ausblick

Die Schätzung von Konstanten, die wir hier diskutiert haben, ist in der Praxis nicht sehr häufig. Das gleiche Modell tritt aber auf, wenn wir zwei Behandlungen an gleichen Versuchseinheiten vergleichen (durch Bildung von Differenzen). Schliesslich lassen sich viele Resultate dieses einfachen Lokationsmodells auf das lineare Modell übertragen, bei dem
$$Y_i = \sum_{j=1}^p x_{ij}\beta_j + \varepsilon_i \quad (1 \leq i \leq n)$$

wobei Y_i beobachtete Werte einer Zielvariablen sind, die x_{ij} bekannte erklärende Variablen und β_j unbekannte Parameter. Dieses Modell ist in der Praxis sehr verbreitet und wird in der Vorlesung Regression ausführlich behandelt.

1.2 Zwei-Stichproben-Tests

Um den Effekt von negativer Beeinflussung auf das Abschneiden in Intelligenztests zu untersuchen, wurden 10 Versuchspersonen in eine Kontroll- und eine Behandlungsgruppe eingeteilt. Alle Personen machten 2 Tests im Abstand von 2 Wochen. In der Kontrollgruppe herrschten beide Male neutrale Bedingungen, während die Personen in der Behandlungsgruppe vor dem zweiten Test kritisiert und entmutigt wurden. Es ergaben sich die folgenden Unterschiede in den Resultaten der beiden Tests:

Kontrolle	5	0	16	2	9
Behandlung	6	-5	-6	1	4.

Da man Hypothesen nur widerlegen, aber nicht beweisen kann, testet man die Nullhypothese “Die Entmutigung hat keine Auswirkungen”.

1.2.1 Zufallsmodelle und Teststatistiken

Wir bezeichnen die Werte der Kontrollgruppe mit $X_1, \ldots, X_n$ und die Werte der Behandlungsgruppe mit $Y_1, \ldots, Y_m$. Wir nehmen an, dass alle Zufallsvariablen unabhängig sind und dass alle X_i die Verteilung F und alle Y_i die Verteilung G haben. Im einfachsten Fall postuliert man ferner, dass die Behandlung höchstens eine Verschiebung ausmacht, d.h. $G(x) = F(x - \delta)$ für ein $\delta \in \mathbb{R}$. Das Modell ist dann also parametrisiert durch $\theta = (\delta, F)$. Bezüglich F können wir wieder mehr oder weniger restriktiv sein: $F \in \mathcal{F} = \{\Phi((. - \mu)/\sigma); \mu \in \mathbb{R}, \sigma \in \mathbb{R}_+\}$ oder $F \in \mathcal{F} = \{$ Alle Verteilungen $\}$. Die Nullhypothese ist $H_0 : \delta = 0$. Ein Test ist eine Abbildung $\varphi : \mathbb{R}^{n+m} \to \{0, 1\}$, und unter allen Tests mit

$$P_{(0,F)}[\varphi(X_1, \ldots, X_n, Y_1, \ldots, Y_m) = 1] \leq \alpha \quad \forall F \in \mathcal{F}$$

(d.h. mit Niveau α) möchten wir die Macht

$$P_{(\delta,F)}[\varphi(X_1, \ldots, X_n, Y_1, \ldots, Y_m) = 1]$$

maximieren für $\delta < 0$ (Wegen der Fragestellung sind wir nur an solchen Alternativen interessiert). Im allgemeinen geht das nicht simultan für alle δ und F, weshalb das Problem schwierig ist.

In der Praxis ist der Test durch eine Teststatistik $T : \mathbb{R}^{n+m} \to \mathbb{R}$ gegeben via

$$\varphi(x) = 1 \Leftrightarrow T(x) > c$$

mit $c = c(\alpha)$. Die beiden bekanntesten Test- sind der t-Test und der Wilcoxon-Test mit

$$T_t = (n^{-1} + m^{-1})^{-1/2}(\overline{X} - \overline{Y})/S,$$

wobei

$$S^2 = (n + m - 2)^{-1}(\sum_{i=1}^{n}(X_i - \overline{X})^2 + \sum_{i=1}^{m}(Y_i - \overline{Y})^2),$$

bzw.

$$T_W = \sum_{i=1}^{n} \mathrm{Rang}(X_i),$$

wobei $\mathrm{Rang}(X_i)$ angibt, die wievielt-kleinste Beobachtung das X_i in der kombinierten Stichprobe $(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ ist. Eine alternative Form ist

$$T_W = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{1}_{[Y_j < X_i]} + \frac{n(n+1)}{2},$$

falls alle Beobachtungen verschieden sind (nachrechnen !).

Zur Bestimmung von c brauchen wir die Verteilung von T unter der Nullhypothese. Im Fall der Normalverteilung ist die Teststatistik des t-Tests t-verteilt mit $n+m-2$ Freiheitsgraden (analoger Beweis wie in Satz A.5). Bei andern Verteilungen erhalten wir wenigstens asymptotisch die gleiche Verteilung aufgrund des zentralen Grenzwertsatzes (Existenz der zweiten Momente vorausgesetzt).

Für die Statistik des Wilcoxon-Tests benützt man folgendes Lemma (wir beweisen etwas mehr als wir jetzt eigentlich brauchen – für später).

Lemma 1.2. *Falls $X_1, \ldots, X_n$ i.i.d. $\sim F$ sind mit F stetig, dann sind die Ränge $(R_1, \ldots, R_n)$ unabhängig von den Werten der geordneten Stichprobe und bilden eine Zufallspermutation.*

Beweis. Zuerst zeigen wir, dass $X_1, \ldots, X_n$ mit Wahrscheinlichkeit 1 alle verschieden sind; die Ränge sind also wohldefiniert. Das ist klar für X_i uniform, und ein allgemeines X_i erhält man als $F^{-1}(U_i)$ mit U_i uniform. Wenn F stetig ist, ist F^{-1} strikt monoton.

Wir müssen zeigen, dass für ein beliebiges $A \subseteq \mathbb{R}^n$ und eine beliebige Permutation π gilt

$$\begin{aligned} &P[(X_{(1)}, \ldots, X_{(n)}) \in A, R_1 = \pi_1, \ldots, R_n = \pi_n] \\ &= P[(X_{(1)}, \ldots, X_{(n)}) \in A] P[R_1 = \pi_1, \ldots, R_n = \pi_n]. \end{aligned}$$

Da $R_j = k$ äquivalent ist zu $X_{(k)} = X_j$ und da sich die Verteilung von $(X_1, \ldots, X_n)$ bei Permutationen der Indizes nicht ändert, haben wir

$$\begin{aligned} &P[(X_{(1)}, \ldots, X_{(n)}) \in A, R_1 = \pi_1, \ldots, R_n = \pi_n] \\ &= P[(X_{(\pi^{-1})_1}, \ldots X_{(\pi^{-1})_n}) \in A, X_{(\pi^{-1})_1} < \cdots < X_{(\pi^{-1})_n}] \\ &= P[(X_1, \ldots, X_n) \in A, X_1 < \cdots < X_n]. \end{aligned}$$

Im letzten Ausdruck tritt π nicht mehr auf. Durch Summation über π folgt daher

$$n! P[(X_{(1)}, \ldots, X_{(n)}) \in A, R_1 = \pi_1, \ldots, R_n = \pi_n] = P[(X_{(1)}, \ldots, X_{(n)}) \in A].$$

Mit $A = \mathbb{R}^n$ ergibt sich $P[R_1 = \pi_1, \ldots, R_n = \pi_n] = 1/n!$ und damit auch die Behauptung. □

Insbesondere also unter H_0 und stetigem F

$$\begin{aligned} P_{0,F}[\sum_{i=1}^{n} \mathrm{Rang}(X_i) = k] &= \frac{|\{\pi = (\pi_1, \ldots, \pi_{n+m}); \quad \sum_{i=1}^{n} \pi_i = k\}|}{(n+m)!} \\ &= \frac{|\{1 \le k_1 < k_2 < \cdots < k_n \le n+m; \quad \sum_{i=1}^{n} k_i = k\}|}{\binom{n+m}{n}}. \end{aligned}$$

Damit kann die Verteilung von T (im Prinzip) durch Abzählen bestimmt werden. Für n und m gross muss man sich aber auf einen zentralen Grenzwertsatz abstützen, um $c(\alpha)$ wenigstens näherungsweise berechnen zu können. Für die Berechnung der Macht ist man noch stärker auf asymptotische Näherungen angewiesen. Diese Näherungen sind ähnlich wie die Näherungen, die wir in Kapitel 6 näher erläutern werden. Wir erwähnen hier schon einige Folgerungen aus diesen Näherungen:

- Bei Normalverteilung braucht der Wilcoxon-Test etwa 5 % mehr Beobachtungen als der t-Test für das gleiche Niveau und die gleiche Macht.
- Bei beliebigem F braucht der Wilcoxon-Test im schlimmsten Fall 16 % mehr Beobachtungen als der t-Test für das gleiche Niveau und die gleiche Macht.
- Umgekehrt kann der zusätzliche Anteil an Beobachtungen, den der t-Test braucht, um das gleiche Niveau und die gleiche Macht wie der Wilcoxon-Test zu erreichen, beliebig gross sein.

Aus diesen Gründen zieht man den Wilcoxon-Test meist dem t-Test vor.

1.2.2 Randomisierung

Bei der Planung dieses Versuches ist es entscheidend, wie man die Einteilung in die beiden Gruppen vornimmt. Die Versuchspersonen in den beiden Gruppen sollten nach allen andern Kriterien wie Intelligenz, psychische Verfassung etc. übereinstimmen. Sonst weiss man nämlich im Fall eines signifikanten Resultates nicht, ob die negative Beeinflussung oder ein anderer Gesichtspunkt die Ursache war (sogenanntes "confounding"). Man wird natürlich bei der Auswahl der Versuchsteilnehmer versuchen, eine möglichst grosse Homogenität zu erhalten, aber man ist nie sicher, ob man alles gedacht hat. Die beste Art, solche Probleme zu vermeiden, ist aber die sogenannte Randomisierung, bei der man die Einteilung zufällig macht, d.h. jede der $\binom{n+m}{n}$ Einteilungen in eine Kontrollgruppe der Grösse n und eine Behandlungsgruppe der Grösse m ist gleich wahrscheinlich.

Diese Randomisierung erlaubt auch eine modifizierte Auswertung der Resultate, die nicht mehr von irgendwelchen Verteilungsannahmen abhängt. Bisher haben wir die Ergebnisse der Versuchspersonen als Realisierungen von Zufallsvariablen angeschaut. Das kann man auf zwei Arten begründen. Erstens hängen die Ergebnisse von zufälligen Faktoren wie Wetter, Tagesform und Interesse an den gerade gestellten Aufgaben ab. Zweitens kann man die Versuchspersonen als Zufallsauswahl aus einer grösseren Population anschauen. Gerade das letztere ist aber oft heikel, denn Freiwillige unterscheiden sich manchmal stark von andern Personen. Auch die Unabhängigkeit und die gleichen Verteilungen sind diskutabel.

Für das Folgende verschärfen wir die Nullhypothese noch etwas: Ob eine Person der Kontroll- oder der Behandlungsgruppe zugeteilt wird, spielt keine Rolle; die Differenz der beiden Testergebnisse ist stets die gleiche. Wenn diese Nullhypothese stimmt, wissen wir, was wir bei einer andern Zuordnung erhalten hätten. Wenn ausserdem randomisiert wurde, dann können wir die Verteilung einer Teststatistik $T = T(x_1, \ldots x_n; y_1, \ldots y_m)$ unter den möglichen Zuordnungen (bei festen Ergebnissen $z_1 = x_1, \ldots, z_n = x_n, z_{n+1} = y_1, \ldots, z_{n+m} = y_m$) berechnen als

$$P_0[T = t] = \frac{|\{A \subset \{1, \ldots, n+m\};\ |A| = n,\ T((z_i)_{i \in A}; (z_i)_{i \notin A}) = t\}|}{\binom{n+m}{m}}.$$

Damit können wir sofort durch Abzählen einen Test konstruieren, indem wir verwerfen, falls der beobachtete Wert von T im Schwanz dieser Verteilung liegt. Dann haben wir nämlich entweder durch Zufall eine aussergewöhnliche Einteilung erwischt, oder unsere Nullhypothese ist falsch.

Wenn T nur von den Rängen abhängt, dann erhalten wir wegen Lemma 1.2 mit der Randomisierung genau den gleichen Test wie mit dem Modell, wo die X_i und Y_j zufällig sind.

1.3 Konstruktion von Schätzern, Vertrauensintervallen und Tests

Zum Abschluss der Einführung stellen wir noch die gebräuchlichsten Methoden zur Konstruktion von Schätzern vor. Die Likelihood-Methode liefert ebenfalls Vertrauensintervalle und Tests.

1.3.1 Momentenmethode

Seien $X_1, X_2, \ldots X_n$ i.i.d. , reellwertig mit Verteilung $F_\theta, \theta \in \mathbb{R}^p$. Wenn $E_\theta[|X_i|^r] < \infty$, dann ist wegen des Gesetzes der grossen Zahlen

$$\frac{1}{n}\sum_{i=1}^{n} X_i^r$$

ein „natürlicher" Schätzer für $E_\theta[X_i^r]$. Falls ausserdem noch die Umkehrung der Abbildung

$$g : \Theta \longrightarrow \mathbb{R}^p \ , \ g(\theta) = (E_\theta[X_i], E_\theta[X_i^2], \ldots, E_\theta[X_i^p])$$

existiert und stetig ist, können wir aufgrund der geschätzten Momente auch den Parameter θ schätzen:

$$\hat{\theta} = g^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i, \ldots, \frac{1}{n}\sum_{i=1}^{n} X_i^p\right)$$

Dieser Schätzer heisst der *Momentenschätzer* für θ.

Beispiel 1.1. *Sei X negativ-Binomial(γ, p)-verteilt. Die ersten beiden Momente sind in Kapitel A.3 gegeben, und man sieht sofort dass die Voraussetzungen erfüllt sind. Man hat*

$$p = \frac{E[X]}{\mathrm{Var}[X]}, \quad \gamma = \frac{E[X]^2}{\mathrm{Var}[X] - E[X]}$$

Im allgemeinen ist der Momentenschätzer leicht zu rechnen, er kann dafür aber unter Umständen auch ziemlich ungenau sein.

1.3.2 Das Einsetzprinzip: Schätzer als Funktionale

Die Momentenmethode beruht auf zwei Überlegungen: Identifikation der Parameter über die Momente und Ersetzen der theoretischen durch die empirischen Momente. Wir können die Parameter aber auch über andere Kenngrössen identifizieren. Bei der $\mathcal{N}(\mu, \sigma^2)$-Verteilung können wir zum Beispiel auch μ als den Median interpretieren und σ als

MAD/$\Phi^{-1}(0.75) \approx 1.48\cdot$MAD, wobei MAD den Median von $|X-\mu|$ bezeichnet. Wenn wir nun den theoretischen Median durch den Stichprobenmedian und den theoretischen MAD durch den Stichproben-MAD schätzen (diese Grössen wurden im Unterabschnitt 1.1.2 eingeführt), so erhalten wir eine Schätzung von μ und σ, die sich von der Momentenmethode unterscheidet.

Dieses Vorgehen können wir wie folgt verallgemeinern: Seien $X_1, X_2, \ldots X_n$ i.i.d. mit Werten in $\mathbb{X}$ und Verteilung $F_\theta, \theta \in \Theta$, so dass $F_\theta \neq F_{\theta'}$ für $\theta \neq \theta'$. Dann definiert $q : F_\theta \to \theta$ eine Abbildung vom Raum der Modellverteilungen in den Parameterraum Θ. Für das Einsetzprinzip brauchen wir nun eine Erweiterung Q von q auf eine Teilmenge $\mathcal{F}$ von Wahrscheinlichkeitsverteilungen, die alle diskreten Wahrscheinlichkeitsverteilungen enthält. Dann können wir nämlich θ schätzen, indem wir anstelle von F_θ die empirische Verteilung F_n einsetzen:

$$\hat{\theta} = Q(F_n).$$

Die **empirische Verteilung** F_n von n Beobachtungen $X_1, X_2, \ldots, X_n$ ist konzentriert auf diesen n Beobachtungen und gibt jeder das Gewicht $\frac{1}{n}$, d.h.

$$F_n = \frac{1}{n}\sum_{i=1}^{n} \Delta_{X_i},$$

wobei Δ_x die Punktmasse eins (das Dirac-Mass) in $x \in \mathbb{X}$ ist. Für jedes $A \subset \mathbb{X}$ ist also $F_n(A)$ die relative Häufigkeit der Beobachtungen, die in A liegen. Für $\mathbb{X} = \mathbb{R}$ ist die zu F_n gehörige Verteilungsfunktion (die wir wieder mit F_n bezeichnen) eine Treppenfunktion mit Stufen der Höhe $\frac{1}{n}$ an den Beobachtungen.

Das Integral einer Funktion u bezüglich F_n ist nichts anderes als ein arithmetisches Mittel

$$\int u(x) F_n(dx) = \frac{1}{n}\sum_{i=1}^{n} u(X_i).$$

Daher ist der Momentenschätzer ein Spezialfall des Einsetzprinzips, wenn wir setzen

$$Q(F) = g^{-1}\left(\int x dF(x), \ldots, \int x^p dF(x)\right).$$

Die Rechtfertigung des Einsetzprinzips beruht darauf, dass für grosses n die empirische Verteilung gegen die wahre Verteilung konvergiert. Dies werden wir im Abschnitt 6.2.4 noch genauer diskutieren.

1.3.3 Likelihoodmethoden

Für diese Methode müssen wir annehmen, dass Dichten bezüglich eines σ-endlichen Bezugsmasses μ auf $\mathbb{X}$ existieren:

$$P_\theta(dx) = p_\theta(x)\mu(dx)$$

In der Praxis sind die folgenden 2 Fälle am häufigsten:

- $\mathbb{X} \subset \mathbb{R}^n$ offen ; $\mu =$ Lebesguemass, $p_\theta(x) =$ übliche Dichte.
- $\mathbb{X}$ diskret ; $\mu =$ Zählmass, $p_\theta(x) = P_\theta[X = x]$.

Wir betrachten nun $p_\theta(x)$ nicht mehr als Funktion von x, sondern als Funktion von θ für festes x. Um die unterschiedliche Betrachtungsweise klar zu machen, führen wir dafür eine neue Bezeichnung ein, nämlich die *Likelihood-Funktion* $L_x(\theta) = p_\theta(x)$.

Der *Maximum-Likelihood-Schätzer* (MLE) ist nun definiert als

$$\hat{\theta} = \arg\max_\theta L_x(\theta) = \arg\max_\theta \log L_x(\theta).$$

Man schätzt also θ so, dass die beobachteten Werte möglichst wahrscheinlich werden.

Beispiel 1.2. *Sei*

$$X_i = \mu + \varepsilon_i \ , \ \varepsilon_i \ \ i.i.d. \ \sim f(x)dx.$$

Dann ist

$$\log L_x(\mu) = \sum_{i=1}^n \log f(x_i - \mu).$$

Speziell erhalten wir für

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2}x^2/\sigma^2)$$

als MLE das arithmetische Mittel $\hat{\theta} = \overline{X}$*, und für*

$$f(x) = \frac{1}{2\sigma} \exp(-|x|/\sigma)$$

den Median $\hat{\theta} = median(X_1, \ldots, X_n)$

Die effektive Berechnung von $\hat{\theta}$ kann aufwendig sein, und es können viele Komplikationen entstehen. So kann L unbeschränkt sein oder mehrere lokale Maxima haben. Wenn L differenzierbar ist, bestimmt man meist den MLE durch Ableiten und Nullsetzen, aber es gibt durchaus praktisch relevante Fälle, wo $\frac{d}{d\theta} \log L_x(\theta) = 0$ mehrere Lösungen hat. Trotzdem ist der MLE aber fast universell anwendbar. Wegen seiner guten asymptotische Eigenschaften (siehe Kapitel 6) ist er auch sehr beliebt.

Allgemein gibt die Likelihood-Funktion an, welche Parameterwerte mit den Daten verträglich sind. Wir werden z.B. im Kapitel 6 sehen, dass in i.i.d. Situationen mit $\theta \in \mathbb{R}^p$

$$I = \{\theta;\ 2(\log L_x(\theta) - \sup_\theta \log L_x(\theta)) \geq -\chi^2_{p,1-\alpha}\}$$

genähert einen Vertrauensbereich mit Niveau $1 - \alpha$ ergibt. Ein universeller Test für die Hypothese $\theta \in \Theta_0 \subset \Theta$ ist der verallgemeinerte *Likelihoodquotienten-Test*, der diese Hypothese verwirft, wenn die Teststatistik

$$T(x) = \sup_{\theta\in\Theta} L_x(\theta) \ / \ \sup_{\theta\in\Theta_0} L_x(\theta)$$

zu gross ist. Wenn $\Theta \subseteq \mathbb{R}^p$ offen, $\Theta_0 \subseteq \mathbb{R}^q$ offen ist, dann ist $2\log T(x)$ in „regulären“ Fällen genähert χ^2_{p-q}-verteilt falls $\theta \in \Theta_0$.

Kapitel 2

Entscheidungstheorie

2.1 Formalisierung eines allgemeinen statistischen Problems

Zur Beantwortung einer wissenschaftlichen Fragestellung oder als Grundlage für eine Entscheidung in der Politik oder in der Wirtschaft stützt man sich auf Daten ab. Wir nehmen hier an, dass bei diesen Daten auch der Zufall mitspielt. Man beobachtet also eine Realisierung einer Zufallsvariablen X mit Werten in $\mathbb{X}$ (häufig hat man $\mathbb{X} = \mathbb{R}^n$ für n Einzelbeobachtungen). Die Verteilung von X ist unbekannt, sie gehöre zu einer Familie $\mathcal{P} = \{P_\theta; \theta \in \Theta\}$. Der Parameter θ kann endlich- oder unendlich-dimensional sein. Er enthält alle Information, die wir für die Beantwortung der Fragestellung oder für unsere Entscheidung benötigen. Ausserdem enthält er oft noch eine sogenannten Störkomponente, die nicht direkt von Interesse ist.

In den Beispielen von Kapitel 1 hatten wir

- Bei der Schätzung der Lichtgeschwindigkeit
  $\Theta = \{(\mu, \sigma^2)\} = \mathbb{R} \times \mathbb{R}_+$ falls $X_i \sim \mathcal{N}(\mu, \sigma^2)$, oder
  $\Theta = \{(\mu, F); \mu \in \mathbb{R}, \int xF(dx) = 0\}$ falls ε_i beliebige Verteilung mit $E[\varepsilon_i] = 0$.

- Beim Vergleich zweier Stichproben
  $\Theta = \{(\mu_1, \mu_2, \sigma^2)\}$ falls $X_i \sim \mathcal{N}(\mu_1, \sigma^2), Y_i \sim \mathcal{N}(\mu_2, \sigma^2)$, oder
  $\Theta = \{(F, G); F, G,$ stetig $\}$.

Die Menge der möglichen Antworten oder Entscheidungen des Statistikers nennen wir $\mathbb{A}$ (Aktionsraum). Beispiele sind

- $\mathbb{A} = \mathbb{R}$ bei Schätzung der Lichtgeschwindigkeit

- $\mathbb{A} = \{0, 1\}$ beim Testen

- $\mathbb{A} = [0, 1]$ beim randomisierten Testen ($d(x) = p$ heisst "Verwirf mit Wahrscheinlichkeit p")

- $\mathbb{A} = \{$Intervalle auf $\mathbb{R}\}$ bei Vertrauensintervallen

- $\mathbb{A} = \{$Permutationen$\}$ beim Erstellen einer Rangliste

Aufgrund der Beobachtung x muss man sich für eine Aktion a entscheiden. Da x zufällig ist, soll man angeben, wie man bei jedem x entscheiden würde, d.h. ein statistisches Verfahren ist dann eine Abbildung

$$d : \mathbb{X} \longrightarrow \mathbb{A}.$$

Die Bewertung der verschiedenen möglichen Aktionen in Abhängigkeit vom wahren Parameter erfolgt durch eine *Verlustfunktion*

$$L : \Theta \times \mathbb{A} \longrightarrow \mathbb{R}_+.$$

$L(\theta, a)$ ist der Verlust bei Aktion a und wahrem Parameter θ. Falls möglich, sollte man die Verlustfunktion spezifisch dem behandelten Problem anpassen. Oft kann man aber die Konsequenzen einer Fehlentscheidung nicht genau quantifizieren. Man behilft sich dann mit einer konventionellen Verlustfunktion wie in den folgenden Beispielen.

Beispiel 2.1. *Schätzprobleme. Die zu schätzende Grösse $g(\theta)$ sei reell. Dann ist $\mathbb{A} = \mathbb{R}$. Als Verlust wählt man meist*

$$L(\theta, a) = w(\theta)\, |g(\theta) - a|^r$$

Mit $w(\theta)$ gewichtet man die verschiedenen Parameterwerte.

Beispiel 2.2. *Bei einem Testproblem mit der Nullhypothese $\theta \in \Theta_0$ ist $\mathbb{A} = \{0, 1\}$. Als Verlustfunktion wählt man meist*

$$L(\theta, a) = \begin{cases} 1 & \textit{falls} \quad \theta \in \Theta_0, \quad a = 1 \quad \textit{(Verlust bei Fehler 1. Art)} \\ c & \textit{falls} \quad \theta \notin \Theta_0, \quad a = 0 \quad \textit{(Verlust bei Fehler 2. Art)} \\ 0 & \textit{sonst} \end{cases}$$

Der Wert von c gibt also an, wie schlimm man einen Fehler 2. Art im Vergleich zu einem Fehler 1. Art einstuft. Bei dieser Wahl ist der Verlust also nicht abhängig davon, welches θ aus der Nullhypothese oder der Alternative vorliegt.

Der Verlust bei Anwendung eines Verfahrens d ist $L(\theta, d(x))$. Er hängt also davon ab, welche Beobachtung x gerade angefallen ist. Zur Bewertung eines Verfahrens betrachtet man daher den *mittleren* Verlust.

Definition 2.1. Risiko = *Erwarteter Verlust eines statistischen Verfahrens:*

$$R(\theta, d) = \int_{\mathbb{X}} L(\theta, d(x)) P_\theta(dx).$$

Gute Verfahren sollten kleines Risiko haben. Weil man aber mehr als ein mögliches θ hat, kann man zwei verschiedene Verfahren i.A. nicht vergleichen: Meist gibt es zwei Werte θ_1 und θ_2 derart dass

$$R(\theta_1, d_1) < R(\theta_1, d_2), \text{ aber } R(\theta_2, d_1) > R(\theta_2, d_2).$$

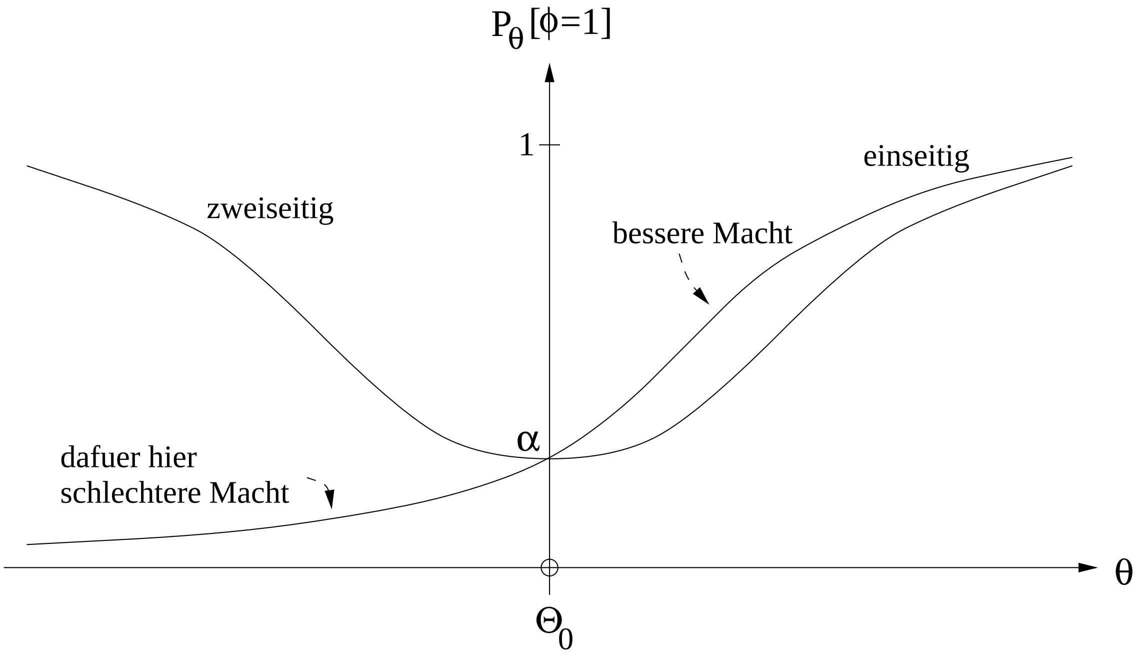
Das heisst, dass d_1 besser als d_2 ist, falls der wahre Parameter θ_1 ist, während bei wahrem Parameter θ_2 d_2 besser ist. Wir wissen aber ja gerade nicht, welches das wahre θ ist, also ist die Wahl des "besten" Verfahrens ein schwieriges Problem.

Beispiel 2.3. *Bei einem Schätzproblem mit $\mathbb{A} = \mathbb{R}$ und $L(\theta, a) = |g(\theta) - a|^r$ können wir insbesondere die konstante Entscheidung $d(x) = g(\theta_0)$ für alle x mit einem beliebigen, aber festen θ_0 betrachten. Dann ist $R(\theta, d) = |g(\theta) - g(\theta_0)|^r$, also ist das Risiko minimal für $\theta = \theta_0$, aber dafür sehr gross für andere θ's.*

Beispiel 2.4. *Beim Testproblem von oben ist*

$$R(\theta, d) = \begin{cases} P_\theta[H_0 \text{ verwerfen }] & \text{falls } \theta \in \Theta_0 \\ cP_\theta[H_0 \text{ beibehalten }] & \text{falls } \theta \notin \Theta_0 \end{cases}$$

d.h. das Risiko steht in direkter Beziehung zu den Irrtumswahrscheinlichkeiten. Offensichtlich kann man die Macht vergrössern auf Kosten des Niveaus, oder bei zusammengesetzten Alternativen kann man die Macht an einer bestimmten Alternative vergrössern auf Kosten der Macht bei einer andern Alternative. Das passiert z.B. wenn man einen zweiseitigen Test durch einen einseitigen ersetzt, vgl. die Figur.



Es besteht also ein Konflikt zwischen dem Risiko bei verschiedenen Parameterwerten. Verschiedene Lösungen dieses Konflikts führen zu verschiedenen Optimalitätsbegriffen. Wir behandeln die drei wichtigsten.

2.1.1 Zulässigkeit

Man begnügt sich damit, die offensichtlich schlechten Verfahren zu eliminieren. Das sind diejenigen, die man an einem Ort verbessern kann ohne Verschlechterung an einem andern Ort.

Definition 2.2. *Ein Verfahren d' ist strikt besser als ein anderes Verfahren d falls*

$$R(\theta, d') \leq R(\theta, d) \quad \forall \theta$$

und

$$R(\theta, d') < R(\theta, d) \text{ für mindestens ein } \theta.$$

Ein Verfahren d heisst unzulässig, *falls ein strikt besseres Verfahren d' existiert. Andernfalls heisst d* zulässig.

Beispiel 2.5. *Sei $\mathbb{A} = \mathbb{R}, L(\theta, a) = |g(\theta) - a|^r$. Das konstante Verfahren $d(x) = g(\theta_0)\ \forall x$ ist zulässig, falls P_θ absolut stetig ist bezüglich P_{θ_0}, $\forall \theta \in \Theta$. Denn aus $R(\theta_0, d') \leq R(\theta_0, d)$ folgt $R(\theta_0, d') = 0$, also $d'(x) = g(\theta_0)$ f.s. bezüglich P_{θ_0}. Nach Voraussetzung ist $P_\theta[B] = 0$ für jedes B mit $P_{\theta_0}[B] = 0$. Also ist $d'(x) = g(\theta_0)$ f.s. bezüglich jedem P_θ, d.h. $d' = d$.*

Beispiel 2.6. *Wir betrachten randomisierte Tests bei einfacher Hypothese und Alternative, d.h. $\Theta = \{\theta_0, \theta_1\}$ und $\mathbb{A} = [0, 1]$. Als Verlust wählen wir $L(\theta_i, a) = |a - i|$. Aus dem Neyman Pearson Lemma folgt, dass alle Likelihoodquotiententests zulässig sind, sofern die Macht kleiner als 1 ist. Wenn die Macht 1 ist, muss das Niveau minimal sein, sonst ist der Test unzulässig.*

Es gibt also im allgemeinen viele zulässige Verfahren, darunter auch solche, die trotzdem unbrauchbar sind. Umgekehrt werden wir später sehen, dass gewisse weit verbreitete Verfahren unzulässig sind. Trotz dieser Vorbehalte ist Zulässigkeit ein nützlicher Begriff. Sehr viel Arbeit wurde in die Entwicklung von Kriterien gesteckt, mit denen man entscheiden kann, ob ein Verfahren zulässig ist.

2.1.2 Minimax-Verfahren

Man versucht, im schlimmsten Fall noch möglichst gut zu sein (pessimistischer Standpunkt), d.h. man minimiert das maximale Risiko.

Definition 2.3. *d heisst* minimax *falls für alle d' gilt:*

$$\sup_{\theta \in \Theta} R(\theta, d) \leq \sup_{\theta \in \Theta} R(\theta, d')$$

Beispiel 2.7. *(Fortsetzung) Randomisierte Tests bei einfacher Hypothese und Alternative. Beim Verlust $L(\theta_i, a) = |a - i|$ ist derjenige Likelihoodquotiententest minimax, bei dem $R(\theta_0, d) = R(\theta_1, d)$ (d.h. das Niveau ist gleich eins minus der Macht). Das beweist man indirekt: Wäre d' ein anderer Test mit kleinerem maximalen Risiko, dann müsste sowohl $R(\theta_0, d') < R(\theta_0, d)$ als auch $R(\theta_1, d') < R(\theta_1, d)$ gelten, und dies ist ein Widerspruch zum Neyman-Pearson Lemma.*

2.1.3 Bayes-Verfahren

Man führt eine Gewichtung der Risiken für verschiedene $\theta's$ ein und erhält so eine Zahl anstelle einer Funktion, die man minimieren kann.

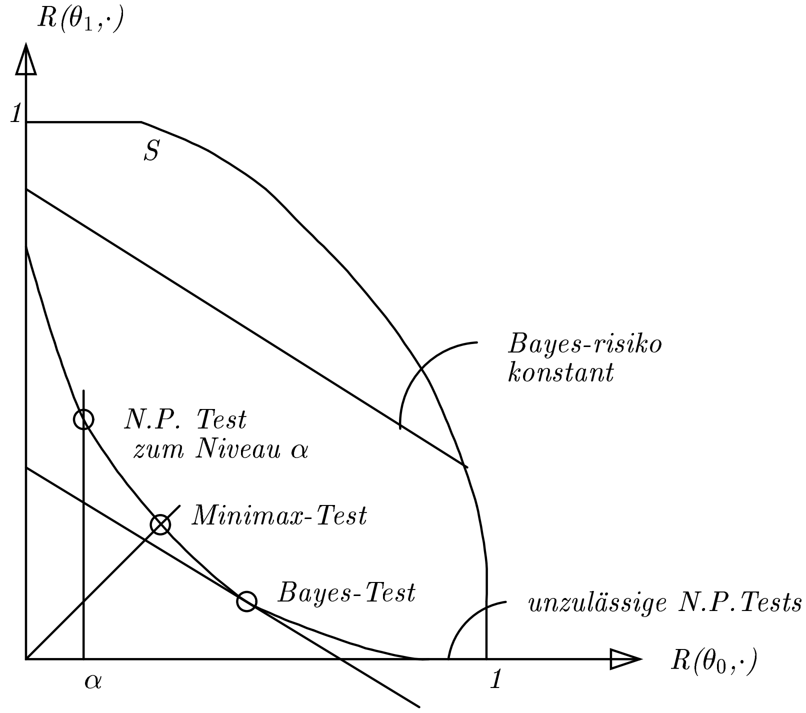
Definition 2.4. *Sei α eine Wahrscheinlichkeitsverteilung auf Θ. Dann heisst*

$$r(\alpha, d) = \int_\Theta R(\theta, d)\alpha(d\theta)$$

das Bayesrisiko *(bezüglich α). d heisst* Bayes *bezüglich α, falls für alle d'*

$$r(\alpha, d) \leq r(\alpha, d').$$

Beispiel 2.8. *(Fortsetzung) Randomisierte Tests bei einfacher Hypothese und Alternative. In diesem Fall ist $\alpha = (p, 1-p)$ und daher $r(\alpha, d) = pR(\theta_0, d) + (1-p)R(\theta_1, d)$. Also ist das Bayesrisiko konstant für alle d's, deren Risikovektor $(R(\theta_0, d), R(\theta_1, d))$ auf einer Geraden mit Steigung $-p/(1-p)$ liegt, und das Bayesrisiko ist proportional zum Achsenabschnitt. Daher erhält man den Bayes-Test, indem man eine Tangente mit Steigung $-p/(1-p)$ an den unteren Rand der Menge S aller Risikovektoren legt (Man kann leicht nachprüfen, dass S konvex ist, so dass Tangenten definiert werden können). Das ganze wird in folgender Figur illustriert.*



2.2 Bayes-Methoden

Wir hatten die Wahrscheinlichkeitsverteilung α als eine Gewichtung der verschiedenen Parameterwerte eingeführt. In der Bayes'schen Statistik ist aber eine andere Interpretation üblich, die jedoch nicht von allen StatistikerInnen akzeptiert wird. Das Bayesrisiko ist ja ein Doppelintegral

$$r(\alpha, d) = \int_\Theta \int_{\mathbb{X}} L(\theta, d(x)) P_\theta(dx) \alpha(d\theta),$$

das wir auch interpretieren können als den Erwartungswert des Verlustes, wobei sowohl X als auch θ zufällig sind mit der gemeinsamen Verteilung $P_\theta(dx)\alpha(d\theta)$. α ist dann die Randverteilung von θ und $P_\theta(dx)$ die bedingte Verteilung von X gegeben θ. Man hat also ein zweistufiges Zufallsexperiment, bei dem zuerst θ gemäss α bestimmt wird und danach X gemäss $P_\theta(dx)$. Man nennt daher α auch die *a-priori Verteilung*.

Die bedingte Verteilung von θ gegeben $X = x$, heisst die *a-posteriori Verteilung*. Wenn $P_\theta(dx) = p_\theta(x)\mu(dx)$, dann folgt aus dem Satz von Bayes (vgl. Kapitel A.1.2)

$$\alpha(d\theta|x) = \frac{p_\theta(x)\alpha(d\theta)}{\int_\Theta p_{\theta'}(x)\alpha(d\theta')}.$$

Der Nenner ist nur eine Normierung. Daher kann man diese Formel auch in Worten formulieren: Die a-posteriori Verteilung ist proportional zum Produkt aus Likelihood und a-priori Verteilung.

Die Randverteilung von X in diesem zweistufigen Zufallsversuch nennen wir Q. Es gilt:

$$Q(dx) = \int_\Theta P_\theta(dx)\alpha(d\theta).$$

Wenn die Verteilungen $P_\theta(dx)$ Dichten haben, dann auch Q, und der Nenner in der Bayesformel ist gerade die Dichte von Q.

Warum ist die Interpretation von θ als Zufallsvariable umstritten ? Wenn wir den Wahrscheinlichkeitsbegriff frequentistisch interpretieren, können wir ihn nur verwenden, wenn Wiederholungen zumindest denkbar sind. In vielen Situationen können wir uns aber keine Wiederholungen, die zu andern Werten von θ führen, vorstellen (Beispiel Lichtgeschwindigkeit). Es gibt im Allgemeinen kein Zufallsexperiment, das den Wert θ festgelegt hat.

Wenn man Wahrscheinlichkeiten aber versteht als subjektive Gradmesser für den Glauben daran, dass gewisse Ereignisse eintreten, gibt es diese Schwierigkeit nicht. Wir können dann Aussagen "θ liegt in einem Bereich B" Wahrscheinlichkeiten zuordnen aufgrund unserer Vorinformationen und (subjektiven) Meinungen. Auf diese Weise legen wir die a-priori Verteilung α fest. Die a-posteriori Verteilung misst dann, wie gross unser Glauben für die gleichen Aussagen nach der Beobachtung von X ist.

Abgesehen von diesen mehr philosophischen Aspekten hat die a-posteriori Verteilung auch konkrete Anwendungen. Mit ihr kann man nämlich das Bayes-Verfahren durch punktweises Minimieren finden.

Satz 2.1. *Falls für fast alle x $d(x) = \arg\min_a E[L(\theta, a)|X = x]$ existiert, dann ist d Bayes bezüglich α.*

Beweis. Nach Voraussetzung gilt f.s. für jedes andere Verfahren d'

$$E[L(\theta, d(X))|X = x] \leq E[L(\theta, d'(X))|X = x]$$

also nach dem Satz vom iterierten Erwartungswert (Satz A.2)

$$r(\alpha, d) = E[L(\theta, d(X))] \leq E[L(\theta, d'(X))] = r(\alpha, d').$$

□

Beispiel 2.9. *Testen ohne Randomisierung. Dann ist $\mathbb{A} = \{0, 1\}$. Für*

$$L(\theta, a) = \begin{cases} 1 & \textit{falls } \theta \in \Theta_0, \quad a = 1 \\ c & \textit{falls } \theta \notin \Theta_0, \quad a = 0 \\ 0 & \textit{sonst} \end{cases}$$

ist

$$E[L(\theta, a)|X = x] = aP[\Theta_0|X = x] + (1 - a)cP[\Theta_0^c|X = x].$$

Also ist der Bayestest gegeben durch

$$d(x) = \begin{cases} 1 & \textit{falls} \quad P[\Theta_0|X = x] < cP[\Theta_0^c|X = x] \\ 0 & \textit{falls} \quad P[\Theta_0|X = x] > cP[\Theta_0^c|X = x] \end{cases}$$

Man entscheidet also je nach der a posteriori Wahrscheinlichkeit der Nullhypothese.

Beim Schätzen eines reellwertigen Parameters ist $\mathbb{A} = \Theta = \mathbb{R}$. Also folgt aus Satz A.8, dass für

$$L(\theta, a) = (\theta - a)^2$$

der Bayesschätzer gleich

$$d(x) = E[\theta | X = x]$$

ist, während für

$$L(\theta, a) = |\theta - a|$$

der Bayesschätzer der Median der a-posteriori-Verteilung ist. Ferner ist für

$$L(\theta, a) = \mathbf{1}_{[|\theta - a| > c]}$$

der Bayesschätzer gleich

$$d(x) = \arg\max_a P[a - c \leq \theta \leq a + c | X = x].$$

Insbesondere erhält man für $\alpha(d\theta) = \alpha(\theta)d\theta$ unter Stetigkeitsbedingungen im Grenzwert $c \to 0$ als Bayes-schätzer

$$\arg\max_\theta p_\theta(x)\alpha(\theta).$$

Für $\alpha(\theta) = const.$ erhält man den MLE, aber dieses α ist i.A. keine Wahrscheinlichkeitsdichte mehr.

Beispiel 2.10. *Sei $X \sim$ Poisson(λ) und $\alpha =$ Gamma(γ, η) (die Gamma-Verteilung ist im Kapitel A.3 besprochen). Dann ist*

$$\alpha(\lambda | x) = \frac{\lambda^{\gamma - 1} e^{-\eta\lambda} e^{-\lambda} \lambda^x}{Z(x)}$$

wobei $Z(x)$ eine von x abhängige Normierung bedeutet. Also ist die a-posteriori-Verteilung wieder eine Gamma-Verteilung, aber mit neuen Parametern $\gamma' = \gamma + x, \eta' = \eta + 1$. Bei quadratischem Verlust ist der Bayesschätzer daher

$$d(x) = E[\lambda | X = x] = \frac{\gamma + x}{\eta + 1} = \frac{\gamma}{\eta} \frac{\eta}{\eta + 1} + x \frac{1}{\eta + 1}.$$

Dies ist nichts anderes als eine konvexe Kombination des Erwartungswertes der a-priori Verteilung und des MLE $\hat{\lambda} = x$. Man sieht hier auch, dass der Bayesschätzer tatsächlich von der Verlustfunktion abhängt, denn

$$\arg\max_\lambda \alpha(\lambda | x) = \frac{\gamma + x - 1}{\eta + 1} \neq E[\lambda | X = x].$$

Beispiel 2.11. *Rekonstruktion verrauschter Bilder. In diesem Beispiel bezeichnet der Parameter ein wahres, unbekanntes Schwarz-Weiss-Bild auf einem quadratischen Raster*

$$\theta_{ij} \in \{-1, +1\} \quad 1 \leq i \leq n, \quad 1 \leq j \leq n.$$

Das beobachtete Bild (X_{ij} ist verrauscht, d.h. mit Wahrscheinlichkeit p bleibt der Wert in jedem Rasterpunkt (Pixel) fest und mit Wahrscheinlichkeit $q = 1 - p$ wird er umgekehrt, unabhängig von andern Rasterpunkten. In Formeln heisst das

$$P[X_{ij} = \theta_{ij}] = p = \frac{p}{p + q} = \frac{\sqrt{p/q}}{\sqrt{p/q} + \sqrt{q/p}} = \frac{e^\gamma}{e^\gamma + e^{-\gamma}}$$

wobei $\gamma = \frac{1}{2}\log(p/q)$, *und analog*

$$P[X_{ij} = -\theta_{ij}] = q = \frac{e^{-\gamma}}{e^{\gamma} + e^{-\gamma}}.$$

Wir wollen (θ_{ij}) *aufgrund von* (X_{ij}) *schätzen, während wir* p *als bekannt voraussetzen. Zuerst bestimmen wir den MLE. Aus der Unabhängigkeit der Rasterpunkte und obigen Formel folgt, dass*

$$p_\theta(x) = \prod_{ij} p_{\theta_{ij}}(x_{ij}) = \exp(\gamma \sum_{i,j} x_{ij}\theta_{ij})(e^{\gamma} + e^{-\gamma})^{-n^2}.$$

(Man beachte, dass $\theta_{ij}x_{ij}$ *gleich eins ist bei Übereinstimmung und sonst gleich minus eins). Daher ist der MLE* $\hat{\theta}_{ij} = x_{ij}$ *falls* $p > q$, *bzw.* $\hat{\theta}_{ij} = -x_{ij}$ *falls* $p < q$. *Dieses Verhalten ist typisch für den MLE, sobald die Anzahl Parameter vergleichbar ist mit der Anzahl Beobachtungen.*

Etwas interessanter ist der Bayesschätzer, wobei wir in die a-priori Verteilung unser Wissen stecken wollen, dass Werte in benachbarten Rasterpunkten üblicherweise übereinstimmen. Dies leistet

$$\alpha(\theta) = \exp(\beta_1 \sum \theta_{ij}\theta_{i+1,j} + \beta_2 \sum \theta_{ij}\theta_{i,j+1})\frac{1}{Z(\beta)}.$$

Die Parameter β_1 *und* β_2 *regeln, wie ausgeprägt die Bevorzugung gleicher Werte bei horizontalen und vertikalen Nachbarn ist.* $Z(\beta)$ *ist eine Normierungskonstante. Die a-posteriori Verteilung ist dann*

$$\alpha(\theta|x) = \exp(\gamma \sum x_{ij}\theta_{ij} + \beta_1 \sum \theta_{ij}\theta_{i+1,j} + \beta_2 \sum \theta_{ij}\theta_{i,j+1})\frac{1}{Z(x,\beta)}.$$

Für den Bayesschätzer müssen wir noch eine Verlustfunktion wählen. Falls

$$L(\theta, a) = 1 - \prod_{ij} \mathbf{1}_{[a_{ij}=\theta_{ij}]}$$

(d.h. Verlust 1 ausser wenn alles richtig), dann

$$\hat{\theta} = \arg\max_{\theta}(\gamma \sum x_{ij}\theta_{ij} + \beta_1 \sum \theta_{ij}\theta_{i+1,j} + \beta_2 \sum \theta_{ij}\theta_{i,j+1})$$

Die Berechnung dieses $\hat{\theta}$ *ist schwierig, da i.A. mehr als ein lokales Maximum existiert. Falls*

$$L(\theta, a) = \sum_{i,j} \mathbf{1}_{[\theta_{ij} \neq a_{ij}]}$$

(Anzahl fehlerhafter Rasterpunkte), dann ist

$$\hat{\theta}_{ij} = \begin{cases} +1 & \textit{falls } P[\theta_{ij} = 1|x] > \frac{1}{2}. \\ -1 & \textit{falls } P[\theta_{ij} = 1|x] < \frac{1}{2}. \end{cases}$$

Diese bedingten Wahrscheinlichkeiten können nicht explizit berechnet werden, sondern müssen mit Simulation bestimmt werden, was ebenfalls nicht leicht ist. Beide Bayesschätzer geben aber recht gute Resultate, obwohl die erste Verlustfunktion unsinnig erscheint.

Bei der Diskussion der Bayesmethoden ergeben sich die folgenden Gesichtspunkte.

1. Der Vorteil der Bayes-Verfahren ist ihre relativ explizite Form. Ihr Nachteil ist die Abhängigkeit von der a-priori-Verteilung α. Für einen extremen Subjektivisten ist das kein Problem. Jede Person hat ihre eigene a-priori Verteilung und damit unterscheiden sich auch die a-posteriori Verteilungen. Es gibt auch Versuche, objektive, nichtinformative a-priori-Verteilungen zu konstruieren. Naheliegend wäre die Gleichverteilung, es ergeben sich aber Probleme (Θ unbeschränkt, Möglichkeit von Parametertransformationen).

2. Auch wenn man θ nicht als Zufallsvariable interpretiert, kann man Bayes-verfahren benutzen (Gewichtung des Risikos). Der Satz 2.1 ist dann einfach ein Trick zur Bestimmung des Bayesverfahrens.

3. Bei Konfidenzintervallen und Tests unterscheiden sich die Bayes'schen und frequentistischen Konzepte. Ein Bayes'sches Konfidenzintervall $I(x)$ ist z.B. gegeben durch $P[\theta \in I(x)|X = x] = \alpha$, d.h. das Intervall ist fest und der Parameter ist zufällig !

In gewissen Situationen hat man Daten aus früheren Versuchen zur Hand, aus denen man eine a-priori Verteilung schätzen kann (sogenannte *empirische Bayesverfahren*). Insgesamt seien Daten aus n Versuchen zur Verfügung, und der Parameter variiere von Versuch zu Versuch, ist also wirklich zufällig im frequentistischen Sinn. Der Einfachheit halber nehmen wir an, dass

$$\theta_1, \ldots, \theta_n \text{ i.i.d. } \sim \alpha_\eta(d\theta),$$

wobei η unbekannt ist, und gegeben $\theta_1, \ldots, \theta_n$ seien die X_i unabhängig voneinander mit

$$X_i \sim P_{\theta_i}(dx).$$

Sei $T(X_i, \eta) =$ der Bayes-schätzer von θ_i bei a-priori Verteilung α_η. Die Idee besteht nun darin, den unbekannten Parameter η der a-priori Verteilung aus der Randverteilung Q_η der X_i zu schätzen. Es gilt

$$Q_\eta(dx_1, \ldots dx_n) = \int_\Theta \cdots \int_\Theta \prod P_{\theta_i}(dx_i)\alpha_\eta(d\theta_i)$$

Die X_i sind also bezüglich der Randverteilung i.i.d., während sie *bedingt auf* die θ_i's unabhängig, aber nicht identisch verteilt sind. Wenn die Verteilungen $P_\theta(dx)$ absolut stetig sind bezüglich einem Bezugsmass, dann gilt das auch für Q_η. Zur Bestimmung des MLE $\hat{\eta}$ muss man das Integral

$$\int p_\theta(x)\alpha_\eta(d\theta)$$

berechnen. Der empirische Bayes-Schätzer ist dann $\hat{\theta}_i = T(X_i, \hat{\eta})$. Oft werden die Formeln einfacher, wenn man anstelle des MLE den Momentenschätzer nimmt.

Beispiel 2.12. *Sei $X_i \sim Poisson(\lambda_i)$ und $\alpha_{\gamma,\eta} = Gamma(\gamma, \eta)$. Die X_i können z.B. die Anzahl Unfälle des i-ten Kunden einer Versicherungsgesellschaft sein, und α beschreibt die Variation der durchschnittlichen Unfallhäufigkeit über die Gesamtheit der Kunden. Die Randverteilung der X_i ist negativ binomial $(\gamma, \eta/(\eta + 1))$ (siehe A.3). Da es für den MLE keine geschlossene Formel gibt, nehmen wir den Momentenschätzer. Dieser ist gemäss 1.3 gegeben als*

$$\hat{\gamma} = \frac{\overline{X}^2}{S_n^2 - \overline{X}}, \quad \hat{\eta} = \frac{\overline{X}}{S_n^2 - \overline{X}},$$

wobei $\overline{X} = \sum X_i/n$ *und* $S_n^2 = \sum(X_i - \overline{X})^2/(n-1)$. *Damit ist der empirische Bayes-schätzer für* λ_i

$$\hat{\lambda}_i = \overline{X}\frac{\overline{X}}{S_n^2} + X_i(1 - \frac{\overline{X}}{S_n^2}).$$

Der empirische Bayes-schätzer $\hat{\lambda}_i$ *verwendet also zur Schätzung des Erwartungswertes* λ_i *von* X_i *auch die Grössen* X_j , $j \neq i$, *obwohl* X_i *und* X_j *unabhängig sind. Auf den ersten Blick erscheint das paradox. Wir kommen im Abschnitt 5.3 darauf zurück.*

2.3 Suffizienz

Bei den Lichtgeschwindigkeitsmessungen von Kapitel 1 waren alle besprochenen Schätzer symmetrisch in den X_i, d.h. die Reihenfolge der Beobachtungen scheint irrelevant zu sein. Das ist auch intuitiv plausibel, denn wenn alle Beobachtungen i.i.d. sind, spielt die Nummerierung keine Rolle. Wir wollen jetzt das dahinter stehende allgemeine Konzept herausarbeiten.

Es ist klar, dass im allgemeinen Information verloren geht, wenn wir die Beobachtungen mit einer nichtumkehrbaren Funktion $S : \mathbb{X} \to \mathbb{Y}$ transformieren. Wie wir soeben gesehen haben, gibt es aber Situationen, wo $S(X)$ gleich viel Information enthält wie die ursprüngliche Beobachtung X. Dann nennen wir S suffizient, d.h. (informations-) ausschöpfend. Wie können wir aber feststellen, ob Information verloren gegangen ist oder nicht ? Wir geben zunächst eine mathematische Definition und überlegen nachher, warum diese die Idee "Kein Informationsverlust" formalisiert.

Definition 2.5. *Sei* $P = \{P_\theta; \theta \in \Theta\}$ *eine Familie von Wahrscheinlichkeitsverteilungen auf* $\mathbb{X}$ *und* S *eine Statistik, d.h. eine Funktion auf dem Stichprobenraum* $S : \mathbb{X} \to \mathbb{Y}$. *Dann heisst* S suffizient *für* P, *falls die bedingte Verteilung* $P_\theta[X \in . \,|S(X) = s]$ *nicht von* θ *abhängt.*

Beispiel 2.13.

- *Seien* $X_1, \ldots, X_n$ *i.i.d.* $\sim$ *Poisson*(λ). *Dann ist* $S = X_1 + \cdots + X_n$ *suffizient, denn (vgl. Satz A.6)*

$$P_\lambda[X_1 = k_1, \ldots, X_n = k_n | S = k] = \frac{k!}{k_1! \ldots k_n!}\left(\frac{1}{n}\right)^k,$$

  was nicht von λ *abhängt.*

- *Seien* $X_1, \ldots, X_n$ *i.i.d.* $\sim F$ *mit* F *stetig, sonst beliebig (*Θ *ist also unendlich-dimensional). Dann sind die geordneten Beobachtungen* $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ *suffizient; denn die ursprünglichen Beobachtungen sind eine Permutation der geordneten, und aus Symmetriegründen hat jede mögliche Permutation die Wahrscheinlichkeit* $1/n!$ *(Davon haben wir schon im Lemma 1.2 Gebrauch gemacht.)*

- *Seien* X_1, X_2 *i.i.d. , gleichverteilt auf* $[0, \theta]$. *Aus Beispiel A.2 folgt, dass die Statistik* $S = \max(X_1, X_2)$ *suffizient ist. Dies gilt auch für mehr als 2 Variablen.*

Diejenigen, die sich in Wahrscheinlichkeitstheorie auskennen, wissen, dass bedingte Verteilungen verschiedene Pathologien haben können (nur f.s. Eindeutigkeit, i.A. nicht regulär). Obige Definition müsste man daher etwas präziser formulieren, nämlich es muss eine Funktion $r(A|s)$ geben, so dass erstens $r(.|s)$ eine Wahrscheinlichkeit ist für jedes s, zweitens

$r(A|.)$ messbar ist für jedes A und drittens $r(A|s)$ eine Version von $P_\theta[X \in A|S(X) = s]$ ist für alle θ.

Wenn wir nur den Wert einer suffizienten Statistik S kennen und S nicht umkehrbar ist, können wir den ursprünglich beobachteten Wert x nicht mehr rekonstruieren, wir wissen nur, dass x im Urbild $S^{-1}(s) = \{x; S(x) = s\}$. Der genaue Wert von x ist aber irrelevant, wenn wir etwas über den Parameter θ lernen wollen. Wir können nämlich einen Wert x^* aus dem Urbild $S^{-1}(s)$ zufällig mit der Verteilung $P_\theta[X \in . \,|S(X) = s]$ auswählen, weil wir dazu θ nicht zu kennen brauchen. Wenn wir dieses x^* anstelle von x verwenden, ändert sich zwar im Allgemeinen die Aktion: $d(x^*) \neq d(x)$, aber die *Verteilung* des Verfahrens $d(X)$ ändert sich nicht (weil X^* die gleiche Verteilung hat wie X). Also haben wir nichts Wesentliches verloren, wenn wir nur $S(X)$ kennen.

Bei der Verwendung von $d(x^*)$ ist die gewählte Aktion nicht nur abhängig von der Beobachtung, sondern auch noch von einem Zufallszahlengenerator, wir haben also eine sogenannte randomisierte Entscheidung. Bei der Berechnung des Risikos einer randomisierten Entscheidung integriert man sowohl über die Beobachtung als auch über die verwendeten Zufallszahlen.

Satz 2.2. *Sei S eine suffiziente Statistik für $\{P_\theta; \theta \in \Theta\}$. Dann existiert zu jedem Verfahren $d : \mathbb{X} \to \mathbb{A}$ ein randomisiertes Verfahren δ basierend auf $S(X)$ derart, dass d und δ das gleiche Risiko haben.*

Beweis. In Worten wurde schon alles gesagt. In Formeln setzen wir

$$\delta(s) = d(X^*)$$

wobei wir X^* mit einem Zufallsgenerator gemäss der Verteilung $P_\theta[X \in .|S = s]$ erzeugen. (Wegen der Suffizienz, brauchen wir dabei θ nicht zu kennen !). Dann gilt

$$R(\theta, \delta) = E_\theta[E[L(\theta, d(X^*))|S]] = E_\theta[E_\theta[L(\theta, d(X))|S]] = R(\theta, d)$$

(Die erste Gleichung ist die Definition des Risikos, die zweite benutzt die Definition der Verteilung von X^*, und die dritte ist der Satz vom iterierten Erwartungswert). □

Wenn der Aktionsraum $\mathbb{A}$ und die Verlustfunktion konvex sind (also z.B. bei Schätzproblemen), kommt man ohne Randomisierung aus. Anstatt ein x^* zufällig auszuwählen, kann man über die möglichen Entscheidungen $d(x^*)$ mitteln. Dann wird das Risiko sogar reduziert.

Satz 2.3 (Rao-Blackwell). *Sei $\mathbb{A} \subseteq \mathbb{R}^p$ konvex, sei $L(\theta, .)$ konvex für alle θ und sei S eine suffiziente Statistik für $\{P_\theta; \theta \in \Theta\}$. Ferner sei $d : \mathbb{X} \to \mathbb{A}$ ein Schätzer mit $R(\theta, d) < \infty$ und $E_\theta[\|d(X)\|] < \infty$. Wir setzen $d'(s) = E[d(X)|S(X) = s]$. Dann ist $R(\theta, d') \leq R(\theta, d)$ für alle θ. Wenn $L(\theta, .)$ strikt konvex ist, gilt sogar $R(\theta, d') < R(\theta, d)$, ausser falls $d = d'$ f.s. (bez. P_θ.)*

Beweis. Wegen der Jensen'schen Ungleichung gilt

$$E[L(\theta, d(X))|S(X) = s] \geq L(\theta, E[d(X)|S(X) = s]).$$

Jetzt bildet man den Erwartungswert auf beiden Seiten. Bei strikter Konvexität ist obige Ungleichung sogar strikt ausser falls $d(X) = E[d(X)|S(X)]$ f.s.. □

Bemerkung 2.1. *Die Suffizienz wird benötigt, damit d' nicht von θ abhängt. Andrenfalls wäre d' gar kein Schätzer.*

Beispiel 2.14. *Seien X_1, X_2 i.i.d. , gleichverteilt auf $[0, \theta]$ und $d(X) = X_1 + X_2$. Gemäss obigem Beispiel ist $S = \max(X_1, X_2)$ suffizient, und gegeben $S = s$ ist eine Variable gleich s und die andere ist uniform auf $[0, s]$. Also liefert das Rezept von Satz 2.2:*

$$\delta(\max(X_1, X_2)) = \max(X_1, X_2)(1 + U),$$

wobei U uniform auf $[0, 1]$ ist. Das Rezept von Satz 2.3 liefert hingegen:

$$d'(X) = E[X_1 + X_2 | \max(X_1, X_2)] = \max(X_1, X_2)(1 + E[U]) = \frac{3}{2} \max(X_1, X_2).$$

Man rechnet nach, dass bei quadratischem Verlust $R(\theta, d) = R(\theta, \delta) = \theta^2/6$ und $R(\theta, d') = \theta^2/8$.

Der folgende Satz erleichtert das Auffinden einer suffizienten Statistik.

Satz 2.4 (Faktorisierungskriterium von Neyman). *Sei P_θ eine Familie von Verteilungen mit Dichten bezüglich eines σ-endlichen Bezugsmasses μ : $P_\theta(dx) = p_\theta(x)\mu(dx)$. Eine Statistik S ist genau dann suffizient, wenn sich $p_\theta(x)$ schreiben lässt als*

$$p_\theta(x) = g_\theta(S(x))h(x).$$

Beweis. Wir betrachten den diskreten Fall. Sei also $\mathbb{X}$ abzählbar und μ das Zählmass. Es gilt

$$P_\theta[X = x | S = s] = \begin{cases} P_\theta[X = x]/P_\theta[S = s] & \text{falls } S(x) = s \\ 0 & \text{sonst} \end{cases}$$

Für die Richtung „$\Rightarrow$“ beachtet man, dass die linke Seite gemäss Voraussetzung nicht von θ abhängt. Also folgt die Behauptung, wenn wir setzen $g_\theta(s) = P_\theta[S(X) = s]$ und $h(x) = P[X = x | S(X) = S(x)]$.

Für die Umkehrung geht man aus von

$$P_\theta[S = s] = g_\theta(s) \sum_{x, S(x)=s} h(x).$$

Also gilt

$$P_\theta[X = x | S = S(x)] = \frac{h(x)}{\sum_{x', S(x')=S(x)} h(x')}$$

Der stetige Fall erfordert Masstheorie, siehe z.B. Lehmann, Testing ...., Kap. 2.6. □

Korollar 2.1. *Falls eine suffiziente Statistik existiert, dann hängen Likelihood- und Bayes-Verfahren nur von dieser ab.*

Beispiel 2.15. *Seien $X_1, X_2, \ldots X_n$ i.i.d. gleichverteilt auf $(0, \theta)$. Dann können wir schreiben*

$$p_\theta(x_1, \ldots, x_n) = \begin{cases} \theta^{-n} \text{ falls } 0 \le \min(x_i) \le \max(x_i) \le \theta \\ 0 \text{ sonst.} \end{cases}$$

Also ist nach obigem Kriterium $\max(X_i)$ suffizient.

Eine ganze Klasse von Beispielen erhalten wir mit folgendem Korollar.

Korollar 2.2. *Seien $X_1, \ldots, X_n$ i.i.d. $\sim g_\theta(x)h(x)\mu(dx)$ wobei $g_\theta(x)$ die Form*

$$\exp(\sum_{j=1}^{k} c_j(\theta)T_j(x) + d(\theta))$$

hat. Dann ist

$$S(x_1, \ldots, x_n) = (\sum_{i=1}^{n} T_j(x_i); j = 1, \ldots, k)$$

eine suffiziente Statistik.

Hier ist die Dimension der suffizienten Statistik also gleich für alle Stichprobenumfänge. Da Verteilungen dieser Form häufig vorkommen, führen wir einen Namen ein:

Definition 2.6. *Eine Familie von Wahrscheinlichkeitsverteilungen $P_\theta(dx) = g_\theta(x)h(x)\mu(dx)$ mit*

$$g_\theta(x) = \exp(\sum_{j=1}^{k} c_j(\theta)T_j(x) + d(\theta))$$

heisst eine k-dimensionale exponentielle Familie.

Sehr viele Standardmodelle sind exponentielle Familien, wie die folgende Aufzählung zeigt. Im diskreten Fall, wo die Dichten bezüglich des Zählmasses genommen werden, haben wir zum Beispiel

Poisson Wir schreiben

$$p_\lambda(x) = \exp(x\log(\lambda) - \lambda)\frac{1}{x!}.$$

Also ist $T(x) = x$, $c(\lambda) = \log(\lambda)$ und $d(\lambda) = -\lambda$.

Binomial Wir schreiben

$$p_\theta(x) = \exp\left(x\log\left(\frac{\theta}{1-\theta}\right) + n\log(1-\theta)\right)\binom{n}{x}.$$

Also ist $T(x) = x$, $c(\theta) = \log(\theta/(1-\theta))$ und $d(\theta) = n\log(1-\theta)$.

Negativ Binomial Es sei $\theta = p$, während γ bekannt ist. Dann ist

$$p_\theta(x) = \exp(x\log(1-\theta) + \gamma\log\theta)(-1)^x\binom{-\gamma}{x}.$$

Also ist $T(x) = x$, $c(\theta) = \log(1-\theta)$ und $d(\theta) = \gamma\log(\theta)$.

Im stetigen Fall, wo die Dichten bezüglich des Lebesguemasses genommen werden, haben wir zum Beispiel

Normal Es sei $\theta = (\mu, \sigma)$. Dann können wir schreiben

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

Also ist $T(x) = (x^2, x)$, $c(\theta) = (-\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2})$ und $d(\theta) = -\frac{\mu^2}{2\sigma^2} - \log\sigma$.

Gamma Es sei $\theta = (\gamma, \lambda)$. Dann gilt

$$p_\theta(x) = \exp(-x\lambda + (\gamma - 1)\log x)\frac{\lambda^\gamma}{\Gamma(\gamma)}\mathbf{1}_{[x>0]}.$$

Also ist $T(x) = (x, \log x)$, $c(\theta) = (-\lambda, \gamma - 1)$ und $d(\theta) = \log(\lambda^\gamma/\Gamma(\gamma))$.

Beta Es sei $\theta = (\gamma, \delta)$. Dann gilt

$$p_\theta(x) = \exp((\gamma - 1)\log(x) + (\delta - 1)\log(1 - x))\frac{\Gamma(\gamma + \delta)}{\Gamma(\gamma)\Gamma(\delta)}\mathbf{1}_{[0<x<1]}.$$

Also ist $T(x) = (\log(x), \log(1 - x))$, $c(\theta) = (\gamma - 1, \delta - 1)$ und $d(\theta) = \log(\Gamma(\gamma + \delta)/(\Gamma(\gamma)\Gamma(\delta)))$.

Bivariat normal Es sei $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$. Ich überlasse es als Übung, die Dichte

$$\begin{aligned} p_\theta(x, y) &= \frac{1}{2\pi\sigma_1\sigma_2(1 - \rho^2)^{1/2}} \cdot \\ &\exp\left(-\frac{1}{2(1 - \rho^2)}\left(\frac{(x - \mu_1)^2}{\sigma_1^2} + \frac{(y - \mu_2)^2}{\sigma_2^2} - \frac{2\rho(x - \mu_1)(y - \mu_2)}{\sigma_1\sigma_2}\right)\right) \end{aligned}$$

in die Form der exponentiellen Familie zu bringen und $T(x)$, $c(\theta)$ und $d(\theta)$ zu bestimmen.

Eine suffiziente Statistik ist natürlich nicht eindeutig. Wenn S suffizient ist und g umkehrbar, dann ist auch $S' = g(S)$ suffizient. Ausserdem wird man S minimal wählen, d.h. so dass keine weitere Datenreduktion mehr möglich ist. Wegen Satz 2.4 könnte man versuchen, eine *minimale suffiziente* Statistik S durch die Forderung

$$S(x) = S(x') \Leftrightarrow L_x(\theta) \propto L_{x'}(\theta)$$

zu definieren, wobei $\propto$ "proportional zu" bedeutet. („$\Rightarrow$“ wäre suffizient, „$\Leftarrow$“ wäre minimal).

Beispiel 2.16. *Sei (X_i) i.i.d mit Dichte $f_\theta(x) = \frac{1}{2}\exp(-|x - \theta|)$. Dann sind die Ordnungsgrössen $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ minimal suffizient nach obigem Kriterium, d.h. wir können nur die Reihenfolge ignorieren, brauchen aber alle Werte. Dies ist leicht einzusehen. Wenn*

$$\sum_i |x_{(i)} - \theta| = \sum_i \left|x'_{(i)} - \theta\right| + const.$$

(wobei const. von den $x_{(i)}$ und $x'_{(i)}$ abhängen kann, aber nicht von θ), dann muss $x_{(i)} = x'_{(i)}$ sein (betrachte z.B. die Stellen, an denen die beiden Seiten nicht differenzierbar sind).

Im Allgemeinen gibt es jedoch Probleme mit dieser Definition, da Dichten nicht eindeutig sind. Ein sauberer Weg führt über suffiziente σ-Algebren, wir verzichten aber darauf.

Die praktische Bedeutung der Suffizienz kann man wie folgt zusammenfassen

- Man kann Schätzer verbessern (Satz 2.3).
- Bei der Suche nach optimalen Verfahren kann man annehmen, dass diese nur von der suffizienten Statistik abhängen.
- Man soll aber beobachtete Daten nicht auf eine suffiziente Statistik reduzieren, um z. B. Speicherplatz zu sparen. Wenn man die Familie $\{P_\theta; \theta \in \Theta\}$ erweitert, braucht man die ursprünglichen Daten.

Kapitel 3

Unverfälschte (erwartungstreue) Verfahren

In Kapitel 2 hatten wir gesehen, dass es meist nicht möglich ist, das Risiko simultan für alle Parameter zu minimieren. Auswege waren die schwächeren Optimalitätsbegriffe Zulässigkeit, Minimax und Bayes, die wir im 5. Kapitel näher besprechen werden. In diesem und dem nächsten Kapitel untersuchen wir einen anderen Ausweg: Wir schränken uns auf Verfahren ein, die in einem gewissen Sinn alle Parameter gleich behandeln, und suchen dann Verfahren, die in dieser kleineren Klasse das Risiko simultan für alle Parameter minimieren.

3.1 Definitionen

Definition 3.1. *Ein Schätzer $T : \mathbb{X} \to \mathbb{R}$ heisst* erwartungstreu *für $g(\theta)$, falls*

$$E_\theta[T(X)] = g(\theta) \ \forall \theta.$$

Ein Schätzer T heisst erwartungstreu mit überall kleinster Varianz *(UMVU= uniformly minimum variance unbiased), falls er erwartungstreu ist und*

$$\mathrm{Var}_\theta[T(X)] \leq \mathrm{Var}_\theta[T'(X)] \ \forall \theta$$

für alle erwartungstreuen T'.

Zur Erinnerung: Für beliebige Schätzer gilt

$$E_\theta[(T(X) - g(\theta))^2] = \mathrm{Var}_\theta[T(X)] + (E_\theta[T(X)] - g(\theta))^2.$$

Also ist bei erwartungstreuen Schätzern und quadratischem Verlust das Risiko gleich der Varianz.

Auf den ersten Blick scheint Erwartungstreue ein vernünftiges Konzept zu sein, das z.B. alle konstanten Schätzer ausschliesst. Es hat jedoch die folgenden Schwachpunkte

- Es gibt nicht immer erwartungstreue Schätzer. Sei z.B. $X \sim \mathrm{Bin}(n, \theta)$ und $g(\theta) = \theta/(1-\theta)$. Da $E_\theta[T]$ ein Polynom in θ vom Grad $\leq n$ und $g(\theta)$ eine gebrochen rationale Funktion ist, können sie nicht auf einem Intervall übereinstimmen.

- UMVU-Schätzer können unsinnig sein. Wenn $X \sim \text{Poisson}(\theta)$ ist und $g(\theta) = \exp(-3\theta)$, dann ist $T(x) = (-2)^x$ der einzige erwartungstreue Schätzer (Beweis als Übung).
- UMVU-Schätzer können unzulässig sein (Ein Beispiel folgt am Ende von Kapitel 5.2).
- Erwartungstreue ist nicht invariant gegenüber Parametertransformationen: Wenn $T(x)$ erwartungstreu ist für θ, dann ist $g(T(x))$ i. A. nicht erwartungstreu für $g(\theta)$.

Bei Tests ist Erwartungstreue weniger problematisch als bei Schätzern. Wir verlangen, dass die Wahrscheinlichkeit, die Nullhypothese zu verwerfen, kleiner ist unter der Nullhypothese als unter einer Alternative.

Definition 3.2. *Ein Test* $\varphi : \mathbb{X} \to [0,1]$ *für* $H_0 : \theta \in \Theta_0$ *heisst* unverfälscht*, wenn*

$$E_\theta[\varphi] \leq E_{\theta'}[\varphi] \ \forall\theta \in \Theta_0, \forall\theta' \notin \Theta_0.$$

Ein unverfälschter Test φ *zum Niveau* α *für* $H_0 : \theta \in \Theta_0$ *heisst* gleichmässig mächtigst unverfälscht *(UMPU = uniformly most powerful unbiased) zum Niveau* α*, falls* $E_\theta[\varphi]$ *für alle* $\theta \notin \Theta_0$ *maximal ist unter allen unverfälschten Tests zum Niveau* α*.*

Man sieht leicht, dass jeder gleichmässig mächtigste (UMP) Test φ unverfälscht sein muss (man vergleicht ihn mit dem Test $\varphi' \equiv \sup_{\theta\in\Theta_0} E_\theta[\varphi]$). Daher ist UMPU eine schwächere Optimalität als UMP (es sind weniger Konkurrenten zugelassen) und man kann dafür eher erwarten, dass UMPU-Tests existieren.

3.2 UMVU-Schätzer mit Suffizienz und Vollständigkeit

Wenn S suffizient ist für $\{P_\theta; \theta \in \Theta\}$, kann man sich bei der Suche nach UMVU-Schätzern auf solche beschränken, die nur von S abhängen: Wenn T erwartungstreu ist für $g(\theta)$, dann ist wegen des Satzes von der iterierten Erwartung auch $T' = E[T|S]$ erwartungstreu, und T' hat kleinere Varianz als T (siehe Satz 2.3). Durch einen Glücksfall gibt es in vielen Situationen nur einen erwartungstreuen Schätzer, der nur von S abhängt. Er ist dann automatisch UMVU.

Definition 3.3. *Eine Statistik* $S : \mathbb{X} \to \mathbb{Y}$ *heisst* vollständig *für* $\{P_\theta; \theta \in \Theta\}$*, wenn für jede Funktion* $f : \mathbb{Y} \to \mathbb{R}$ *aus*

$$E_\theta[f(S)] = 0 \ \forall\theta$$

folgt, dass

$$f(S) = 0 \ P_\theta - f.s. \ \forall\theta.$$

Satz 3.1. *Wenn* S *vollständig ist für* $\{P_\theta; \theta \in \Theta\}$*, dann existiert höchstens ein erwartungstreuer Schätzer, der nur von* S *abhängt.*

Beweis. $E_\theta[T(S)] = E_\theta[T'(S)] = g(\theta) \ \forall\theta$ impliziert

$$E_\theta[T(S) - T'(S)] = 0 \ \forall\theta.$$

Nach Definition der Vollständigkeit ist also $T(S) = T'(S) P_\theta - \text{f.s.} \ \forall\theta$. □

Korollar 3.1 (Lehmann-Scheffé). *Sei S suffizient und vollständig für $\{P_\theta; \theta \in \Theta\}$ und sei $T(X)$ ein erwartungstreuer Schätzer von $g(\theta)$. Dann ist $E[T|S]$ UMVU.*

Beweis. Sei T' ein beliebiger erwartungstreuer Schätzer für $g(\theta)$. Wegen des Satzes von Rao-Blackwell (Satz 2.3) hat dann $T'' = E[T'|S]$ eine kleinere oder gleich grosse Varianz wie T', und wegen Satz 3.1 ist $T'' = E[T|S]$. □

Beispiel 3.1. *Seien $X_1, \dots, X_n$ i.i.d. $, \sim$ Poisson (λ) und $g(\lambda) = P_\lambda[X_i = 0] = e^{-\lambda}$. Der Schätzer $T(x_1, \dots, x_n) = \mathbf{1}_{[x_1=0]}$ ist erwartungstreu. Wir verbessern ihn mithilfe der suffizienten Statistik $S = \sum_{i=1}^n X_i$:*

$$T'(s) = E[T|S = s] = P[X_1 = 0|S = s] = (1 - \frac{1}{n})^s$$

(vgl. Kapitel 2.3). Wenn wir noch zeigen, dass S auch vollständig ist, dann folgt, dass $T' = (1 - \frac{1}{n})^{X_1 + \dots + X_n}$ UMVU ist. Sei

$$0 = E_\lambda[f(S)] = e^{-n\lambda} \sum_{k=0}^{\infty} f(k) \frac{(n\lambda)^k}{k!} \ \forall \lambda.$$

Die rechte Seite ist eine Potenzreihe in λ, also $\equiv 0$ nur falls $f(k) = 0 \ \forall k$.

Für exponentielle Familien gibt es ein Kriterium für Vollständigkeit:

Satz 3.2. *Sei $P_\theta(dx) = \exp(\sum_{j=1}^k c_j(\theta) T_j(x) + d(\theta)) h(x) \mu(dx)$. Dann ist $(T_1(x), \dots, T_k(x))$ vollständig, falls $\{(c_1(\theta), \dots, c_k(\theta)); \theta \in \Theta\}$ eine offene Kugel im $\mathbb{R}^k$ enthält (d.h. die suffiziente Statistik und der Parameter müssen die gleiche Dimension haben).*

Beweis. siehe Lehmann, Testing Statistical Hypotheses, Kap. 4.3. □

Beispiel 3.2. *Seien*

$$X_1, \dots, X_n \text{ i.i.d. } \sim \mathcal{N}(\mu_1, \sigma_1^2), \quad Y_1, \dots, Y_m \text{ i.i.d. } \sim \mathcal{N}(\mu_2, \sigma_2^2).$$

Ferner seien alle Variablen unabhängig. Dies ist eine exponentielle Familie mit $\theta = (\mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$, $T = (\sum X_i, \sum Y_i, \sum X_i^2, \sum Y_i^2)$ und $c_1 = \mu_1/\sigma_1^2, c_2 = \mu_2/\sigma_2^2, c_3 = -1/(2\sigma_1^2), c_4 = -1/(2\sigma_2^2)$. Offenbar ist T suffizient und vollständig, wenn $\theta \in \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R} \times \mathbb{R}_+$. Wenn man nun das Untermodell $\mu_1 = \mu_2$ betrachtet, bleibt T natürlich suffizient. Die Bedingung von Satz 3.2 ist aber nicht mehr erfüllt, und T ist auch nicht mehr vollständig, denn

$$E_\theta[\frac{1}{n} \sum X_i - \frac{1}{m} \sum Y_i] = 0 \ \forall \theta.$$

Man kann zeigen, dass es in diesem Fall keine Statistik gibt, die gleichzeitig suffizient und vollständig ist.

3.3 UMVU-Schätzer via Cramér-Rao-Ungleichung

Wir leiten hier eine untere Schranke für $\mathrm{Var}_\theta[T]$ her, die für alle erwartungstreuen Schätzer gilt. Falls dann ein Schätzer diese Schranke erreicht, ist er offensichtlich UMVU. Diese Schranke wird auch in der Asymptotik (siehe Kapitel 6) eine wichtige Rolle spielen.

Für eine beliebige Funktion $\psi : \Theta \times \mathbb{X} \to \mathbb{R}$ mit $0 < \mathrm{Var}_\theta[\psi(\theta, X)] < \infty$ besagt die Schwarz'sche Ungleichung, dass

$$\mathrm{Var}_\theta[T] \geq \frac{\mathrm{Cov}_\theta[T, \psi(\theta, X)]^2}{\mathrm{Var}_\theta[\psi(\theta, X)]} \tag{3.1}$$

Wenn man ψ so wählen kann, dass $\mathrm{Cov}_\theta[T, \psi(\theta, X)]$ unabhängig von T wird, hat man eine untere Schranke für $\mathrm{Var}_\theta[T]$. Wie wir sehen werden, ist dies der Fall für

$$\psi(\theta, x) = \frac{p_{\theta+\Delta}(x) - p_\theta(x)}{\Delta p_\theta(x)}$$

mit beliebigem festem $\Delta > 0$, sowie für die Funktion, die wir als Grenzwert für $\Delta \to 0$ erhalten:

$$\psi(\theta, x) = \frac{\frac{\partial}{\partial\theta} p_\theta(x)}{p_\theta(x)} = \frac{\partial}{\partial\theta} \log p_\theta(x).$$

Zuerst betrachten wir den Fall, wo Θ eindimensional ist.

Satz 3.3. *Sei $P_\theta(dx) = p_\theta(x)\mu(dx)$ mit $\theta \in \Theta$, Θ ein offenes Intervall, so dass $A = \{x | p_\theta(x) > 0\}$ unabhängig von θ ist. Ferner sei T erwartungstreu für $g(\theta)$. Dann gilt die* Chapman-Robbins Ungleichung

$$\mathrm{Var}_\theta[T] \geq \sup_\Delta \frac{(g(\theta + \Delta) - g(\theta))^2}{\mathrm{Var}_\theta[(p_{\theta+\Delta}(X) - p_\theta(X))/p_\theta(X)]}.$$

Falls zusätzlich

$$\frac{\partial}{\partial\theta} \log p_\theta(x) = \lim_{\Delta\to 0} \frac{p_{\theta+\Delta}(x) - p_\theta(x)}{\Delta p_\theta(x)}$$

existiert in $L_2(P_\theta)$, dann ist $g(\theta)$ differenzierbar, $E_\theta[\frac{\partial}{\partial\theta} \log p_\theta(X)] = 0$ und es gilt die Cramér-Rao Ungleichung

$$\mathrm{Var}_\theta[T] \geq \frac{g'(\theta)^2}{I(\theta)}.$$

Dabei ist

$$I(\theta) = E_\theta[(\frac{\partial}{\partial\theta} \log p_\theta(X))^2]$$

die sogenannte Fisher-Information.

Beweis. Da $P_\theta(A) = 0$, spielt es für die Berechnung von $E_\theta[\psi(\theta, X)]$ keine Rolle, wie wir $\psi(\theta, x)$ auf A definieren. Wir setzen daher ohne Beschränkung der Allgemeinheit $\psi(\theta, x) = 0$ für $x \in A$. Dann erhalten wir

$$E_\theta[\psi(\theta, X)] = \int_A \frac{p_{\theta+\Delta}(x) - p_\theta(x)}{\Delta p_\theta(x)} p_\theta(x)\mu(dx) = \frac{1}{\Delta} \int_A (p_{\theta+\Delta}(x) - p_\theta(x))\mu(dx) = 0.$$

(Im letzten Schritt wurde benutzt, dass A nicht von θ abhängt.) Ebenso zeigt man, dass

$$\mathrm{Cov}_\theta[T, \psi(\theta, X)] = E_\theta[T(X) \frac{p_{\theta+\Delta}(X) - p_\theta(X)}{\Delta p_\theta(X)}] = \frac{g(\theta + \Delta) - g(\theta)}{\Delta}.$$

Also folgt die Chapman-Robbins Ungleichung aus (3.1) (das Supremum ist gratis).

Als nächstes beweisen wir die Cramér-Rao Ungleichung. Aus der Schwarz'schen Ungleichung folgt, dass für jedes $U \in L_2(P_\theta)$

$$\left(E_\theta[U\frac{p_{\theta+\Delta}(X)-p_\theta(X)}{\Delta p_\theta(X)}]-E_\theta[U\frac{\partial}{\partial\theta}\log p_\theta(X)]\right)^2$$
$$\leq E[U^2]\ E\left[\left(\frac{p_{\theta+\Delta}(X)-p_\theta(X)}{\Delta p_\theta(X)}-\frac{\partial}{\partial\theta}\ \log p_\theta(X)\right)^2\right]\to 0.$$

Wählt man $U \equiv 1$, erhält man $E_\theta[\frac{\partial}{\partial\theta}\log p_\theta(X)]=0$. Wählt man $U=T(X)$, so folgt, dass g differenzierbar ist und

$$\mathrm{Cov}_\theta(T,\frac{\partial}{\partial\theta}\log p_\theta(X))=E_\theta[T(X)\frac{\partial}{\partial\theta}\log p_\theta(X)]=g'(\theta).$$

Die Behauptung folgt also wieder aus (3.1). (Wenn T nicht in $L_2(P_\theta)$ ist, ist $\mathrm{Var}_\theta[T]$ unendlich, und die Ungleichung ist trivialerweise erfüllt.) □

Bemerkung 3.1. *Die Chapman-Robbins Schranke ist zwar schärfer als die Cramér-Rao Schranke, aber schwieriger explizit zu berechnen.*

Lemma 3.1. *a) Sei $p_\theta(x)$ eine Dichte, welche die Bedingungen von Satz 3.3 erfüllt. Dann sind diese Bedingungen auch erfüllt für $p_\theta(x_1)\cdots p_\theta(x_n)$, und es gilt*

$$I_n(\theta)=nI_1(\theta).$$

b) Unter stärkeren Regularitätsbedingungen als im Satz 3.3 gilt

$$I(\theta)=-E_\theta[\frac{\partial^2}{\partial\theta^2}\log p_\theta(X)].$$

"Beweis"

a) Dass $I_n(\theta)=nI_1(\theta)$, ist klar. Der Rest ist eine mühsame Rechnung.

b) Durch Differenzieren von $E_\theta[\frac{\partial}{\partial\theta}\log p_\theta(X)]=0$ unter dem Integral folgt die Behauptung.

Beispiel 3.3. *Seien $X_1,\ldots,X_n$ i.i.d. ,$\sim$ Poisson (λ). Für $n=1$ haben wir $\log p_\lambda(x)=-\lambda+x\log(\lambda)-\log(x!)$. Also ist*

$$\frac{\partial}{\partial\lambda}\log p_\lambda(x)=-1+\frac{x}{\lambda},\quad \frac{\partial^2}{\partial\lambda^2}\log p_\lambda(x)=-\frac{x}{\lambda^2},$$

und daher $I_1(\lambda)=\frac{1}{\lambda}$ und $I_n(\lambda)=\frac{n}{\lambda}$. Wir betrachten nun zwei verschiedene Schätzgrössen $g(\lambda)$. Für $g(\lambda)=\lambda$ ist $\bar{X}$ erwartungstreu, und

$$Var_\lambda(\bar{X})=\frac{\mathrm{Var}_\lambda(X_1)}{n}=\frac{\lambda}{n}=\frac{1}{I_n(\lambda)}.$$

Also erreicht $\bar{X}$ die Cramér-Rao Schranke und ist UMVU für λ.

Für $g(\lambda)=e^{-\lambda}$ ist $T=(1-\frac{1}{n})^{n\bar{X}}$ UMVU, siehe 3.2. Es gilt aber

$$\begin{aligned}\mathrm{Var}_\lambda(T) &= E_\lambda[T^2]-E_\lambda[T]^2=e^{-n\lambda}\sum_{k=0}^{\infty}(1-\tfrac{1}{n})^{2k}\frac{(n\lambda)^k}{k!}-e^{-2\lambda}\\ &= \exp((1-\tfrac{1}{n})^2 n\lambda-n\lambda)-\exp(-2\lambda)=e^{-2\lambda}(e^{\lambda/n}-1)>e^{-2\lambda}\frac{\lambda}{n}=\frac{g'(\lambda)^2}{I_n(\lambda)},\end{aligned}$$

d.h. obwohl T UMVU ist, wird die Cramér-Rao Schranke nicht erreicht!

Allgemein wird die Cramér-Rao Schranke nur für exponentielle Familien und – abgesehen von linearen Transformationen – nur für eine Schätzgrösse $g(\theta)$ und nur für einen Schätzer erreicht.

Satz 3.4. *T ist erwartungstreu für $g(\theta)$ und erreicht die Cramér-Rao-Schranke genau dann, wenn zwei differenzierbare Funktionen $c(\theta)$ und $d(\theta)$ existieren, so dass $p_\theta(x) = \exp(c(\theta)T(x) + d(\theta))h(x)$ und $g(\theta) = -d'(\theta)/c'(\theta)$.*

Beweis. Gleichheit in der Schwarz'schen Ungleichung gilt genau dann, wenn (P_θ − f.s.)

$$\frac{\partial}{\partial\theta}\log p_\theta(x) = a(\theta)T(x) + b(\theta),$$

was äquivalent ist zu

$$p_\theta(x) = \exp(c(\theta)T(x) + d(\theta))h(x).$$

Ferner folgt aus

$$0 = E_\theta[\frac{\partial}{\partial\theta}\log p_\theta(X)] = \int [c'(\theta)T(x) + d'(\theta)]p_\theta(x)\mu(dx),$$

dass T erwartungstreu ist für $-d'(\theta)/c'(\theta)$. □

3.3.1 Mehrdimensionale Erweiterungen

Es sei jetzt θ p-dimensional, aber die zu schätzende Grösse $g(\theta)$ sei immer noch eindimensional. Die Chapman-Robbins Ungleichung lässt sich direkt übertragen. Für die Cramér-Rao Ungleichung wählen wir

$$\psi(\theta, x) = \sum_{i=1}^{p} a_i \frac{\partial}{\partial\theta_i}\log p_\theta(x)$$

mit zunächst beliebigen Koeffizienten a_i. Die Schwarz'sche Ungleichung impliziert dann unter analogen Bedingungen wie bei Satz 3.3, dass für jeden für $g(\theta)$ erwartungstreuen Schätzer T gilt

$$\mathrm{Var}_\theta[T(X)] \geq \frac{\left(\sum_{i=1}^{p} a_i \frac{\partial}{\partial\theta_i} g(\theta)\right)^2}{\sum_{i,j=1}^{p} a_i a_j I(\theta)_{ij}}.$$

Dabei ist $I(\theta)$ die sogenannte Fisher-Informationsmatrix, die definiert ist als

$$I(\theta)_{ij} = E_\theta[\frac{\partial}{\partial\theta_i}\log p_\theta(X)\ \frac{\partial}{\partial\theta_j}\log p_\theta(X)] = -E_\theta[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p_\theta(X)].$$

Die mehrdimensionale Cramér-Rao Ungleichung erhalten wir, indem wir in obiger Ungleichung die a_i so wählen, dass die rechte Seite maximal wird.

Satz 3.5. *Sei $P_\theta(dx) = p_\theta(x)\mu(dx), \theta \in \Theta \subset \mathbb{R}^p$ offen, und sei $T(X)$ erwartungstreu für $g(\theta) \in \mathbb{R}$. Unter Regularitätsbedingungen gilt:*

$$\mathrm{Var}_\theta[T] \geq (\frac{\partial}{\partial\theta}g(\theta))^T I(\theta)^{-1}(\frac{\partial}{\partial\theta}g(\theta)).$$

Beweis. Sei V eine positiv definite $p \times p$-Matrix und c ein Vektor im $\mathbb{R}^p$. Mit einem Lagrangemultiplikator sieht man leicht, dass unter der Nebenbedingung $a^T V a = 1$ der Ausdruck $(a^T c)^2$ maximal wird, wenn a ein Vielfaches von $V^{-1}c$ ist. Dieses Resultat wendet man nun an für $V = I(\theta)$ und $c = \frac{\partial}{\partial\theta} g(\theta)$. □

Daraus erhalten wir auch eine Ungleichung, wenn wir alle p Komponenten θ_i simultan schätzen.

Korollar 3.2. *Sei $P_\theta(dx)$ wie oben und sei $T_j(X)$ erwartungstreu für θ_j $(j = 1, \ldots, p)$. Dann gilt $\mathrm{Cov}_\theta[T] \geq I(\theta)^{-1}$ im Sinne dass $\mathrm{Cov}_\theta[T] - I(\theta)^{-1}$ positiv semidefinit ist. ($\mathrm{Cov}_\theta[T]$ ist die Matrix mit Elementen $\mathrm{Cov}_\theta[T_i, T_j]$.)*

Beweis. Aus Satz 3.5 angewandt für $a^T T(X)$ folgt

$$a^T \mathrm{Cov}_\theta[T] a = \mathrm{Var}_\theta[a^T T(X)] \geq a^T I(\theta)^{-1} a.$$

□

3.4 UMPU-Tests

3.4.1 1-dimensionale exponentielle Familien

Sei $P_\theta(dx) = p_\theta(x)\mu(dx)$ mit $p_\theta(x) = \exp(c(\theta)T(x) + d(\theta))h(x)$ und $c(.)$ stetig und monoton wachsend.

Satz 3.6. *i) Nullhypothese $H_0 : \theta \leq \theta_0$ und Alternative $H_A : \theta > \theta_0$. Es gibt einen Test φ mit*

$$\varphi(x) = \begin{cases} 1 \text{ falls} & T(x) > c_0 \\ 0 \text{ falls} & T(x) < c_0 \end{cases}$$

und $E_{\theta_0}[\varphi] = \alpha$. Dieser Test hat dann Niveau α und ist gleichmässig mächtigst (UMP).

ii) Nullhypothese $H_0 : \theta_1 \leq \theta \leq \theta_2$ und Alternative $H_A : \theta < \theta_1$ oder $\theta > \theta_2$. Es gibt einen Test mit

$$\varphi(x) = \begin{cases} 1 \text{ falls} & T(x) < c_1 \text{ oder } T(x) > c_2 \\ 0 \text{ falls} & c_1 < T(x) < c_2 \end{cases}$$

und $E_{\theta_1}[\varphi] = E_{\theta_2}[\varphi] = \alpha$. Dieser Test hat dann Niveau α und ist gleichmässig mächtigst unverfälscht (UMPU).

iii) Nullhypothese $H_0 : \theta = \theta_0$ und Alternative $H_A : \theta \neq \theta_0$. Es gibt einen Test mit

$$\varphi(x) = \begin{cases} 1 \text{ falls} & T(x) < c_1 \text{ oder } T(x) > c_2 \\ 0 \text{ falls} & c_1 < T(x) < c_2 \end{cases}$$

und $E_{\theta_0}[\varphi] = \alpha$, $\frac{d}{d\theta} E_\theta[\varphi]|_{\theta=\theta_0} = 0$. Dieser Test hat dann Niveau α und ist UMPU.

Beweis. i) Betrachte zuerst einfache Hypothesen $\Theta_0 = \{\theta\}$, $\Theta_A = \{\theta'\}$, $\theta < \theta'$. Der Neyman-Pearson Test für Θ_0 gegen Θ_A hat die Form

$$\varphi(x) = \begin{cases} 1, & \text{falls } p_{\theta'}(x)/p_\theta(x) > const \\ 0, & \text{falls } p_{\theta'}(x)/p_\theta(x) < const. \end{cases}$$

Nun ist aber

$$\frac{p_{\theta'}(x)}{p_\theta(x)} > const \Longleftrightarrow T(x) > \frac{\log(const) - [d(\theta') - d(\theta)]}{c(\theta') - c(\theta)} := c_0.$$

Der kritische Wert c_0 wird aber bestimmt durch die Forderung $E_\theta[\varphi] \leq \alpha$, also hängt c_0 und damit der ganze Test *nicht* von der Alternative θ' ab !

Wir wählen jetzt φ wie oben mit c_0 so dass $E_{\theta_0}[\varphi] = \alpha$ (dies ist möglich aufgrund des Neyman-Pearson Lemmas). Ferner setzen wir $\beta(\theta) = E_\theta[\varphi]$. Dann ist nach obiger Überlegung für jedes $\theta > \theta_0$ und für jeden Test φ' mit $E_{\theta_0}[\varphi'] \leq \alpha$

$$\beta(\theta) = E_\theta[\varphi] \geq E_\theta[\varphi'].$$

Wir müssen also nur noch zeigen, dass $\beta(\theta) \leq \alpha$ für alle $\theta < \theta_0$. Fixiere ein $\theta < \theta_0$. Dann ist nach obiger Überlegung φ auch der Neyman-Pearson Test für θ gegen θ_0 mit Niveau $\beta(\theta)$. Also ist φ mächtiger als $\varphi'(x) \equiv \beta(\theta)$, d.h.

$$\beta(\theta) = E_{\theta_0}[\varphi'] \leq E_{\theta_0}[\varphi] = \alpha.$$

ii) und iii) Man sieht leicht, dass $\beta(\theta) = E_\theta[\varphi]$ stetig ist für jeden Test φ. Jeder unverfälschte Test mit Niveau α muss daher erfüllen $\beta(\theta_1) = \beta(\theta_2) = \alpha$ bzw. $\beta(\theta_0) = \alpha, \beta$ ist minimal für $\theta = \theta_0$. Dies erklärt die Bedingungen für c_1 und c_2. Für den Rest des Beweises siehe Lehmann, Testing Statistical Hypotheses, Kap. 4.2 oder Ferguson, Kap. 5.3 .

□

3.4.2 Mehrdimensionale exponentielle Familien

Viele wichtige Testprobleme haben die folgende Gestalt. Das Modell ist

$$P_\theta(dx) = \exp(\sum_{j=1}^{k} \theta_j T_j(x) + d(\theta))\mu(dx),$$

und man testet $H_0 : \theta_1 \leq 0$ gegen $H_A : \theta_1 > 0$. Wir diskutieren hier das spezielle Problem des Vergleichs zweier Poissonverteilungen, bei dem man bereits das Wesentliche sieht. Seien also $X_1, \ldots, X_n \sim \text{Poisson}(\lambda), Y_1, \ldots, Y_m \sim \text{Poisson}(\mu)$, wobei alle Variablen unabhängig sind, und $H_0 : \lambda \leq \mu$. Das Modell lässt sich schreiben als

$$\begin{aligned} & P_{\lambda,\mu}[X_1 = x_1, \ldots, X_n = x_n, Y_1 = y_1, \ldots, Y_m = y_m] \\ = \ & \exp(-n\lambda - m\mu) \frac{\lambda^{x_1 + \cdots + x_n} \mu^{y_1 + \cdots + y_m}}{x_1! \cdots y_m!} \\ = \ & \frac{\exp\left(\sum x_i \log(\lambda/\mu) + (\sum x_i + \sum y_i)\log(\mu) - (n\lambda + m\mu)\right)}{x_1! \cdots y_m!}. \end{aligned}$$

Dies ist von der obigen Form mit $\theta_1 = \log(\lambda/\mu)$, $\theta_2 = \log(\mu)$, $T_1 = \sum x_i$, $T_2 = \sum x_i + \sum y_i$ und $-d(\theta) = ne^{\theta_1+\theta_2} + me^{\theta_2}$.

Die Hauptidee ist, durch Bedingen auf T_2 eine eindimensionale exponentielle Familie zu bekommen, auf die man Satz 3.6 anwenden kann. Mithilfe von Satz A.6 folgt sofort

$$\begin{aligned} P_\theta[T_1 = t_1 | T_2 = t_2] &= \binom{t_2}{t_1} \left(\frac{n\lambda}{n\lambda + m\mu}\right)^{t_1} \left(\frac{m\mu}{n\lambda + m\mu}\right)^{t_2 - t_1} \\ &= \exp\left(t_1 \log(\frac{n\lambda}{m\mu}) + t_2 \log(\frac{m\mu}{n\lambda + m\mu})\right) \binom{t_2}{t_1} \\ &= \exp\left(t_1\theta_1 + t_2 \log(\frac{m}{ne^{\theta_1} + m})\right) \left(\frac{n}{m}\right)^{t_1} \binom{t_2}{t_1}. \end{aligned}$$

Wir haben also tatsächlich eine exponentielle Familie mit Parameter θ_1 (unabhängig vom Störparameter θ_2 !). Bedingt auf $T_2 = t_2$ ist der UMP-Test zum Niveau α von der Form

$$\varphi(t_1, t_2) = \begin{cases} 1 & \text{falls} \quad t_1 > k(t_2) \\ 0 & \text{falls} \quad t_1 < k(t_2) \\ \gamma(t_2) & \text{falls} \quad t_1 = k(t_2). \end{cases}$$

Dabei ist $k(t_2)$ der kritische Wert für einen Test bei der Binomial-(t_2, p) Verteilung auf die Nullhypothese $p \leq n/(n+m)$. Diesen Test können wir nun auch für zufälliges t_2, d.h. im ursprünglichen Modell, benützen. Es gilt

Satz 3.7. *Der obige Test ist UMPU zum Niveau α im ursprünglichen Problem.*

Beweis. Zunächst gilt wegen des Satzes von der iterierten Erwartung

$$E_\theta[\varphi(T_1, T_2)] = E_\theta[E_{\theta_1}[\varphi(T_1, T_2)|T_2]] \leq E_\theta[\alpha] = \alpha$$

für $\theta_1 \leq 0$ und

$$E_\theta[\varphi(T_1, T_2)] = E_\theta[E_{\theta_1}[\varphi(T_1, T_2)|T_2]] \geq E_\theta[\alpha] = \alpha$$

für $\theta_1 \geq 0$, d.h. φ ist unverfälscht zum Niveau α.

Sei jetzt $\varphi'(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ ein anderer unverfälschter Test zum Niveau α. Wegen Suffizienz können wir annehmen, dass $\varphi' = \varphi'(T_1, T_2)$. Der entscheidende Punkt besteht darin, zu zeigen, dass

$$E_0[\varphi'(T_1, T_2)|T_2] \leq \alpha. \tag{3.2}$$

Das heisst also, dass $\varphi'(T_1, t_2)$ auch das Niveau α hat im bedingten Testproblem. Wir beenden zunächst den Beweis unter Benutzung dieser Ungleichung. Da φ UMP ist im bedingten Problem, muss für $\theta_1 > 0$ gelten

$$E_{\theta_1}[\varphi(T_1, T_2)|T_2] \geq E_{\theta_1}[\varphi'(T_1, T_2)|T_2].$$

Wieder mit dem Satz von der iterierten Erwartung folgt daher $E_\theta[\varphi] \geq E_\theta[\varphi']$ für $\theta_1 > 0$, d.h. φ ist ein gleichmässig mächtigster Test.

Für die entscheidende Ungleichung (3.2) gehen wir aus von

$$\alpha' := \sup_{\theta_1 \leq 0} E_\theta[\varphi'] \leq \inf_{\theta_1 \geq 0} E_\theta[\varphi'], \ \alpha' \leq \alpha.$$

Da $E_\theta[\varphi']$ stetig ist in θ (was man leicht einsehen kann), folgt $E_{(0,\theta_2)}[\varphi'] = \alpha'$. Nun ist aber T_2 vollständig für das Modell $(P_{(0,\theta_2)})$ (siehe Satz 3.2). Also folgt aus

$$E_{(0,\theta_2)}[E_0[\varphi'|T_2] - \alpha'] = E_{(0,\theta_2)}[\varphi'] - \alpha' = 0 \quad \forall \theta_2,$$

dass $E_0[\varphi'|T_2] - \alpha' = 0$. □

Der allgemeine Fall geht analog. Einzig der Nachweis, dass die bedingte Verteilung von T_1 gegeben $T_2 = t_2, \ldots, T_k = t_k$ die Form

$$\exp(\theta_1 t_1 + d(\theta_1, t_2, \ldots, t_k))\nu(dt_1, t_2, \ldots, t_k)$$

hat, erfordert etwas Masstheorie. Weitere wichtige Beispiele sind der t-Test für X_i i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$ sowie der exakte Test von Fisher für 2×2 Kontingenztafeln.

Kapitel 4

Äquivariante Verfahren

Wir schränken uns hier auf Verfahren ein, die die Symmetrie eines Problems erhalten. Unter diesen gibt es solche, die das Risiko gleichzeitig für alle Parameter minimieren.

4.1 Äquivariante Lokationsschätzer

Ein Lokationsmodell hat die Form

$$X_i = \theta + \varepsilon_i \quad (i = 1, \ldots, n),$$

wobei die gemeinsame Verteilung der ε_i bekannt ist und bezüglich des Lebesguemasses die Dichte $p(u_1, \ldots, u_n)$ hat. Damit ist die Dichte der X_i gleich $p(x_1 - \theta, \ldots, x_n - \theta)$.

Definition 4.1. *Ein Schätzer $T : \mathbb{R}^n \to \mathbb{R}$ heisst (lokations-)* äquivariant *falls*

$$T(x_1 + c, \ldots, x_n + c) = T(x_1, \ldots, x_n) + c \quad \forall x_1, \ldots, x_n, \ \forall c.$$

Eine Verlustfunktion L heisst (lokations-) invariant *falls*

$$L(\theta + c, a + c) = L(\theta, a) \quad \forall c,$$

oder äquivalent dazu

$$L(\theta, a) = L(a - \theta).$$

Der folgende Satz zeigt, warum unter den äquivarianten Schätzern üblicherweise einer existiert, der das Risiko simultan für alle Parameter minimiert. Diese Eigenschaft nennen wir "uniformly minimum risk equivariant"(UMRE).

Satz 4.1. *Wenn T äquivariant ist und L invariant, dann ist das Risiko konstant.*

Beweis.

$$R(\theta, T) = E_\theta[L(T(X_1, \ldots, X_n) - \theta)] = E[L(T(\varepsilon_1, \ldots, \varepsilon_n))] = R(0, T).$$

□

Man muss also nur das Risiko für $\theta = 0$ minimieren. Als ersten Schritt dazu bestimmen wir die allgemeine Form eines äquivarianten Schätzers.

Lemma 4.1. *T ist äquivariant genau dann, wenn es eine Funktion $v : \mathbb{R}^{n-1} \to \mathbb{R}$ gibt so dass gilt $T = x_n + v(y_1, \ldots, y_{n-1})$ mit $y_i = x_i - x_n$.*

Beweis. „$\Leftarrow$" ist klar. Für „$\Rightarrow$" benutzt man, dass

$$T(x_1, \ldots, x_n) = x_n + T(x_1 - x_n, \ldots, x_{n-1} - x_n, 0).$$

Also setzt man einfach $v(y_1, \ldots, y_{n-1}) = T(y_1, \ldots, y_{n-1}, 0)$. □

Den UMRE Schätzer kann man nun durch punktweises Minimieren finden, indem man auf $Y = (Y_1, \ldots, Y_{n-1})$ konditioniert (analog wie beim Bayes-schätzer).

Satz 4.2. *Wenn für fast alle y $v^*(y) = \arg\min_v E_0[L(X_n + v)|Y = y]$, dann ist $T^* = X_n + v^*(Y)$ UMRE.*

Beweis. Sei $T = X_n + v(Y)$ äquivariant. Dann ist nach Voraussetzung

$$R(0,T) = E_0[L(T)] = E_0[E_0[L(X_n + v(Y))|Y]] \geq E_0[E_0[L(X_n + v^*(Y))|Y]] = R(0, T^*)$$

(wenn Y gegeben ist, ist $v(Y)$ eine Konstante). Also folgt die Behauptung mit Satz 4.1. □

Um diesen Satz anwenden zu können, müssen wir die bedingte Verteilung von X_n gegeben Y berechnen.

Lemma 4.2. *i) Y hat die Dichte $\int_{\mathbb{R}} p(y_1 + u, \ldots, y_{n-1} + u, u)du$, unabhängig von θ.*

ii) Für $\theta = 0$ ist die bedingte Dichte von X_n gegeben $Y = y$ gleich

$$\frac{p(y_1 + x_n, \ldots, y_{n-1} + x_n, x_n)}{\int_{\mathbb{R}} p(y_1 + u, \ldots, y_{n-1} + u, u)du}.$$

Beweis. Die Abbildung $g : (x_1, \ldots, x_n) \to (y_1, \ldots, y_{n-1}, x_n)$ ist umkehrbar mit Funktionaldeterminante 1. Das Lemma folgt daher leicht aus A.1.1 und A.1.2. □

Korollar 4.1 (Pitman). *Für $L(\theta, a) = (\theta - a)^2$ ist der UMRE-Schätzer gleich*

$$T^* = \frac{\int zp(x_1 - z, \ldots, x_n - z)dz}{\int p(x_1 - z, \ldots, x_n - z)dz} = x_n - E_0[X_n|Y = y].$$

Beweis. Gemäss Satz 4.2 und Satz A.8 ist $v^*(y) = -E_0[X_n|Y = y]$. Wegen Lemma 4.2, ii) gilt

$$\begin{aligned} E_0[X_n|Y = y] &= \frac{\int up(y_1 + u, \ldots, y_{n-1} + u, u)du}{\int p(y_1 + u, \ldots, y_{n-1} + u, u)du} \\ &= x_n - \frac{\int (x_n - u)p(x_1 - x_n + u, \ldots, x_{n-1} - x_n + u, u)du}{\int p(x_1 - x_n + u, \ldots, x_{n-1} - x_n + u, u)du} \\ &= x_n - \frac{\int zp(x_1 - z, \ldots, x_{n-1} - z, x_n - z)dz}{\int p(x_1 - z, \ldots, x_{n-1} - z, x_n - z)dz} \end{aligned}$$

(im letzten Schritt wurde $z = x_n - u$ substituiert). □

Beispiel 4.1. *Sei $L(\theta, a) = (\theta - a)^2$ und ε_i i.i.d. d.h. $p(u_1, \ldots, u_n) = p(u_1) \cdots p(u_n)$.*

i) *Gleichverteilung:* $p(u) = \mathbf{1}_{[-\frac{1}{2},\frac{1}{2}]}(u)$. *Dann:*

$$p(x_1 - z, \ldots, x_n - z) = \begin{cases} 1 & \text{falls } \max(x_i) - \frac{1}{2} \leq z \leq \min(x_i) + \frac{1}{2} \\ 0 & \text{sonst.} \end{cases}$$

Daher

$$T^* = \frac{(\min(X_i) + \frac{1}{2})^2 - (\max(X_i) - \frac{1}{2})^2}{2(\min(X_i) + \frac{1}{2} - (\max(X_i) - \frac{1}{2}))} = \frac{\min(X_i) + \max(X_i)}{2}.$$

ii) *Normalverteilung: Direkte Rechnung ist mühsam. Besser ersetzt man* X_n *in Lemma 4.1 durch* $\overline{X}$, *d.h. man schreibt einen äquivarianten Schätzer* T *als* $T = \overline{X} + v(Y_1, \ldots, Y_{n-1})$. *Analog zu Satz 4.2 erhält man dann den UMRE-Schätzer als*

$$\overline{X} + \arg\min_v E_0[(\overline{X} + v)^2 | Y = y] = \overline{X} - E_0[\overline{X} | Y = y].$$

Nun ist aber $\overline{X}$ *unkorreliert mit* $Y_1, \ldots, Y_{n-1}$, *und bei Normalverteilung impliziert Unkorreliertheit auch Unabhängigkeit. Also ist* $E_0[\overline{X}|Y = y] = E_0[\overline{X}] = 0$, *und das arithmetische Mittel ist UMRE.*

Im allgemeinen muss man die Integrale für den UMRE-Schätzer numerisch berechnen. Dies ist aber ein kleines Problem. Schlimmer ist es, dass der optimale Schätzer stark von der Dichte p abhängt, welche in der Praxis meist unbekannt ist.

Der UMRE-Schätzer mit quadratischem Verlust steht in Beziehung zu früheren Optimalitätsbegriffen. Die Formel in Korollar 4.1 zeigt, dass T^* formal nichts anderes ist als der Bayes-schätzer mit der Gleichverteilung auf $\mathbb{R}$ als „a-priori-Verteilung“ (dies ist natürlich keine Wahrscheinlichkeitsverteilung mehr). In Kapitel 5 werden wir sehen, dass der UMRE-Schätzer minimax und im allgemeinen auch zulässig ist. Ausserdem ist der UMRE-Schätzer stets erwartungstreu: Analog wie in Satz 4.1 folgt, dass der Bias $E_\theta[T] - \theta$ eines äquivarianten Schätzers konstant sein muss. Also hat $T - E_0[T]$ überall kleineres Risiko als T, wenn $E_0[T] \neq 0$. Das impliziert aber nicht, dass T^* UMVU ist, denn es kann Schätzer geben, die erwartungstreu, aber nicht äquivariant sind.

4.2 Äquivariante Schätzer in Transformationsmodellen

Wenn

$$X = u_\theta(\varepsilon),$$

wobei die Verteilung P von ε bekannt ist und $(u_\theta)_{\theta \in \Theta}$ eine Gruppe von Transformationen $\mathbb{X} \to \mathbb{X}$ bildet, sprechen wir von einem Transformationsmodell. Die Verteilung P_θ von X ist dann gegeben durch $P_\theta(A) = P(u_\theta^{-1}(A))$. Der Parameterraum Θ ist ebenfalls eine Gruppe, wenn man $\theta_2 \circ \theta_1$ definiert durch: $u_{\theta_2} \circ u_{\theta_1} = u_{\theta_2 \circ \theta_1}$ (wir setzen voraus, dass $u_\theta \neq u_{\theta'}$ falls $\theta' \neq \theta$). Das neutrale Element in Θ bezeichnen wir mit 0, d.h. $u_0(x) = x, P_0 = P$.

Beispiel 4.2. *Lokation und Skala. Sei* $\theta = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+$, $\mathbb{X} = \mathbb{R}^n$ *und*

$$u_\theta(x_1, \ldots, x_n) = (\mu + \sigma x_1, \ldots, \mu + \sigma x_n).$$

Dann ist $\theta_2 \circ \theta_1 = (\mu_2 + \sigma_2 \mu_1, \sigma_2 \sigma_1)$. *Falls* $(\varepsilon_1, \ldots, \varepsilon_n)$ *eine Dichte* $p(u_1, \ldots, u_n)$ *hat, dann ist*

$$p_\theta(x_1, \ldots, x_n) = \sigma^{-n} p\left(\frac{x_1 - \mu}{\sigma}, \ldots, \frac{x_n - \mu}{\sigma}\right).$$

Definition 4.2. *Ein Schätzer $T : \mathbb{X} \to \Theta$ heisst äquivariant falls*

$$T(u_\theta(x)) = \theta \circ T(x) \quad \forall\theta\ \forall x.$$

Eine Verlustfunktion $L : \Theta \times \Theta \to \mathbb{R}_+$ heisst invariant falls

$$L(\theta' \circ \theta, \theta' \circ a) = L(\theta, a) \quad \forall\theta' \Leftrightarrow L(\theta, a) = L(\theta^{-1} \circ a).$$

Beispiel 4.3. *(Fortsetzung). T äquivariant heisst*

$$T(\mu + \sigma x_1, \dots, \mu + \sigma x_n) = (\mu + \sigma T_1(x_1, \dots, x_n), \sigma T_2(x_1, \dots, x_n)).$$

L invariant heisst $L(\theta, a) = L(\frac{a_1-\mu}{\sigma}, \frac{a_2}{\sigma})$.

Satz 4.3. *: Falls T äquivariant ist und L invariant, dann ist das Risiko konstant.*

Beweis.

$$R(\theta, T) = E_\theta[L(\theta, T(X))] = E[L(\theta, T(u_\theta(\varepsilon)))] = E[L(\theta, \theta \circ T(\varepsilon))] = E[L(T(\varepsilon))] = R(0, T).$$

□

Als nächstes brauchen wir eine Statistik Y, welche die Rolle von $X_1 - X_n, \dots, X_{n-1} - X_n$ im Lokationsmodell übernimmt.

Definition 4.3. *$Y : \mathbb{X} \to \mathbb{Y}$ heisst* maximal invariant *falls*

$$Y(x) = Y(x') \Leftrightarrow x' = u_\theta(x) \text{ für ein } \theta \in \Theta.$$

Beispiel 4.4. *(Fortsetzung).*

$$Y = \left(\frac{x_1 - x_n}{|x_{n-1} - x_n|}, \dots, \frac{x_{n-2} - x_n}{|x_{n-1} - x_n|},\ sign\ (x_{n-1} - x_n) \right)$$

ist maximal invariant.

Jetzt gehen die Überlegungen gleich wie im Lokationsmodell.

Lemma 4.3. *Sei T_0 ein beliebiger äquivarianter Schätzer und Y maximal invariant. Dann ist T äquivariant genau dann, wenn eine Abbildung $v : \mathbb{Y} \to \Theta$ existiert mit $T(x) = T_0(x) \circ v(Y(x))$.*

Beweis. „$\Leftarrow$“ ist klar. Für „$\Rightarrow$“ setzt man $v(y) = T_0(x)^{-1} \circ T(x)$ für ein x mit $Y(x) = y$. Das ist wohldefiniert, denn aus $Y(x) = Y(x')$ folgt $x' = u_\theta(x)$ und damit

$$T_0(x')^{-1} \circ T(x') = (\theta \circ T_0(x))^{-1} \circ (\theta \circ T(x)) = T_0(x)^{-1} \circ T(x).$$

□

Satz 4.4. *Wenn $v^*(y) = \arg\min_v E_0[L(T_0 \circ v)|Y = y]$, dann ist $T^* = T_0 \circ v^*$ UMRE.*

Beweis. analog wie für Satz 4.2. □

Die Berechnung der bedingten Verteilung von T_0 gegeben Y ist i.A. schwierig und hängt vom speziellen Problem ab. Oft hilft eine geschickte Wahl von T_0 (vgl. das Beispiel mit der Normalverteilung oben).

Beispiel 4.5. *(Fortsetzung). Seien speziell $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. ,$\sim$Exp (1)und $L(\theta, a) = (\frac{a_1-\mu}{\sigma})^2$. Dann ist*

$$T_0 = (X_{(1)}, \sum_{i=1}^{n}(X_i - X_{(1)}))$$

äquivariant (klar), suffizient (leicht) und vollständig (kompliziert, siehe Lehmann, Point Estimation, Ex. 1.5.14). Das Lemma 4.4 unten zeigt, dass T_0 und die maximale Invariante Y unabhängig sind. Also ist

$$E_0[L(T_0 \circ v)|Y = y] = E_0[L(T_0 \circ v)] = E_0[(X_{(1)} + v_1 \sum(X_i - X_{(1)}))^2].$$

Minimierung bezüglich v_1 ergibt

$$v_1^* = -\frac{E[X_{(1)} \sum(X_i - X_{(1)})]}{E[(\sum(X_i - X_{(1)}))^2]} = \cdots = -\frac{1}{n^2}.$$

Damit ist der optimale Schätzer der Lokation $X_{(1)} - n^{-2}\sum(X_i - X_{(1)})$.

Lemma 4.4 (Basu). *Sei T eine suffiziente und vollständige Statistik für $(P_\theta, \theta \in \Theta)$ und Y eine beliebige Statistik, deren Verteilung nicht von θ abhängt. Dann sind Y und T unabhängig.*

Beweis. Zu zeigen ist $P[Y \in A|T] = P[Y \in A]$ (beide Seiten sind unabhängig von θ). Mit dem Satz von der iterierten Erwartung ist $E_\theta[P[Y \in A|T] - P[Y \in A]] = 0$, also folgt die Behauptung aus der Vollständigkeit. □

4.3 Invariante Tests

Wir begnügen uns hier, die Grundidee an folgendem Beispiel darzulegen.

$$X_1, \ldots, X_n \text{ i.i.d. } \sim \mathcal{N}(\mu, \sigma^2), \; H_0 : \mu \leq 0, \; H_A : \mu > 0.$$

Das Modell ist ein Transformationsmodell gemäss dem vorhergehenden Kapitel, mit $u_\theta(x) = (\mu + \sigma x_1, \ldots, \mu + \sigma x_n)$. Bei einer solchen Transformation ändern sich aber im allgemeinen H_0 und H_A. Sie bleiben nur invariant bei Transformationen aus der Untergruppe mit $\mu = 0$. Daher kann man auch nur verlangen, dass sich ein Test nicht ändert bei Transformationen aus dieser Untergruppe: Ein Test φ heisst *invariant*, falls

$$\varphi(\sigma x_1, \ldots, \sigma x_n) = \varphi(x_1, \ldots, x_n) \quad \forall \sigma > 0.$$

Damit ist aber auch das Risiko nicht mehr konstant: Niveau und Macht eines invarianten Tests sind unabhängig von σ, aber abhängig von μ. Also ist es nicht klar, ob gleichmässig mächtigste invariante Tests überhaupt existieren.

In unserem Beispiel kommt man weiter, wenn man Invarianz mit Suffizienz kombiniert. Die Statistik $(\overline{X}, S_n^2)$ ist suffizient. Ein invarianter Test, der nur davon abhängt, hat die Form $\varphi(\overline{X}/S_n)$. Man erhält damit also das folgende, einfachere Testproblem: Maximiere $E_{\mu,1}[\varphi(\overline{X}/S_n)]$ simultan für alle $\mu > 0$ unter der Nebenbedingung $E_{\mu,1}[\varphi(\overline{X}/S_n)] \leq$ „ für alle $\mu \leq 0$. Die Lösung ergibt den t-Test.

Für eine allgemeine Theorie siehe Lehmann, Testing Statistical Hypotheses, Kap. 6.

Kapitel 5

Nachweis von Zulässigkeit und Minimaxeigenschaft

5.1 Minimax-Schätzer

Typischerweise haben Minimax-Schätzer konstantes Risiko. Die Umkehrung gilt unter Bedingungen, die man relativ leicht nachprüfen kann. Wie wir im Fall von äquivarianten Lokationsschätzern gesehen haben, sind Schätzer mit konstantem Risiko nicht Bayes, wenn wir nur Wahrscheinlichkeitsverteilungen als a-priori Verteilungen zulassen. Wir definieren daher noch, was man als Grenzwert von Bayesverfahren versteht.

Definition 5.1. *T heisst* **extended Bayes** *wenn eine Folge von a-priori Verteilungen (α_n) existiert derart, dass $r(\alpha_n, T) - \inf_{T'} r(\alpha_n, T') \longrightarrow 0$ wenn $n \to \infty$.*

Satz 5.1. *Ein Schätzer T mit konstantem Risiko ist minimax, falls eine der folgenden Bedingungen erfüllt ist:*

i) T ist zulässig,

ii) T ist Bayes,

iii) T ist extended Bayes.

Beweis. iii) indirekt: Sei T' mit $\sup R(\theta, T') < \sup R(\theta, T) = r$. Da für alle a-priori Verteilungen α gilt $r(\alpha, T') = \int R(\theta, T')\alpha(d\theta) \leq \sup R(\theta, T') < r = r(\alpha, T)$, erhält man sofort einen Widerspruch.

i) und ii) gehen analog. □

Beispiel 5.1. *Sei $X \sim$ Binomial(n, θ) und $L(\theta, a) = (a - \theta)^2$. Gesucht ist ein Minimax-Schätzer. Wir betrachten Bayes-Schätzer mit der Beta(r, s)-Verteilung als a-priori-Verteilung und versuchen, das Risiko konstant zu machen. Mit der Bayes'schen Formel folgt*

$$\alpha(\theta|x) \propto \alpha(\theta)p(x|\theta) \propto \theta^{r-1}(1-\theta)^{s-1}\theta^x(1-\theta)^{n-x},$$

also ist die a-posteriori-Verteilung = Beta$(r + x, s + n - x)$. Ferner ist der Bayes-schätzer

$$T_{r,s}(X) = E[\theta|X] = \frac{X + r}{r + s + n}.$$

Damit erhalten wir

$$R(\theta, T_{r,s}) = \mathrm{Var}_\theta[T_{r,s}] + (E_\theta[T_{r,s}] - \theta)^2 = \frac{n\theta(1-\theta) + (n\theta + r - \theta(r+s+n))^2}{(r+s+n)^2}.$$

Damit dies unabhängig von θ ist, müssen die Koeffizienten von θ^2 und von θ im Zähler beide null sein. Das heisst $(r+s)^2 - n = 0$ und $n - 2r(r+s) = 0$, was äquivalent ist zu $r = s = \sqrt{n}/2$. Also ist $T = \frac{x+\sqrt{n}/2}{n+\sqrt{n}}$ minimax, und das Minimax-Risiko ist

$$\frac{r^2}{(r+s+n)^2} = \frac{n}{4(n+\sqrt{n})^2} = \frac{1}{4n}\,\frac{1}{(1+1/\sqrt{n})^2}.$$

Zum Vergleich: Der übliche Schätzer $T = X/n$ hat das Risiko $\frac{1}{4n}\,4\theta(1-\theta)$. Der Minimax-Schätzer gewinnt also nur wenig für $\theta \approx 0.5$ und verliert viel in allen anderen Fällen. Der pessimistische Standpunkt zahlt sich hier nicht aus.

Äquivariante Schätzer haben konstantes Risiko, und sind also Kandidaten für Minimax-Schätzer.

Satz 5.2. *Betrachte das Lokationsmodell und den Pitman-Schätzer T^* von Kapitel 4. 1. Falls T^* endliche Varianz hat, ist T^* minimax.*

Beweis. Formal ist der Pitman-Schätzer Bayes'sch bezüglich des Lebesguemasses auf $\mathbb{R}$ als a-priori-Verteilung. Man vermutet also, dass T^* extended Bayes ist für $\alpha_n \to$ Lebesguemass. Wir wählen $\alpha_n =$ Gleichverteilung auf $[-n, n]$ und beweisen das.

Sei $T_{a,b}$ der Bayes-schätzer für $\alpha =$ Gleichverteilung auf $[a, b]$, d.h.

$$T_{a,b} = E[\theta | X_1, \ldots, X_n] = \frac{\int_a^b p(x_1 - \theta, \ldots, x_n - \theta)\theta d\theta}{\int_a^b p(x_1 - \theta, \ldots, x_n - \theta) d\theta}.$$

Man sieht, dass $T_{a,b}$ gegen T^* konvergiert für $a \to -\infty$ und $b \to \infty$.

Wir wollen das Bayesrisiko von $T_{a,b}$ abschätzen. Da

$$T_{a,b}(x_1 + c, \ldots, x_n + c) = T_{a-c,b-c}(x_1, \ldots, x_n) + c,$$

folgt $R(\theta, T_{-n,n}) = E_\theta[(T_{-n,n} - \theta)^2] = E_0[T^2_{-n-\theta,n-\theta}] = R(0, T_{-n-\theta,n-\theta})$, und somit

$$\begin{aligned} r(\alpha_n, T_{-n,n}) &= \frac{1}{2n}\int_{-n}^{n} R(\theta, T_{-n,n}) d\theta \geq (1-\varepsilon) \inf_{|\theta| \leq (1-\varepsilon)n} R(\theta, T_{-n,n}) \\ &= (1-\varepsilon) \inf_{|\theta| \leq (1-\varepsilon)n} R(0, T_{-n-\theta,n-\theta}) \geq (1-\varepsilon) \inf_{a \leq -\varepsilon n, b \geq \varepsilon n} R(0, T_{a,b}) \end{aligned}$$

(für die letzte Ungleichung beachte man, dass aus $|\theta| \leq (1-\varepsilon)n$ folgt $-n - \theta \leq -\varepsilon n$ und $n - \theta \geq \varepsilon n$).

Da $R(0, T^*) = r(\alpha_n, T^*) \geq r(\alpha_n, T_{-n,n})$, genügt es zu zeigen, dass $\liminf r(\alpha_n, T_{-n,n}) \geq R(0, T^*)$. Aus obigen Ungleichungen folgt aber mit dem Lemma von Fatou

$$\begin{aligned} \liminf r(\alpha_n, T_{-n,n}) &\geq (1-\varepsilon) \lim_n \inf_{a \leq -\varepsilon n, b \geq \varepsilon n} R(0, T_{a,b}) = (1-\varepsilon) \liminf_{a,b \to \infty} R(0, T_{-a,b}) \\ &\geq (1-\varepsilon) E_0[\liminf T^2_{-a,b}] = (1-\varepsilon) R(0, T^*). \end{aligned}$$

Da $\varepsilon > 0$ beliebig ist, folgt die Behauptung. □

5.2 Zulässigkeit

Der Nachweis von Zulässigkeit ist meist schwierig. Ein erster Zugang gibt Bedingungen für die Zulässigkeit von Bayes- und extended Bayes-Schätzern.

Satz 5.3. *Sei T Bayes bezüglich α und $R(\theta, T) < \infty \quad \forall\theta$. Dann ist T zulässig, falls eine der folgenden Bedingungen erfüllt ist:*

i) Der Bayes-Schätzer bez. α ist P_θ-f.s. eindeutig, d.h. aus $r(\alpha, T') = r(\alpha, T)$ folgt $P_\theta[T \neq T'] = 0 \quad \forall\theta$.

ii) $\alpha(G) > 0$ für alle $G \subset \Theta$ offen und $R(\theta, T')$ ist stetig in θ für alle T', für die $R(\theta, T') < \infty \quad \forall\theta$.

Beweis. i): Sei $R(\theta, T') \leq R(\theta, T) \quad \forall\theta$. Dann ist $r(\alpha, T') \leq r(\alpha, T)$. Da T Bayes ist, muss Gleichheit gelten, also ist nach Voraussetzung $T = T' \quad P_\theta -$ f.s. $\forall\theta$. Dann hat T' aber auch das gleiche Risiko wie T.

ii): Indirekt. Wir nehmen an, dass ein T' existiert mit $R(\theta, T') \leq R(\theta, T)\ \forall\theta$ und $R(\theta_0, T') < R(\theta_0, T)$ für ein θ_0. Dann gilt wegen der Stetigkeit von R sogar $R(\theta, T') \leq R(\theta, T) - \varepsilon$ für ein $\varepsilon > 0$ in einer offenen Umgebung U von θ_0. Daraus folgt aber $r(\alpha, T') \leq r(\alpha, T) - \varepsilon\alpha(U) < r(\alpha, T)$, was unmöglich ist, wenn T Bayes ist bezüglich α. □

Beispiel 5.2. *Seien $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(\theta, 1)$ und $L(\theta, a) = (\theta - a)^2$. Wegen Suffizienz genügt es, den Fall $n = 1$ zu betrachten. Für $\alpha = \mathcal{N}(\xi, \tau^2)$ kann man leicht nachprüfen, dass der Bayes-schätzer wie folgt aussieht:*

$$T = E[\theta|X] = \frac{1}{1+\tau^2}\xi + \frac{\tau^2}{1+\tau^2}X.$$

Beide Bedingungen von Satz 5.3 sind hier erfüllt:

i) Aus dem Beweis von Satz 2.1 folgt, dass T Bayes ist, genau dann, wenn $T = E[\theta|X]$ f.s. bezüglich der Randverteilung $Q(dx) = \int P_\theta(dx)\alpha(d\theta)$. Für obiges α ist $Q = \mathcal{N}(\xi, 1+\tau^2)$ und damit ist P_θ absolut stetig bezüglich Q für alle θ.

ii) Die Stetigkeit von $\int (T'(x) - \theta)^2 \phi(x - \theta) dx$ ist leicht zu zeigen.

Man kann auch leicht zeigen, dass $T = aX + b$ unzulässig ist falls $a < 0$ oder $a > 1$ oder $a = 1, b \neq 0$. Für die Frage der Zulässigkeit von linearen Schätzern bleibt also nur noch der Fall $T = X$ zu untersuchen. Dazu ist der nächste Satz nützlich.

Satz 5.4. *Sei T extended Bayes bez. (α_n), $R(\theta, T) < \infty \quad \forall\theta$ und $R(\theta, T')$ stetig in θ für alle T' mit $R(\theta, T') < \infty \quad \forall\theta$. Dann ist T zulässig, falls für alle $G \subset \Theta$ offen*

$$\frac{r(\alpha_n, T) - \inf_{T^*} r(\alpha_n, T^*)}{\alpha_n(G)} \longrightarrow 0.$$

Beweis. Analog wie bei Satz 5.3, ii) erhält man

$$\inf_{T^*} r(\alpha_n, T^*) \leq r(\alpha_n, T') \leq r(\alpha_n, T) - \varepsilon\alpha_n(U).$$

Es ist klar, dass dies ein Widerspruch zur Bedingung des Satzes darstellt. □

Beispiel 5.3. *(Fortsetzung von oben). Wähle* $\alpha_n = \mathcal{N}(0,n)$. *Dann ist* $R(\alpha_n, X) = 1$ *und* $\inf_{T^*} R(\alpha_n, T^*) = \frac{n}{n+1}$. *Also ist* X *extended Bayes und minimax. Ferner ist für* $U = [x_1, x_2]$ $\alpha_n(U) = \Phi(x_2/\sqrt{n}) - \Phi(x_1/\sqrt{n}) = O(1/\sqrt{n})$. *Also ist* X *auch zulässig.*

Wir haben hier die Zulässigkeit des UMRE-Schätzers im Lokationsmodell mit Normalverteilung bewiesen. Allgemein ist die Zulässigkeit von äquivarianten Schätzern selbst im Lokationsmodell schwierig. Es gilt aber: Falls ein äquivarianter Schätzer T existiert mit $E_0[|T|^3] < \infty$ (was meistens erfüllt ist), dann ist der Pitman-Schätzer zulässig (Stein, Ann. Math. Statist. 30 (1959), 970-979).

Man kann auch die Zulässigkeit mit Hilfe der Cramér-Rao-Ungleichung beweisen. Wir betrachten exponentielle Familien, bei denen ja die Cramér-Rao Schranke erreicht wird. Es geht also darum, ob nicht erwartungstreue Schätzer überall kleineres Risiko haben können. Wir wählen die Parametrisierung so, dass $c(\theta) \equiv \theta$. Gemäss Satz 3.4 ist dann T erwartungstreu für $g(\theta) = -d'(\theta)$.

Satz 5.5. *Sei* $P_\theta(dx) = \exp(\theta T(x) + d(\theta))h(x)\mu(dx)$ *und* $L(\theta, a) = (a + d'(\theta))^2$. *Wenn* $\Theta = \mathbb{R}$, *dann ist* T *zulässig.*

Beweis. Von Satz 3.4 wissen wir, dass T die Cramér-Rao-Schranke erreicht, d.h.

$$R(\theta, T) = \mathrm{Var}_\theta[T] = \frac{d''(\theta)^2}{I(\theta)}.$$

Ferner ist

$$I(\theta) = -E_\theta[\frac{\partial^2}{\partial\theta^2} \log p_\theta(x)] = -d''(\theta).$$

Also ist auch $I(\theta) = R(\theta, T)$.

Nun sei T' mit $R(\theta, T') \leq R(\theta, T)$ für alle θ. Wir wollen zeigen, dass T' erwartungstreu sein muss für $g(\theta) = -d'(\theta)$. Dann folgt nämlich aus der Cramér-Rao Ungleichung $R(\theta, T') \geq R(\theta, T)$. Dazu betrachten wir den Bias $b(\theta) = E_\theta[T'] + d'(\theta)$ und zeigen zunächst $b(\theta)^2 + 2b'(\theta) \leq 0$. Mit der Cramér-Rao Ungleichung bekommen wir

$$I(\theta) = R(\theta, T) \geq R(\theta, T') = b(\theta)^2 + \mathrm{Var}_\theta[T'] \geq b(\theta)^2 + \frac{(b'(\theta) - d''(\theta))^2}{I(\theta)}.$$

Wenn wir beide Seiten mit $I(\theta) > 0$ multiplizieren und $d''(\theta) = -I(\theta)$ benützen, erhalten wir

$$I(\theta)^2 \geq b(\theta)^2 I(\theta) + b'(\theta)^2 + 2b'(\theta)I(\theta) + I(\theta)^2,$$

oder äquivalent dazu

$$I(\theta)(b(\theta)^2 + 2b'(\theta)) \leq -b'(\theta)^2 \leq 0.$$

Da $I(\theta) > 0$, ergibt sich das Gewünschte.

Aus $b(\theta)^2 + 2b'(\theta) \leq 0$ folgt insbesondere $b'(\theta) \leq 0$, also ist b monoton fallend. Wenn $b(\theta_0) < 0$ wäre, dann wäre auch $b(\theta) \leq b(\theta_0) < 0$ für $\theta \geq \theta_0$, und damit:

$$\frac{1}{b(\theta)} = \frac{1}{b(\theta_0)} + \int_{\theta_0}^{\theta} (\frac{1}{b})' d\theta \longrightarrow \infty \text{ für } \theta \longrightarrow \infty,$$

weil $(1/b)' = -b'/b^2 \geq 1/2$. Dann muss aber $b(\theta)$ gegen 0 gehen für θ gegen ∞, was einen Widerspruch ergibt. Analog argumentiert man falls $b(\theta_0) > 0$. $\square$

Die Bedingung des Satzes kann auf zwei Arten verletzt sein: Entweder sind gewisse Parameter prinzipiell unmöglich (weil z.B. eine Varianz stets positiv ist), oder aufgrund der speziellen Fragestellung können gewisse Parameterwerte von vorneherein ausgeschlossen sein. Die folgenden Beispiele betrachten nur die erste Möglichkeit.

Beispiel 5.4.

- *Falls* $X \sim \mathcal{N}(\mu, 1)$, *dann ist* $\theta = \mu$, *also ist* X *zulässig.*
- *Falls* $X \sim \mathrm{Bin}(n, p)$, *dann ist* $\theta = n \log(p/(1-p))$, *also ist* X/n *zulässig.*
- *Falls* $X \sim$ *Poisson*(λ), *dann ist* $\theta = \log \lambda$, *also ist* X *zulässig.*
- *Falls* $X \sim \mathcal{N}(0, \sigma^2)$, *dann ist* $\theta = -1/(2\sigma^2)$, *also ist die Bedingung des Satzes nicht erfüllt. Man kann leicht nachrechnen, dass* $T' = \frac{1}{3}X^2$ *überall kleineres Risiko hat als* $T = X^2$.

5.3 Ein erstaunliches Beispiel von Unzulässigkeit

Wir betrachten folgendes Problem. Man hat von p Variablen jeweils n Beobachtungen gemacht, die wir als normalverteilt mit bekannter Varianz annehmen: $\mathbf{X}_1, \ldots, \mathbf{X}_n$ i.i.d. $\sim \mathcal{N}_p(\boldsymbol{\mu}, I)$. In diesem Kapitel kennzeichnen wir Vektoren speziell durch Fettdruck. Wir suchen einen Schätzer $\mathbf{T} : \mathbb{R}^{np} \longrightarrow \mathbb{R}^p$, der gut ist bezüglich der Verlustfunktion $L(\boldsymbol{\mu}, \mathbf{a}) = \sum_{i=1}^p (a_i - \mu_i)^2 = \|\boldsymbol{\mu} - \mathbf{a}\|^2$. Man kann z.B. an die Bestimmung des Blutdrucks von p Personen anhand von je n Messungen denken. $(\mu_1, \ldots, \mu_p)$ wären dann die wahren Werte dieser p Personen.

Wegen Suffizienz betrachten wir nur $n = 1$. Der übliche Schätzer ist $\mathbf{T} = \mathbf{X}$, denn die p Komponenten $X_1, \ldots, X_p$ von $\mathbf{X}$ sind ja unabhängig. Charles Stein hat aber 1956 gezeigt, dass dieses $\mathbf{T}$ unzulässig ist für $p > 2$. Dies ist überraschend, weil der Schätzer $\mathbf{T}$ sonst viele schöne Eigenschaften hat: Er ist UMRE, extended Bayes, minimax und erreicht die mehrdimensionale Cramér-Rao-Schranke (direkte Verallgemeinerungen früherer Resultate).

Der Beweis der Unzulässigkeit geht so, dass man explizit einen anderen Schätzer mit überall kleinerem Risiko angibt. Diesen verbesserten Schätzer kann man als empirischen Bayes-schätzer plausibel machen. Als a priori Verteilung wählen wir $\mu_1, \ldots, \mu_p$ i.i.d. $\sim \mathcal{N}(\xi, \tau^2)$. Dann ist der Bayes-schätzer für μ_i gleich

$$E[\mu_i | X_i] = \frac{1}{1+\tau^2}\xi + (1 - \frac{1}{1+\tau^2})X_i$$

und unter der Randverteilung Q sind $X_1, \ldots, X_p$ i.i.d. $\sim \mathcal{N}(\xi, 1+\tau^2)$, vgl. das Beispiel in 5.2.

Wenn wir ξ bekannt voraussetzen und τ^2 aus der Randverteilung als $\frac{1}{p}\|\mathbf{X} - \xi \cdot \mathbf{1}\|^2 - 1$ schätzen, erhalten wir als empirischen Bayes-schätzer

$$\mathbf{T}^{(1)} = \frac{p}{\|\mathbf{X} - \xi \cdot \mathbf{1}\|^2}\xi \cdot \mathbf{1} + (1 - \frac{p}{\|\mathbf{X} - \xi \cdot \mathbf{1}\|^2}) \cdot \mathbf{X}.$$

($\mathbf{1}$ ist natürlich der Vektor bestehend aus lauter Einsen). Wenn wir sowohl ξ und τ^2 aus der Randverteilung schätzen als $\overline{X} = \frac{1}{p}\sum_{j=1}^p X_j$ und $\frac{1}{p}\left\|\mathbf{X} - \overline{X} \cdot \mathbf{1}\right\|^2 - 1$, erhalten wir als empirischen Bayesschätzer

$$\mathbf{T}^{(2)} = \frac{p}{\left\|\mathbf{X} - \overline{X} \cdot \mathbf{1}\right\|^2}\overline{X} \cdot \mathbf{1} + (1 - \frac{p}{\left\|\mathbf{X} - \overline{X} \cdot \mathbf{1}\right\|^2}) \cdot \mathbf{X}.$$

Wenn $\mu_1 \approx \mu_2 \approx \cdots \approx \mu_p$ ($\approx \xi$), dann sollte $\mathbf{T}^{(2)}$ (bzw. $\mathbf{T}^{(1)}$) besser sein als $\mathbf{T}$, da ja dann unsere a-priori Annahmen zutreffen. Das Erstaunliche ist aber, dass geringfügige Modifikationen von $\mathbf{T}^{(1)}$ und $\mathbf{T}^{(2)}$ für alle $\mu_1, \dots \mu_p$ besser sind. Zur Vereinfachung der Notation wählen wir $\xi = 0$.

Satz 5.6. *i)* $\mathbf{T} = \mathbf{X}$ *ist zulässig für* $p = 1, 2$.

ii) Für $p \geq 3$ *hat der sogenannte Steinschätzer*

$$\mathbf{T}^* = (1 - \frac{b}{\|\mathbf{X}\|^2}) \cdot \mathbf{X}$$

überall kleineres Risiko als $\mathbf{T} = \mathbf{X}$ *für* $0 < b < 2(p-2)$. *Für* $b = p - 2$ *ist das Risiko minimal.*

iii) Für $p \geq 4$ *und* $0 < b < 2(p-3)$ *hat der Schätzer*

$$\mathbf{T}^{**} = \frac{b}{\left\|\mathbf{X} - \overline{X} \cdot \mathbf{1}\right\|^2} \overline{X} \cdot \mathbf{1} + (1 - \frac{b}{\left\|\mathbf{X} - \overline{X} \cdot \mathbf{1}\right\|^2}) \cdot \mathbf{X}$$

überall kleineres Risiko als $\mathbf{T} = \mathbf{X}$. *Für* $b = p - 3$ *ist das Risiko minimal.*

Beweis. i) Für $p = 1$ siehe Satz 5.4 und das anschliessende Beispiel. Der Fall $p = 2$ ist kompliziert, wir lassen ihn daher weg.

ii)

$$R(\boldsymbol{\mu}, \mathbf{T}^*) = E_{\boldsymbol{\mu}}[\left\|\mathbf{X} - \boldsymbol{\mu} - \mathbf{X}\frac{b}{\|\mathbf{X}\|^2}\right\|^2] = p + b^2 E_{\boldsymbol{\mu}}[\frac{1}{\|\mathbf{X}\|^2}] - 2b E_{\boldsymbol{\mu}}[\frac{\mathbf{X}^T(\mathbf{X} - \boldsymbol{\mu})}{\|\mathbf{X}\|^2}].$$

Den letzten Erwartungswert auf der rechten Seite können wir schreiben als

$$\sum_{i=1}^{p} E_{\boldsymbol{\mu}} \left[\frac{X_i(X_i - \mu_i)}{X_i^2 + \sum_{j \neq i} X_j^2} \right],$$

und der i-te Summand ist gleich

$$\int \cdots \int \int \frac{x_i}{x_i^2 + \sum_{j \neq i} x_j^2} (x_i - \mu_i) \phi(x_i - \mu_i) dx_i \prod_{j \neq i} \phi(x_j - \mu_j) dx_j.$$

Mit partieller Integration können wir das Integral bezüglich x_i umformen zu

$$\int \frac{\partial}{\partial x_i} \left(\frac{x_i}{x_i^2 + \sum_{j \neq i} x_j^2} \right) \phi(x_i - \mu_i) dx_i = \int \left(\frac{1}{\sum_{j=1}^{p} x_j^2} - \frac{2x_i^2}{(\sum_{j=1}^{p} x_j^2)^2} \right) \phi(x_i - \mu_i) dx_i.$$

Durch Aufsummieren erhalten wir

$$\sum_{i=1}^{p} E_{\boldsymbol{\mu}} \left[\frac{X_i(X_i - \mu_i)}{X_i^2 + \sum_{j \neq i} X_j^2} \right] = (p-2) E_{\boldsymbol{\mu}} \left[\frac{1}{\|\mathbf{X}\|^2} \right].$$

Also ist

$$R(\boldsymbol{\mu}, \mathbf{T}^*) = p - b(2(p-2) - b) E_{\boldsymbol{\mu}}[\frac{1}{\|\mathbf{X}\|^2}] < p = R(\boldsymbol{\mu}, \mathbf{T}).$$

iii) analog: Ersetze $\mathbf{X}$ durch $\mathbf{Y} = A\mathbf{X}$ wobei A orthogonal ist mit erster Zeile $= (1, \ldots, 1)/\sqrt{p}$.

$\square$

Die folgenden Bemerkungen sollen helfen, die Bedeutung dieses Satzes besser zu verstehen:

- Man kann zeigen, dass $E_{\boldsymbol{\mu}}[\|\mathbf{X}\|^{-2}]$ eine monoton fallende Funktion von $\|\boldsymbol{\mu}\|^2$ ist, die für $\|\boldsymbol{\mu}\| = 0$ den Wert $1/(p-2)$ hat und für $\|\boldsymbol{\mu}\| \to \infty$ gegen 0 konvergiert. Der Stein-Schätzer ergibt eine substantielle Verbesserung nur für $\boldsymbol{\mu} \approx \mathbf{0}$. Analog folgt, dass $\mathbf{T}^{**}$ nur wesentlich besser ist für $\mu_1 \approx \mu_2 \approx \cdots \approx \mu_p$.

- T_j^* und T_j^{**} verwenden zur Schätzung von μ_j auch Beobachtungen, die einen andern Mittelwert haben und unabhängig von X_j sind. Bei der Blutdruckmessung würde der Wert für eine Person auch mithilfe der Messungen bei andern Personen geschätzt, was unsinnig erscheint. Die Erklärung dieses Paradoxes ist die Verlustfunktion, welche die quadratischen Abweichungen addiert. Es gilt zwar

$$\sum_{i=1}^{p} E_{\boldsymbol{\mu}}[(T_i^* - \mu_i)^2] < \sum_{i=1}^{p} E_{\boldsymbol{\mu}}[(T_i - \mu_i)^2],$$

  aber im allgemeinen gibt es ein i mit

$$E_{\boldsymbol{\mu}}[(T_i^* - \mu_i)^2] > E_{\boldsymbol{\mu}}[(T_i - \mu_i)^2] = 1.$$

  Dies ist eine Konsequenz daraus, dass für festes i der Schätzer T_i zulässig ist.

- Die Schätzer $\mathbf{T}^*$ und $\mathbf{T}^{**}$ sind auch unzulässig. Die möglichen Verbesserungen sind aber minimal. Die Bestimmung von einem zulässigen Minimax-Schätzer ist schwierig (gelöst für $p \geq 6$).

- Analoge Phänomene treten bei vielen Problemen mit mehrdimensionalem Parameter auf, insbesondere bei der Regression (sogenannte Ridge Regression).

5.4 Rückblick

Wir beschränken uns hier darauf, einige Fakten zusammenzutragen und zu kommentieren. Die Illustration durch Beispiele, bzw. Gegenbeispiele ist der Leserin oder dem Leser überlassen.

Wir haben fünf verschiedene Optimalitätsbegriffe angetroffen: Zulässig, Minimax, Bayes, UMVU (bzw. UMPU bei Tests) und UMRE. Die ersten drei lassen alle Verfahren zur Konkurrenz zu, die beiden andern schränken die betrachteten Verfahren ein. Äquivarianz setzt ein spezielles Modell voraus. Zulässigkeit ist eine schwache Form von Optimalität (keine Verbesserung an einer Stelle ohne eine Verschlechterung an einer andern Stelle). Minimax und Bayes reduzieren die Risikofunktion auf eine Zahl und machen so alle Verfahren vergleichbar. Die Definition von Äquivarianz impliziert, dass alle Verfahren vergleichbar sind. UMVU bedeutet Optimalität simultan für alle Parameter. Extended Bayes können wir als Grenzwert von Bayes-Verfahren auffassen.

Es gibt viele zulässige und viele Bayes-Verfahren (wenn man die a priori Verteilung variiert). Diese beiden Klassen haben einen sehr grossen Durchschnitt, sind jedoch nicht ganz deckungsgleich. Wenn der Träger der a priori Verteilung nicht der ganze Parameterraum ist, ist das Bayes-Verfahren im allgemeinen nicht eindeutig, und gewisse Bayes-Verfahren sind unzulässig. Umgekehrt sind gewisse (aber leider nicht alle) Grenzwerte von Bayes-Verfahren zulässig.

UMVU- und UMRE-Schätzer sind im allgemeinen eindeutig. In vielen, aber nicht in allen Fällen, sind sie wenigstens zulässig. UMVU- und UMRE-Schätzer sind praktisch nie Bayes (da eine a priori Verteilung bedeutet, dass nicht alle Parameter gleichwertig sind), aber oft extended Bayes. UMRE-Schätzer sind zwar erwartungstreu (bei quadratischem Verlust), aber selten UMVU, mit der Normalverteilung als Ausnahme.

Minimax-Verfahren können mehrdeutig sein. In diesem Fall gibt es im allgemeinen unzulässige Minimax-Verfahren.

Exponentielle Familien haben viele Vorteile. Man kann im allgemeinen UMVU-Schätzer und UMPU-Tests bestimmen. Ausserdem gibt es eine Klasse von a priori Verteilungen, für die auch die a posteriori Verteilung wieder in der gleichen Klasse liegt (sogenannte **konjugierte** a priori Verteilungen). Dies haben wir in den Spezialfällen von Normal-, Poisson- und Binomial-Verteilung gesehen, wo die konjugierten a priori Verteilungen die Normal-, bzw. Gamma- bzw. Beta-Verteilungen waren. Wie man zu einer allgemeinen exponentiellen Familie die konjugierten a priori Verteilungen bestimmt, bleibt als Übung überlassen.

Kapitel 6

Asymptotik von Schätzern

6.1 Einführung

Wir betrachten in diesem Kapitel nur unabhängige, identisch verteilte Zufallsgrössen. Die verwendeten Modelle haben stets die Form

$$X_1, X_2, \ldots \text{ i.i.d. } \sim P_\theta, \theta \in \Theta,$$

und die zu schätzende Grösse sei $g(\theta)$, wobei $g : \Theta \to \mathbb{R}^p$. Ferner nehmen wir an, dass die Beobachtungen erzeugt werden gemäss

$$X_1, X_2, \ldots \text{ i.i.d. } \sim P_0.$$

Im einfachsten Fall ist $\Theta \subset \mathbb{R}^p$, $g(\theta) = \theta$ und $P_0 = P_{\theta_0}$ für ein $\theta_0 \in \Theta$. Das heisst, dass das Modell korrekt ist und wir am ganzen Parameter interessiert sind. Es ist aber auch wichtig zu untersuchen, was passiert, wenn das Modell nicht korrekt ist (d.h. $P_0 \neq P_\theta$ für alle θ). Mit der Einführung von g lassen wir ferner die Möglichkeit von Störparametern offen, die auch unendlich dimensional sein können. Insbesondere kann Θ die Menge aller Verteilungen auf $\mathbb{X}$ und $g(\theta)$ ein gewisses Funktional der Verteilung θ sein, z.B. der Median. In diesem Fall ist natürlich das Modell automatisch korrekt.

Der Stichprobenumfang ist variabel, so dass wir eine Folge von Schätzern $T_n : \mathbb{X}^n \to \mathbb{R}^p$ haben. Wir untersuchen das Verhalten, wenn n gegen unendlich geht. Zunächst ist es einmal plausibel, dass man mit immer mehr Beobachtungen $g(\theta)$ immer genauer schätzen kann, wenn das Modell korrekt ist. Dies führt zu folgender Definition

Definition 6.1. $(T_n)_{n\in\mathbb{N}}$ *heisst* **konsistent** *für $g(\theta)$, falls für alle $\varepsilon > 0$ und für alle $\theta \in \Theta$ $\lim_n P_\theta[\|T_n - g(\theta)\| > \varepsilon] = 0$. In andern Worten: T_n konvergiert gegen $g(\theta)$ in P_θ-Wahrscheinlichkeit.*

Wenn das Modell nicht korrekt ist, ist weniger klar, wie das typische Verhalten aussieht für wachsendes n. Wir werden sehen, dass auch dann T_n üblicherweise konvergiert, also nicht z. B. zwischen mehreren Werten oszilliert. Der Grenzwert t_0 hängt natürlich ab von der wahren Verteilung P_0, und wir werden diesen Grenzwert für bestimmte Klassen von Schätzfolgen auch identifizieren.

Konsistenz ist eine gewisse Minimalforderung für vernünftige Folgen von Schätzern (die konstanten Schätzer sind z.B. ausgeschlossen). Es gibt aber im allgemeinen viele konsistente Schätzfolgen. Um zu sehen, wie rasch T_n gegen $g(\theta)$ konvergiert, und damit auch zwischen verschiedenen konsistenten Schätzfolgen zu differenzieren, blasen wir $T_n - g(\theta)$ so mit a_n auf, dass im Grenzwert eine nichtausgeartete Verteilung entsteht.

Definition 6.2. *Eine Folge von p-dimensionalen Zufallsvariablen* (Z_n) **konvergiert in Verteilung** *gegen eine p-dimensionale Zufallsvariable* Z *(in Formeln* $Z_n \xrightarrow{d} Z$*) falls*

$$E[f(Z_n)] \to E[f(Z)]$$

für alle stetigen und beschränkten Funktionen f *auf* $\mathbb{R}^p$*.*

Wir suchen also eine Folge (a_n) und eine Zufallsvariable $Z \neq 0$ derart, dass unter P_θ $a_n(T_n - g(\theta)) \xrightarrow{d} Z$. (Die Grenzverteilung, d.h. die Verteilung von Z, hängt im Allgemeinen auch von θ ab). Je rascher a_n gegen unendlich geht, desto genauer ist die Schätzfolge. Bei gleichem a_n kommt es dann darauf an, wie stark die Verteilung von Z um null herum konzentriert ist. In den meisten Fällen ist $a_n = \sqrt{n}$ und Z $\mathcal{N}(0, V(\theta))$-verteilt. Dann sagen wir, dass T_n unter P_θ asymptotisch $\mathcal{N}(g(\theta), \frac{1}{n}V(\theta))$-verteilt ist. Auch im Fall, wo das Modell nicht korrekt ist, haben wir üblicherweise asymptotisch eine Normalverteilung mit Standardabweichung proportional zu $1/\sqrt{n}$: T_n ist unter P_0 asymptotisch $\mathcal{N}(t_0, \frac{1}{n}V_0)$-verteilt, wobei t_0 und V_0 Funktionale von P_0 sind.

Das folgende elementare Beispiel illustriert dieses Verhalten.

Beispiel 6.1. *Sei* $X_i = \mu + \varepsilon_i$ *wobei* $E[\varepsilon_i] = 0, \sigma^2 = E[\varepsilon_i^2] < \infty$*. Wir setzen* $\theta = (\mu, F_\varepsilon)$ *und* $g(\theta) = \mu$*. Dann ist aufgrund des Gesetzes der grossen Zahlen* $T_n = \frac{1}{n}\sum_{i=1}^n X_i = \overline{X}_n$ *konsistent, und aufgrund des zentralen Grenzwertsatzes ist* T_n *asymptotisch* $\mathcal{N}(\mu, \frac{1}{n}\sigma^2)$*-verteilt.*

Für allgemeinere Schätzer sind der Beweis der Konsistenz sowie die Bestimmung von a_n und der Grenzverteilung nicht so offensichtlich. Unser Vorgehen wird darin bestehen, T_n durch ein arithmetisches Mittel anzunähern:

$$T_n = g(\theta) + \frac{1}{n}\sum_{i=1}^n IF(X_i, \theta) + R_n, \tag{6.1}$$

mit

$$E_\theta[IF(X_i, \theta)] = 0$$

und

$$\sqrt{n}R_n \to 0 \text{ in } P_\theta\text{-Wahrscheinlichkeit.}$$

Dabei ist IF die sogenannte **Einflussfunktion** (auf englisch influence function). Der Name rührt davon her, dass mit (6.1)

$$T_{n+1} - T_n \approx \frac{1}{n+1}\left(IF(X_{n+1}, \theta) - \frac{1}{n}\sum_{i=1}^n IF(X_i, \theta)\right) \approx \frac{1}{n+1}IF(X_{n+1}, \theta),$$

sofern wir die Restterme vernachlässigen dürfen (vgl. aber das Unterkapitel 6.3.3 unten). Die Approximation auf der rechten Seite von (6.1) hängt vom wahren Parameter θ ab und kann daher nicht aus den Daten berechnet werden. Sie ist aber nützlich, um das asymptotische Verhalten des Schätzers zu verstehen. Im Fall, wo das Modell nicht korrekt ist, werden wir eine analoge Darstellung bekommen, wobei die Grössen rechts dann von P_0 statt von θ abhängen.

Wenn die Approximation (6.1) gilt, dann folgt die asymptotische Normalität von T_n leicht. Auf das arithmetische Mittel $\frac{1}{n}\sum IF(X_i,\theta)$ können wir den zentralen Grenzwertsatz anwenden (die benötigte multivariate Version ist in Korollar 6.2 unten), und aufgrund von Satz 6.4 unten ist der Rest R_n wie erwartet vernachlässigbar. Das heisst also, dass T_n asymptotisch $\mathcal{N}(g(\theta), \frac{1}{n}V(\theta))$-verteilt ist mit der asymptotischen Kovarianzmatrix

$$V(\theta) = E_\theta[IF(X_i,\theta)IF(X_i,\theta)^T].$$

In Kapitel 6.2 werden wir die Approximation (6.1) schrittweise für recht allgemeine Schätzfolgen beweisen und dabei auch Formeln für die Einflussfunktion IF erhalten. Statistische Anwendungen der asymptotischen Normalität bilden dann das Thema von Kapitel 6.3.

6.2 Nachweis von Konsistenz und asymptotischer Normalität

6.2.1 Konvergenz in Verteilung

Viele Argumente werden hier nur kurz gestreift. Für ausführlichere Angaben verweise ich auf die Bücher von Serfling und van der Vaart (siehe Anhang B).

Wir haben die Konvergenz in Verteilung von Z_n gegen Z oben bereits definiert als die Konvergenz der Erwartungswerte $E[f(Z_n)]$ gegen $E[f(Z)]$ für f stetig und beschränkt. Es zeigt sich, dass es genügt, die Konvergenz für eine kleinere Klasse von Funktionen nachzuweisen, und dass sie dann automatisch in einer grösseren Klasse von Funktionen gilt:

Satz 6.1. *Die folgenden Aussagen sind äquivalent:*

i) $Z_n \xrightarrow{d} Z$.

ii) $E[f(Z_n)] \to E[f(Z)]$ *für alle Funktionen f auf $\mathbb{R}^p$, die beschränkt und Lipschitz stetig sind (d.h. $|f(x) - f(y)| \le c||x-y||$).*

iii) $E[f(Z_n)] \to E[f(Z)]$ *für alle Funktionen f auf $\mathbb{R}^p$, die beschränkt und bezüglich der Grenzverteilung f.s. stetig sind (d.h. $P[Z \in \{z; f \text{ ist stetig im Punkt } z\}] = 1$).*

iv) $P[Z_n \le z] \to P[Z \le z]$ *für alle z, bei denen die rechte Seite stetig ist (Kleiner gleich in $\mathbb{R}^p$ ist komponentenweise definiert).*

Beweis. siehe van der Vaart, Lemma 2.2, oder Serfling, Kap. 1.5. □

Korollar 6.1. *Wenn $Z_n \xrightarrow{d} Z$ und f bezüglich der Grenzverteilung f.s. stetig ist, dann $f(Z_n) \xrightarrow{d} f(Z)$.*

Wenn f unbeschränkt ist, gilt die Konvergenz von $E[f(Z_n)]$ nicht mehr automatisch. Wir geben ein Resultat für Konvergenz von Momenten.

Satz 6.2. *Wenn Z_n in Verteilung gegen Z konvergiert, dann gilt* $\liminf E|Z_n|^r \ge E|Z|^r$. *Falls zusätzlich* $\sup E|Z_n|^s < \infty$ *für ein $s > r$, dann konvergiert $E|Z_n|^r$ gegen $E|Z|^r$.*

Beweis. siehe van der Vaart, Kap. 2.5, oder Serfling, Kapitel 1.4. □

Das nächste Resultat zeigt, wie man den Fall $p > 1$ auf den Fall $p = 1$ reduzieren kann.

Satz 6.3 (Cramér-Wold). *Z_n konvergiert genau dann in Verteilung gegen Z, wenn $a^T Z_n$ in Verteilung gegen $a^T Z$ konvergiert für alle $a \in \mathbb{R}^p$.*

Beweis. siehe van der Vaart, Kap. 2.3, oder Serfling, Kap. 1.5. □

Korollar 6.2 (Multivariater ZGS). *Wenn X_i eine i.i.d. Folge von p-dimensionalen Zufallsvektoren ist mit $E[X_i] = 0$ und $E[X_i X_i^T] = V$, dann gilt*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \xrightarrow{d} Z$$

wobei $Z \sim \mathcal{N}_p(0, V)$.

Der folgende Satz ist zur Behandlung von Resttermen äusserst wichtig.

Satz 6.4 (Slutsky). *Seien (A_n) und (B_n) Folgen von zufälligen $q \times p$-Matrizen bzw. q-Vektoren derart, dass A_n, bzw. B_n, in Wahrscheinlichkeit gegen A, bzw. b, konvergieren (elementweise) mit A und b nicht zufällig. Ferner konvergiere Z_n in Verteilung gegen Z. Dann konvergiert $A_n Z_n + B_n$ in Verteilung gegen $AZ + b$.*

Beweis. Zur Vereinfachung betrachten wir den Fall wo $p = q, b = 0$ und A_n konstant gleich der Einheitsmatrix ist. Wir wenden ii) von Satz 6.1 an. Sei also f Lipschitz stetig und beschränkt. Wir müssen zeigen, dass

$$E[f(Z_n + B_n)] - E[f(Z)] = E[f(Z_n + B_n)] - E[f(Z_n)] + E[f(Z_n)] - E[f(Z)]$$

gegen null konvergiert. Die zweite Differenz konvergiert nach Voraussetzung gegen null. Die Annahmen für f implizieren ferner, dass für jedes $\varepsilon > 0$

$$|f(Z_n + B_n) - f(Z_n)| \leq c\varepsilon + 2 \sup_x |f(x)| \mathbf{1}_{[||B_n|| > \varepsilon]}.$$

Weil $P[||B_n|| > \varepsilon]$ gegen null konvergiert, folgt daraus

$$\limsup_n |E[f(Z_n + B_n)] - E[f(Z_n)]| \leq c\varepsilon.$$

Weil ε beliebig war, haben wir also gezeigt, dass auch die erste Differenz in obiger Zerlegung gegen null konvergiert.

Der Fall, wo A_n mit n variiert geht analog. □

Beispiel 6.2. *Sei (X_i) i.i.d. mit $E[X_i] = \mu$, $E[(X_i - \mu)^2] = \sigma^2$ und und $E[(X_i - \mu)^4] = \kappa < \infty$. Die Stichproben-Varianz S_n^2 ist*

$$S_n^2 = \frac{1}{n-1} \sum (X_i - \overline{X})^2 = \frac{1}{n-1} \sum (X_i - \mu)^2 - \frac{n}{n-1} (\overline{X} - \mu)^2$$

(man beachte, dass $\sum (X_i - \overline{X}) = 0$). Damit kann man schreiben

$$\sqrt{n}(S_n^2 - \sigma^2) = \frac{n}{n-1} \frac{1}{\sqrt{n}} \sum ((X_i - \mu)^2 - \sigma^2) + R_n$$

wobei

$$R_n = -\frac{\sqrt{n}}{n-1}\sigma^2 - \frac{n}{n-1}\sqrt{n}(\overline{X}-\mu)^2.$$

Da $E[\sqrt{n}(\overline{X}-\mu)^2] = \sigma^2/\sqrt{n}$, *konvergiert* $\sqrt{n}(\overline{X}-\mu)^2$ *in Wahrscheinlichkeit gegen null (Chebyshev-Ungleichung). Das heisst,* R_n *konvergiert in Wahrscheinlichkeit gegen null und mit Satz 6.4 folgt, dass* S_n^2 *asymptotisch* $\mathcal{N}(\sigma^2, \frac{1}{n}(\kappa-\sigma^4))$*–verteilt ist.*

Zur Vereinfachung der Notation in vielen Beweisen führen wir noch die folgenden stochastischen Verallgemeinerungen der $O-$ und $o-$Notation aus der Analysis ein. Sei Z_n eine Folge von Zufallsvariablen und a_n eine Folge von positiven reellen Zahlen.

Definition 6.3. *Die Schreibweise* $Z_n = O_P(a_n)$ *bedeutet, dass für jedes* $\varepsilon > 0$ *ein* $M < \infty$ *existiert derart, dass* $P[\|Z_n\| > Ma_n] < \varepsilon$, *falls* n *genügend gross ist.*

Die Schreibweise $Z_n = o_P(a_n)$ *bedeutet, dass für jedes* $\varepsilon > 0$ *und jedes* $\delta > 0$ $P[\|Z_n\| > \varepsilon a_n] < \delta$ *ist, falls* n *genügend gross ist.*

$Z_n = o_P(1)$ heisst also nichts anderes, als dass Z_n stochastisch gegen null konvergiert. Es ist leicht zu sehen, dass $Z_n \xrightarrow{d} Z$ impliziert, dass $Z_n = O_P(1)$. Alle Regeln, die man für diese Symbole erwarten würde, gelten tatsächlich, z.B. $o(O_P(a_n)) = o_P(a_n)$, $\quad O_P(a_n) \cdot O_P(b_n) = O_P(a_n b_n)$ etc. .

6.2.2 Funktionen von arithmetischen Mitteln

Manche Schätzer sind gegeben als eine glatte Funktion von arithmetischen Mitteln:

$$T_n = h\left(\frac{1}{n}\sum_{i=1}^{n} u(X_i)\right), \tag{6.2}$$

mit $u : \mathbb{X} \to \mathbb{R}^q$ und $h : \mathbb{R}^q \to \mathbb{R}^p$. Dazu gehört zum Beispiel der Momentenschätzer aus Kapitel 1.3.1. Zwei wichtige weitere Beispiele sind

Beispiel 6.3. Stichprobenkorrelation *Die Beobachtungen* X_i *seien bivariat mit* $E_\theta[X_{i,1}] = E_\theta[X_{i,2}] = 0$. *Die zu schätzende Funktion ist die Korrelation*

$$g(\theta) = \frac{E_\theta[X_{i,1}, X_{i,2}]}{\sqrt{E_\theta[X_{i,1}^2]E_\theta[X_{i,2}^2]}}.$$

Als Schätzer betrachten wir

$$T_n = \frac{\sum_{i=1}^n X_{i,1}X_{i,2}}{\sqrt{\sum_{i=1}^n X_{i,1}^2 \sum_{i=1}^n X_{i,2}^2}}.$$

Also ist hier $q = 3$, $u(X_i) = (X_{i,1}^2, X_{i,2}^2, X_{i,1}X_{i,2})^T$ *und* $h(u) = u_3/\sqrt{u_1 u_2}$.

Beispiel 6.4 (Diskrete Beobachtungen). *Wenn die Beobachtungen* X_i *nur* M *verschiedene Werte* $\{1, 2, \ldots, M\}$ *annehmen können und* T_n *nicht von der Reihenfolge der Beobachtungen abhängt, dann ist* T_n *offensichtlich eine Funktion der relativen Häufigkeiten*

$$\widehat{p}_{n,k} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[X_i=k]} \quad (k = 1, 2, \ldots, M).$$

In den meisten Fällen hat man für jedes n sogar die gleiche Funktion, d.h.

$$T_n = h(\widehat{p}_{n,1}, \ldots, \widehat{p}_{n,M})$$

und wir haben somit die Form (6.2). Weil die Summe der $\widehat{p}_{n,k}$ immer gleich eins ist, ist h nicht eindeutig bestimmt: Ausserhalb vom $(M-1)$-dimensionalen Simplex kann man h beliebig fortsetzen, zum Beispiel mit $h(cp_1, \ldots, cp_M) = h(p_1, \ldots, p_M)$ für alle $c > 0$.

Ein Schätzer (T_n) von der Form (6.2) ist offensichtlich konsistent, wenn $E_\theta[|u_j(X_i)|] < \infty$, $h(E_\theta[u(X_i)]) = g(\theta)$ und h stetig ist an den Stellen $E_\theta[u(X_i)]$. Der folgende Satz beantwortet die Frage nach der asymptotischen Normalität.

Satz 6.5 (Delta-Technik). *Wenn $a_n(U_n - u) \xrightarrow{d} Z$ und die Abbildung h differenzierbar ist an der Stelle u, dann $a_n(h(U_n) - h(u)) \xrightarrow{d} \frac{\partial h}{\partial u}(u)Z$, wobei $\frac{\partial h}{\partial u}$ die $p \times q$ Matrix der Ableitungen $(\partial h_i/\partial u_j)$ von h bezeichnet.*

Beweis. Differenzierbarkeit an der Stelle u heisst

$$h(v) - h(u) = \frac{\partial h}{\partial u}(u)(v-u) + o(\|v-u\|).$$

Also ist

$$a_n(h(U_n) - h(u)) = \frac{\partial h}{\partial u}(u)a_n(U_n - u) + o_P(1). \tag{6.3}$$

Satz 6.4 zeigt, dass der erste Term rechts in Verteilung gegen $\frac{\partial h}{\partial u}(u)Z$ konvergiert und dass der zweite Term vernachlässigbar ist. □

Daraus erhält man nicht nur die asymptotische Normalität von (T_n), sondern aus der Formel (6.3) auch die Einflussfunktion.

Korollar 6.3. *Sei T_n von der Form (6.2), $E_\theta[u_j(X_i)^2] < \infty$, $h(E_\theta[u(X_i)]) = g(\theta)$ und h differenzierbar an den Stellen $E_\theta[u(X_i)]$. Dann*

i) *Die Approximation (6.1) gilt mit*

$$IF(X_i, \theta) = \frac{\partial h}{\partial u}(E_\theta[u(X_i)])(u(X_i) - E_\theta[u(X_i)]).$$

ii) *Mit $C(\theta)_{jk} = Cov_\theta(u_j(X_i), u_k(X_i))$ gilt*

$$T_n \overset{as}{\sim} \mathcal{N}(g(\theta), \frac{1}{n}\frac{\partial h}{\partial u}(E_\theta[u(X_i)])C(\theta)\frac{\partial h}{\partial u}(E_\theta[u(X_i)])^T).$$

Beispiel 6.5. *(Stichprobenkorrelation, Fortsetzung). Ohne Beschränkung der Allgemeinheit dürfen wir annehmen, dass $E_\theta[X_{i,1}^2] = E_\theta[X_{i,2}^2] = 1$. Wenn die 4. Momente existieren, können wir obigen Satz anwenden und erhalten mit einer kleinen Rechnung, dass $T_n \overset{as}{\sim} \mathcal{N}(g(\theta), \frac{1}{n}V(\theta))$ wobei*

$$V(\theta) = E_\theta[X_{i,1}^2 X_{i,2}^2] - g(\theta)(E_\theta[X_{i,1}^3 X_{i,2}] + E_\theta[X_{i,1}X_{i,2}^3]) + \frac{g(\theta)^2}{4}(E_\theta[(X_{i,1}^2 + X_{i,2}^2)^2] - 1).$$

Wenn die Erwartungswerte von X_i unbekannt sind, können wir sie mit dem arithmetischen Mittel schätzen und von den Beobachtungen subtrahieren. Das ändert nichts an *der asymptotischen Verteilung von T_n (analoge Überlegung wie bei der Stichprobenvarianz). Wenn (X_1, X_2) bivariat normalverteilt ist, dann ergibt sich mit einiger Rechnung $V(\theta) = (1 - g(\theta)^2)^2$. Man benützt dazu die Formel*

$$E[X_1X_2X_3X_4] = E[X_1X_2]E[X_3X_4] + E[X_1X_3]E[X_2X_4] + E[X_1X_4]E[X_2X_3],$$

welche für beliebige normalverteilte Zufallsvariable mit Erwartungswert null gilt.

Beispiel 6.6. *(Diskrete Beobachtungen, Fortsetzung.) Die Einflussfunktion in diesem Beispiel ist*

$$IF(x, \mathbf{p}) = \sum_{k=1}^{M} \frac{\partial h}{\partial p_k}(\mathbf{p})(\mathbf{1}_{[x=k]} - p_k).$$

Dabei kommt es nicht darauf an, wie wir h ausserhalb des Simplexes $\{\mathbf{p}; p_k \geq 0, \sum p_k = 1\}$ definiert haben: Weil $\sum(\mathbf{1}_{[x=k]} - p_k) = 0$ ist, spielt die Komponente des Gradienten orthogonal zum Simplex keine Rolle. Dies wird viel klarer sichtbar, wenn wir die Verteilung, die auf die Stelle x konzentriert ist, als Δ_x bezeichnen und damit schreiben

$$IF(x, \mathbf{p}) = \frac{d}{d\varepsilon} h((1-\varepsilon)\mathbf{p} + \varepsilon\Delta_x) \mid_{\varepsilon=0} = \lim_{\varepsilon \to 0} \frac{h((1-\varepsilon)\mathbf{p} + \varepsilon\Delta_x) - h(\mathbf{p})}{\varepsilon}.$$

Die Einflussfunktion ist also die Ableitung von h in die Richtung von p zu den Extremalpunkten des Simplex. Für stetige Beobachtungen werden wir eine analoge Formel für die Einflussfunktion im Abschnitt 6.2.4 antreffen.

6.2.3 M-Schätzer

Definition

Unter einem M-Schätzer verstehen wir einen Schätzer, der definiert ist entweder durch ein Minimierungsproblem der Form

$$T_n = \arg\min_t \sum_{i=1}^{n} \rho(X_i, t), \tag{6.4}$$

wobei $\rho : \mathbb{X} \times \mathbb{R}^p \to \mathbb{R}$, oder als Lösung eines impliziten Gleichungssystems der Form

$$\sum_{i=1}^{n} \psi(X_i, t) = 0, \tag{6.5}$$

wobei $\psi : \mathbb{X} \times \mathbb{R}^p \to \mathbb{R}^p$. (Man hat gleich viele Gleichungen wie Unbekannte). Wenn man $\arg\min$ durch Ableiten und Nullsetzen bestimmt, kommt man von der Form (6.4) zur Form (6.5). Die Gleichung (6.5) kann aber mehrere Lösungen haben, und für $p > 1$ muss ψ nicht ein Gradient einer Funktion ρ sein. Zu den M-Schätzern gehört insbesondere der Maximum Likelihood Schätzer (MLE), bei dem $\rho(x, \theta) = -\log p_\theta(x)$ ist, siehe Kapitel 1.3.2, (die Terminologie soll an diesen wichtigen Spezialfall erinnern). Ein anderes Beispiel ist der Huber-Schätzer aus Kapitel 1.1.2. Um die beiden Formen zu unterscheiden, nennt man einen Schätzer, der als Lösung von (6.5) definiert ist, auch einen Z-Schätzer (Z für zero).

Den Schätzer (6.4) können wir auch schreiben als

$$T_n = \arg\min_t \frac{1}{n} \sum_{i=1}^n \rho(X_i, t), \tag{6.6}$$

und ebenso können wir in (6.5) die Summe durch ein arithmetisches Mittel ersetzen. Das heisst, T_n ist eine Funktion von arithmetischen Mitteln, aber im Unterschied zu Abschnitt 6.2.2 sind es jetzt unendlich viele arithmetische Mittel (eines für jedes t).

Konsistenz

Wir gehen von der Form (6.6) aus. Wenn die X_i i.i.d sind mit Verteilung P_0, dann folgt aus dem Gesetz der grossen Zahlen, dass für jedes feste t $\frac{1}{n}\sum_{i=1}^n \rho(X_i, t)$ gegen

$$r(t) = r(t, P_0) = E_0[\rho(X_i, t)] = \int \rho(x, t) P_0(dx)$$

konvergiert. Wir erwarten also, dass T_n gegen $\arg\min_t r(t)$ konvergiert, unabhängig davon, ob das Modell korrekt ist oder nicht. Für einen Beweis dieser Konvergenz müssen wir die Operationen $\arg\min_t$ und $\lim_n$ vertauschen können, und dazu brauchen wir Regularitätsbedingungen, die wir unten diskutieren. Wenn das Modell korrekt ist, brauchen wir neben den mehr technischen Regularitätsbedingungen die folgende wesentliche Bedingung für Konsistenz

$$\arg\min_t r(t, P_\theta) = g(\theta) \quad \forall \theta. \tag{6.7}$$

Im Fall des MLE zeigt das folgende Lemma, dass die Bedingung (6.7) automatisch erfüllt ist für $g(\theta) = \theta$, wenn zu verschiedenen Parametern auch verschiedene Modelle gehören (sonst kann man keine Konsistenz erwarten) :

Lemma 6.1. *Seien $F_i(dx) = f_i(x)\mu(dx)$ für $i = 1, 2$. Dann gilt*

$$E_1[\log f_1(X)] \geq E_1[\log f_2(X)],$$

und Gleichheit gilt genau dann, wenn $F_1 = F_2$.

Beweis. Da $\log y \leq y - 1$, ist

$$E_1[\log f_2(X)] - E_1[\log f_1(X)] = E_1[\log(\frac{f_2(X)}{f_1(X)})] \leq E_1[\frac{f_2(X)}{f_1(X)} - 1] = 1 - 1 = 0.$$

Zudem ist die Ungleichung strikt, ausser falls $f_2(x) = f_1(x)$ f.s.. □

Wenn das Modell nicht korrekt ist, die wahre Verteilung aber auch eine Dichte p_0 bezüglich μ hat, dann konvergiert der MLE gegen

$$\arg\min_\theta \int -\log p_\theta(x) P_0(dx) = \arg\min_\theta \int \log(p_0(x)/p_\theta(x)) p_0(x) d\mu(x). \tag{6.8}$$

Das letzte Integral spielt eine Rolle in der Informationstheorie und heisst die Kullback-Leibler Divergenz oder relative Entropie $H(P_0 : P_\theta)$. Gemäss obigem Lemma ist diese stets positiv und wird daher manchmal auch als Abstand interpretiert (allerdings gilt weder

Symmetrie noch Dreiecksungleichung). Der MLE bestimmt also asymptotisch diejenige Modellverteilung, die im Sinne der Kullback-Leibler Divergenz am nächsten bei der wahren Verteilung ist.

Wir kommen nun zu den Regularitätsbedingungen, unter denen die Vertauschung von Grenzwert und Bilden des Minimums gerechtfertigt werden kann. Der folgende Satz leistet dies unter minimalen Glattheitsbedingungen an ρ, nimmt aber dafür an, dass T_n nur in einem kompakten Bereich variiert.

Satz 6.6. *Sei T_n definiert durch (6.4). Für jedes t und jedes x gelte*

$$\liminf_{t_k \to t} \rho(x, t_k) \geq \rho(x, t)$$

(d.h. $\rho(x, .)$ ist unten halbstetig) und für eine genügend kleine Umgebung U von t sei

$$E_0[\inf_{t' \in U} \rho(X_i, t')] > -\infty.$$

Ferner soll $\arg\min_t r(t)$ aus genau einem Punkt t_0 bestehen und es soll eine kompakte Menge $K \ni t_0$ existieren, so dass mit Wahrscheinlichkeit 1 T_n schliesslich in K liegt. Dann konvergiert T_n f.s. gegen t_0.

Beweis. Zur Abkürzung setzen wir für ein offenes U

$$\rho_U(x) := \inf_{t \in U} \rho(x, t).$$

Wir betrachten ein $t \in K$ und eine abnehmende Folge von Umgebungen U_k von t, deren Durchschnitt gleich t ist. Dann ist ρ_{U_k} eine monoton wachsende Folge, die wegen der ersten Voraussetzung gegen $\rho(x, t)$ konvergiert. Aus der zweiten Voraussetzung und dem monotonen Konvergenzsatz folgt ferner, dass

$$E_0[\rho_{U_k}(X_i)] \to E_0[\rho(X_i, t)] = r(t).$$

Für $t \neq t_0$ ist nach Voraussetzung $r(t) > r(t_0)$. Aus der Überlegung von oben sehen wir daher, dass eine offene Umgebung $U = U_t$ von t existiert mit $E_0[\rho_{U_t}(X_i)] > r(t_0)$. Die Menge $B = \{t \in K; ||t - t_0|| \geq \varepsilon\}$ ist kompakt und wird überdeckt von der Vereinigung aller U_t, $t \in B$. Also können wir B mit endlich vielen dieser Umgebungen überdecken:

$$B \subseteq \cup_{j=1}^{M} U_{t_j}.$$

Mit dem Gesetz der grossen Zahlen folgt daher, dass f.s.

$$\frac{1}{n} \inf_{t \in B} \sum_{i=1}^{n} \rho(X_i, t) \geq \frac{1}{n} \inf_{j=1,\ldots M} \sum_{i=1}^{n} \rho_{U_{t_j}}(X_i) \to \inf_{j=1,\ldots M} E_0[\rho_{U_{t_j}}(X_i)] > r(t_0).$$

Nach Definition von T_n ist ferner

$$\frac{1}{n} \sum_{i=1}^{n} \rho(X_i, T_n) \leq \frac{1}{n} \sum_{i=1}^{n} \rho(X_i, t_0) \to r(t_0).$$

Daher kann T_n nicht mehr zu B gehören für n gross genug. Nach Voraussetzung ist aber $T_n \in K$ und damit muss $||t - t_0|| < \varepsilon$ gelten für n gross genug. □

Wenn der Schätzer die Form (6.5) hat, dann erwarten wir aus analogen Überlegungen, dass T_n gegen eine Lösung von

$$\int \psi(x,t)P_0(dx) = 0$$

konvergiert. Damit ist die wesentliche Bedingung für Konsistenz offenbar

$$\int \psi(x, g(\theta))P_\theta(dx) = 0.$$

In Abbildung 6.1 sind die beiden Funktionen $n^{-1}\sum \psi(x_i, .)$ und $\int \psi(x, .)P_0(dx)$ gezeichnet für das ψ, das zum Huber-Schätzer mit $c = 1$ gehört, und $n = 10$ Beobachtungen, die gemäss $P_0 = \mathcal{N}(0,1)$ erzeugt wurden. Hier ist $\int \psi(x,0)P_0(dx) = 0$ aus Symmetriegründen erfüllt.

Für $p = 1$ kann man die Vertauschung von Grenzwert und Lösen der impliziten Gleichung leicht rechtfertigen:

Satz 6.7. *Sei $\psi(x, .)$ stetig für alle x und $E_0[|\psi(X_i,t)|] < \infty$ für alle t. Wenn $E_0[\psi(X_i,t)] > 0$ für $t_0 < t < t_0 + \delta$ und $E_0[\psi(X_i,t)] < 0$ für $t_0 - \delta < t < t_0$, dann existiert eine Folge T_n, die f.s. gegen t_0 konvergiert und die Gleichung $\sum \psi(X_i,t) = 0$ löst.*

Beweis. Sei $0 < \varepsilon < \delta$ gegeben. Wegen des Gesetzes der grossen Zahlen ist für n gross genug $\sum \psi(X_i, t_0 + \varepsilon) > 0$ und $\sum \psi(X_i, t_0 - \varepsilon) < 0$. Aus dem Zwischenwertsatz folgt also, dass es eine Lösung von $\sum \psi(X_i,t) = 0$ im Intervall $(t_0 - \varepsilon, t_0 + \varepsilon)$ gibt. ☐

Wenn (6.5) mehrere Lösungen hat, dann sagt uns der Satz leider nicht, welche Lösung man nehmen muss, um Konsistenz zu erhalten.

In der Literatur gibt es noch weitere Varianten von Regularitätsbedingungen, die eine Vertauschung von Limes und arg min, bzw. Limes und Nullstellensuche erlauben. Wir verzichten auf die Details.

Asymptotische Normalität

Wir gehen von der Form (6.5) aus und nehmen an, dass $\psi(x, .)$ differenzierbar ist für alle x. Dann folgt aus $\sum_{i=1}^n \psi(X_i, T_n) = 0$ mit einer Taylorentwicklung an der Stelle t_0

$$0 = \frac{1}{n}\sum_{i=1}^n \psi(X_i, t_0) + \frac{1}{n}\sum_{i=1}^n \frac{\partial\psi}{\partial t}(X_i, \tau_n)\ (T_n - t_0).$$

(ψ und t sind Vektoren, die partiellen Ableitungen bilden eine Matrix). Dabei ist $\|\tau_n - t_0\| \leq \|T_n - t_0\|$. Wenn T_n gegen t_0 konvergiert, dann ist es plausibel, dass $M_n = \frac{1}{n}\sum_{i=1}^n \frac{\partial\psi}{\partial t}(X_i, \tau_n)$ stochastisch gegen

$$M_0 = E_0\left[\frac{\partial\psi}{\partial t}(X_i, t)\bigg|_{t=t_0}\right] \tag{6.9}$$

konvergiert. Wenn M_0 noch invertierbar ist, dann gilt die Approximation (6.1) mit

$$IF(X_i, P_0) = -M_0^{-1}\psi(X_i, t_0), \tag{6.10}$$

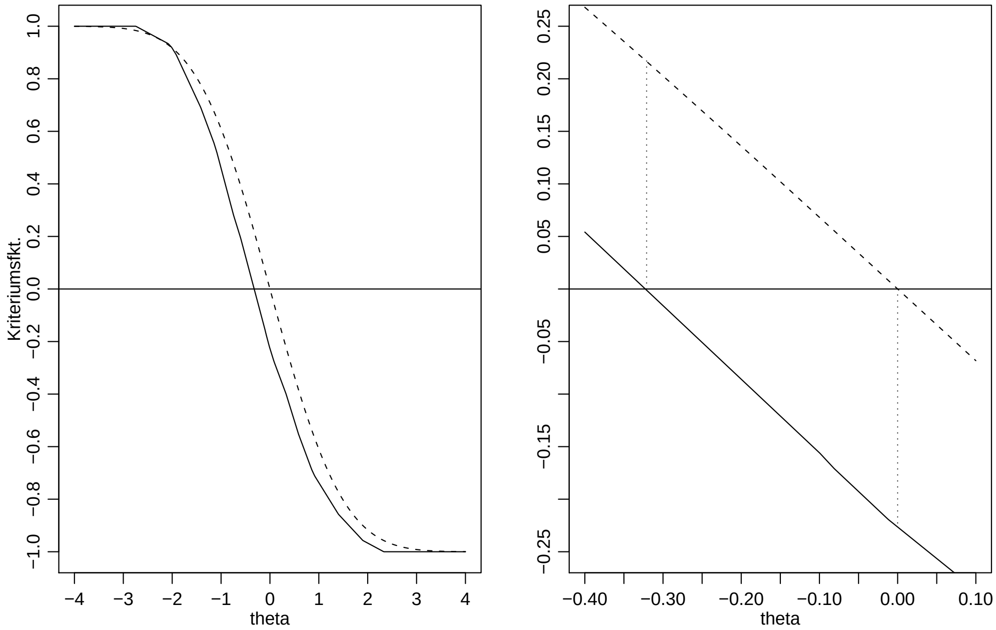


Abbildung 6.1: Illustration der Asymptotik des M-Schätzers. Es sind die beiden Funktionen $n^{-1}\sum\psi(x_i,.)$ (durchgezogen) und $\int\psi(x,.)P_0(dx)$ (gestrichelt) gezeichnet für das ψ, das zum Huber-Schätzer mit $c=1$ gehört, und $n=10$ Beobachtungen, die gemäss $P_0=\mathcal{N}(0,1)$ erzeugt wurden. Die Figur rechts ist ein vergrösserter Ausschnitt der Figur links.

und demnach

$$T_n \overset{as}{\sim} \mathcal{N}(t_0, \frac{1}{n}M_0^{-1}E_0[\psi(X_i,t_0)\psi(X_i,t_0)^T]M_0^{-T}). \tag{6.11}$$

Diese heuristischen Überlegungen sind im rechten Teil von Abbildung 6.1 illustriert.

Ein einfacher Beweis dieser Vermutung ergibt sich unter den folgenden Bedingungen.

Satz 6.8. *Sei $t\to\psi(x,t_0)$ differenzierbar für alle x in einer Umgebung von t_0 mit integrierbaren partiellen Ableitungen, sei M_0 invertierbar und es existiere ferner eine Funktion H mit $E_0[H(X_i)]<\infty$ so dass für alle i,j*

$$\left|\frac{\partial\psi_i}{\partial t_j}(x,t)-\frac{\partial\psi_i}{\partial t_j}(x,t')\right|\le H(x)\left\|t-t'\right\|.$$

Dann gilt (6.11) für jede Folge (T_n), die $\sum\psi(X_i,T_n)=0$ erfüllt und gegen t_0 konvergiert. Ferner ist

$$M_0=\frac{\partial}{\partial t}E_0[\psi(X_i,t)]\bigg|_{t=t_0} \tag{6.12}$$

Beweis.

$$|(M_n)_{kj}-(M_0)_{kj}|\le\left|\frac{1}{n}\sum_{i=1}^n\frac{\partial\psi_k}{\partial t_j}(X_i,t_0)-(M_0)_{kj}\right|+\|T_n-t_0\|\,\frac{1}{n}\sum_{i=1}^n H(X_i).$$

Die rechte Seite geht stochastisch gegen null wegen des Gesetzes der grossen Zahlen und der Konvergenz von T_n. Die Behauptung folgt also aus dem Satz von Slutsky und der obigen Taylorapproximation. Die Vertauschung von Erwartungswert und Ableitung folgt mit dem Konvergenzsatz von Lebesgue. □

Bemerkung 6.1. *Die Form (6.12) für M_0 macht auch Sinn, wenn $\psi(x,.)$ an ein paar wenigen Stellen nicht differenzierbar ist (z.B. beim Huber-Schätzer), weil der Erwartungswert diese Stellen ausglättet. Tatsächlich gilt die asymptotische Normalität auch unter weniger Glattheitsbedingungen, siehe z.B. Kap. 7.2.2 in Serfling, oder Kap. 5.3 und 19 in van der Vaart.*

Im Fall, wo $P_0 = P_\theta(dx) = p_\theta(x)\mu(dx)$ und $g(\theta) = \theta$ ist, erhalten wir eine dritte Form für $M_0 = M(\theta)$ durch Ableiten von $\int \psi(x,\theta)p_\theta(x)\mu(dx) = 0$:

$$M(\theta) = -\int \psi(x,\theta)(\frac{\partial}{\partial\theta}p_\theta(x))^T\mu(dx) = -E_\theta[\psi(X_i,\theta)s(X_i,\theta)^T], \tag{6.13}$$

wobei s die sogenannte **score-Funktion** *ist*

$$s(x,\theta) = \frac{\partial}{\partial\theta}\log(p_\theta(x)).$$

Beispiel 6.7 (Der Maximum Likelihood Schätzer). *Wenn das Modell korrekt ist, ist er konsistent für $g(\theta) = \theta$ und $\psi(x,\theta) = \frac{\partial}{\partial\theta}\log p_\theta(x) = s(x,\theta)$. Also gilt (vgl. Kapitel 3.3)*

$$M(\theta) = -E_\theta[\psi(X_i,\theta)\psi(X_i,\theta)^T] = -I(\theta),$$

wobei $I(\theta)$ die Fisher-Informationsmatrix ist. Daraus ergibt sich das folgende zentrale Resultat dieses Kapitels

Korollar 6.4. *Wenn (X_i) i.i.d $\sim p_\theta(x)\mu(dx)$, dann ist unter Regularitätsbedingungen an p_θ der Maximum Likelihood Schätzer asymptotisch $\mathcal{N}(\theta, \frac{1}{n}(I(\theta))^{-1})$-verteilt.*

Für präzise Bedingungen und Beweise verweise ich auf van der Vaart, Kap. 5.5-5.6.

Wir können aber auch sagen, was passiert, wenn das Modell nicht zutrifft: Sei die wahre Verteilung $P_0(dx) = p_0(x)\mu(dx)$. Dann konvergiert ja T_n gegen dasjenige θ_0, welches die Kullback-Leibler Divergenz $H(P_\theta : P_0)$ minimiert, siehe (6.8). Unter Regularitätsbedingungen ist ferner

$$T_n \overset{as}{\sim} \mathcal{N}(\theta_0, \frac{1}{n}M(P_0)^{-1}J(P_0)M(P_0)^{-T}),$$

wobei

$$\begin{aligned} M(P_0)_{jk} &= E_0\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log p_\theta(X_i)\Big|_{\theta=\theta_0}\right] \\ J(P_0)_{jk} &= E_0\left[\frac{\partial}{\partial\theta_j}\log p_\theta(X_i)\Big|_{\theta=\theta_0}\frac{\partial}{\partial\theta_k}\log p_\theta(X_i)\Big|_{\theta=\theta_0}\right]. \end{aligned}$$

Jetzt sind $M(P_0)$ und $J(P_0)$ im allgemeinen verschieden.

Beispiel 6.8 (Quantile). *Wenn wir setzen*

$$\psi(x,t) = \begin{cases} 1-\alpha & \text{für} \quad x < t, \\ -\alpha & \text{für} \quad x > t, \end{cases}$$

erhalten wir als zugehörigen M-Schätzer das empirische α-Quantil, der für n gegen unendlich gegen das α-Quantil $t_0 = P_0^{-1}(\alpha)$ konvergiert. Wenn P_0 bezüglich des Lebesguemasses eine Dichte $p_0(x)$ hat, dann ergibt die Formel (6.12)

$$M_0 = \frac{\partial}{\partial t}E_0[\psi(X_i,t)]\Big|_{t=t_0} = \frac{\partial}{\partial t}(P_0(t) - \alpha)\Big|_{t=t_0} = p_0(P_0^{-1}(\alpha)).$$

Die Einflussfunktion des α-Quantils ist also

$$IF(x, P_0^{-1}(\alpha), P_0) = \frac{1}{p_0(P_0^{-1}(\alpha))} \, (\alpha - \mathbf{1}_{[x < P_0^{-1}(\alpha)]}).$$

Ferner gilt

$$E_0[IF(x, P_0^{-1}(\alpha), P_0)^2] = \frac{\alpha(1-\alpha)}{p_0(P_0^{-1}(\alpha))^2}.$$

Das empirische α-Quantil ist tatsächlich asymptotisch normal mit dieser Varianz, siehe z.B. Serfling, Kap. 2.3.3.

6.2.4 Schätzer als Funktionale der empirischen Verteilung

Definition

Wir hatten im Abschnitt 1.3.2 bereits gesehen, wie man mit dem Einsetzprinzip auf natürliche Weise zu Schätzern kommt, die als Funktionale der empirischen Verteilung gegeben sind. Wir repetieren noch einmal die Definition: Die **empirische Verteilung** F_n von n Beobachtungen $X_1, X_2, \ldots, X_n$ ist konzentriert auf diesen n Beobachtungen und gibt jeder das Gewicht $\frac{1}{n}$, d.h.

$$F_n = \frac{1}{n} \sum_{i=1}^{n} \Delta_{X_i},$$

wobei Δ_x die Punktmasse eins (das Dirac-Mass) in $x \in \mathbb{X}$ ist. Für jedes $A \subset \mathbb{X}$ ist also $F_n(A)$ die relative Häufigkeit der Beobachtungen, die in A liegen.

Wir betrachten also in diesem Unterkapitel Schätzfolgen von der Form

$$T_n = Q(F_n),$$

wobei Q ein Funktional auf der Menge $\mathcal{M}$ aller Verteilungen auf $\mathbb{X}$ ist, d.h. $Q : \mathcal{M} \to \mathbb{R}^p$. Da F_n ein unendlich dimensionales arithmetisches Mittel ist, kann jedes solche T_n als eine Funktion von unendlich vielen arithmetischen Mitteln aufgefasst werden.

Die empirische Verteilung ändert sich nicht, wenn man die beobachteten Werte permutiert, und daher gilt das gleiche auch für $Q(F_n)$. Wegen der Suffizienz der geordneten Stichprobe haben auch alle vernünftigen Schätzer in i.i.d. Modellen diese Eigenschaft. In der Tat lassen sich auch fast alle der bisher vorgekommenen Schätzer als ein Funktional der empirischen Verteilung schreiben (ev. bis auf eine für grosse n vernachlässigbare Modifikation):

- Schätzer der Form (6.2): Wähle $Q(F) = h(\int u(x) F(dx))$.
- M-Schätzer: Wähle $Q(F) = \arg\min \int \rho(x,t) F(dx)$, bzw. implizit $Q(F)$ Lösung von $\int \psi(x,t) F(dx) = 0$.
- Gestutztes Mittel: Wähle

$$Q(F) = \frac{1}{1-2\alpha} \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} x F(dx).$$

$Q(F_n)$ ist bis auf Diskretisierungseffekte gleich dem in I.1.2 definierten Schätzer. Wenn F eine strikt positive Dichte hat, dann erhält man mit der Substitution $u = F(x)$ die alternative Form

$$Q(F) = \frac{1}{1-2\alpha} \int_{\alpha}^{1-\alpha} F^{-1}(u)du$$

(im Allgemeinen unterscheiden sich die beiden Formen leicht). Dies ist ein Beispiel, das weder zu den M-Schätzern noch zu den Funktionen von endlich vielen arithmetischen Mitteln gehört.

Konsistenz

Wegen des Gesetzes der grossen Zahlen konvergiert $F_n(A)$ für jedes feste A f.s. gegen $P_0[A]$. Wenn wir sogar zeigen können, dass F_n gegen P_0 konvergiert in einer geeigneten Metrik auf dem Raum der Verteilungen und wenn das Funktional Q stetig ist bezüglich dieser Metrik, dann folgt sofort, dass $T_n = Q(F_n)$ gegen $Q(P_0)$ konvergiert. Der folgende Satz zeigt die Konvergenz von F_n gegen P_0 bezüglich der Supremum-Norm im Fall $\mathbb{X} = \mathbb{R}$.

Satz 6.9 (Glivenko-Cantelli). *Seien $X_i, i = 1, 2\dots$ reellwertig, i.i.d. $\sim F$ und F_n sei die empirische Verteilungsfunktion. Dann gilt* $\sup_x |F_n(x) - F(x)| \to 0$ *f.s..*

Beweis. Wir schreiben $F(b-)$ für $\lim_{h\downarrow 0} F(b-h) = P[X_i < b]$. Sei $a < b$ fest. Wegen der Monotonie von F_n und F gilt für $a \leq x < b$

$$\begin{aligned} F_n(x) - F(x) &\leq (F_n(b-) - F(b-)) + (F(b-) - F(a)), \\ F_n(x) - F(x) &\geq (F_n(a) - F(a)) - (F(b-) - F(a)). \end{aligned}$$

Also für $a \leq x < b$

$$|F_n(x) - F(x)| \leq \max(|F_n(a) - F(a)|, |F_n(b-) - F(b-)|) + (F(b-) - F(a)).$$

Für gegebenes $\varepsilon > 0$ kann man $a_1 = -\infty < a_2 < \dots < a_M = \infty$ so wählen, dass $F(a_i-) - F(a_{i-1}) \leq \varepsilon$. Jetzt wenden wir obige Ungleichung an mit $a = a_i, b = a_{i+1}$. Dies ergibt mit dem Gesetz der grossen Zahlen, dass f.s. für n gross genug

$$\begin{aligned} \sup_x |F_n(x) - F(x)| &\leq \max_{1<i<M} |F_n(a_i) - F(a_i)| \\ &+ \max_{1<i<M} |F_n(a_i-) - F(a_i-)| + \varepsilon \leq 3\varepsilon. \end{aligned}$$

□

Als Anwendung zeigen wir die Konsistenz des Stichprobenmedians:

Beispiel 6.9. *Sei $Q(F)$ = Median von F (Definition beliebig, falls mehrdeutig). Wir zeigen, dass Q stetig an der Stelle F ist, wenn der Median von F eindeutig ist, d.h. wenn $F(x) < 0.5$ für $x < Q(F)$ und $F(x) > 0.5$ für $x > Q(F)$: Wenn*

$$\|G - F\|_\infty < \min(F(Q(F) + \varepsilon) - 0.5, 0.5 - F(Q(F) - \varepsilon)),$$

dann ist $G(Q(F) + \varepsilon) > 0.5$ und $G(Q(F) - \varepsilon) < 0.5$, also $|Q(G) - Q(F)| \leq \varepsilon$. Also konvergiert der Stichprobenmedian gegen den Median von P_0, falls dieser eindeutig ist.

Wenn das parametrische Modell korrekt ist, brauchen wir für die Konsistenz von $Q(F_n)$ für $g(\theta)$ die folgende wesentliche Bedingung, die sogenannte **Fisher-Konsistenz**

$$Q(P_\theta) = g(\theta)\ \forall\theta.$$

Asymptotische Normalität

Wenn $T_n = Q(F_n)$ gegen $Q(P_0)$ konvergiert, ist es naheliegend die Differenz $Q(F_n) - Q(P_0)$ mit Hilfe einer Taylorentwicklung 1. Ordnung zu untersuchen:

$$Q(F_n) - Q(P_0) \approx (\text{Ableitung von } Q \text{ an der Stelle } P_0) \cdot (F_n - P_0).$$

Genau dies war auch das Vorgehen im Fall, wo T_n eine Funktion von endlich vielen arithmetischen Mitteln war, vgl. Satz 6.5.

Dazu müssen wir uns Gedanken machen über Ableitungen in unendlich dimensionalen Räumen. In der abstrakten Analysis ist die Ableitung ein Element aus dem Dualraum, und der Punkt in der obigen Formel soll das innere Produkt andeuten. Unsere Funktionale sind definiert auf dem Raum $\mathcal{M}$ der Wahrscheinlichkeitsverteilungen auf $\mathbb{X}$, und ein lineares Funktional auf diesem Raum hat die Form $P \to \int f(x)P(dx)$ mit einem festen f. Damit die Addition und Subtraktion von Argumenten der Funktion definiert ist, betrachtet man Ableitungen normalerweise für Funktionen auf einem Vektorraum. Wir können zwei Wahrscheinlichkeitsverteilungen zwar nicht addieren, aber konvex kombinieren, was genügt, um Ableitungen zu definieren.

Definition 6.4. *Sei Q ein Funktional auf der Menge der Wahrscheinlichkeitsverteilungen auf $\mathbb{X}$ mit Werten in $\mathbb{R}^p$. Die* **Einflussfunktion** *von Q an der Stelle F, $IF(.,Q,F) : \mathbb{X} \to \mathbb{R}^p$, ist definiert als*

$$IF(x,Q,F) = \lim_{\varepsilon \downarrow 0} \frac{Q((1-\varepsilon)F + \varepsilon\Delta_x) - Q(F)}{\varepsilon},$$

falls der Limes existiert. Q heisst **Gâteaux-ableitbar** *an der Stelle F falls*

$$\lim_{\varepsilon \downarrow 0} \frac{Q((1-\varepsilon)F + \varepsilon G) - Q(F)}{\varepsilon}$$

exisitiert und gleich $\int IF(x,Q,F)G(dx)$ ist für alle G. Q heisst **Fréchet-ableitbar** *an der Stelle F bez. einer Metrik d falls*

$$Q(G) - Q(F) = \int IF(x,Q,F)G(dx) + o(d(F,G))$$

Bemerkung 6.2.

- *Im $\mathbb{R}^2$ entspricht die Einflussfunktion den partiellen Ableitungen (die Punktmassen entsprechen also den Basisrichtungen). Gâteaux-Ableitbarkeit entspricht der Existenz von allen Richtungsableitungen (welche sich in $\mathbb{R}^2$ als Linearkombinationen der beiden partiellen Ableitungen darstellen lassen), und Fréchet-Ableitbarkeit der Differenzierbarkeit.*
- *Falls Q Gâteaux-ableitbar ist, dann existiert die Einflussfunktion.*
- *Falls $d((1-\varepsilon)F + \varepsilon G, F) = O(\varepsilon)$, dann impliziert Fréchet-Ableitbarkeit Gâteaux-Ableitbarkeit.*
- *Wenn Q Gâteaux-ableitbar ist, dann $\int IF(x,Q,F)F(dx) = 0$.*

Wir werden gleich sehen, dass die hier definierte Einflussfunktion tatsächlich auftritt in der Approximation (6.1). Wir wollen vorher aber noch die Berechnung an Beispielen illustrieren.

Beispiel 6.10 (M-Schätzer). *Wir betrachten einen Schätzer der Form (6.5) und setzen*

$$h(\varepsilon,t) = (1-\varepsilon)\int \psi(x,t)F(dx) + \varepsilon \int \psi(x,t)G(dx).$$

Wir nehmen an, dass $h(0,t)=0$ *eine Lösung* $t=Q(F)$ *hat und*

$$M(\psi,F) = \left(\frac{\partial}{\partial t_j} \int \psi_i(x,t)F(dx)\Big|_{t=Q(F)} \right)_{1\le i,j\le p}$$

nicht singulär ist. Der Satz über implizite Funktionen zeigt, dass dann für ε *klein genug* $h(\varepsilon,t)=0$ *eine Lösung hat, und wir definieren* $Q((1-\varepsilon)F+\varepsilon G)$ *als diese Lösung. Mit dem Satz über implizite Funktionen folgt zudem noch*

$$\frac{d}{d\varepsilon}Q((1-\varepsilon)F+\varepsilon G) = -M(\psi,F)^{-1}\int \psi(x,Q(F))G(dx).$$

Damit ist also

$$IF(x,Q,F) = -M(\psi,F)^{-1}\psi(x,Q(F)),$$

in Übereinstimmung mit (6.10) und (6.12).

Beispiel 6.11 (Gestutztes Mittel). *Wenn wir von der Form*

$$Q(F) = \frac{1}{1-2\alpha}\int_\alpha^{1-\alpha} F^{-1}(u)du,$$

ausgehen, dann erhalten wir durch Vertauschen von Integral und Ableitung

$$\frac{d}{d\varepsilon}Q((1-\varepsilon)F+\varepsilon\Delta_x) = \frac{1}{1-2\alpha}\int_\alpha^{1-\alpha} \frac{d}{d\varepsilon}((1-\varepsilon)F+\varepsilon\Delta_x)^{-1}(u)du.$$

Der Integrand ist gerade die Einflussfunktion des u*-Quantils, welche im Fall, dass* F *eine strikt positive Dichte* f *hat, die folgende Form hat (vgl. Beispiel 6.8)*

$$\frac{1}{f(F^{-1}(u))}(u - \mathbf{1}_{[x<F^{-1}(u)]}) = \frac{1}{f(F^{-1}(u))}(u - \mathbf{1}_{[F(x)<u]}),$$

Daraus folgt

$$(1-2\alpha)IF(x,Q,F) = \int_\alpha^{1-\alpha} \frac{u}{f(F^{-1}(u))}du - \mathbf{1}_{[F(x)<1-\alpha]}\int_{\max(F(x),\alpha)}^{1-\alpha} \frac{1}{f(F^{-1}(u))}du,$$

bzw. mit der Substitution $y=F^{-1}(u)$

$$(1-2\alpha)IF(x,Q,F) = \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} F(y)dy + \mathbf{1}_{[F(x)<1-\alpha]}(\max(x,F^{-1}(\alpha)) - F^{-1}(1-\alpha)).$$

Mit partieller Integration erhalten wir für das Integral rechts

$$\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} F(y)dy = (1-\alpha)F^{-1}(1-\alpha) - \alpha F^{-1}(\alpha) - \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} yf(y)dy,$$

während der zweite Term rechts gleich

$$\min(F^{-1}(1-\alpha), \max(x,F^{-1}(\alpha))) - F^{-1}(1-\alpha)$$

ist. Damit haben wir gezeigt, dass

$$IF(x,Q,F) = \begin{cases} (F^{-1}(\alpha) - W(F))/(1-2\alpha) & \text{für} \quad x < F^{-1}(\alpha) \\ (x - W(F))/(1-2\alpha) & \text{für} \quad F^{-1}(\alpha) < x < F^{-1}(1-\alpha) \\ (F^{-1}(1-\alpha) - W(F))/(1-2\alpha) & \text{für} \quad x > F^{-1}(1-\alpha) \end{cases}$$

ist, wobei $W(F) = (1-2\alpha)Q(F) + \alpha F^{-1}(\alpha) + \alpha F^{-1}(1-\alpha)$. *Insbesondere ist für symmetrisches* F *(d.h.* $F(\mu + x) = 1 - F(\mu - x)$ *für ein* μ*) das gestutzte Mittel konsistent für* μ *und hat die gleiche Einflussfunktion wie der Huberschätzer mit* $c = F^{-1}(1-\alpha)$.

Wenn wir von der Form

$$Q(F) = \frac{1}{1-2\alpha} \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} xF(dx)$$

ausgehen, dann können wir schreiben

$$Q((1-\varepsilon)F + \varepsilon G) = h(((1-\varepsilon)F + \varepsilon G)^{-1}(\alpha), ((1-\varepsilon)F + \varepsilon G)^{-1}(1-\alpha), \varepsilon),$$

wobei

$$h(u,v,\varepsilon) = \frac{1}{1-2\alpha}((1-\varepsilon)\int_u^v xF(dx) + \varepsilon \int_u^v xG(dx).$$

Wenn man für G *eine absolut stetige Verteilung statt einer Punktmasse wählt, ist* h *differenzierbar, so dass man die Kettenregel anwenden kann. Man erhält man dann das gleiche Ergebnis wie oben.*

Wir wenden nun die Ableitungsbegriffe an auf das Problem der asymptotischen Normalität. Wir nehmen an, dass Q stetig ist und eine Einflussfunktion besitzt an der Stelle P_0, und dass bezüglich einer geeigneten Norm d F_n gegen P_0 konvergiert. Dann können wir schreiben

$$T_n - Q(P_0) = Q(F_n) - Q(P_0) = \int IF(x,Q,P_0)F_n(dx) + R_n = \frac{1}{n}\sum_{i=1}^n IF(X_i,Q,P_0) + R_n.$$

Wir haben also genau die Approximation (6.1): Man beachte, dass $E_0[IF(X_i,Q,P_0)] = 0$ gilt, und den Restterm R_n kann man abschätzen mit Fréchet-Ableitbarkeit und Grenzwertsätzen für $d(F_n, P_\theta)$.

Satz 6.10. *Falls* Q *Fréchet-ableitbar an der Stelle* P_0 *ist,* $E_0[\|IF(X_i,Q,P_0\|^2] < \infty$ *und* $d(F_n,P_0) = O_p(1/\sqrt{n})$, *dann ist* $T_n = Q(F_n) \overset{as}{\sim} \mathcal{N}(Q(P_0), \frac{1}{n}V_0)$, *wobei*

$$V_0 = E_0[IF(X_i,Q,P_0)IF(X_i,Q,P_0)^T].$$

Beweis. Nach Definition von Fréchet-ableitbar gilt für den Restterm R_n oben $R_n = o(d(F_n,P_0)) = o(O_P(1/\sqrt{n})) = o_P(1/\sqrt{n})$. Also folgt die Behauptung aus dem Zentralen Grenzwertsatz und dem Satz von Slutsky. □

Wann sind die Voraussetzungen dieses Satzes erfüllt ? Im Fall $\mathbb{X} = \mathbb{R}$ weiss man, dass $\sqrt{n}\sup_x |F_n(x) - P_0(x)|$ in Verteilung konvergiert und man kennt auch die Grenzverteilung (welche für uns hier nicht wichtig ist). Daraus folgt, dass $d(F_n,P_0) = O_p(1/\sqrt{n})$ ist für den zur supremum-Norm gehörenden Abstand.

Satz 6.11 (Kolmogorov). *Seien $X_1, X_2, \ldots$ i.i.d. $\sim F$ mit F stetig. Dann gilt für alle $a > 0$*

$$\lim P[\sqrt{n} \sup_x |F_n(x) - F(x)| \le a] = 1 - 2\sum_{j=1}^{\infty}(-1)^{j+1}\exp(-2j^2a^2).$$

Beweis. Da $F(X_i)$ uniform verteilt ist, genügt es den Fall zu betrachten, wo F die uniforme Verteilung ist. Der Beweis beruht auf der Konvergenz des "stochastischen Prozesses" $n^{1/2}(F_n(t) - t)$ in Verteilung gegen $W(t) - tW(1)$, wobei $W(.)$ die "Brown'sche Bewegung" ist. Die Argumente sind aber zu lang für dieses Skript. Ich verweise auf van der Vaart, Kapitel 19. □

Die Fréchet-Differenzierbarkeit ist leider oft schwierig nachzuweisen oder gar nicht erfüllt. Es gibt noch den Begriff der kompakten Differenzierbarkeit, der zwischen Gâteaux- und Fréchet-Differenzierbarkeit liegt und für viele Probleme besser geeignet ist, siehe L. Fernholz, von Mises Calculus for Statistical Functionals, Springer Lecture Notes in Statistics 19, 1983. Der Zugang über Funktionale und deren Ableitung ist aber vor allem als heuristisches Werkzeug wichtig. Die Einflussfunktion lässt sich meist leicht berechnen, und damit hat man zumindest eine Vermutung, wie die asymptotische Varianz aussehen wird. Oft ist es dann einfacher, direkt die Approximation (6.1) statt der Differenzierbarkeit zu beweisen. Im Fall der M-Schätzer hatten wir das in Satz 6.8 getan. Die verschiedenen Formeln für die asymptotische Varianz stimmen überein, wenn man die partielle Ableitung und den Erwartungswert bei der Definition von M_0 vertauscht.

6.3 Anwendungen der asymptotischen Normalität

6.3.1 Effizienz von Schätzern

Relative Effizienz

Zunächst betrachten wir den Fall $p = 1$. Wir hatten bereits früher gesagt, dass ein asymptotisch normaler Schätzer umso besser ist, je kleiner seine asymptotische Varianz ist. Wir können das noch etwas präziser fassen, indem wir das Risiko anschauen. Wenn $L_n(\theta, a) = L(\sqrt{n}(a - g(\theta)))$ mit L beschränkt und f.s stetig bezüglich des Lebesguemasses, dann konvergiert $R(\theta, T_n)$ gegen $\int L(z\sqrt{V(\theta)})\phi(z)dz$ (Satz 6.1). Für L symmetrisch und monoton wachsend auf der positiven Halbachse ist das asymptotische Risiko daher monoton wachsend in $V(\theta)$.

Die asymptotischen Varianzen bilden also die Basis für den Vergleich von 2 Schätzfolgen $(T_n^{(1)})$ und $(T_n^{(2)})$ mit $T_n^{(i)} \overset{as}{\sim} \mathcal{N}(g(\theta), \frac{1}{n}V_i(\theta))$ $(i = 1, 2)$, unabhängig von der Verlustfunktion. Etwas genauer können wir sagen, dass $T_{n_1}^{(1)}$ und $T_{n_2}^{(2)}$ ungefähr gleich genau sind, wenn $V_1(\theta)/n_1 = V_2(\theta)/n_2$. In andern Worten $T^{(2)}$ braucht $V_2(\theta)/V_1(\theta)$-mal soviele Beobachtungen wie $T^{(1)}$, wenn der wahre Parameter θ ist. Daher definiert man die **asymptotische relative Effizienz** von $T_n^{(2)}$ bezüglich $T_n^{(1)}$ als

$$e_{2:1}(\theta) = \frac{V_1(\theta)}{V_2(\theta)}.$$

Beispiel 6.12. *Sei $X_i = \theta + \varepsilon_i$ mit ε_i i.i.d. $\sim (1-\eta)\mathcal{N}(0,1) + \eta\mathcal{N}(0,9)$, d.h. mit Wahrscheinlichkeit $1-\eta$ hat man einen normalen Fehler, und mit Wahrscheinlichkeit η einen Ausreisser mit dreimal so grosser Standardabweichung. Wir wollen das gewöhnliche arithmetische Mittel $T_n^{(1)} = \overline{X}$ mit dem α-gestutzten Mittel $T_n^{(2)}$ vergleichen. Die Formeln für die asymptotischen Varianzen haben wir im vorhergehenden Abschnitt hergeleitet, so dass es klar sein sollte, was man rechnen muss für die Effizienz $e_{2:1}(\eta)$. Das Ergebnis ist in der folgenden Tabelle zusammengefasst.*

	$\alpha = 0.05$	0.125	0.5
$\eta = 0.00$	0.99	0.94	0.64
0.05	1.20	1.19	0.83
0.25	1.40	1.66	1.33

Mit einem Wert $\alpha = 0.125$ verliert man also wenig und gewinnt unter Umständen viel.

Für $p > 1$ hat man die bekannte Schwierigkeit, dass zwei positiv definite Matrizen nicht vergleichbar sind. In diesem Fall lässt sich die asymptotische relative Effizienz im Allgemeinen nicht definieren.

Asymptotische Cramér-Rao-Ungleichung

Wir betrachten den Fall $g(\theta) = \theta,\ p \geq 1$ und suchen den besten Schätzer überhaupt, d.h. wir wollen $V(\theta)$ minimieren, simultan für alle θ's (wobei a priori nicht klar ist, ob das geht). Wenn $T_n \overset{as}{\sim} \mathcal{N}(\theta, \frac{1}{n}V(\theta))$, so legt das nahe zu vermuten, dass

$$E_\theta[T_n] \approx \theta, \quad Var_\theta(T_n) \approx \frac{V(\theta)}{n}.$$

Aus der Cramér-Rao-Ungleichung würde dann folgen

$$V(\theta) \geq I(\theta)^{-1} \tag{6.14}$$

(im Sinne, dass die Differenz $V(\theta) - I(\theta)^{-1}$ positiv semidefinit ist). Einen Schätzer, bei dem Gleichheit gilt für alle θ's, nennen wir **asymptotisch effizient**. Unter Regularitätsbedingungen ist der MLE also asymptotisch effizient.

Obiges Argument für (6.14) ist aber nicht stichhaltig. Erstens folgt aus der Konvergenz in Verteilung nicht immer die Konvergenz der Momente (vgl. Satz 6.2), und zweitens müsste man noch die Vertauschung von Ableitung und Grenzwert rechtfertigen. Es gibt auch tatsächlich Beispiele, wo (6.14) verletzt ist (sogenannte **Supereffizienz**).

Beispiel 6.13 (Hodges-Lehmann). *Sei (X_i) i.i.d. $\sim \mathcal{N}(\theta, 1)$ und*

$$T_n = \begin{cases} \overline{X}_n & \textit{falls } |\overline{X}_n| \geq n^{-1/4} \\ \frac{1}{2}\overline{X}_n & \textit{sonst.} \end{cases}$$

Wenn wir $U(n,\theta) = \sqrt{n}(\overline{X}_n - \theta)$ und $A(n,\theta) = [-\theta n^{1/2} - n^{1/4}, -\theta n^{1/2} + n^{1/4}]$ setzen, dann können wir schreiben

$$\sqrt{n}(T_n - \theta) = \begin{cases} U(n,\theta) & \textit{falls} \quad U(n,\theta) \notin A(n,\theta) \\ \frac{1}{2}U(n,\theta) - \frac{1}{2}\sqrt{n}\theta & \textit{falls} \quad U(n,\theta) \in A(n,\theta). \end{cases}$$

Da $U(n,\theta) \sim \mathcal{N}(0,1)$ und $A(n,\theta) \to \emptyset$ $(\theta \neq 0)$, bzw. $A(n,\theta) \to \mathbb{R}$ $(\theta = 0)$, folgt mit Satz 6.4, dass $\sqrt{n}(T_n - \theta) \xrightarrow{d} \mathcal{N}(0, V(\theta))$, wobei

$$V(\theta) = \begin{cases} 1 & \textit{falls} \quad \theta \neq 0 \\ \frac{1}{4} & \textit{falls} \quad \theta = 0. \end{cases}$$

Man kann sich fragen, wie relevant dieses Beispiel ist. Weil $\overline{X}$ ist zulässig ist, muss es für jedes n ein θ geben, so dass

$$E_\theta[(T_n - \theta)^2] > E_\theta[(\overline{X}_n - \theta)^2] = \frac{1}{n}.$$

In der Tat, wenn wir $\theta_n = h/\sqrt{n}$ wählen, dann $A(n,\theta_n) \to \mathbb{R}$ und damit ist

$$nE_{\theta_n}((T_n - \theta_n)^2) \geq \frac{1}{4} \int_{A(n,\theta_n)} (z-h)^2 \varphi(z) dz \to \frac{1+h^2}{4}.$$

Man erkauft sich also die Verbesserung für $\theta = 0$ mit einer (drastischen) Verschlechterung in der Nähe von $\theta = 0$ und darum sollte dieser Schätzer nicht verwendet werden. Andrerseits ist er ähnlich wie der Steinschätzer, der für $p > 2$ durchaus von praktischer Bedeutung ist.

Es gibt eine ganze Reihe von mathematischen Theorien, die die Gültigkeit von (6.14) unter zusätzlichen Regularitätsbedingungen an (T_n) nachweisen. Wir nehmen an, dass eine Approximation (6.1) existiert. Eine hinreichende Bedingung ist dann

$$E_\theta[IF(X_i,\theta) s(X_i,\theta)^T] = Id_p \tag{6.15}$$

($s(x,\theta)$ ist die score-Funktion $\frac{\partial}{\partial\theta} \log p_\theta(x)$):

Lemma 6.2. *Wenn (6.15) erfüllt ist, dann gilt (6.14).*

Beweis. Aus der Schwarz'schen Ungleichung folgt für beliebige $a, b \in \mathbb{R}^p$:

$$\begin{aligned} (a^T b)^2 &= \operatorname{Cov}_\theta(a^T IF(X_i,\theta), b^T s(X_i,\theta))^2 \\ &\leq \operatorname{Var}_\theta(a^T IF(X_i,\theta)) \operatorname{Var}_\theta(b^T s(X_i,\theta)) = a^T V(\theta) a \;\; b^T I(\theta) b. \end{aligned}$$

Das Lemma folgt, wenn wir $b = I(\theta)^{-1} a$ wählen, vgl. die analogen Argumenten wie bei der mehrdimensionalen Cramér-Rao-Ungleichung. □

Wann ist aber (6.15) erfüllt ? Bei einem konsistenten M-Schätzer folgt dies sofort aus den Formeln (6.10) und (6.13). Wenn ein Funktional Q Fisher-konsistent und Fréchet-differenzierbar ist und wenn $\|P_{\theta'} - P_\theta\| = O(\|\theta' - \theta\|)$, dann gilt

$$h = Q(P_{\theta+h}) - Q(P_\theta) = \int IF(x,Q,P_\theta) \frac{p_{\theta+h}(x) - p_\theta(x)}{p_\theta(x)} p_\theta(x) \mu(dx) + o(\|h\|).$$

Daraus folgt wieder (6.15).

Eine andere Bedingung, die (6.15) impliziert, lautet, dass (T_n) bei festem Parameter θ und bei einem mit dem Stichprobenumfang n variierenden Parameter $\theta_n = \theta + h/\sqrt{n}$ das gleiche asymptotische Verhalten haben soll. Im Beispiel von Hodges-Lehmann haben wir gesehen, dass der supereffiziente Schätzer in der Nähe von $\theta = 0$ einen beliebig grossen systematischen Fehler hat. Der nächste Satz zeigt, dass dies ganz allgemein gilt, wenn 6.15 verletzt ist. In diesem Sinne ist (6.15) eine einleuchtende Regularitätsbedingung.

Satz 6.12 (Le Cam's 3. Lemma). *Sei $\theta_n = \theta + h/\sqrt{n}$ mit beliebigem $h \in \mathbb{R}^p$. Wenn für (T_n) eine Approximation der Form (6.1) gilt und wenn die Likelihoodfunktion die Approximation (6.16) unten erfüllt, dann konvergiert unter P_{θ_n} der standardisierte Schätzer $\sqrt{n}(T_n - \theta_n)$ in Verteilung gegen eine Normalverteilung mit Erwartungswert $(E_\theta[IF(X_i,\theta)s(X_i,\theta)^T] - Id_p)h$ und Varianz $V(\theta)$.*

Beweis. Zur Abkürzung verwenden wir die Notation $\beta = E_\theta[IF(X_i,\theta)s(x_i,\theta)^T]h$, und $\delta^2 = h^T I(\theta)h$. Wegen $\sqrt{n}(T_n - \theta_n) = \sqrt{n}(T_n - \theta) - h$, genügt es zu zeigen, dass für eine stetige und beschränkte Funktion f

$$E_{\theta_n}[f(\sqrt{n}(T_n - \theta))] \to E[f(Z)],$$

wobei Z $\mathcal{N}_p(\beta, V(\theta))$-verteilt ist.

Die Grundidee ist nun die, dass wir den Erwartungswert bezüglich P_{θ_n} umschreiben als Erwartungswert bezüglich P_θ:

$$E_{\theta_n}[f(\sqrt{n}(T_n - \theta))] = E_\theta[f(\sqrt{n}(T_n - \theta))\exp(\Lambda_n)],$$

wobei

$$\Lambda_n = \log\left(\prod_{i=1}^n p_{\theta_n}(X_i)/p_\theta(X_i)\right)$$

der Logarithmus des Likelihoodquotienten ist. Mit einer Taylor-Approximation zweiter Ordnung erhält man

$$\Lambda_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n s(X_i,\theta)^T h + h^T \frac{1}{2n}\sum_{i=1}^n \frac{\partial^2}{\partial\theta\partial\theta^T}\log p_{\tau_n}(X_i)h,$$

wobei $\|\tau_n - \theta\| \le \|h\|/\sqrt{n}$. Unter geeigneten Regularitätsbedingungen folgt daraus, dass

$$\Lambda_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n s(X_i,\theta)^T h - \frac{1}{2}\delta^2 + o_P(1). \tag{6.16}$$

(Die Bedingungen von Satz 6.8 sind hinreichend, aber es gilt auch unter wesentlich schwächeren Bedingungen.)

Mit den Sätzen 6.3 und 6.4 folgt nun aus (6.1) und (6.16), dass unter P_θ der Vektor $(\sqrt{n}(T_n - \theta), \Lambda_n)$ in Verteilung konvergiert gegen einen Zufallsvektor

$$(Y,V) \sim \mathcal{N}_{p+1}\left(\begin{pmatrix} 0 \\ -\delta^2/2 \end{pmatrix}, \begin{pmatrix} V(\theta) & \beta \\ \beta^T & \delta^2/2 \end{pmatrix}\right).$$

Die Funktion $h(y,v) = f(y)\exp(v)$ ist zwar stetig, aber nicht beschränkt. Mit einer kleinen Zusatzüberlegung kann man trotzdem zeigen, dass

$$E_\theta[f(\sqrt{n}(T_n - \theta))\exp(\Lambda_n)] \to E[f(Y)\exp(V)].$$

(Man schneidet die Exponentialfunktion zunächst ab an einer Stelle M und lässt dann M gegen unendlich gehen.) Indem man nun Eigenschaften der mehrdimensionalen Normalverteilung ausnutzt, kann man nachrechnen, dass wie behauptet

$$E[f(Y)\exp(V)] = E[f(Z)].$$

$\square$

Zum Schluss weisen wir noch darauf hin, dass jeder andere Schätzer, der sich vom MLE nur um einen $o_P(1/\sqrt{n})$-Term unterscheidet, ebenfalls asymptotisch effizient ist. Insbesondere kann man zeigen, dass auch der Bayesschätzer asymptotisch effizient ist für jede a priori Verteilung, die eine stetige und überall positive Dichte bezüglich des Lebesguemasses hat (siehe Lehmann, Point Estimation, Kap. 6.7).

6.3.2 Asymptotische Konfidenzbereiche und Tests

Wir konstruieren uns hier drei asymptotische Pivots, mit denen wir dann wie in Kapitel 1.1.3 Tests und Vertrauensbereiche für $g(\theta)$ konstruieren können. Das erste Pivot ist das allgemeinste.

Lemma 6.3. *Sei $T_n \overset{as}{\sim} \mathcal{N}(g(\theta), \frac{1}{n}V(\theta))$ und sei V_n eine konsistente Schätzung für $V(\theta)$. Dann gilt*

$$n(T_n - g(\theta))^T V_n^{-1}(T_n - g(\theta)) \xrightarrow{d} Z \sim \chi_p^2.$$

Beweis. Sei $A(\theta)$ eine "Quadratwurzel von $V(\theta)^{-1}$", d.h. $V(\theta)^{-1} = A(\theta)^T A(\theta)$. Ferner sei A_n die entsprechende "Quadratwurzel von V_n", die dann konsistent für $A(\theta)$ ist. Aus Satz 6.4 folgt, dass $\sqrt{n}A_n(T_n - g(\theta))$ in Verteilung gegen $Z \sim \mathcal{N}(0, Id_p)$ konvergiert. Die Behauptung ergibt sich daher aus Satz 6.1. □

Für konsistente Schätzer V_n gibt es meist mehrere Möglichkeiten, z.B.

- Wenn $\hat{\theta}_n$ ein konsistenter Schätzer von θ ist und $V(\theta)$ stetig in θ, dann ist $V_n = V(\hat{\theta}_n)$ konsistent.
- Wenn T_n ein M-Schätzer ist, dann können wir $V_n = M_n^{-1} J_n M_n^{-T}$ nehmen, wobei $M_n = n^{-1} \sum \frac{\partial}{\partial t}\psi(X_i, T_n)$ und $J_n = n^{-1} \sum \psi(X_i, T_n)\psi(X_i, T_n)^T$. Dies ist konsistent unter geeigneten Regularitätsbedingungen an ψ.
- Das Jackknife (siehe Kapitel 6.3.3).

Das zweite Pivot setzt voraus, dass $g(\theta) = \theta$. Dann folgt nämlich aus $T_n \overset{as}{\sim} \mathcal{N}(\theta, V(\theta)/n)$

$$n(T_n - \theta)^T V(\theta)^{-1}(T_n - \theta) \xrightarrow{d} Z \sim \chi_p^2.$$

Wenn T_n der MLE ist, dann steht noch ein drittes Pivot zur Verfügung, das wir bereits im Kapitel 1.3.2 erwähnt haben.

Lemma 6.4. *Sei $T_n = \arg\max \log L_n(\theta)$ der Maximum Likelihood Schätzer. Dann gilt unter Regularitätsbedingungen*

$$2(\log L_n(T_n) - \log L_n(\theta)) \xrightarrow{d} Z \sim \chi_p^2.$$

Beweis. Mit einer Taylorentwicklung von $\log L_n$ um T_n folgt

$$\log L_n(\theta) = \log L_n(T_n) + \frac{1}{2}\sqrt{n}(\theta - T_n)^T \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta^T} \log p_\theta(X_i)|_{\theta=\tau_n} \sqrt{n}(\theta - T_n)$$

mit $\|\tau_n - T_n\| \leq \|\theta - T_n\|$; vgl. die Herleitung von (6.16) und beachte, dass

$$\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log p_{T_n}(X_i) = 0.$$

Unter Regularitätsbedingungen konvergiert das arithmetische Mittel der zweiten partiellen Ableitungen gegen $-I(\theta)$. Also ist

$$2(\log L_n(T_n) - \log L_n(\theta)) = \sqrt{n}(T_n - \theta)^T I(\theta)\sqrt{n}(T_n - \theta) + o_P(1), \qquad (6.17)$$

woraus das Lemma folgt. □

Beispiel 6.14. *Multinomialverteilung. Sei* X_i *diskret mit möglichen Werten* $1, 2, \ldots, p+1$ *und Verteilung* $P[X_i = j] = \pi_j$*. Der unbekannte Parameter sei* $\theta_j = \pi_j \quad (j = 1, \ldots, p)$*. Es ist also* $\pi_{p+1} = 1 - \theta_1 - \ldots - \theta_p$*. Ferner seien*

$$N_j = \sum_{i=1}^{n} \mathbf{1}_{[X_i = j]} \quad (j = 1, \ldots, p+1)$$

die beobachteten Häufigkeiten der verschiedenen Werte. $(N_1, \ldots, N_{p+1})$ *ist dann multinomialverteilt. Man kann leicht nachrechnen, dass der MLE*

$$T_n = \frac{1}{n}(N_1, \ldots N_p)^T$$

ist und

$$I(\theta)_{jk} = \delta_{jk}\frac{1}{\theta_j} + \frac{1}{\pi_{p+1}}.$$

Die drei asymptotischen Pivots sind dann

$$\sum_{j=1}^{p+1} \frac{(N_j - n\pi_j)^2}{N_j}, \quad \sum_{j=1}^{p+1} \frac{(N_j - n\pi_j)^2}{n\pi_j}, \quad 2\sum_{j=1}^{p+1} N_j \log(\frac{N_j}{n\pi_j}).$$

Die zweite Form ist die übliche Chiquadrat-Statistik von Pearson.

Alle drei Pivots unterscheiden sich nur um $o_p(1)$-Terme. Unsere Theorie kann also nicht sagen, welches besser ist. Beim ersten Pivot sind die zugehörigen Vertrauensbereiche leicht zu bestimmen (es sind offensichtlich Ellipsoide). Die ersten zwei Pivots haben den Nachteil, dass die zugehörigen Vertrauensbereiche nicht äquivariant sind unter umkehrbaren Transformationen $\tau = h(\theta)$. Die Delta-Technik zeigt, dass dann $h(T_n)$ ebenfalls asymptotisch normal ist. Die beiden Bereiche $\{\theta; (T_n - \theta)^T V_n^{-1}(T_n - \theta) \leq c\}$ und $\{\theta; (h(T_n) - h(\theta))^T V_n'^{-1}(h(T_n) - h(\theta)) \leq c\}$ können aber stark verschieden sein ($V_n{}'$ ist natürlich eine geeignete Schätzung der asymptotischen Kovarianz von $h(T_n)$). Das dritte Pivot hat diesen Nachteil nicht.

Man kann sich auch fragen, ob es eine ausgezeichnete Transformation gibt, für die die Normalapproximation besonders gut ist. Dies ist jedoch ein sehr schwieriges Problem. Ein weniger ambitiöses Ziel einer Transformation wäre, dass die asymptotische Kovarianzmatrix nicht mehr von θ abhängt. Dann muss man wenigstens diese nicht mehr schätzen, und die ersten zwei Pivots unterscheiden sich nicht mehr.

Definition 6.5. *Eine Transformation $h : \mathbb{R}^p \to \mathbb{R}^p$ heisst* **varianzstabilisierend** *falls*

$$(\frac{\partial h(\theta)}{\partial \theta})V(\theta)(\frac{\partial h(\theta)}{\partial \theta})^T = Id_p.$$

Für $p = 1$ lautet die Bedingung $h'(\theta) = 1/\sqrt{V(\theta)}$, also ist die folgende Transformation varianzstabilisierend

$$h(\theta) = \int_{\theta_0}^{\theta} \frac{1}{\sqrt{V(t)}} dt.$$

Beispiel 6.15. *Sei X_i i.i.d. ,$P_\theta[X_i = 1] = \theta$, $P_\theta[X_i = 0] = 1 - \theta$. Die relative Häufigkeit von Einsen, $T_n = \sum X_i/n$, ist asymptotisch normal mit $V(\theta) = \theta(1-\theta)$. Als varianzstabilisierende Transformation erhält man daher $h(\theta) = \arcsin(2\theta - 1)$.*

Für $p > 1$ schreiben wir $V(\theta)^{-1} = A(\theta)^T A(\theta)$. Die Bedingung lautet dann $\frac{\partial h_i(\theta)}{\partial \theta_j} = A(\theta)_{ij}$. Diese Gleichung hat i.A. keine Lösung, d.h. eine varianzstabilisierende Transformation existiert nicht immer.

6.3.3 Sensitivität und Jackknife

Wir hatten bereits ganz am Anfang dieses Kapitels erwähnt, dass die Einflussfunktion eine Näherung für das Hinzufügen einer Beobachtung an der Stelle x ergibt:

$$T_{n+1} - T_n \approx \frac{1}{n+1} IF(x, P_0).$$

Daraus erhält man insbesondere eine Näherung für die Sensitivität (vgl. das Einführungskapitel):

$$\sup |T_{n+1}(x_1, \dots x_n, x) - T_n(x_1, \dots x_n)| \approx \frac{1}{n+1} \sup_x |IF(x, P_0)|.$$

Diese Approximationen sind als Heuristik äusserst nützlich. Das folgende Beispiel zeigt aber, dass man nicht ohne zusätzliche Regularitätsbedingungen auskommt.

Beispiel 6.16. *Median. Für $n = 2m$ bekommt man durch direkte Rechnung*

$$T_{n+1} - T_n = \begin{cases} -(x_{(m+1)} - x_{(m)})/2 & x \le x_{(m)} \\ x - (x_{(m+1)} + x_{(m)})/2 & x_{(m)} \le x \le x_{(m+1)} \\ (x_{(m+1)} - x_{(m)})/2 & x \ge x_{(m+1)}. \end{cases}$$

Im wesentlichen ist also $T_{n+1} - T_n$ proportional zu $sign(x - T_n)$, genau wie die Einflussfunktion des Medians. Der Proportionalitätsfaktor verhält sich aber nicht korrekt. Für $n \to \infty$ konvergiert nämlich $n(x_{(m+1)} - x_{(m)})/2$ nicht gegen den Proportionalitätsfaktor der Einflussfunktion, $1/(2p_0(P_0^{-1}(0.5)))$, sondern ist asymptotisch exponential-verteilt (siehe z.B. H.A. David, Order Statistics, Wiley 1981).

Statt eine Beobachtung hinzuzufügen, können wir natürlich auch eine weglassen. Sei $T_{n-1}^{(j)} = T_{n-1}(X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_n)$ der Schätzer ohne die j-te Beobachtung. Dann gilt $T_{n-1}^{(j)} - T_n \approx -\frac{1}{n} IF(X_j, P_0)$, also ist

$$V_n = n \sum_{j=1}^{n} (T_{n-1}^{(j)} - T_n)(T_{n-1}^{(j)} - T_n)^T$$

eine plausible Schätzung der asymptotischen Kovarianzmatrix $V(P_0)$. Traditionellerweise nimmt man die folgende äquivalente Form

$$V_n = \frac{n^2}{n-1} \sum_{j=1}^{n} (T_{n-1}^{(j)} - T_{n-1}^{(\cdot)})(T_{n-1}^{(j)} - T_{n-1}^{(\cdot)})^T,$$

wobei $T_{n-1}^{(\cdot)} = \frac{1}{n} \sum_{j=1}^{n} T_{n-1}^{(j)}$. Dieser Schätzer der asymptotischen Kovarianzmatrix heisst **Jackknife**.

6.3.4 Robustheit

Bei der robusten Statistik, deren Grundlagen in den 60-er und 70-er Jahren unter anderem von P. Huber und F. Hampel an der ETH Zürich gelegt wurden, möchte man Schätzer konstruieren, die wenig empfindlich sind auf Abweichungen vom Modell. Insbesondere möchte man die Möglichkeit von Ausreissern zulassen, die einer andern, beliebigen Verteilung G folgen. Wir nehmen also an, dass die Beobachtungen verteilt sind gemäss

$$(1-\varepsilon)P_\theta + \varepsilon G,$$

wobei ε den Anteil der Ausreisser bezeichnet. Wir suchen Schätzfolgen $T_n = Q(F_n)$, welche unter dem strikten Modell (d.h. $\varepsilon = 0$) nicht viel Effizienz gegenüber dem MLE verlieren und im Fall von Ausreissern dennoch brauchbare Resultate liefern. Zur Vereinfachung sei $p = 1$. Die erste Forderung soll bedeuten, dass Q Fisher-konsistent ist, d.h. $Q(P_\theta) = \theta$, und dass die asymptotische Varianz $V(\theta) = E_\theta[IF(X_i, Q, P_\theta)^2]$ klein ist. Für die zweite Forderung beachte man, dass unter Regularitätsbedingungen $\lim T_n = Q((1-\varepsilon)P_\theta + \varepsilon G)$. Daher verlangt man, dass der „asymptotische Bias“

$$Q((1-\varepsilon)P_\theta + \varepsilon G) - Q(P_\theta) \approx \varepsilon \int IF(x, Q, P_\theta) G(dx)$$

klein ist. Wir formulieren dazu das folgende Optimalitätsproblem: Für gegebenes $b(\theta)$ finde man Q so, dass $E_\theta[IF(X, Q, P_\theta)^2]$ minimal ist unter den folgenden Nebenbedingungen

$$\sup_x |IF(x, Q, P_\theta)| \leq b(\theta), \quad Q(P_\theta) = \theta \ \ \forall \theta.$$

Wir setzen also eine Schranke für den asymptotischen Bias, und tun dann unser bestes bei der asymptotischen Varianz am Modell.

Wenn wir noch annehmen, dass (6.15) erfüllt ist, dann können wir das obige Problem auf das folgende Problem der gewöhnlichen Variationsrechnung reduzieren: Minimiere

$$E_\theta[\psi(X, \theta)^2]$$

unter den Nebenbedingungen

$$E_\theta[\psi(X, \theta)] = 0, \quad E_\theta[\psi(X, \theta) \frac{\partial}{\partial \theta} \log p_\theta(X)] = 1, \quad \sup_x |\psi(x, \theta)| \leq b(\theta).$$

Jedes Funktional Q gibt über seine Einflussfunktion IF eine Funktion ψ, die zur Konkurrenz zugelassen ist, und umgekehrt liefert jedes zugelassene ψ ein Funktional Q, nämlich das Funktional, das durch den zu ψ gehörigen M-Schätzer definiert ist.

Dieses Problem kann man nun für jedes θ separat lösen. Mit den üblichen Argumenten der Variationsrechnung sieht man, dass man für das optimale ψ das ψ des MLE bei $\pm c(\theta)$ stutzen sollte. Um wieder Fisher-Konsistenz zu erhalten, muss man das ψ des MLE aber zuerst neu zentrieren. Für die zweite Nebenbedingung muss man dann noch mit einer Konstanten multiplizieren. Dadurch bleibt die Einflussfunktion zwar beschränkt, aber die Schranke $b(\theta)$ ist verschieden vom Stutzungspunkt $c(\theta)$. Genauer erhält man:

Satz 6.13 (F. Hampel). *Zu jedem $c(\theta)$ existiert eine Funktion $a(\theta)$ derart, dass*

$$\varphi_0(x,\theta) = \max(-c(\theta), \min(\frac{\partial}{\partial\theta}\log p_\theta(x) - a(\theta), c(\theta))$$

$E_\theta[\varphi_0(X,\theta)] = 0$ erfüllt. Ferner seien

$$d(\theta) = E_\theta[\varphi_0(X,\theta)\frac{\partial}{\partial\theta}\log p_\theta(X)], \quad \psi_0(x,\theta) = \frac{\varphi_0(x,\theta)}{d(\theta)}, \quad b(\theta) = \frac{c(\theta)}{d(\theta)}.$$

Dann gilt für jedes $\psi(x,\theta)$, welches die drei Nebenbedingungen erfüllt,

$$E_\theta[\psi(X,\theta)^2] \geq E_\theta[\psi_0(X,\theta)^2].$$

Für einen Beweis verweise ich auf Hampel et al., Robust Statistics, Wiley 1986, p. 117-119. Für das ursprünglich formulierte Problem müsste man noch untersuchen, für welche Schranken $b(\theta)$ ein Stutzungspunkt $c(\theta)$ existiert. Wir verzichten darauf.

Beispiel 6.17. *Sei $P_\theta = \mathcal{N}(\theta,1)$. Dann ist $\frac{\partial}{\partial\theta}\log p_\theta(x) = \theta - x$. Aus Symmetriegründen ist $a(\theta) = 0$ für beliebiges $c(\theta)$. Um einen äquivarianten Schätzer zu erhalten, nimmt man ferner $c(\theta) = c$ unabhängig von θ. Damit ist der optimale robuste Schätzer der Huber-Schätzer.*

6.3.5 Verallgemeinerter Likelihoodquotiententest

Wir betrachten wie immer das Modell X_i i.i.d. $\sim P_\theta, \theta \in \mathbb{R}^p$. Wir wollen nun die Nullhypothese testen, dass θ q Restriktionen erfüllt:

$$R(\theta) = (R_1(\theta), \ldots, R_q(\theta))^T = 0.$$

Unter der Nullhypothese haben wir also nur $p-q$ freie Parameter. Die Maximum-Likelihood-Schätzer im vollen Modell, bzw. unter der Nullhypothese sind

$$\begin{aligned} T_n &= \arg\max\{\log L_n(\theta) \mid \theta \in \mathbb{R}^p\}, \\ T_n^0 &= \arg\max\{\log L_n(\theta) \mid R(\theta) = 0\}. \end{aligned}$$

Die Teststatistik des verallgemeinerten Likelihoodquotiententests lautet dann

$$2(\log L_n(T_n) - \log L_n(T_n^0)).$$

Deren asymptotische Verteilung und somit auch einen genäherten kritischen Bereich erhält man mit Hilfe des folgenden Satzes.

Satz 6.14. *Die Matrix der partiellen Ableitungen $\frac{\partial}{\partial\theta}R(\theta)$ habe maximalen Rang q und P_θ erfülle geeignete Regularitätsbedingungen. Wenn die Restriktionen $R(\theta) = 0$ erfüllt sind, dann konvergiert $2(\log L_n(T_n) - \log L_n(T_n^0))$ in Verteilung gegen $Z \sim \chi^2_q$.*

Beweis. Mit den üblichen Argumenten erhält man für T_n und T_n^0 eine Approximation (6.1). Wir können daher den Ansatz

$$T_n^0 = T_n + \frac{Z_n}{\sqrt{n}} + o_P(\frac{1}{\sqrt{n}})$$

mit $Z_n = O_P(1)$ machen. Mit einer Taylorentwicklung von L_n an der Stelle T_n (vgl. (6.17)) folgt daher

$$\begin{aligned} 2(\log L_n(T_n) - \log L_n(T_n^0)) &= n(T_n^0 - T_n)^T I(\theta)(T_n^0 - T_n) + o_P(1) \\ &= Z_n^T I(\theta) Z_n + o_P(1). \end{aligned}$$

Ferner gilt

$$0 = R(T_n^0) = R(T_n) + \frac{\partial}{\partial\theta}R(\theta)\frac{Z_n}{\sqrt{n}} + o_P(\frac{1}{\sqrt{n}}).$$

Daraus folgt aber, dass wir Z_n erhalten, indem wir $z^T I(\theta) z$ bezüglich z minimieren unter der Nebenbedingung $\frac{\partial}{\partial\theta}R(\theta)z = -\sqrt{n}R(T_n)$. Eine einfache Rechnung unter Benutzung der Annahme, dass $\frac{\partial}{\partial\theta}R(\theta)$ maximalen Rang hat, ergibt

$$Z_n = -I(\theta)^{-1}\frac{\partial}{\partial\theta}R(\theta)^T W(\theta)^{-1}\sqrt{n}R(T_n),$$

wobei

$$W(\theta) = \frac{\partial}{\partial\theta}R(\theta)I(\theta)^{-1}\frac{\partial}{\partial\theta}R(\theta)^T.$$

Durch Einsetzen ergibt sich

$$2(\log L_n(T_n) - \log L_n(T_n^0)) = nR(T_n)^T(W(\theta)^{-1})R(T_n) + o_P(1).$$

Da $R(\theta) = 0$ und $T_n \overset{as}{\sim} \mathcal{N}(\theta, \frac{1}{n}I(\theta)^{-1})$, folgt aus der Delta-Technik, dass $R(T_n) \overset{as}{\sim} \mathcal{N}(0, \frac{1}{n}W(\theta))$. Also

$$nR(T_n)^T W(\theta)^{-1} R(T_n) \xrightarrow{d} Z \sim \chi_q^2.$$

□

Beispiel 6.18. *Kontingenztafeln (vgl. das Beispiel mit der Multinomialverteilung im Unterkapitel 6.3.2). X_i nehme Werte an in $\{1,\dots,r\}\times\{1,\dots,s\}$, d.h. wir betrachten zwei Merkmale mit r, bzw. s Ausprägungen (z.B. Haar- und Augenfarbe). Im vollen Modell sind die Wahrscheinlichkeiten $P[X_i = (j,k)] = \pi_{jk}$ beliebig bis auf $\sum_{j,k}\pi_{jk} = 1$. Das heisst, dass der Parameter θ $p = (rs-1)$-dimensional ist. Unter der Nullhypothese sind die beiden Merkmale unabhängig, d.h. wir haben $q = (r-1)(s-1)$ Restriktionen*

$$\pi_{jk} = \sum_{m=1}^{s}\pi_{jm}\sum_{l=1}^{r}\pi_{lk} \quad (j < r, k < s).$$

Ferner seien $N_{jk} = \sum_{i=1}^n \mathbf{1}_{[X_i=(j,k)]}$ die beobachteten Häufigkeiten der verschiedenen Werte. Die Matrix (N_{jk}) heisst Kontingenztafel. Der MLE ohne Einschränkungen ist gleich

$$T_n = \frac{1}{n}(N_{11},\dots,N_{r,s-1})^T,$$

während unter den Einschränkungen der Nullhypothese

$$T_n^0 = \frac{1}{n^2}(N_{1\cdot}N_{\cdot 1},\dots N_{r\cdot}N_{\cdot,s-1})^T,$$

wobei $N_{j.} = \sum_k N_{jk}$ *und* $N_{.k} = \sum_j N_{jk}$ *die Zeilen-, bzw. Spaltensummen in der Kontingenztafel sind. Schliesslich lautet die Teststatistik des verallgemeinerten Likelihoodquotiententests*

$$2\sum_{j=1}^{r}\sum_{k=1}^{s} N_{jk}\log(\frac{N_{jk}n}{N_{j.}N_{.k}}).$$

Meist benützt man aber eine andere Teststatistik, die unter dem Teilmodell asymptotisch die gleiche Verteilung hat. Der Beweis von Satz 6.14 zeigt, dass

$$2(\log L_n(T_n) - \log L_n(T_n^0)) = n(T_n^0 - T_n)^T I(T_n^0)(T_n^0 - T_n) + o_P(1).$$

Da die Fisher-Information

$$I(\theta)_{jk,lm} = \delta_{jl}\delta_{km}\frac{1}{\pi_{jk}} + \frac{1}{\pi_{rs}},$$

ist, können wir als Teststatistik auch

$$\sum_{j=1}^{s}\sum_{k=1}^{r}\frac{(N_{jk} - N_{j.}N_{.k}/n)^2}{N_{j.}N_{.k}/n}$$

wählen. Dies ist die üblicherweise verwendete χ^2*-Teststatistik von Pearson.*

6.3.6 Akaike-Kriterium

In den Unterkapiteln 6.3.2 und 6.3.5 hatten wir gesehen, wie man mithilfe von log likelihood bestimmen kann, wie gut eine Verteilung oder ein ganzes Modell zu gegebenen Daten passt. Wir wollen hier diese Idee nochmals anwenden, um zwei Modelle $(P_\tau; \tau \in \mathbb{R}^p)$ und $(Q_\eta; \eta \in \mathbb{R}^q)$ zu vergleichen. Wir nehmen aber nicht mehr an, dass die wahre Verteilung P_0 zu einer der beiden Modellfamilien gehört. Auch muss im Unterschied zum Unterkapitel 6.3.5 nicht mehr ein Modell im andern enthalten sein.

Sei T_n bzw. U_n der MLE im ersten, bzw. zweiten Modell. Im Unterkapitel 6.2.3 haben wir gesehen, dass T_n konvergiert gegen

$$g(F) = \arg\min_\tau H(P_0 : P_\tau) =: \tau_0,$$

wobei

$$H(P_0 : P_\tau) = E_0\left[\log\frac{p_0(X_i)}{p_\tau(X_i)}\right]$$

die relative Entropie ist. Wenn wir die Güte eines Modells mit dieser relativen Entropie messen, dann ziehen wir das erste Modell dem zweiten vor falls

$$H(P_0 : P_{T_n}) < H(P_0 : Q_{U_n}) \quad \Longleftrightarrow \quad \int \log\frac{p_{T_n}(x)}{q_{U_n}(x)}P_0(dx) > 0.$$

Wenn wir das Integral durch ein arithmetisches Mittel ersetzen, ergibt sich das folgende Kriterium: Ziehe das erste Modell vor, falls

$$\sum_{i=1}^{n}(\log p_{T_n}(X_i) - \log q_{U_n}(X_i)) > 0.$$

Dies führt jedoch zu einer ungewünschten Bevorzugung von komplizierten Modellen mit vielen Parametern. Wenn z.B. das erste Modell im zweiten enthalten ist, dann wählt dieses Kriterium automatisch immer das zweite, umfassendere Modell ! Der Grund ist, dass T_n durch Maximieren von $\sum \log p_\tau(X_i)$ bestimmt wurde, d.h. diese Summe ist systematisch zu gross. Wir berechnen daher eine heuristische Näherung für

$$E_0[\sum_{i=1}^n \log p_{T_n}(X_i) - n\int \log p_{T_n}(x)P_0(dx)],$$

um nachher korrigieren zu können.

Wir können schreiben

$$\begin{aligned}\sum_{i=1}^n \log p_{T_n}(X_i) - n\int \log p_{T_n}(x)P_0(dx) &= \sum_{i=1}^n(\log p_{T_n}(X_i) - \log p_{\tau_0}(X_i))\\ &+ \sum_{i=1}^n(\log p_{\tau_0}(X_i) - E_0[\log p_{\tau_0}(X_i)])\\ &+ \int(\log p_{\tau_0}(x) - \log p_{T_n}(x))P_0(dx)\\ &=: Z_{n,1} + Z_{n,2} + Z_{n,3}.\end{aligned}$$

Offensichtlich ist $E_0[Z_{n,2}] = 0$. Um $E_0[Z_{n,1}]$ zu approximieren, machen wir eine Taylorentwicklung an der Stelle T_n:

$$Z_{n,1} \approx -\frac{1}{2}\sqrt{n}(\tau_0 - T_n)^\intercal M(P_0)\sqrt{n}(\tau_0 - T_n),$$

wobei

$$M(P_0) = (E_0[\frac{\partial^2}{\partial\tau_j\partial\tau_k}\log p_\tau(X_i)]).$$

Indem wir die zweiten Momente von $\sqrt{n}(\tau_0 - T_n)$ durch die entsprechenden Momente der Grenzverteilung ersetzen, erhalten wir dann

$$E[Z_{n,1}] \approx -\frac{1}{2}\mathrm{spur}(J(P_0)M(P_0)^{-1})$$

wobei

$$J(P_0) = (E_0[\frac{\partial}{\partial\tau_j}\log p_\tau(X_i)\frac{\partial}{\partial\tau_k}\log p_\tau(X_i)]).$$

Nach Definition von τ_0 ist auch $Z_{n,3}$ stets positiv, und mit einer Taylorentwicklung an der Stelle τ_0 erhalten wir analog

$$E_0[Z_{n,3}] \approx -\frac{1}{2}\mathrm{spur}(J(P_0)M(P_0)^{-1}).$$

Damit wissen wir, um wieviel etwa $\sum_{i=1}^n \log p_{T_n}(X_i)$ überschätzt. Wir kennen aber die Matrizen $J(P_0)$ und $M(P_0)$ nicht, und können daher nicht sofort korrigieren. Ein Ausweg wäre, diese Matrizen zu schätzen, indem wir die Erwartungswerte durch arithmetische Mittel ersetzen. Ein anderer Weg geht davon aus, dass im Fall $P_0 = P_\tau$ gilt $J(P_0) = I(\tau)$ und $M(P_0) = -I(\tau)$. Also ist in diesem Fall $\mathrm{spur}(J(P_0)M(P_0)^{-1}) = -p$. Wenn das Modell einigermassen passt, sollte sich aber P_0 nicht allzu stark von P_τ unterscheiden, und damit ist auch

$$\mathrm{spur}(J(P_0)M(P_0)^{-1}) \approx -p.$$

Diese gleichen Argumente kann man natürlich auch für das zweite Modell anstellen. Somit ergibt sich das Akaike-Informations-Kriterium (AIC), welches besagt, dass wir das erste Modell vorziehen sollen, wenn

$$-2\sum_{i=1}^{n}\log p_{T_n}(X_i) + 2p < -2\sum_{i=1}^{n}\log q_{U_n}(X_i) + 2q.$$

Mit der Anzahl verwendeter Parameter (p, bzw. q) bestrafen wir komplizierte Modelle.

Anhang A

Einige Resultate aus der Analysis und Wahrscheinlichkeitstheorie

A.1 Verteilung von Zufallsvektoren

A.1.1 Transformation von Zufallsvektoren

Satz A.1. *Sei $g : G \subseteq \mathbb{R}^n \to G' \subseteq \mathbb{R}^n$ eine stetig differenzierbare, umkehrbare Abbildung, deren Funktionaldeterminante $D(x) = \det(\frac{\partial g_i}{\partial x_j})(x)$ nirgends verschwindet in G. Ferner sei X ein Zufallsvektor mit Werten in G, dessen Verteilung die Dichte $f_X(x)$ bezüglich des Lebesguemasses hat. Dann hat auch $Y = g(X)$ eine Dichte, und zwar*

$$f_Y(y) = \frac{f_X(g^{-1}(y))}{|D(g^{-1}(y))|}.$$

Beweis. Sei $h : \mathbb{R}^n \to \mathbb{R}$ stetig und beschränkt. Dann gilt (siehe Analysis II)

$$E[h(Y)] = E[h(g(X))] = \int_G h(g(x)) f_X(x) dx = \int_{G'} \frac{h(y) f_X(g^{-1}(y))}{|D(g^{-1}(y))|} dy.$$

Da dies richtig ist für alle h's, folgt die Behauptung aus der Masstheorie. □

A.1.2 Bedingte Verteilung und bedingte Erwartung

Wir behandeln folgendes Problem: Für zwei Zufallsgrössen $X : \Omega \to \mathbb{X}$ und $Y : \Omega \to \mathbb{Y}$ wollen wir die bedingte Verteilung von X gegeben $Y = y$ berechnen. Wir beginnen mit dem Fall, wo $\mathbb{X} = \mathbb{R}^n$, $\mathbb{Y} = \mathbb{R}^m$ und (X, Y) eine gemeinsame Dichte $f_{X,Y}(x, y)$ bezüglich des Lebesguemasses haben. Dann ist die *Randdichte* von Y gleich

$$f_Y(y) = \int_{\mathbb{R}^n} f_{X,Y}(x, y) dx,$$

und man definiert die *bedingte Dichte* von X gegeben $Y = y$ als

$$f_{X|Y=y}(x) = \frac{f_{X,Y}(x, y)}{f_Y(y)}.$$

Bemerkung A.1. *Das Ereignis $\{Y(\omega) = y\}$ hat Wahrscheinlichkeit 0. Für stetige Dichten erhält man aber obige Formel, indem man die bedingte Verteilung von X gegeben $\{|Y(\omega) - y| \leq h\}$ betrachtet und h gegen null gehen lässt.*

Aus dieser Definition ergibt sich durch Vertauschen der Rollen von Y und X die *Bayes'sche Formel*:

$$f_{Y|X=x}(y) = \frac{f_{X|Y=y}(x)f_Y(y)}{\int_{\mathbb{R}^m} f_{X|Y=u}(x)f_Y(u)du}.$$

Mithilfe der bedingten Dichte definieren wir die *bedingte Erwartung.* Für $g : \mathbb{R}^{n+m} \to \mathbb{R}$ integrierbar, setzen wir

$$E[g(X,Y)|Y = y] := \int_{\mathbb{R}^n} g(x,y)f_{X|Y=y}(x)dx.$$

Dies gibt den erwarteten Wert von $g(X,Y)$, wenn der Wert y von Y schon bekannt ist, der Wert von X aber nicht. Man kann diese bedingte Erwartung aber auch als Funktion der Zufallsvariablen $Y(\omega)$ auffassen (z.B. wenn der Wert von Y im Moment noch unbekannt ist, aber vor dem Wert von X ermittelt werden wird). Die bedingte Erwartung wird dann eine Funktion von Y und damit auch wieder eine Zufallsvariable, die wir mit $E[g(X,Y)|Y]$ bezeichnen.

Satz A.2 (Satz vom iterierten Erwartungswert). $E[g(X,Y)] = E[E[g(X,Y)|Y]]$.

Beweis. Die rechte Seite ist $= \int_{\mathbb{R}^m} E[g(X,Y)|Y = y]f_Y(y)dy$. Jetzt setzt man die Definition von $E[g(X,Y)|Y = y]$ ein und wendet Fubini an. □

Etwas schwieriger wird es, wenn Y eine (nicht-umkehrbare) Funktion von X ist, also

$$Y = h(X), \quad h : \mathbb{R}^n \to \mathbb{R}^m.$$

Dann hat die gemeinsame Verteilung von X und Y weder eine Dichte im $\mathbb{R}^{n+m}$, noch ist sie diskret. Wir können also die obigen Überlegungen nicht direkt anwenden. In vielen Fällen können wir aber eine zweite Funktion $g : \mathbb{R}^n \to \mathbb{R}^{n-m}$ finden derart, dass $x \in \mathbb{R}^n \to (g(x), h(x))$ umkehrbar und stetig differenzierbar ist mit Funktionaldeterminante $\neq 0$. Wenn X zudem eine Dichte $f_X(x)$ hat, dann hat das Paar $(Z = g(X), Y = h(X))$ gemäss Satz A.1 ebenfalls eine Dichte, und damit ist die bedingte Verteilung von Z gegeben $h(X) = y$ wie vorher definiert. Aus den Werten $Z = z$ und $Y = y$ können wir dann den Wert von X berechnen.

Beispiel A.1. *Sei $m = 1, h(x) = x_1 + \cdots + x_n$. Dann wählen wir $g(x) = (x_1, \ldots, x_n)$. Gemäss Satz A.1 ist die gemeinsame Dichte von $(Z = g(X), Y = h(X))$ gleich*

$$f_X(z_1, \ldots, z_{n-1}, y - z_1 - \cdots - z_{n-1}).$$

Also ist die bedingte Dichte von Z gegeben $h(X) = y$ gleich

$$\frac{f_X(z_1, \ldots, z_{n-1}, y - z_1 - \cdots - z_{n-1})}{\int_{\mathbb{R}^{n-1}} f_X(u_1, \ldots, u_{n-1}, y - u_1 - \cdots - u_{n-1})du}.$$

Mit diesem Trick kann man aber nicht alle Situationen abdecken.

Beispiel A.2. *Es sei* $n = 2, m = 1$ *und* $h(x_1, x_2) = \max(x_1, x_2)$. *Um* x *zu bestimmen, brauchen wir zusätzlich zu* $h(x)$ *den Wert von* $\min(x_1, x_2)$ *sowie die Information, ob* $x_1 < x_2$ *oder* $x_2 < x_1$ *ist. Damit ist die bedingte Verteilung eine Mischung von stetigen und diskreten Verteilungen. Diese können wir heuristisch bestimmen, indem wir infinitesimal kleine Quadrate betrachten: Wenn* (X_1, X_2) *die Dichte* f *hat, dann hat* $(Z = \min(X_1, X_2), Y = \max(X_1, X_2))$ *die Dichte* $(f(z,y) + f(y,z))1_{\{z<y\}}$. *Daraus bekommen wir die bedingte Dichte von* Z *gegeben* $Y = y$. *Bedingt darauf, dass* $Z = z$ *und* $Y = y$ *gilt dann*

$$P(X_1 = z, X_2 = y \mid Z = z, Y = y) = \frac{f(z,y)}{f(z,y) + f(y,z)} = 1 - P(X_1 = y, X_2 = z \mid Z = z, Y = y).$$

Falls X_1 *und* X_2 *unabhängig und gleichverteilt auf* $[0, \theta]$ *sind, dann gilt bedingt auf* $Y = y$ *mit Wahrscheinlichkeit 0.5* $X_1 = y$ *und* X_2 *ist uniform auf* $[0, y]$ *(und analog mit Wahrscheinlichkeit 0.5* $X_2 = y$ *und* X_1 *ist uniform auf* $[0, y]$*).*

Auf ähnliche Art und Weise kann man die meisten auftretenden Spezialfälle wenigstens heuristisch lösen. Für eine zufriedenstellende Behandlung aller Fälle benötigt man aber eine neue Definition, auf der man nachher mathematisch sauber aufbauen kann. Dies wird in der Wahrscheinlichkeitstheorie gemacht. Wir wollen darauf nicht weiter eingehen. Das einzige, was wir konkret brauchen werden, ist der Satz von der iterierten Erwartung, der ganz allgemein gilt, nicht nur in der von Satz A.2 abgedeckten Situation.

A.2 Mehrdimensionale Normalverteilung

Definition A.1. *Sei* μ *ein* $(n \times 1)$*-Vektor und* Σ *eine positiv definite* $(n \times n)$*-Matrix. Ein* n*-dimensionaler Zufallsvektor* X *heisst dann normalverteilt mit Parametern* μ *und* Σ *(in Formeln* $X \sim \mathcal{N}_n(\mu, \Sigma)$*), falls* X *die Dichte hat*

$$f(x) = (2\pi)^{-n/2} (\det \Sigma)^{-1/2} \exp(-\tfrac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)).$$

(Die nächsten Überlegungen zeigen, dass dies wirklich eine Dichte ist). Insbesondere folgt also aus $X \sim \mathcal{N}_n(0, I)$, dass $X_1, \dots X_n$ unabhängig und standard-normalverteilt sind. Der folgende Satz ergibt, wie die allgemeine n-dimensionale Normalverteilung aus diesem Spezialfall hervorgeht. Man beachte, dass sich jede positiv definite Matrix Σ als AA^T schreiben lässt (auf mehrere Arten).

Satz A.3. $Y \sim \mathcal{N}(\mu, \Sigma)$ *genau dann, wenn* $Y = AX + \mu$ *wobei* $X \sim \mathcal{N}(0, I)$ und $AA^T = \Sigma$.

Beweis. Die Richtung " $\Leftarrow$ " folgt direkt aus Satz A.1. Für die andere Richtung beachtet man zuerst, dass A nicht singulär sein kann, wenn Σ strikt positiv definit ist. Aus Satz A.1 folgt dann, dass $X = A^{-1}Y - A^{-1}\mu$ $\mathcal{N}_n(0, I)$ -verteilt ist. □

Korollar A.1. *Die* n*-dimensionale Normalverteilung hat Erwartungswert* $E[X_i] = \mu_i$ *und Kovarianzen* $\mathrm{Cov}(X_i, X_j) = \Sigma_{ij}$.

Satz A.4. *Sei* $Y \sim \mathcal{N}_n(\mu, \Sigma)$, A *eine* $k \times n$*-Matrix mit Rang* $k \leq n$ *und* $b \in \mathbb{R}^k$ *beliebig. Dann ist* $X = AY + b$ $\mathcal{N}_k(A\mu + b, A \sum A^T)$*-verteilt.*

Beweis. Wegen Satz A.3 können wir annehmen, dass $b = 0$ und $Y \sim \mathcal{N}_n(0, I)$. Wir wählen eine orthogonale $n \times n$-Matrix D, deren letzte $n - k$ Zeilen senkrecht auf allen Zeilen von A stehen. Gemäss Satz A.3 ist DY immer noch $\mathcal{N}_n(0, I)$-verteilt, und nach Konstruktion sind die letzten $n - k$ Spalten von $AD^T = 0$. Wegen $X = AY = AD^T DY$ ist also X eine lineare Transformation von $(DY)_1, \ldots, (DY)_k$. Damit folgt die Behauptung aus Satz A.3. □

Korollar A.2. *Die Randverteilung normalverteilter Zufallsvariablen ist normal. Für normalverteilte Zufallsvariable folgt aus Unkorreliertheit Unabhängigkeit.*

Bemerkung A.2. *Normale Randverteilungen genügen aber nicht für gemeinsame Normalverteilung: Sei $Y_1 \sim \mathcal{N}(0, 1)$ und Z unabhängig von Y_1 mit $P[Z = 1] = P[Z = -1] = \frac{1}{2}$. Setze $Y_2 = Z \cdot Y_1$. Dann ist $Y_2 \sim \mathcal{N}(0, 1)$, aber $Y_1 + Y_2$ ist nicht normalverteilt.*

Aus der Normalverteilung abgeleitete Verteilungen spielen an verschiedenen Stellen eine wichtige Rolle. Seien $X_0, X_1, X_2, \ldots$ i.i.d. $\sim \mathcal{N}(0, 1)$. Dann definieren wir

- Die *Chi-Quadrat-Verteilung* mit n Freiheitsgraden ist die Verteilung von $X_1^2 + X_2^2 + \cdots + X_n^2$. Ihre Dichte berechnet sich als
$$2^{-n/2}\Gamma(\frac{n}{2})^{-1} e^{-z/2} z^{n/2-1} \quad (z \geq 0).$$
- Die *t-Verteilung* mit n Freiheitsgraden ist die Verteilung von $X_0/\sqrt{\sum_{i=1}^n X_i^2/n}$. Ihre Dichte berechnet sich als
$$\frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{\pi n}}(1 + t^2/n)^{-(n+1)/2}.$$

Der folgende Satz ist fundamental für die Statistik normalverteilter Zufallsvariablen.

Satz A.5. *Seien $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}_1(\mu, \sigma^2)$. Dann sind*
$$\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$$
und
$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$$
unabhängig, $\overline{X}$ ist $\mathcal{N}(\mu, \sigma^2/n)$- und $(n-1)S_n^2/\sigma^2$ χ_{n-1}^2-verteilt. Insbesondere hat also
$$\frac{\sqrt{n}(\overline{X} - \mu)}{S_n}$$
eine t-Verteilung mit $n - 1$ Freiheitsgraden.

Beweis. Sei A orthogonal mit 1. Zeile $(1, \ldots, 1)/\sqrt{n}$. Dann ist $Y = A(X_1, \ldots X_n)^T$ $\mathcal{N}_n(\nu, \sigma^2 I)$-verteilt, wobei $\nu = (\mu\sqrt{n}, 0, \ldots, 0)^T$. Insbesondere sind also Y_1 und$(Y_2, \ldots, Y_n)$ unabhängig. Die Behauptung folgt dann aus
$$\sum_{i=1}^n (X_i - \overline{X})^2 = \sum_{i=1}^n X_i^2 - n\overline{X}^2 = \sum_{i=1}^n Y_i^2 - Y_1^2 = \sum_{i=2}^n Y_i^2.$$
□

A.3 Weitere wichtige Verteilungen

A.3.1 Multinomial-Verteilung:

Wir betrachten n unabhängige Wiederholungen eines Experiments mit k möglichen Ausgängen, welche mit Wahrscheinlichkeiten $p_1, \ldots, p_k$ eintreten (wobei natürlich $p_1 + p_2 + \cdots + p_k = 1$ gilt). Dann ist die Wahrscheinlichkeit, dass Ausgang j n_j-mal eintritt für $j = 1, \ldots, k$, gegeben durch

$$p(n_1, n_2, \ldots, n_k) = \frac{n!}{n_1! n_2! \ldots n_k!} p_1^{n_1} \ldots p_k^{n_k} \quad (n_1 + n_2 + \cdots + n_k = n).$$

Dies ist die sogenannte Multinomialverteilung. Für $k = 2$ ergibt sich die Binomialverteilung. Es besteht folgender Zusammenhang zwischen der Multinomial- und der Poisson-Verteilung.

Satz A.6. *Seien $X_1, X_2, \ldots, X_k$ Zufallsvariablen mit Werten in $\mathbb{N}_0$, und sei $X = \sum_{i=1}^k X_i$. Dann sind äquivalent:*

i) $X_1, \ldots, X_k$ sind unabhängig und Poisson-verteilt mit Parametern $\lambda_1, \ldots, \lambda_k$.

ii) X ist Poisson-verteilt mit Parameter $\lambda = \sum \lambda_i$, und gegeben $X = n$ ist $(X_1, \ldots, X_k)$ multinomialverteilt mit Parametern $p_i = \lambda_i / \lambda$.

Beweis. Wir beginnen mit dem Teil "i) $\Rightarrow$ ii)". Sei $\pi_n = P[X = n]$. Dann ist

$$P[X_1 = n_1, \ldots, X_k = n_k | X = n] = e^{-(\lambda_1 + \cdots + \lambda_k)} \frac{1}{\pi_n} \frac{\lambda_1^{n_1} \ldots \lambda_k^{n_k}}{n_1! \ldots n_k!}.$$

Summation über alle $(n_1, \ldots, n_k)$ mit $n_1 + \cdots + n_k = n$ ergibt

$$1 = e^{-(\lambda_1 + \cdots + \lambda_k)} \frac{1}{\pi_n} \frac{1}{n!} (\lambda_1 + \cdots + \lambda_k)^n.$$

Daraus folgt die Behauptung.

Die Umkehrung folgt aus

$$\begin{aligned} P[X_1 = n_1, \ldots, X_k = n_k] &= P[X_1 = n_1, \ldots, X_k = n_k | X = n] P[X = n] \\ &= \frac{n!}{n_1! \ldots n_k!} p_1^{n_1} \ldots p_k^{n_k} e^{-\lambda(p_1 + \cdots + p_k)} \lambda^{n_1} \lambda^{n_2} \ldots \lambda^{n_k} \frac{1}{n!}. \end{aligned}$$

□

A.3.2 Gamma- und Beta-Verteilung

Die Gamma-Verteilung mit Parametern $\gamma > 0, \quad \lambda > 0$ ist konzentriert auf $\mathbb{R}_+$ und hat dort die folgende Dichte

$$f(x) = \frac{\lambda^\gamma}{\Gamma(\gamma)} x^{\gamma - 1} e^{-\lambda x}.$$

Der Parameter γ bestimmt die Form der Verteilung, während λ nur ein Skalenparameter ist. Für $\gamma = 1$ erhält man die Exponential-Verteilungen und für $\lambda = 1/2$, $\gamma = n/2$ die χ^2-Verteilungen. Die ersten zwei Momente einer Gamma-verteilten Zufallsvariable X sind

$$E[X] = \frac{\gamma}{\lambda}, \quad \mathrm{Var}[X] = \frac{\gamma}{\lambda^2}.$$

Für Gamma-Verteilungen gilt ein Additionssatz: Wenn X_1 und X_2 unabhängig und Gamma(γ_i, λ)-verteilt sind, dann ist $X_1 + X_2$ Gamma$(\gamma_1 + \gamma_2, \lambda)$-verteilt.

Die Beta-verteilung mit Parametern $\gamma > 0, \quad \delta > 0$ ist konzentriert auf $[0,1]$ und hat dort die Dichte

$$f(x) = \frac{\Gamma(\gamma+\delta)}{\Gamma(\gamma)\Gamma(\delta)} x^{\gamma-1}(1-x)^{\delta-1}.$$

Die ersten zwei Momente sind

$$E[X] = \frac{\gamma}{\gamma+\delta}, \quad \mathrm{Var}[X] = \frac{\gamma\delta}{(\gamma+\delta)^2(1+\gamma+\delta)}.$$

Wenn X_1 und X_2 unabhängig und Gamma(γ_i, λ)-verteilt sind, dann ist $X_1/(X_1+X_2)$ Beta(γ_1, γ_2)-verteilt.

A.3.3 Negative Binomialverteilung

Dies ist eine diskrete Verteilung auf $\mathbb{N}_0$ mit Parametern $\gamma > 0$ und $0 < p < 1$. Sie ist gegeben durch

$$P[X=k] = \frac{\gamma(\gamma+1)\cdots(\gamma+k-1)}{k!} p^\gamma (1-p)^k = (-1)^k \binom{-\gamma}{k} p^\gamma (1-p)^k.$$

Die ersten zwei Momente dieser Verteilung sind

$$E[X] = \frac{\gamma(1-p)}{p}, \quad \mathrm{Var}[X] = \frac{\gamma(1-p)}{p^2}.$$

Beim Münzwurf mit Erfolgsparameter p ist die Anzahl Misserfolge bis zum r-ten Erfolg negativ-binomialverteilt mit $\gamma = r$. Ferner erhält man diese Verteilung aus der Poisson-Verteilung mit zufälligem Parameter:

Satz A.7. *Sei Λ Gamma(γ, η)-verteilt und gegeben $\Lambda = \lambda$ sei X Poisson(λ)-verteilt. Dann ist X negativ-binomial$(\gamma, \frac{\eta}{\eta+1})$-verteilt.*

Beweis.

$$\begin{aligned} P[X=k] &= E[P[X=k|\Lambda]] = \frac{\eta^\gamma}{\Gamma(\gamma)} \int_0^\infty \lambda^{\gamma-1} e^{-\eta\lambda} e^{-\lambda} \frac{\lambda^k}{k!} d\lambda \\ &= \eta^\gamma (\eta+1)^{-(k+\gamma)} \frac{\Gamma(k+\gamma)}{k!\Gamma(\gamma)} = \frac{\gamma\cdots(\gamma+k-1)}{k!} (\frac{\eta}{\eta+1})^\gamma (\frac{1}{\eta+1})^k. \end{aligned}$$

□

A.4 Einige Resultate aus der Analysis

Satz A.8. *Falls $E[X^2] < \infty$, dann* $\arg\min_c E[(X-c)^2] = E[X]$.

Falls $E[|X|] < \infty$, dann $\arg\min_c E[|X-c|] = \{c | P[X < c] \leq \frac{1}{2} \leq P[X \leq c]\}$.

Beweis. Die erste Behauptung folgt aus

$$\begin{aligned} E[(X-c)^2] &= E[(X-E[X])^2]+(E[X]-c)^2+2(E[X]-c)E[X-E[X]] \\ &= E[(X-E[X])^2]+(E[X]-c)^2. \end{aligned}$$

Heuristisch folgt die zweite Behauptung durch Ableiten unter dem Erwartungswert und Nullsetzen. Das ergibt $0 = E[sign(X-c)] = P[X > c] - P[X < c]$. Für einen exakten Beweis benützen wir, dass $F^{-1}(U)$, wobei U uniform auf $[0,1]$ ist, die gleiche Verteilung wie X hat. Also

$$\begin{aligned} E[|X-c|] &= E[\left|F^{-1}(U)-c\right|] = \int_0^1 \left|F^{-1}(u)-c\right| du \\ &= \int_0^{1/2} (\left|F^{-1}(u)-c\right| + \left|F^{-1}(1-u)-c\right|)du. \end{aligned}$$

Der Integrand wird für jedes feste u bezüglich c genau dann minimiert, wenn $F^{-1}(u) \leq c \leq F^{-1}(1-u)$. Man kann sich leicht überlegen, dass die letzte Ungleichung genau dann erfüllt ist für jedes $0 < u < 0.5$, wenn

$$\lim_{h \downarrow 0} F(c-h) \leq 0.5 \leq F(c).$$

□

Korollar A.3. *Seien $a_1 < a_2 \ldots < a_n$ und $w_i > 0$ $(i = 1, \ldots, n)$. Dann ist*

$$\arg\min_c \sum_{i=1}^n w_i\,|a_i - c| = [a_{k_1}, a_{k_2}],$$

wobei $k_1 = \inf\{k | \sum_{i=1}^k w_i \geq \frac{1}{2}\sum_{i=1}^n w_i\}$ und $k_2 = \sup\{k | \sum_{i=1}^{k-1} w_i \leq \frac{1}{2}\sum_{i=1}^n w_i\}$.

Beweis. Ohne Beschränkung der Allgemeinheit sei $\sum_{i=1}^n w_i = 1$ (Die Stelle, wo das Minimum angenommen wird, ändert sich nicht, wenn wir eine Funktion mit einer Konstanten multiplizieren). Dann betrachten wir die Zufallsvariable X, welche den Wert a_i mit Wahrscheinlichkeit w_i annimmt. Es gilt $E\,|X-c| = \sum_{i=1}^n w_i\,|a_i - c|$ und $P[X \leq c] = \sum_{i=1}^k w_i$ für $a_k \leq c < a_{k+1}$. □

Die folgenden Ungleichungen gehören zum Standardstoff der Funktionalanalysis und der Wahrscheinlichkeitstheorie.

Jensen'sche Ungleichung Falls $E[|X|] < \infty$ und g konvex, dann

$$E[g(X)] \geq g(E[X]).$$

Falls g strikt konvex ist und X nicht *f.s.* konstant, dann gilt sogar die strikte Ungleichung.

Hölder-Ungleichung Falls $E[|X|^p] < \infty$, $E[|Y|^q] < \infty$ und $\frac{1}{p} + \frac{1}{q} = 1$, dann ist $E[|XY|] < \infty$ und

$$|E[XY]| \leq E[|X|^p]^{1/p} E[|Y|^q]^{1/q}.$$

Für $p = q = 2$ ergibt sich die Schwarz'sche Ungleichung

$$E[XY]^2 \leq E[X^2]E[Y^2].$$

Anhang B

Literatur

[1] J.O. Berger: *Statistical Decision Theory and Bayesian Analysis.* 2. Auflage, Springer 1985.
Behandelt Bayes-Theorie.

[2] P.J. Bickel, K.A. Doksum: *Mathematical Statistics: Basic Ideas and Selected Topics.* Vol. I, Second Edition. Prentice Hall, 2001.
Einfach und gut lesbar.

[3] D.R. Cox, D.V. Hinkley: *Theoretical Statistics.* Chapman and Hall, 1974.
Viele gute Diskussionen über Begriffe und deren praktische Bedeutung. Mathematik skizzenhaft.

[4] J.G. Kalbfleisch: *Probability and Statistical Inference.* Bd. 2, 2. Auflage, Springer 1985.
Behandelt Likelihood-methoden.

[5] L.M. Le Cam: *Asymptotic Methods in Statistical Decision Theory.* Springer, 1986.
Behandelt asymptotische Entscheidungstheorie, auf sehr abstraktem Niveau.

[6] E.L. Lehmann: *Theory of Point Estimation.* Wiley, 1983.
Ein Klassiker. Das Skript lehnt sich ziemlich eng an dieses Buch an.

[7] E.L. Lehmann: *Testing Statistical Hypotheses.* 2. Auflage, Wiley, 1986.
Gehört zusammen mit dem vorherigen Buch.

[8] Mark J. Schervish: *Theory of Statistics.* Springer, 1995.
Mathematisch exakt und ziemlich umfassend. Geeignet als Nachschlagewerk.

[9] R.J. Serfling: *Approximation Theorems of Mathematical Statistics.* Wiley, 1980.
Behandelt Asymptotik. Gut geeignet als Ergänzung zum Kapitel 6 des Skripts.

[10] A. W. van der Vaart: *Asymptotic Statistics.* Cambridge University Press, 1998.
Moderne, elegante Behandlung der Asymptotik.