Vorlesung über Mathematische Statistik
Sommersemester 2006

Wolfgang Wefelmeyer

Version vom 27. Juni 2014

Inhaltsverzeichnis

1 Exponentielle Familien

Unter den parametrischen Verteilungsfamilien sind exponentielle Familien im wesentlichen die einzigen, für die nichtasymptotisch optimale Schätzer existieren. Für nicht-exponentielle Familien und für nichtparametrische und semiparametrische Modelle lassen sich nur Schätzer finden, die asymptotisch, also mit gegen unendlich wachsendem Stichprobenumfang optimal sind. (Das geht auch nur dann, wenn solche Modelle in einem geeigneten Sinne gegen exponentielle Familien konvergieren.)

Gegeben sei ein meßbarer Raum $(\Omega, \mathcal{F})$ und eine Familie $P_\vartheta|\mathcal{F}$, $\vartheta \in \Theta$, von Wahrscheinlichkeitsmaßen.

Definition. Eine Familie P_ϑ, $\vartheta \in \Theta$, heißt *exponentielle Familie* in $\eta(\vartheta)$ und T, wenn sie bezüglich eines dominierenden Maßes $\mu|\mathcal{F}$ Dichten folgender Form hat:

$$f_\vartheta(x) = c(\vartheta) \exp(\eta(\vartheta)^\top T(x)),$$

wobei sowohl $1, \eta_1, \ldots, \eta_d$ als auch $1, T_1, \ldots, T_d$ linear unabhängig sind. Sie heißt *kanonisch*, wenn $\eta(\vartheta) = \vartheta$ gilt. Dann besteht der *natürliche Parameter-Raum* (das heißt: der maximale Parameter-Raum) aus den ϑ mit

$$\int \exp(\vartheta^\top T(x))\mu(dx) < \infty.$$

Damit sich die Dichten zu 1 integrieren, muß gelten:

$$c(\vartheta) = \Big(\int \exp(\eta(\vartheta)^\top T(x))\mu(dx) \Big)^{-1}.$$

Nützlich ist die Struktur einer exponentiellen Familie insbesondere, wenn die Verteilung von T einfacher als P_ϑ ist; zum Beispiel, wenn T eine kleinere Dimension als x hat.

Gelegentlich wird die affine Unabhängigkeit der η_j und der T_j nicht in die Definition genommen. Dann läßt sich im allgemeinen die Dimension reduzieren. Trotzdem ist die Darstellung nicht eindeutig.

Bemerkung. In der Darstellung für $f_\vartheta(x)$ darf auch ein Faktor $h(x)$ vorkommen: Hat P_ϑ die ν-Dichte $g_\vartheta(x) = h(x)f_\vartheta(x)$, so hat P_ϑ die μ-Dichte $f_\vartheta(x)$ für $\mu(dx) = h(x)\nu(dx)$.

Beispiel. Die Binomialverteilungen $Bi_{n,p}$, $p \in (0,1)$, haben in $k = 0, \ldots, n$ die Zähldichte

$$g_p(k) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{n}{k} (1-p)^n \exp(k \log(p/(1-p))).$$

Das ist eine exponentielle Familie in $\eta(p) = \log(p/(1-p))$ und $T(k) = k$.

Beispiel. Die Normalverteilungen N_{μ,σ^2}, $\mu \in \mathbb{R}$, $\sigma^2 > 0$, haben in $x \in \mathbb{R}$ die Lebesgue-Dichte

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\mu^2/(2\sigma^2)} e^{x\mu/\sigma^2 - x^2/(2\sigma^2)}.$$

Das ist eine exponentielle Familie in $\eta(\mu, \sigma^2) = (\mu/\sigma^2, 1/(2\sigma^2))$ und $T(x) = (x, -x^2)$.

Beispiel. Für eine Verteilung $P|\mathcal{B}^d$ ist die um $\vartheta \in \mathbb{R}^d$ verschobene Verteilung durch $P_\vartheta(A) = P(A - \vartheta)$, $A \in \mathcal{B}$, definiert. Die Familie P_ϑ, $\vartheta \in \mathbb{R}^d$, heißt *von P erzeugte Lageparameter-Familie.*

Für die Verteilung $P|\mathcal{B}$ mit Lebesgue-Dichte $f(x) = C \exp(-x^4)$ hat die um $\vartheta \in \mathbb{R}$ verschobene Verteilung die Dichte

$$f(x - \vartheta) = Ce^{-(x-\vartheta)^4} = Ce^{-x^4} e^{-\vartheta^4} e^{4x^3\vartheta - 6x^2\vartheta^2 + 4x\vartheta^3}.$$

Das ist eine exponentielle Familie mit $\eta(\vartheta) = (4\vartheta, -6\vartheta^2, \vartheta^3)$ und $T(x) = (x^3, x^2, x)$. Hier ist η ein Pfad in $\mathbb{R}^3$.

Exponentielle Familien, bei denen ϑ durch η in einen höherdimensionalen Raum eingebettet wird, heißen *gekrümmt.*

Satz 1 *Sei P_ϑ, $\vartheta \in \Theta$, eine exponentielle Familie in ϑ und T, und sei Θ offen. Sei h eine Funktion, für die $H(\xi) = \int h \exp(\xi^\top T) d\mu$ für $\xi = (\xi_1, \ldots, \xi_d)$ mit $\xi_j = \vartheta_j + i\tau_j$ und $\vartheta \in \Theta$ existiert und endlich ist. Dann ist H analytisch in jedem ξ_j, und die Ableitungen können unter das Integral gezogen werden.*

Beweis. Schreibe $h \exp(\xi^\top T) = \exp(\xi_1 T_1) h \exp(\sum_{j=2}^d \xi_j T_j)$. Zerlege das Produkt aus den letzten beiden Faktoren in Real- und Imaginärteil und dann jeweils in Positiv- und Negativteil und schlage diese Funktionen zu μ. Dann läßt sich H schreiben als $H(\xi) = \int \exp(\xi_1 T_1) d(\mu_1 - \mu_2 + i(\mu_3 - \mu_4))$. Es reicht also, ein Integral der Form $H(\xi) = \int \exp(\xi T) d\mu$ zu betrachten. Schreibe den Differenzenquotienten als

$$\frac{H(\zeta) - H(\xi)}{\zeta - \xi} = \int \frac{e^{\zeta T} - e^{\xi T}}{\zeta - \xi} d\mu.$$

Wir zeigen, daß wir unter dem Integral differenzieren dürfen. Für $|z| \leq \delta$ gilt

$$\Big|\frac{e^{az}-1}{z}\Big| = \Big|\sum_{m=1}^{\infty} \frac{z^{m-1}a^m}{m!}\Big| \leq \frac{e^{\delta|a|}}{\delta}.$$

Für $|\zeta - \xi| \leq \delta$ läßt sich also der Integrand abschätzen durch

$$e^{\xi T}\Big|\frac{e^{(\zeta-\xi)T}-1}{\zeta-\xi}\Big| \leq \frac{1}{\delta}e^{\xi T+\delta|T|} \leq \frac{1}{\delta}\big|e^{(\xi+\delta)T} + e^{(\xi-\delta)T}\big|.$$

Die rechte Seite ist integrierbar. Mit dem Satz von der dominierten Konvergenz folgt für $\zeta \to \xi$

$$\frac{H(\zeta)-H(\xi)}{\zeta-\xi} \to \int Te^{\xi T}d\mu.$$

Die höheren Ableitungen erhält man durch Induktion.

2 Suffiziente Statistiken

Gegeben sei ein meßbarer Raum $(\Omega, \mathcal{F})$ und eine Familie $P_\vartheta|\mathcal{F}$, $\vartheta \in \Theta$, von Wahrscheinlichkeitsmaßen. Sei $T : (\Omega, \mathcal{F}) \to (E, \mathcal{E})$ eine Statistik. Bezeichne P_ϑ^T die induzierte Verteilung von T unter P_ϑ. Es existiere eine reguläre bedingte Verteilung $P_\vartheta(\cdot|T)$ gegeben T. Dann gilt

$$P_\vartheta A = \int P_\vartheta(A|T=t)P_\vartheta^T(dt), \quad A \in \mathcal{F}.$$

Definition. Die Statistik T heißt *suffizient* für ϑ, wenn es für alle $A \in \mathcal{F}$ eine von ϑ unabhängige Version der bedingten Wahrscheinlichkeit $P_\vartheta(A|T)$ gibt.

Wir werden in Kapitel 4 sehen, daß dann die besten Schätzer Funktionen von T sein müssen. Das vereinfacht die Suche nach guten Schätzern.

Beispiel. Seien $X_1, \ldots, X_n$ unabhängig und $Bi_{1,p}$-verteilt. Dann ist $T = \sum_{i=1}^n X_i$ suffizient für p.

Setze $X = (X_1, \ldots, X_n)$. Es gilt

$$P(X=x|T=t) = \frac{P(X=x, T=t)}{P(T=t)} = \frac{p^t(1-p)^{n-t}}{\binom{n}{t}p^t(1-p)^{n-t}} = \frac{1}{\binom{n}{t}}.$$

Das ist unabhängig von p.

Satz 2 *Sei $(\Omega, \mathcal{F}) = (\mathbb{R}^k, \mathcal{B}^k)$ und T suffizient für ϑ. Dann existiert eine von ϑ unabhängige reguläre bedingte Verteilung $P(\cdot|T)$.*

Beweis. Für $x \in \mathbb{Q}^k$ existiert eine von ϑ unabhängige Version der bedingten Wahrscheinlichkeit $P(X \leq x|T)$. Dadurch sind die Verteilungsfunktionen $P(X \leq x|T = t)$ für $x \in \mathbb{R}$ festgelegt und von ϑ unabhängig. Es seien $P(\cdot|T = t)$ die zugehörigen Wahrscheinlichkeitsmaße. Sei $\mathcal{L}$ das System der Mengen B, für die $P(B|T)$ eine bedingte Wahrscheinlichkeit von B gegeben T ist. Aus der Linearität und der monotonen Konvergenz bedingter Erwartungswerte folgt, daß $\mathcal{L}$ ein λ-System ist. Es enthält das π-System der Mengen $(-\infty, x]$, $x \in \mathbb{Q}^k$, welches $\mathcal{B}^k$ erzeugt. Nach dem π-λ-Satz von Dynkin gilt also $\mathcal{L} = \mathcal{B}^k$.

Hauptergebnis dieses Kapitels ist das Faktorisierungskriterium von Neyman. Es besagt, daß für dominierte Familien T genau dann suffizient ist, wenn sich die Dichte bis auf einen von ϑ unabhängigen Faktor als Funktion von T schreiben läßt. Wir benötigen zwei Hilfsresultate. Das erste beweisen wir nicht.

Definition. Zwei Familien $\mathcal{M}$ und $\mathcal{N}$ von Maßen auf $(\Omega, \mathcal{F})$ heißen *äquivalent*, wenn $\mu A = 0$ für alle $\mu \in \mathcal{M}$ genau dann gilt, wenn $\nu A = 0$ für alle $\nu \in \mathcal{N}$.

Satz 3 *(Halmos und Savage) Eine Familie von Wahrscheinlichkeitsmaßen ist durch ein σ-endliches Maß dominiert genau dann, wenn sie eine abzählbare äquivalente Teilfamilie besitzt.*

Satz 4 *Sei P_ϑ, $\vartheta \in \Theta$, dominiert durch ein σ-endliches Maß. Sei $Q = \sum_{n=1}^{\infty} c_n P_{\vartheta_n}$ äquivalent zu P_ϑ, $\vartheta \in \Theta$. Dann ist T suffizient für ϑ genau dann, wenn eine nichtnegative, $\mathcal{E}$-meßbare Funktion g_ϑ existiert mit*

$$dP_\vartheta = g_\vartheta \circ T dQ, \quad \vartheta \in \Theta.$$

Beweis. Sei $\mathcal{F}_0$ die von T induzierte σ-Algebra.

Notwendig. Sei T suffizient für ϑ. Nach Definition des bedingten Erwartungswerts gilt für $A \in \mathcal{F}$ und $A_0 \in \mathcal{F}_0$:

$$\int_{A_0} P(A|T) dP_\vartheta = P_\vartheta A \cap A_0, \quad \vartheta \in \Theta.$$

Weil der Integrand nicht von ϑ abhängt, gilt auch

$$\int_{A_0} P(A|T) dQ = QA \cap A_0,$$

also $P(A|T) = Q(A|T)$. Nach dem Satz von Radon–Nikodým und dem Faktorisierungslemma existiert eine $Q|\mathcal{F}_0$-Dichte von $P_\vartheta|\mathcal{F}_0$ der Form $g_\vartheta \circ T$. Es bleibt zu zeigen, daß $g_\vartheta \circ T$ auch $Q|\mathcal{F}$-Dichte von $P_\vartheta|\mathcal{F}$ ist. Weil $Q(A|T)$ $\mathcal{F}_0$-meßbar ist, gilt für $A \in \mathcal{F}$:

$$\begin{aligned} P_\vartheta A &= \int P(A|T)dP_\vartheta = \int Q(A|T)dP_\vartheta = \int Q(A|T)g_\vartheta \circ T dQ \\ &= \int E_Q(1_A g_\vartheta \circ T|T)dQ = \int_A g_\vartheta \circ T dQ. \end{aligned}$$

Hinreichend. Gelte $dP_\vartheta = g_\vartheta \circ T dQ$. Sei $A \in \mathcal{F}$. Wir zeigen: $Q(A|T) = P_\vartheta(A|T)$ Q-f.s. Definiere $d\nu = 1_A dP_\vartheta$. Schreibe

$$\frac{d\nu}{dQ} = \frac{d\nu}{dP_\vartheta}\frac{dP_\vartheta}{dQ} = 1_A g_\vartheta \circ T.$$

Für $A_0 \in \mathcal{F}_0$ gilt einerseits

$$\begin{aligned} \int_{A_0} \frac{d\nu}{dQ} dQ &= \int_{A_0} 1_A \frac{dP_\vartheta}{dQ} dQ = \int_{A_0} 1_A dP_\vartheta \\ &= \int_{A_0} P_\vartheta(A|T)dP_\vartheta = \int_{A_0} P_\vartheta(A|T)g_\vartheta \circ T dQ, \end{aligned}$$

also

$$\frac{d\nu|\mathcal{F}_0}{dQ|\mathcal{F}_0} = P_\vartheta(A|T)g_\vartheta \circ T;$$

andererseits

$$\int_{A_0} \frac{d\nu}{dQ} dQ = \int_{A_0} 1_A g_\vartheta \circ T dQ = \int_{A_0} E_Q(1_A g_\vartheta \circ T|T)dQ,$$

also

$$\frac{d\nu|\mathcal{F}_0}{dQ|\mathcal{F}_0} = E_Q(1_A g_\vartheta \circ T|T) = Q(A|T)g_\vartheta \circ T.$$

Also gilt $Q(A|T) = P_\vartheta(A|T)$ $P_\vartheta|\mathcal{F}_0$-f.s. wegen $g_\vartheta \circ T \neq 0$ $P_\vartheta|\mathcal{F}_0$-f.s. Das heißt: T ist suffizient.

Satz 5 *(Faktorisierungskriterium von Neyman) Sei P_ϑ, $\vartheta \in \Theta$, dominiert durch ein σ-endliches Maß μ. Dann ist T suffizient für ϑ genau dann, wenn eine nichtnegative, $\mathcal{E}$-meßbare Funktion g_ϑ und eine nichtnegative, $\mathcal{F}$-meßbare Funktion h existieren mit*

$$dP_\vartheta = h g_\vartheta \circ T d\mu, \quad \vartheta \in \Theta.$$

Beweis. Notwendig. Seit T suffizient für ϑ. Nach Satz 3 existiert ein $Q = \sum c_n P_{\vartheta_n}$ äquivalent zu P_ϑ, $\vartheta \in \Theta$. Wegen

$$\frac{dP_\vartheta}{d\mu} = \frac{dQ}{d\mu}\frac{dP_\vartheta}{dQ}$$

folgt die Behauptung mit $h = dQ/d\mu$ aus Satz 4.

Hinreichend. Es gilt nach Voraussetzung:

$$\frac{dQ}{d\mu} = \sum c_n \frac{dP_{\vartheta_n}}{d\mu} = h \sum c_n g_{\vartheta_n} \circ T,$$

also

$$\frac{dP_\vartheta}{dQ} = \frac{d\mu}{dQ}\frac{dP_\vartheta}{d\mu} = \frac{g_\vartheta \circ T}{\sum c_n g_{\vartheta_n} \circ T}.$$

Aus Satz 4 folgt, daß T suffizient ist.

Korollar 1 *Bei einer exponentiellen Familie in $\eta(\vartheta)$ und T ist T suffizient für ϑ.*

Beweis. Wir haben Dichten der Form $f_\vartheta(x) = c(\vartheta)\exp(\eta(\vartheta)^\top T(x))$, also wie im Satz 5.

Beispiel. Sind $X_1, \ldots, X_n$ unabhängig und $Bi_{1,p}$-verteilt, so ist $\sum_{i=1}^n X_i$ suffizient für p.

Die Zähldichte von $X = (X_1, \ldots, X_n)$ ist

$$P(X = x) = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}.$$

Das ist von der Form des Faktorisierungskriteriums.

Eine suffiziente Statistik ist nicht eindeutig. Suffizienz bleibt unter allen (meßbaren) eineindeutigen Transformationen erhalten.

3 Vollständige und anzilläre Statistiken

Gegeben sei ein meßbarer Raum $(\Omega, \mathcal{F})$ und eine Familie $P_\vartheta|\mathcal{F}$, $\vartheta \in \Theta$, von Wahrscheinlichkeitsmaßen. Sei $T : (\Omega, \mathcal{F}) \to (E, \mathcal{E})$ eine Statistik. Wir sind an möglichst einfachen suffizienten Statistiken interessiert.

Definition. Eine suffiziente Statistik T heißt *minimal suffizient*, wenn für jede suffiziente Statistik S eine meßbare Funktion h existiert, so daß $T = h \circ S$ P_ϑ-f.s. für $\vartheta \in \Theta$.

Ist $(\Omega, \mathcal{F}) = (\mathbb{R}^k, \mathcal{B}^k)$ und ist P_ϑ, $\vartheta \in \Theta$, durch ein σ-endliches Maß dominiert, so existiert eine minimale suffiziente Statistik (Bahadur, 1957). Das zeigen wir hier nicht.

Um zu beschreiben, was an einer suffizienten Statistik noch überflüssig ist, verwenden wir folgenden Begriff.

Definition. Eine Statistik T heißt *anzillär*, wenn ihre Verteilung P_ϑ^T nicht von ϑ abhängt. Sie heißt *anzillär erster Ordnung*, wenn ihr Erwartungswert $E_\vartheta T$ nicht von ϑ abhängt.

An der Faktorisierung $P_\vartheta A = \int P_\vartheta(A|T = t) P_\vartheta^T(dt)$ sieht man, daß anzillär komplementär zu suffizient ist. Eine anzilläre Statistik enthält keine Information über ϑ. Es ist plausibel, daß eine suffiziente Statistik minimal ist, wenn keine nichtkonstante Funktion von ihr anzillär ist, oder auch nur anzillär erster Ordnung.

Definition. Eine Statistik T heißt *(beschränkt) vollständig* für ϑ, wenn für jede (beschränkte) meßbare Funktion $g : E \to \mathbb{R}$ gilt:

$$E_\vartheta g(T) = 0 \text{ für } \vartheta \in \Theta \text{ impliziert: } g = 0 \; P_\vartheta^T\text{-f.s. für } \vartheta \in \Theta.$$

In anderen Worten: Die Familie P_ϑ^T, $\vartheta \in \Theta$, ist so reichhaltig, daß eine nicht identisch verschwindende Funktion zumindest für ein ϑ einen nicht verschwindenden Erwartungswert unter P_ϑ^T hat.

Es gibt sehr große Familien, die nicht vollständig sind. Zum Beispiel verschwinden für alle um Null symmetrischen Verteilungen die Erwartungswerte aller um Null antisymmetrischen Funktionen.

Eine vollständige suffiziente Statistik ist minimal (Lehmann und Scheffé 1950, Bahadur 1957). Das zeigen wir hier nicht.

Beispiel. Seien $X_1, \ldots, X_n$ unabhängig und verteilt nach der Gleichverteilung auf $(0, \vartheta)$. Dann ist $T = \max X_i$ vollständig und suffizient für $\vartheta > 0$.

Suffizienz. Die Dichte von $(X_1, \ldots, X_n)$ ist

$$\frac{1}{\vartheta^n} \prod_{i=1}^n 1_{(0,\vartheta)}(X_i) = \frac{1}{\vartheta^n} 1_{(0,\vartheta)}(T).$$

Also ist T suffizient nach Satz 5.

Vollständigkeit. Die Verteilungsfunktion von T ist $P(T \le x) = x^n/\vartheta^n$ für $x \in (0, \vartheta)$, die Dichte also nx^{n-1}/ϑ^n. Sei g eine Funktion mit

$$0 = E_\vartheta g(T) = \frac{n}{\vartheta^n} \int_0^\vartheta g(x) x^{n-1} dx, \quad \vartheta > 0.$$

Das Integral läßt sich auffassen als Verteilungsfunktion eines signierten Maßes. Dessen Dichte muß also f.s. verschwinden.

Für das folgende schöne Ergebnis haben wir zunächst keine Verwendung. Es ist aber in der asymptotischen Statistik nützlich.

Satz 6 *(Lemma von Basu) Ist T beschränkt vollständig und suffizient, dann ist jede anzilläre Statistik unabhängig von T.*

Beweis. Sei V anzillär. Setze $a(t) = P_\vartheta(V \in A|T = t)$. Weil T suffizient ist, hängt a nicht von ϑ ab. Weil V anzillär ist, hängt $E_\vartheta a(T) = P_\vartheta(V \in A)$ nicht von ϑ ab. Also gilt $E_\vartheta g(T) = 0$ für $g(t) = a(t) - P(V \in A)$. Weil T beschränkt vollständig ist, gilt also $a = P(V \in A)$ P_ϑ^T-f.s. Das heißt: V und T sind unabhängig unter P_ϑ.

Satz 7 *Sei $P_\vartheta|\mathcal{F}$, $\vartheta \in \Theta$, eine exponentielle Familie in $\eta(\vartheta)$ und T. Ist das Innere von $\eta(\Theta)$ nichtleer, so ist T vollständig für $\vartheta \in \Theta$.*

Beweis. Sei ohne Einschränkung $\eta(\vartheta) = \vartheta$ und $I = (-a, a)^d \subset \Theta$. Sei g eine Funktion mit $E_\vartheta g(T) = 0$ für $\vartheta \in \Theta$. Schreibe

$$E_\vartheta g(T) = \int g(t) \exp(\vartheta^\top t) \mu^T(dt).$$

Es gilt

$$\int g^+(t) \exp(\vartheta^\top t) \mu^T(dt) = \int g^-(t) \exp(\vartheta^\top t) \mu^T(dt). \tag{3.1}$$

Für $\vartheta = 0$ erhält man insbesondere $\int g^+ d\mu^T = \int g^- d\mu^T$. Ohne Einschränkung nehmen wir $\int g^+ d\mu^T = 1$ an. Definiere Wahrscheinlichkeitsmaße $dP^+ = g^+ d\mu^T$ und $dP^- = g^- d\mu^T$. Dann läßt sich (3.1) schreiben als $\int \exp(\vartheta^\top t) P^+(dt) = \int \exp(\vartheta^\top t) P^-(dt)$. Für $\xi_j = \vartheta_j + i\tau_j$ mit $\vartheta_j \in (-a, a)$ sind die Funktionen $\int \exp(\xi^\top t) P^+(dt)$ und $= \int \exp(\xi^\top t) P^-(dt)$ nach Satz 1 analytisch in jedem ξ_j, stimmen also überein. Für $\xi_j = i\vartheta_j$ gilt also insbesondere $\int \exp(i\vartheta^\top t) P^+(dt) = \int \exp(i\vartheta^\top t) P^-(dt)$. Nach dem Eindeutigkeitssatz für charakteristische Funktionen gilt deshalb $P^+ = P^-$, also $g^+ = g^-$ μ^T-f.s., also $g = 0$ μ^T-f.s., also P_ϑ^T-f.s. für $\vartheta \in \Theta$.

4 Konvexe Verlustfunktionen

Gegeben sei wieder ein meßbarer Raum $(\Omega, \mathcal{F})$ und eine Familie $P_\vartheta|\mathcal{F}$, $\vartheta \in \Theta$, von Wahrscheinlichkeitsmaßen. Bezeichne X die Identität auf Ω. Sei

$s : \Theta \to \mathbb{R}$ eine Funktion. Wir beobachten $X = x$ und wollen den Wert $s(\vartheta)$ schätzen.

Definition. Ein *Schätzer* (für s) ist eine meßbare Abbildung $S : \Omega \to \mathbb{R}$. Eine *Verlustfunktion* ist eine Abbildung $L : \Theta \times \mathbb{R} \to [0, \infty)$ mit $L(\vartheta, \cdot)$ meßbar und $L(\vartheta, s(\vartheta)) = 0$ für $\vartheta \in \Theta$. Die Verlustfunktion heißt *(strikt) konvex*, wenn $L(\vartheta, \cdot)$ (strikt) konvex für jedes $\vartheta \in \Theta$ ist. Die Verlustfunktion $L(\vartheta, s) = (s - s(\vartheta))^2$ heißt *quadratisch*.

Es ist kein realistisches Ziel, S so zu wählen, daß $L(\vartheta, S(x))$ für alle ϑ und x minimiert wird. Wie versuchen stattdessen, den mittleren Verlust zu minimieren.

Definition. Die *Risikofunktion* von S ist $R(\vartheta, S) = E_\vartheta L(\vartheta, S(X))$.

Satz 8 *(Rao und Blackwell) Sei T suffizient für ϑ, S ein P_ϑ-integrierbarer Schätzer und L eine konvexe Verlustfunktion für s. Dann gilt $R(\vartheta, S^T) \leq R(\vartheta, S)$ für $S^T = E(S|T)$. Ist die Verlustfunktion strikt konvex und $L(\vartheta, S)$ P_ϑ-integrierbar, so ist die Ungleichung strikt, falls nicht $S = S^T$ f.s.*

Beweis. Nach der Jensenschen Ungleichung gilt

$$L(\vartheta, E(S|t)) \leq E(L(\vartheta, S)|t) \quad P_\vartheta^T\text{-f.s.}$$

Die Ungleichung ist strikt, wenn $L(\vartheta, \cdot)$ strikt konvex ist. Die Behauptung folgt durch Integration nach P_ϑ^T.

5 Erwartungstreue Schätzer

Die Minimierung von $R(\vartheta, S)$ ist immer noch kein realistisches Ziel, wenn man keine Einschränkungen an die konkurrierenden Schätzer macht. Zum Beispiel haben die konstanten Schätzer $S(x) = c$ Risiko Null, falls $s(\vartheta) = c$ für das wahre ϑ gilt. Für einen sinnvollen Optimalitätsbegriff muß man solche "voreingenommenen" Schätzer ausschließen.

Definition. Ein Schätzer S ist *erwartungstreu* (für s), wenn $E_\vartheta S = s(\vartheta)$ für jedes $\vartheta \in \Theta$ gilt.

Jede (meßbare) Funktion S ist ein erwartungstreuer Schätzer ihres Erwartungswerts (falls er existiert).

Beispiel. Sei X verteilt nach $Bi_{n,p}$, $p \in (0,1)$, und sei $s(p) = 1/p$. Ein Schätzer S ist erwartungstreu für s, wenn

$$\sum_{k=0}^{n} S(k)\binom{n}{k}p^k(1-p)^{n-k} = \frac{1}{p}, \quad p \in (0,1).$$

Für $p \to 0$ geht die linke Seite gegen $S(0)$, die rechte gegen unendlich. Also existiert kein erwartungstreuer Schätzer.

Es gibt Funktionen s, die sich auf großen Verteilungsfamilien erwartungstreu schätzen lassen.

Beispiel. Seien $X_1, \ldots, X_n$ unabhängig und verteilt nach P aus einer Familie $\mathcal{P}$ auf $\mathcal{F}$. Sei f integrierbar für $P \in \mathcal{P}$. Setze $\vartheta = P$ und $s(P) = E_P f$. Ein erwartungstreuer Schätzer für s ist der *empirische Schätzer*

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i).$$

Definition. Sei L eine Verlustfunktion für s. Ein erwartungstreuer Schätzer S für s heißt *erwartungstreu mit gleichmäßig kleinstem Risiko* für L, wenn für jeden erwartungstreuen Schätzer S' gilt:

$$R(\vartheta, S) \leq R(\vartheta, S'), \quad \vartheta \in \Theta.$$

Ist die Verlustfunktion die quadratische, so heißt S *UMVU* (*uniformly minimum variance unbiased*).

Definition. Eine Funktion s heißt *erwartungstreu schätzbar*, wenn ein erwartungstreuer Schätzer für s existiert.

Lemma 1 *Sei T vollständig und suffizient für $\vartheta \in \Theta$. Ist s erwartungstreu schätzbar, so existiert (f.s.) genau ein erwartungstreuer Schätzer, der eine Funktion von T ist.*

Beweis. Existenz. Nach Voraussetzung gibt es einen erwartungstreuen Schätzer S. Setze $S^T = E_\vartheta(S|T)$.

Eindeutigkeit. Seien $S \circ T$ und $S' \circ T$ zwei erwartungstreue Schätzer, die Funktionen von T sind. Dann gilt $E_\vartheta(S \circ T - S' \circ T) = 0$ für $\vartheta \in \Theta$. Wegen der Vollständigkeit folgt $S \circ T = S' \circ T$ P_ϑ-f.s. für $\vartheta \in \Theta$.

Satz 9 *Sei T vollständig und suffizient für $\vartheta \in \Theta$. Sei s erwartungstreu schätzbar.*

a) Es existiert ein erwartungstreuer Schätzer S, der gleichmäßig kleinstes Risiko für jede konvexe Verlustfunktion hat.

b) Er ist der eindeutige erwartungstreue Schätzer, der eine Funktion von T ist.

c) Ist $L(\vartheta, \cdot)$ strikt konvex und $L(\vartheta, S)$ integrierbar, so ist S der eindeutige erwartungstreue Schätzer mit minimalem Risiko.

Beweis. a) Sei S erwartungstreu für s. Der Schätzer $S^T = E_\vartheta(S|T)$ ist ebenfalls erwartungstreu und hat gleichmäßig kleineres Risiko nach Satz 8.

b) Obiges Lemma.

c) Nach Zusatz zu Satz 8 wird das minimale Risiko nur von S^T angenommen.

Wenn wir eine vollständige und suffiziente Statistik T haben, finden wir einen gleichmäßig besten erwartungstreuen Schätzer durch Lösen von $E_\vartheta(S \circ T) = s(\vartheta)$, $\vartheta \in \Theta$, nach S. Falls wir schon einen erwartungstreuen Schätzer S gefunden haben, brauchen wir nur noch $S^T = E_\vartheta(S|T)$ zu berechnen. Das folgende Beispiel zeigt, daß UMVU-Schätzer ziemlich schlecht sein können.

Beispiel. Sei X verteilt nach P_λ, $\lambda > 0$. Der (eindeutige) UMVU-Schätzer für $\exp(-2\lambda)$ ist

$$S = \begin{cases} 1 & X \text{ gerade} \\ -1 & X \text{ ungerade.} \end{cases}$$

P_λ, $\lambda > 0$, ist eine exponentielle Familie in $T = X$. Also ist X vollständig nach Satz 7 und dem Korollar zu Satz 5. Nach Satz 9 (Lehmann–Scheffé) ist der UMVU-Schätzer der (eindeutige) erwartungstreue Schätzer der Form $S \circ X$. Nach der obigen Methode finden wir S, indem wir $E_\lambda S = \exp(-2\lambda)$ nach S lösen. Es gilt

$$E_\lambda S = e^{-\lambda} \sum_{k=0}^{\infty} S(k) \frac{\lambda^k}{k!}, \quad e^{-\lambda} = \sum_{k=0}^{\infty} (-1)^k \frac{\lambda^k}{k!}.$$

Die Behauptung folgt durch Koeffizientenvergleich.

6 Cramér–Rao-Ungleichung

Es gilt $E_\vartheta(S - s(\vartheta))^2 = E_\vartheta(S - E_\vartheta S)^2 + (E_\vartheta S - s(\vartheta))^2$. Für die quadratische Verlustfunktion ist das Risiko also die Summe aus der Varianz des Schätzers

und dem Quadrat des *Bias* $E_\vartheta S - s(\vartheta)$. Für erwartungstreue Schätzer verschwindet der Bias, und das Risiko ist gleich der Varianz des Schätzers. In diesem Kapitel geben wir eine untere Schranke dafür an.

Sei $P_\vartheta|\mathcal{F}$, $\vartheta \in \Theta$, eine Familie äquivalenter Wahrscheinlichkeitsmaße und $s : \Theta \to \mathbb{R}$ eine Funktion. Der *Dichtequotient* von P_τ nach P_ϑ ist $L_{\vartheta\tau} = dP_\tau/dP_\vartheta$. Es gilt $E_\vartheta L_{\vartheta\tau} = 1$. Sei $\vartheta \in \Theta$ fest.

Satz 10 *(Hammersley–Chapman–Robbins-Ungleichung) Sei S erwartungstreu für s. Dann gilt (mit $0/0 = 0$)*

$$Var_\vartheta S \geq \sup_{\tau\in\Theta} \frac{(s(\tau) - s(\vartheta))^2}{E_\vartheta(L_{\vartheta\tau} - 1)^2}.$$

Beweis. Mit $E_\vartheta(L_{\vartheta\tau} - 1) = 0$ gilt

$$s(\tau) - s(\vartheta) = E_\tau S - E_\vartheta S = E_\vartheta((L_{\vartheta\tau} - 1)S) = E_\vartheta((L_{\vartheta\tau} - 1)(S - s(\vartheta)).$$

Jetzt die Schwarzsche Ungleichung anwenden.

Im folgenden sei $\Theta \subset \mathbb{R}$, und ϑ liege im Inneren von Θ.

Definition. $L_{\vartheta\tau}$ heißt im *quadratischen Mittel differenzierbar* in $\tau = \vartheta$ mit Ableitung $\dot\ell_\vartheta \in L_2(P_\vartheta)$, wenn

$$\big(E_\vartheta(L_{\vartheta\tau} - 1 - (\tau - \vartheta)\dot\ell_\vartheta)^2\big)^{1/2} = o(\tau - \vartheta).$$

Wegen $E_\vartheta(L_{\vartheta\tau} - 1) = 0$ gilt $E_\vartheta\dot\ell_\vartheta = 0$, denn

$$|E_\vartheta\dot\ell_\vartheta| = \Big|E_\vartheta\Big(\frac{L_{\vartheta\tau} - 1}{\tau - \vartheta} - \dot\ell_\vartheta\Big)\Big| \leq \Big(E_\vartheta\Big(\frac{L_{\vartheta\tau} - 1}{\tau - \vartheta} - \dot\ell_\vartheta\Big)^2\Big)^{1/2} = o(1).$$

Die *Fisher-Information* in ϑ ist $I_\vartheta = E_\vartheta\dot\ell_\vartheta^2 = Var_\vartheta\dot\ell_\vartheta$.

Satz 11 *(Cramér–Rao-Ungleichung) Sei $L_{\vartheta\tau}$ im quadratischen Mittel differenzierbar in $\tau = \vartheta$. Die Fisher-Information in ϑ sei positiv. Sei s differenzierbar in ϑ mit Ableitung $s'(\vartheta)$, und sei S erwartungstreu für s. Dann gilt*

$$Var_\vartheta S \geq \frac{s'(\vartheta)^2}{I_\vartheta}.$$

Beweis. Es gilt nach Voraussetzung für $\tau \to \vartheta$,

$$\Big|\Big(E_\vartheta\Big(\frac{L_{\vartheta\tau}-1}{\tau-\vartheta}\Big)^2\Big)^{1/2} - (E_\vartheta \dot{\ell}_\vartheta^2)^{1/2}\Big| \le \Big(E_\vartheta\Big(\frac{L_{\vartheta\tau}-1}{\tau-\vartheta} - \dot{\ell}_\vartheta\Big)^2\Big)^{1/2} \to 0.$$

Wegen $E_\vartheta \dot{\ell}_\vartheta^2 = I_\vartheta > 0$ folgt die Behauptung aus Satz 10.

Die Cramér–Rao-Schranke ist offensichtlich i.a. schlechter als die Hammersley–Chapman–Robbins-Schranke. Insbesondere ist sie i.a. nicht scharf.

7 Neyman–Pearson-Lemma

Gegeben sei eine Hypothese $H \subset \Theta$. Die Alternative sei $K = \Theta \setminus H$. Wir beobachten $X = x$ und wollen testen, ob H zutrifft oder nicht.

Definition. Ein *(randomisierter) Test* (für H gegen K) ist eine meßbare Abbildung $\varphi : \Omega \to [0,1]$.

Wird x beobachtet, so entscheiden wir uns mit Wahrscheinlichkeit $\varphi(x)$ für K. Ist $\varphi = 1_C$, so heißt φ *nichtrandomisiert* und C *kritischer Bereich.*

Definition. Die *Gütefunktion* von φ ist die Abbildung $\vartheta \to E_\vartheta\varphi$.

Die Alternative beschreibt gewöhnlich die riskanteren Entscheidungen. Eine fälschliche Entscheidung dafür heißt *Fehler erster Art*; der umgekehrte Fehler heißt *Fehler zweiter Art.* Für $\vartheta \in H$ ist $E_\vartheta\varphi$ die Wahrscheinlichkeit für den Fehler erster Art; für $\vartheta \in K$ ist $1 - E_\vartheta\varphi$ die für den Fehler zweiter Art. Wir können nicht beide Wahrscheinlichkeiten gleichzeitig minimieren. Deshalb beschränken wir die Wahrscheinlichkeit für den Fehler erster Art und versuchen einen Test zu finden, der unter dieser Nebenbedingung die Wahrscheinlichkeit für den Fehler zweiter Art minimiert, also die Güte $E_\vartheta\varphi$ für $\vartheta \in K$ maximiert.

Definition. Ein Test φ hat das Niveau α (für H), wenn $E_\vartheta\varphi \le \alpha$ für $\vartheta \in H$.

Definition. Ein Test ψ zum Niveau α (für H) heißt *bester Test* zum Niveau α gegen $\vartheta \in K$, wenn für jeden Test φ zum Niveau α gilt: $E_\vartheta\psi \ge E_\vartheta\varphi$. Der Test ψ heißt *gleichmäßig bester Test* zum Niveau α (gegen K), wenn er optimal gegen alle $\vartheta \in K$ ist.

Das ist im allgemeinen nicht gleichzeitig für alle $\vartheta \in K$ möglich. Zunächst betrachten wir den Fall, daß H und K *einfach*, nämlich einpunktig

sind, und schreiben P und Q für Hypothese und Alternative und p und q für ihre Dichten bezüglich eines dominierenden Maßes μ. Dann läßt sich $E_Q\varphi$ maximieren. Das α-*Quantil* einer Verteilung $P|\mathcal{B}$ mit Verteilungsfunktion $F(x) = P(-\infty, x]$ ist

$$F^{-1}(\alpha) = \inf\{x : F(x) \geq \alpha\}.$$

Satz 12 *(Neyman–Pearson-Lemma) Sei P ein Wahrscheinlichkeitsmaß. Sei Q ein endliches signiertes Maß.*

a) Sei ψ ein Test mit

$$\psi = \begin{cases} 1 & q > cp \\ 0 & q < cp. \end{cases}$$

Dann gilt für jeden Test φ:

$$E_Q\varphi \leq E_Q\psi + cE_P(\varphi - \psi).$$

b) Gilt $E_P\varphi \leq E_P\psi$ und $E_Q\varphi \geq E_Q\psi$, so gilt

$$\varphi = \begin{cases} 1 & q > cp \\ 0 & q < cp. \end{cases}$$

c) Sei $\alpha \in (0, 1)$. Sei c das $(1-\alpha)$-Quantil der Verteilung von q/p unter P, also

$$c = \inf\{y : P(q > yp) \leq \alpha\}.$$

Setze

$$a = \begin{cases} \frac{\alpha - P(q>cp)}{P(q=cp)} & P(q = cp) > 0 \\ 0 & P(q = cp) = 0, \end{cases}$$

und

$$\psi = \begin{cases} 1 & q > cp \\ a & q = cp \\ 0 & q < cp. \end{cases}$$

Dann gilt $E_P\psi = \alpha$.

Beweis. a) Es gilt $(q - cp)(\psi - \varphi) \geq 0$.

b) Nach a) gilt $0 \leq \int (q-cp)(\psi-\varphi)d\mu = E_Q\psi - E_Q\varphi - c(E_P\psi - E_P\varphi) \leq 0$. Da der Integrand nichtnegativ ist, folgt die Behauptung.

c) Es gilt $E_P\psi = P(q > cp) + aP(q = cp) = \alpha$.

Die Randomisierung ist nötig, wenn die Verteilungsfunktion von q/p unter P in c den Wert $1-\alpha$ nicht annimmt. Die Randomisierung ist dieselbe für alle x, für die $q(x) = cp(x)$ gilt.

Ist $\alpha \in (0,1)$ und ψ der *Neyman–Pearson-Test* aus Satz 12c, so gilt $E_Q\varphi \leq E_Q\psi$ für jeden Test φ zum Niveau α. Ist umgekehrt φ ein bester Test zum Niveau α, so ist φ nach Satz 12b $(P+Q)$-f.s. von der Form

$$\varphi = \begin{cases} 1 & q > cp \\ 0 & q < cp. \end{cases}$$

Also unterscheidet sich φ höchstens auf $\{q = cp\}$ von ψ.

8 Monotone Dichtequotienten und gleichmäßig beste Tests

Sei $P_\vartheta|\mathcal{F}$, $\vartheta \in \Theta$, eine Familie von Wahrscheinlichkeitsmaßen. Wir hatten in Kapitel 6 Dichtequotienten für äquivalente Maße eingeführt. Im allgemeinen dominiert P_ϑ nicht P_τ. Für $\mu = P_\vartheta + P_\tau$ setzen wir dann

$$p_\vartheta = \frac{dP_\vartheta}{d\mu}, \quad p_\tau = \frac{dP_\tau}{d\mu}, \quad L_{\vartheta\tau} = \frac{p_\tau}{p_\vartheta}\mathbf{1}(p_\vartheta > 0) + \infty\mathbf{1}(p_\vartheta = 0, p_\tau > 0).$$

Im folgenden nehmen wir $\Theta \subset \mathbb{R}$ an.

Definition. Sei $T : \Omega \in \mathbb{R}$ eine Zufallsvariable. Die Familie P_ϑ, $\vartheta \in \Theta$, hat *monotone Dichtequotienten* in T, wenn für $\vartheta, \tau \in \Theta$ mit $\vartheta < \tau$ eine nichtfallende Funktion $H_{\vartheta\tau}$ existiert mit $L_{\vartheta\tau} = H_{\vartheta\tau} \circ T$ $(P_\vartheta + P_\tau)$-f.s.

Satz 13 *Hat P_ϑ, $\vartheta \in \Theta$, monotone Dichtequotienten in T, und ist $\alpha \in (0,1)$, so existiert ein gleichmäßig bester Test zum Niveau α für $\tau \leq \vartheta$ gegen $\tau > \vartheta$, nämlich*

$$\psi = \begin{cases} 1 & T > b \\ a & T = b \\ 0 & T < b, \end{cases}$$

wobei a und b bestimmt sind durch $E_\vartheta\psi = \alpha$.

Für $\tau < \vartheta$ minimiert $\varphi = \psi$ das Niveau $E_\tau\varphi$ unter allen Tests φ mit $E_\vartheta\varphi = \alpha$.

Beweis. Wie im Beweis von Satz 12b wähle a und b mit $E_\vartheta\psi = \alpha$. Sei $\tau > \vartheta$. Nach Vorausetzung hat ψ die Form

$$\psi = \begin{cases} 1 & L_{\vartheta\tau} > c = H_{\vartheta\tau}(b) \\ 0 & L_{\vartheta\tau} < c = H_{\vartheta\tau}(b). \end{cases}$$

Nach Satz 12a ist ψ bester Test zum Niveau α für ϑ gegen τ. Analog ist $1-\psi$ bester Test zum Niveau $1-\alpha$ für ϑ gegen $\tau < \vartheta$. Insbesondere gilt für $\tau < \vartheta$:

$$E_\tau(1-\psi) \geq E_\tau(1-\alpha) = 1-\alpha,$$

also $E_\tau\psi \leq \alpha$.

Satz 14 *Sei P_ϑ, $\vartheta \in \Theta$, eine eindimensionale exponentielle Familie in $\eta(\vartheta)$ und T. Ist η nichtfallend (nichtwachsend), so hat die Familie monotone Dichtequotienten in T ($-T$).*

Beweis. P_ϑ hat μ-Dichte der Form $c(\vartheta)\exp(\eta(\vartheta)T)$. Also gilt

$$L_{\vartheta\tau} = \frac{c(\tau)}{c(\vartheta)} e^{(\eta(\tau)-\eta(\vartheta))T}.$$

Ist η nichtfallend und $\tau > \vartheta$, so gilt $\eta(\tau) - \eta(\vartheta) \geq 0$. Also ist

$$H_{\vartheta\tau}(t) = \frac{c(\tau)}{c(\vartheta)} e^{(\eta(\tau)-\eta(\vartheta))t}$$

nichtfallend in t.

Beispiel. Seien $X_1, \ldots, X_n$ unabhängig mit Bernoulliverteilung $Bi_{1,p}$, $p \in (0,1)$. Nach Kapitel 1 ist $Bi_{1,p}$, $p \in (0,1)$, eine exponentielle Familie in $\eta(p) = \log(p/(1-p))$ und $T(X) = X$. Also ist $Bi^n_{1,p}$, $p \in (0,1)$, eine exponentielle Familie in $\eta(p)$ und $T(X_1, \ldots, X_n) = \sum_{i=1}^n X_i$. Die Funktion $p \to \log(p/(1-p))$ ist nichtfallend. Nach Satz 14 hat also $(Bi_{1,p})^n$, $p \in (0,1)$, monotone Dichtequotienten in $\sum X_i$. Nach Satz 13 ist dann ein gleichmäßig bester Test zum Niveau α für $q \leq p$ gegen $q > p$ gegeben durch

$$\psi = \begin{cases} 1 & \sum X_i > b \\ a & \sum X_i = b \\ 0 & \sum X_i < b, \end{cases}$$

wobei a und b bestimmt sind durch $E_p\psi = \alpha$. Unter $(Bi_{1,p})^n$ ist $\sum X_i$ verteilt wie $Bi_{n,p}$. Also ergeben sich a und b aus

$$Bi_{n,p}\{b+1, \ldots, n\} + aBi_{n,p}\{b\} = \alpha.$$

9 Lokal beste Tests

Sei $P_\vartheta|\mathcal{F}$, $\vartheta \in \Theta \subset \mathbb{R}$, eine Familie äquivalenter Wahrscheinlichkeitsmaße. Sei ϑ im Inneren von Θ.

Definition. $L_{\vartheta\tau}$ ist *im Mittel differenzierbar* in $\tau = \vartheta$ mit Ableitung $\dot{\ell}_\vartheta \in L_1(P_\vartheta)$, wenn

$$E_\vartheta \big| L_{\vartheta\tau} - 1 - (\tau - \vartheta)\dot{\ell}_\vartheta \big| = o(\tau - \vartheta).$$

Lemma 2 *Sei $L_{\vartheta\tau}$ im Mittel differenzierbar in $\tau = \vartheta$. Dann gilt für jeden Test φ, daß $E_\tau\varphi$ differenzierbar in $\tau = \vartheta$ mit Ableitung $E_\vartheta(\dot{\ell}_\vartheta\varphi)$ ist.*

Beweis. Es gilt

$$\frac{E_\tau\varphi - E_\vartheta\varphi}{\tau - \vartheta} = \frac{E_\vartheta((L_{\vartheta\tau} - 1)\varphi)}{\tau - \vartheta} \to E_\vartheta(\dot{\ell}_\vartheta\varphi).$$

Definition. Ein Test φ heißt *α-ähnlich* für ϑ, wenn $E_\vartheta\varphi = \alpha$. Ein α-ähnlicher Test ψ heißt *lokal bester ähnlicher Test* zum Niveau α gegen $\tau > \vartheta$, wenn für jeden α-ähnlichen Test φ gilt:

$$\partial_{\tau=\vartheta} E_\tau\psi \geq \partial_{\tau=\vartheta} E_\tau\varphi.$$

Satz 15 *Sei $L_{\vartheta\tau}$ im Mittel differenzierbar in $\tau = \vartheta$ mit Ableitung $\dot{\ell}_\vartheta$. Dann existiert ein lokal bester ähnlicher Test zum Niveau α gegen $\tau > \vartheta$, nämlich*

$$\psi = \begin{cases} 1 & \dot{\ell}_\vartheta > c \\ a & \dot{\ell}_\vartheta = c \\ 0 & \dot{\ell}_\vartheta < c, \end{cases}$$

wobei a und c bestimmt sind durch $E_\vartheta\psi = \alpha$.

Beweis. Es ist zu zeigen: $\varphi = \psi$ maximiert $E_\vartheta(\dot{\ell}_\vartheta\varphi)$ unter den α-ähnlichen Tests. Dazu wenden wir Satz 12a (Neyman–Pearson-Lemma) an für $P = P_\vartheta$, $dQ = \dot{\ell}_\vartheta dP_\vartheta$ und $\mu = P_\vartheta$.

10 Konfidenzbereiche

Sei $P_\vartheta|\mathcal{F}$, $\vartheta \in \Theta \subset \mathbb{R}$, eine Familie von Wahrscheinlichkeitsmaßen. Bezeichne X die Identität auf Ω und $\mathfrak{P}$ die Menge der Teilmengen von $\mathbb{R}$.

Definition. Ein *Konfidenzbereich* (für ϑ) ist eine Abbildung $B : \Omega \to \mathfrak{P}$ mit $\{x \in \Omega : \vartheta \in B(x)\} \in \mathcal{F}$ für $\vartheta \in \mathbb{R}$.

Wird $X = x$ beobachtet, so nehmen wir an, daß der wahre Parameter in $B(x)$ liegt. Manchmal akzeptieren wir auch zu große (oder zu kleine) Parameter. Das entspricht den Hypothesen $\tau \leq \vartheta$ (oder $\tau \geq \vartheta$) statt $\{\vartheta\}$. Allgemein bezeichne H_ϑ die Menge der akzeptablen Parameter, wenn ϑ wahr ist.

Definition. Ein Konfidenzbereich B hat das *Niveau* $1 - \alpha$ (für ϑ und H), wenn

$$P_\tau\{x \in \Omega : \vartheta \in B(x)\} \geq 1 - \alpha, \quad \tau \in H_\vartheta, \vartheta \in \Theta.$$

Definition. Ein Konfidenzbereich B^* zum Niveau $1 - \alpha$ heißt *gleichmäßig bester Konfidenzbereich zum Niveau* $1 - \alpha$, wenn für jeden Konfidenzbereich B zum Niveau $1 - \alpha$ gilt:

$$P_\tau(\vartheta \in B^*) \leq P_\tau(\vartheta \in B), \quad \tau \notin H_\vartheta, \vartheta \in \Theta.$$

Satz 16 *a) Hat der Konfidenzbereich B das Niveau $1 - \alpha$ für H, so hat für jedes ϑ der kritische Bereich $C_\vartheta = \{x \in \Omega : \vartheta \notin B(x)\}$ das Niveau α für H_ϑ. Ist B gleichmäßig bester Konfidenzbereich, so ist C_ϑ gleichmäßig bester kritischer Bereich für jedes ϑ.*

b) Hat C_ϑ das Niveau α für H_ϑ, so hat der Konfidenzbereich $B(x) = \{\vartheta : x \notin C_\vartheta\}$ das Niveau $1 - \alpha$ für H. Ist C_ϑ gleichmäßig bester kritischer Bereich für jedes ϑ, so ist B gleichmäßig bester Konfidenzbereich.

Beweis. Es gilt $P_\tau C_\vartheta = P_\tau(\vartheta \notin B(x))$.

11 M-Schätzer und Maximum-Likelihood-Schätzer

Beispiel. Sei $\mathcal{P}$ eine Familie von Wahrscheinlichkeitsmaßen auf $(\mathbb{R}, \mathcal{B})$ mit endlicher Varianz. Seien $X_1, \ldots, X_n$ unabhängig mit Verteilung $P \in \mathcal{P}$. Ein Schätzer für den Erwartungswert $s(P) = E_P(X) = PX$ ist das Stichprobenmittel $\frac{1}{n}\sum_{i=1}^n X_i$. Sei $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ die *empirische Verteilung*. Hier bezeichnet δ_x das *Dirac-Maß* in x. Das Stichprobenmittel läßt sich schreiben als $s(\mathbb{P}_n)$. Ist s geeignet stetig, so ist $s(\mathbb{P}_n)$ konsistent; ist s

geeignet differenzierbar, so ist $n^{1/2}(s(\mathbb{P}_n) - s(P))$ asymptotisch normal. In diesem Beispiel wissen wir das schon:

$$n^{1/2}(s(\mathbb{P}_n) - s(P)) = n^{-1/2} \sum_{i=1}^{n} (X_i - PX).$$

Der Erwartungswert ist das Minimum von $P(X-s)^2$ in s, denn $P(X-s)^2 = P(X - PX)^2 + (PX - s)^2$. Entsprechend ist das Stichprobenmittel $s(\mathbb{P}_n)$ das Minimum der stochastischen Version $\mathbb{P}(X - s)^2$. Ein solcher Schätzer heißt M-Schätzer. Für $s(P)$ muß gelten:

$$\partial_{s=s(P)} P(X - s)^2 = 0.$$

Ebenso muß für $s(\mathbb{P}_n)$ gelten:

$$\partial_{s=s(\mathbb{P}_n)} \mathbb{P}_n(X - s)^2 = 0.$$

In unserem Beispiel wissen wir das schon:

$$P(X - s(P)) = 0, \quad \mathbb{P}_n(X - s(\mathbb{P}_n)) = 0.$$

Das läßt sich verallgemeinern.

Definition. Sei $\Theta \neq \emptyset$ und ψ_ϑ meßbar für $\vartheta \in \Theta$. Für $P \in \mathcal{P}$ habe $\tau \to P\psi_\tau(X)$ ein eindeutiges Minimum $s(P)$ in Θ. Dann heißt $s : \mathcal{P} \to \Theta$ *M-Funktional*. Ist s fortgesetzt auf $\mathbb{P}_n$, so heißt $s(\mathbb{P}_n)$ *M-Schätzer*.

Bemerkung. Sei $\Theta \subset \mathbb{R}^d$ und $s(P)$ im Inneren von Θ. Ist $\tau \to P\psi_\tau(X)$ differenzierbar unter dem Integral in $\tau = s(P)$, so gilt

$$\partial_{\tau=s(P)} P\psi_\tau(X) = P\dot{\psi}_{s(P)}(X) = 0.$$

Hier ist $\dot{\psi}_\tau$ der Vektor der partiellen Ableitungen von ψ_τ nach den Komponenten von τ. Dann bestimmt man den M-Schätzer als Lösung $\tau = \hat{\vartheta}$ der *Schätzgleichung*

$$\mathbb{P}_n\dot{\psi}_\tau(X) = \frac{1}{n} \sum_{i=1}^{n} \dot{\psi}_\tau(X_i) = 0.$$

Es macht asymptotisch i.a. keinen Unterschied und ist einfacher, stattdessen nur einen *asymptotischen* M-Schätzer zu nehmen, d.h. eine Lösung der Schätzgleichung

$$\frac{1}{n} \sum_{i=1}^{n} \dot{\psi}_\tau(X_i) = o_p(n^{-1/2}).$$

Bemerkung. Sei $(\Omega, \mathcal{F}) = (\mathbb{R}^d, \mathcal{B}^d)$. Die von einer Verteilung $P|\mathcal{F}$ *erzeugte Lageparameter-Familie* P_a, $a \in \mathbb{R}$, ist definiert durch $P_a B = P(B - a)$. Eine Familie $\mathcal{P}$ heißt *Lage-abgeschlossen*, wenn $P_a \in \mathcal{P}$ für $a \in \mathbb{R}$ und $P \in \mathcal{P}$. Ein Funktional $s : \mathcal{P} \in \mathbb{R}^d$ heißt *Lage-Funktional*, wenn $s(P_a) = s(P) + a$.

Ist $\psi_\tau(x) = \psi(x - \tau)$, so ist das zugehörige M-Funktional ein Lage-Funktional, denn

$$P_a \psi(X - \tau) = P\psi(X - (\tau - a)).$$

Für $d = 1$ und $\psi(x) = x^2$ ergibt sich das M-Funktional $s(P) = PX$ aus dem obigen Beispiel.

Satz 17 *Sei $P|\mathcal{F}$ ein Wahrscheinlichkeitsmaß, $\Theta \subset \mathbb{R}^d$ und ϑ im Inneren von Θ. Für τ in einer Umgebung von ϑ sei $\dot{\psi}_\tau$ eine d-dimensionale meßbare Funktion und $\dot{\psi}_\tau(x)$ stetig differenzierbar in τ. Für die Matrix $\ddot{\psi}_\tau$ der partiellen Ableitungen von $\dot{\psi}_\tau$ gelte $|\ddot{\psi}_\tau(x)| \le H(x)$ für ein P-integrierbares H, und $P\ddot{\psi}_\vartheta$ sei invertierbar. Sei $\hat{\vartheta} = \vartheta + o_P(1)$ eine (konsistente) Lösung der Schätzgleichung $\frac{1}{n}\sum_{i=1}^n \dot{\psi}_\tau(X_i) = o_P(n^{-1/2})$. Dann gilt*

$$n^{1/2}(\hat{\vartheta} - \vartheta) = -(P\ddot{\psi}_\vartheta)^{-1} n^{-1/2} \sum_{i=1}^n \dot{\psi}_\vartheta(X_i) + o_P(1).$$

Beweis. Mit einer Taylorentwicklung erhalten wir

$$\dot{\psi}_\tau(x) = \dot{\psi}_\vartheta(x) + \ddot{\psi}_\vartheta(x)(\tau - \vartheta) + \int_0^1 \left(\ddot{\psi}_{\vartheta + s(\tau-\vartheta)}(x) - \ddot{\psi}_\vartheta(x)\right) ds\,(\tau - \vartheta).$$

Aus der Stetigkeit von $\tau \mapsto \ddot{\psi}_\tau(x)$ folgt

$$h_a(x) = \sup_{|\tau - \vartheta| \le a} |\ddot{\psi}_\tau(x) - \ddot{\psi}_\vartheta(x)| \downarrow 0, \quad a \downarrow 0.$$

Mit dem Satz von der dominierten Konvergenz also $Ph_a \downarrow 0$ für $a \downarrow 0$. Für $a_n \le a$ gilt mit dem starken Gesetz der großen Zahl

$$\limsup_n \frac{1}{n}\sum_{i=1}^n h_{a_n}(X_i) \le \limsup_n \frac{1}{n}\sum_{i=1}^n h_a(X_i) = Ph_a \quad \text{f.s.},$$

also $(1/n)\sum_{i=1}^n h_{a_n}(X_i) \to 0$ f.s. Nach Voraussetzung gibt es $a_n \downarrow 0$, so daß $P(|\hat{\vartheta} - \vartheta| > a_n) \to 0$. Mit der Taylorentwicklung für $\tau = \hat{\vartheta}$ gilt also

$$o_P(n^{-1/2}) = \frac{1}{n}\sum_{i=1}^n \dot{\psi}_{\hat{\vartheta}}(X_i) = \frac{1}{n}\sum_{i=1}^n \dot{\psi}_\vartheta(X_i) + \Big(\frac{1}{n}\sum_{i=1}^n \ddot{\psi}_\vartheta(X_i) + o_P(1)\Big)(\hat{\vartheta} - \vartheta).$$

Die Behauptung folgt durch Auflösen nach $\hat{\vartheta} - \vartheta$ und Anwendung des schwachen Gesetzes der großen Zahl.

Im folgenden geben wir Bedingungen an, unter denen M-Schätzer stark konsistent sind.

Lemma 3 *Sei $P|\mathcal{F}$ ein Wahrscheinlichkeitsmaß und $\Theta \subset \mathbb{R}^d$ offen. Sei ψ_τ stetig in τ und meßbar in x mit $\psi_\tau \geq H$ für alle τ und ein P-integrierbares H. Dann gilt für jedes kompakte $K \subset \Theta$:*

$$\liminf_n \inf_{\tau \in K} \frac{1}{n} \sum_{i=1}^n \psi_\tau(X_i) \geq \inf_{\tau \in K} P\psi_\tau \quad P\text{-f.s.}$$

Beweis. Sei $K \subset \Theta$ kompakt. Es gilt

$$a = \inf_{\tau \in K} P\psi_\tau \geq PH > -\infty.$$

Für $\varepsilon > 0$ und $\tau \in K$ setze

$$\psi_{\tau\varepsilon}(x) = \inf_{t \in K : |t-\tau| \leq \varepsilon} \psi_t(x).$$

Dann ist $\psi_{\tau\varepsilon}$ meßbar und $\psi_{\tau\varepsilon} \geq H$, und es gilt $\psi_{\tau\varepsilon}(x) \uparrow \psi_\tau(x)$ für $\varepsilon \downarrow 0$. Mit $\psi_{\tau\varepsilon} - H \geq 0$ folgt aus dem Satz von der monotonen Konvergenz, daß $P(\psi_{\tau\varepsilon} - H) \uparrow P(\psi_\tau - H)$, also $P\psi_{\tau\varepsilon} \uparrow P\psi_\tau$.

Sei $b < a$. Für $\tau \in K$ existiert $\varepsilon_\tau > 0$, so daß $P\psi_{\tau\varepsilon_\tau} \geq b$. Die offenen Kugeln $S_\tau = \{t \in K : |t - \tau| < \varepsilon_\tau\}$ überdecken K. Da K kompakt ist, existiert eine endliche Teilüberdeckung $S_{\tau_1}, \ldots, S_{\tau_m}$. Also gilt

$$\inf_{\tau \in K} \frac{1}{n} \sum_{i=1}^n \psi_\tau(X_i) \geq \min_{1 \leq j \leq m} \frac{1}{n} \sum_{i=1}^n \psi_{\tau_j \varepsilon_{\tau_j}}(X_i).$$

Aus dem starken Gesetz der großen Zahl folgt

$$\liminf_n \inf_{\tau \in K} \frac{1}{n} \sum_{i=1}^n \psi_\tau(X_i) \geq \min_{1 \leq j \leq m} P\psi_{\tau_j \varepsilon_{\tau_j}} \geq b.$$

Da $b < a$ beliebig war, folgt die Behauptung.

Lemma 4 *Seien die Voraussetzungen von Lemma 3 erfüllt, und ϑ sei das eindeutige Minimum von $\tau \to P\psi_\tau$ über Θ. Dann gilt für jedes kompakte $K \subset \Theta$ mit $\vartheta \notin K$:*

$$\inf_{\tau \in K} P\psi_\tau > P\psi_\vartheta.$$

Beweis. Mit dem Lemma von Fatou folgt

$$\liminf_{t\to\tau} P\psi_t = PH + \liminf_{t\to\tau} P(\psi_t - H) \geq P\psi_\tau.$$

Das heißt: $\tau \to P\psi_\tau$ ist unterhalbstetig. Für jedes kompakte K nimmt $P\psi_\tau$ das Minimum über K an. Da das Minimum von $P\psi_\tau$ über Θ eindeutig angenommen wird, folgt die Behauptung.

Satz 18 *Sei $P|\mathcal{F}$ ein Wahrscheinlichkeitsmaß und $\Theta \subset \mathbb{R}^d$ offen. Sei ψ_τ stetig in τ und meßbar in x mit $\psi_\tau \geq H$ für alle τ und ein P-integrierbares H. Sei ϑ das eindeutige Minimum von $\tau \to P\psi_\tau$ über Θ. Für eine kompakte Umgebung $K \subset \Theta$ von ϑ gelte*

$$(C) \qquad \liminf_n \inf_{\tau\notin K} \frac{1}{n}\sum_{i=1}^n \psi_\tau(X_i) > P\psi_\vartheta \quad P\text{-f.s.}$$

Erfülle $\hat{\vartheta}_n$ die Ungleichung

$$\frac{1}{n}\sum_{i=1}^n \psi_{\hat{\vartheta}_n}(X_i) \leq \inf_{\tau\in\Theta} \frac{1}{n}\sum_{i=1}^n \psi_\tau(X_i) + \frac{1}{n}.$$

Dann gilt $\hat{\vartheta}_n \to \vartheta$ P-f.s.

Beweis. Sei

$$A = \Big\{ \limsup_n |\hat{\vartheta}_n - \vartheta| > 0, \quad \lim_n \frac{1}{n}\sum_{i=1}^n \psi_\vartheta(X_i) = P\psi_\vartheta \Big\}.$$

Zu zeigen ist $PA = 0$. Sei $\omega = (x_1, x_2, \dots) \in A$. Dann existiert ein $\varepsilon > 0$ und eine wachsende Teilfolge, so daß $|\hat{\vartheta}_{m_n}(\omega) - \vartheta| \geq \varepsilon$ für $n \in \mathbb{N}$. Also gilt

$$\inf_{|\tau-\vartheta|\geq\varepsilon} \frac{1}{m_n}\sum_{i=1}^{m_n} \psi_\tau(x_i) \leq \frac{1}{m_n}\sum_{i=1}^{m_n} \psi_{\hat{\vartheta}_{m_n}(\omega)}(x_i) \leq \frac{1}{m_n}\sum_{i=1}^{m_n} \psi_\vartheta(x_i) + \frac{1}{m_n},$$

also

$$T_\varepsilon(\omega) = \liminf_n \inf_{|\tau-\vartheta|\geq\varepsilon} \frac{1}{n}\sum_{i=1}^n \psi_\tau(x_i) \leq P\psi_\vartheta,$$

das heißt $\omega \in B_\varepsilon = \{T_\varepsilon \leq P\psi_\vartheta\}$. Also gilt $A \subset \bigcup_{\varepsilon>0} B_\varepsilon$, und wegen $B_\varepsilon \uparrow$ für $\varepsilon \downarrow$ genügt zu zeigen, daß $PB_\varepsilon = 0$. Lemmas 3 und 4 implizieren

$$\liminf_n \inf_{\tau\in K, |\tau-\vartheta|\geq\varepsilon} \frac{1}{n}\sum_{i=1}^n \psi_\tau(X_i) \geq \inf_{\tau\in K, |\tau-\vartheta|\geq\varepsilon} P\psi_\tau > P\psi_\vartheta \quad P\text{-f.s.}$$

Nach Voraussetzung gilt

$$\liminf_n \inf_{\tau\notin K,|\tau-\vartheta|\geq\varepsilon} \frac{1}{n}\sum_{i=1}^n \psi_\tau(X_i) > P\psi_\vartheta \quad P\text{-f.s.}$$

Also gilt $PB_\varepsilon = 0$.

Bemerkung. Hinreichend für (C) ist die Existenz einer kompakten Umgebung $K \subset \Theta$ von ϑ mit

$$P \inf_{\tau\notin K} \psi_\tau > P\psi_\vartheta.$$

Im folgenden wenden wir Satz 17 auf eine parametrische Familie an.

Sei $\mathcal{P} = \{P_\vartheta : \vartheta \in \Theta\}$ mit $\Theta \subset \mathbb{R}^d$ eine parametrische Familie mit μ-Dichten f_ϑ für P_ϑ. Die Dichte von $X_1, \ldots, X_n$ ist $\prod_{i=1}^n f_\vartheta(X_i)$. Der *Maximum-Likelihood-Schätzer* $\hat\vartheta$ maximiert diese Dichte. Insbesondere gilt $\sum \partial_{\vartheta=\hat\vartheta} \log f_\vartheta(X_i) = 0$. Wir zeigen, daß die asymptotische Kovarianzmatrix dieses Schätzers gleich der Inversen der Fisher-Information ist.

Definition. Sei P_ϑ, $\vartheta \in \Theta \subset \mathbb{R}^d$, eine Familie von Wahrscheinlichkeitsmaßen und ϑ im Inneren von Θ. Die Familie ist *Cramér-regulär* bei ϑ, wenn P_τ für τ in einer Umgebung von ϑ eine positive μ-Dichte f_τ hat und f_τ und $\ell_\tau = \log f_\tau$ zweimal stetig differenzierbar in τ sind mit $|\dot f_\tau| \leq A$ und $|\ddot\ell_\tau| \leq H$ für μ-integrierbares A und P_ϑ-integrierbares H, und wenn die *Informationsmatrix* $I_\vartheta = P_\vartheta \dot\ell_\vartheta \dot\ell_\vartheta^\top$ positiv definit ist.

Lemma 5 *Für bei ϑ Cramér-reguläre Familien gilt $P_\tau\dot\ell_\tau = 0$ und $P_\tau\ddot\ell_\tau = -I_\tau$ für τ in einer Umgebung von ϑ.*

Beweis. Die Bedingungen an die zweiten Ableitungen implizieren ähnliche Bedingungen an die ersten Ableitungen. Insbesondere gilt

$$\dot f_\tau = \dot f_\vartheta + \int_0^1 \ddot f_{\vartheta+u(\tau-\vartheta)} du(\tau-\vartheta),$$

also sind $\dot f_\tau$ durch eine μ-integrierbare Funktion dominiert. Es gilt $\mu f_\tau = 1$, also

$$0 = \mu(f_t - f_\tau) = \mu\dot f_\tau(t-\tau) + \int_0^1 \mu(\dot f_{\tau+s(t-\tau)} - \dot f_\tau) ds(t-\tau).$$

Mit dem Satz von der dominierten Konvergenz konvergiert das Integral für $t \to \tau$ gegen 0. Also gilt $P_\tau \dot{\ell}_\tau = \mu \dot{f}_\tau = 0$. Mit demselben Argument gilt $\mu \ddot{f}_\tau = 0$. Wegen $\dot{f} = \dot{\ell} f$ gilt $\ddot{f} = \ddot{\ell} f + \dot{\ell} \dot{f}^\top = \ddot{\ell} f + \dot{\ell}\dot{\ell}^\top f$, insbesondere also $P_\tau \ddot{\ell}_\tau + P_\tau \dot{\ell}_\tau \dot{\ell}_\tau^\top = 0$.

Satz 19 *Sei P_ϑ, $\vartheta \in \Theta \subset \mathbb{R}^d$, Cramér-regulär bei ϑ. Sei $\hat{\vartheta} = \vartheta + o_P(1)$ eine Lösung der Schätzgleichung $(1/n)\sum_{i=1}^n \dot{\ell}_\tau(X_i) = o_P(n^{-1/2})$. Dann gilt*

$$n^{1/2}(\hat{\vartheta} - \vartheta) = I_\vartheta^{-1} n^{-1/2} \sum_{i=1}^n \dot{\ell}_\vartheta(X_i) + o_P(1).$$

Insbesondere ist $\hat{\vartheta}$ asymptotisch normal mit Kovarianz-Matrix I_ϑ^{-1}.

Beweis. Wende Satz 17 für $\dot{\psi} = -\dot{\ell}$ und $P = P_\vartheta$ an. Nach Lemma 5 ist $P_\vartheta \ddot{\ell}_\vartheta = -I_\vartheta$.

12 Empirische Schätzer und lineare Regression

Sei $\mathcal{P}$ eine Familie von Wahrscheinlichkeitsmaßen auf $(\Omega, \mathcal{F})$, und $X_1, \ldots, X_n$ seien unabhängig mit Verteilung P. Für $A \in \mathcal{F}$ gilt nach dem Gesetz der großen Zahl

$$\mathbb{P}_n A = \frac{1}{n}\sum_{i=1}^n 1_A(X_i) \to PA \quad \text{f.s.}$$

Nach dem zentralen Grenzwertsatz gilt

$$n^{1/2}(\mathbb{P}_n A - PA) = n^{-1/2}\sum_{i=1}^n (1_A(X_i) - PA) \Rightarrow N(0, PA(1-PA)).$$

Wir sagen (etwas mißverständlich): $\mathbb{P}_n A$ ist *asymptotisch normal* mit Varianz $PA(1-PA)$. Sei $f : \mathbb{R} \to \mathbb{R}^k$ mit $Pf_r^2 = E_P f_r^2 = \int f_r^2 dP < \infty$ für $r = 1, \ldots, k$. Der *empirische Schätzer* für Pf ist

$$\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^n f(X_i).$$

Wieder gilt

$$\mathbb{P}_n f \to Pf \quad \text{f.s.}$$

und

$$n^{1/2}(\mathbb{P}_n f - Pf) = n^{-1/2}\sum_{i=1}^n (f(X_i) - Pf) \Rightarrow N(0, \Sigma)$$

mit Kovarianzmatrix $\Sigma = (\sigma_{rs})$ und

$$\sigma_{rs} = P(f_r - Pf_r)(f_s - Pf_s) = Pf_rf_s - Pf_rPf_s.$$

Bemerkung. Ist $t : \mathbb{R}^k \to \mathbb{R}^m$ stetig differenzierbar mit $m \times k$-Matrix $Dt = (t_a^b)_{a,b}$ von partiellen Ableitungen, so erhält man durch Taylor-Entwicklung

$$\begin{aligned} t(\mathbb{P}_n f) &= t(Pf) + \int_0^1 Dt(Pf + u(\mathbb{P}_n f - Pf))\, du\, (\mathbb{P}_n f - Pf) \\ &= t(Pf) + Dt(Pf)\,(\mathbb{P}_n f - Pf) + o_p(n^{-1/2}). \end{aligned}$$

Also ist $t(\mathbb{P}_n f)$ asymptotisch normal mit Kovarianzmatrix $T\Sigma T^\top$, wobei $T = Dt(Pf)$.

Beispiel. (*Lineare Einschränkung.*) Sei $h : \Omega \to \mathbb{R}^m$ meßbar mit $Phh^\top$ positiv definit. Für $P \in \mathcal{P}$ gelte die lineare Einschränkung $Ph = 0$. Sei $f : \Omega \to \mathbb{R}^k$ meßbar mit $Pf_s^2 < \infty$ für $s = 1, \ldots, k$. Außer dem empirischen Schätzer $\mathbb{P}_n f$ finden wir dann noch andere erwartungstreue Schätzer für Pf: Für jede $k \times m$-Matrix A ergibt sich der erwartungstreue Schätzer

$$\mathbb{P}_n f - A\mathbb{P}_n h = \frac{1}{n}\sum_{i=1}^n (f(X_i) - Ah(X_i)).$$

Seine asymptotische Kovarianzmatrix ist

$$\Sigma(A) = P(f - Pf - Ah)(f - Pf - Ah)^\top.$$

Für $k \times k$-Matrizen A und B schreiben wir $A \le B$, wenn $B - A$ positiv semidefinit ist. Dadurch wird eine partielle Ordnung auf Matrizen eingeführt. Die Kovarianzmatrix $\Sigma(A)$ wird in dieser Ordnung minimiert durch

$$A^* = Pfh^\top (Phh^\top)^{-1}.$$

Denn wegen $Ph = 0$ ist A^*h die Projektion von $f - Pf$ auf den von h erzeugten linearen Raum:

$$P(f - Pf - A^*h)h^\top = P(f - A^*h)h^\top = Pfh^\top - Pfh^\top (Phh^\top)^{-1} Phh^\top = 0.$$

Wir haben die orthogonale Zerlegung

$$f - Pf - Ah = f - Pf - A^*h - (A - A^*)h,$$

also

$$\Sigma(A) - \Sigma(A^*) = (A - A^*)Phh^\top(A - A^*)^\top \geq 0.$$

Die Matrix $\Sigma(A^*)$ hängt vom unbekannten P ab. Wir schätzen A^*, indem wir die Erwartungswerte durch empirische Schätzer ersetzen:

$$\hat{A} = \mathbb{P}_n f h^\top (\mathbb{P}_n h h^\top)^{-1} = \sum_{i=1}^n f(X_i)h^\top(X_i)\Big(\sum_{i=1}^n h(X_i)h^\top(X_i)\Big)^{-1}.$$

Mit dem Gesetz der großen Zahl gilt $\hat{A} = A^* + o_P(1)$. Also ist der Schätzer $\mathbb{P}_n f - \hat{A}\mathbb{P}_n h$ asymptotisch äquivalent zu $\mathbb{P}_n f - A^*\mathbb{P}_n h$. Aus dem zentralen Grenzwertsatz und $Ph = 0$ ergibt sich $n^{1/2}\mathbb{P}_n h = O_P(1)$, also

$$\begin{aligned} n^{1/2}(\mathbb{P}_n f - \hat{A}\mathbb{P}_n h - Pf) &= n^{1/2}(\mathbb{P}_n f - A^*\mathbb{P}_n h - Pf) - (\hat{A} - A^*)n^{1/2}\mathbb{P}_n h \\ &= n^{1/2}(\mathbb{P}_n f - A^*\mathbb{P}_n h - Pf) + o_P(1). \end{aligned}$$

Insbesondere hat $\mathbb{P}_n f - \hat{A}\mathbb{P}_n h - Pf$ dieselbe asymptotische Kovarianzmatrix wie $\mathbb{P}_n f - A^*\mathbb{P}_n h - Pf$, nämlich

$$P(f - Pf - A^*h)(f - Pf - A^*h)^\top = Pff^\top - PfPf^\top - Pfh^\top(Phh^\top)^{-1}Phf^\top.$$

Man kann zeigen, daß $\mathbb{P}_n f - \hat{A}\mathbb{P}_n h$ asymptotisch nicht zu übertreffen ist, es sei denn, man weiß mehr über $\mathcal{P}$.

Problem. Es gelte die Einschränkung $Ph_\vartheta = 0$ mit unbekanntem d-dimensionalen Parameter ϑ. Wir können dann die obige Verbesserung des empirischen Schätzers für Pf vornehmen, müssen aber den unbekannten Parameter ϑ durch einen Schätzer ersetzen, zum Beispiel einen M-Schätzer. Was ist die asymptotische Verteilung des resultierenden Schätzers für Pf? Man kann wiederum zeigen, daß er asymptotisch nicht zu übertreffen ist, es sei denn, man weiß mehr über $\mathcal{P}$.

Beispiel. (*Lineare Regression.*) Sei X eine k-dimensionale Zufallsvariable und Y eine reelle Zufallsvariable. Sei P die Verteilung von (X, Y). Es gelte $P|X|^4 < \infty$ und $PY^4 < \infty$, und $PXX^\top$ sei positiv definit. Das *Kleinste-Quadrate-Funktional* $\vartheta(P)$ minimiert $P(Y - \vartheta^\top X)^2$ in ϑ. Durch Differenzieren erhalten wir $PX(Y - \vartheta^\top(P)X) = 0$, also

$$\vartheta(P) = (PXX^\top)^{-1}PXY.$$

Der *Kleinste-Quadrate-Schätzer* für $\vartheta(P)$ ist

$$\vartheta(\mathbb{P}_n) = \Big(\sum_{i=1}^n X_i X_i^\top\Big)^{-1}\sum_{i=1}^n X_i Y_i.$$

Wir erhalten ihn natürlich auch als Lösung der empirischen Version von $PX(Y - \vartheta^\top(P)X) = 0$, also als Lösung der Schätzgleichung $\mathbb{P}_n X(Y - \vartheta^\top X) = 0$. Da sich das nach ϑ auflösen läßt, brauchen wir die Theorie der M-Schätzer nicht zu bemühen. Der Kleinste-Quadrate-Schätzer ist eine Funktion zweier empirischer Schätzer. Wir könnten deshalb die asymptotische Verteilung von $\vartheta(\mathbb{P}_n)$ aus der obigen Bemerkung herleiten. Hier geht es aber einfacher. Schreibe

$$n^{1/2}(\vartheta(\mathbb{P}_n) - \vartheta(P)) = \Big(\frac{1}{n}\sum_{i=1}^n X_i X_i^\top\Big)^{-1} n^{-1/2}\sum_{i=1}^n X_i(Y_i - \vartheta(P)^\top X_i).$$

Die zweite Summe ist zentriert. Mit dem Gesetz der großen Zahl erhalten wir

$$n^{1/2}(\vartheta(\mathbb{P}_n) - \vartheta(P)) = (PXX^\top)^{-1} n^{-1/2}\sum_{i=1}^n X_i(Y_i - \vartheta(P)^\top X_i) + o_p(1)$$

Also ist $\vartheta(\mathbb{P})$ asymptotisch normal mit Kovarianzmatrix

$$(PXX^\top)^{-1} P[XX^\top(Y - \vartheta(P)^\top X)^2)](PXX^\top)^{-1}.$$

Bis jetzt haben wir keine Modellannahmen gemacht. Nehmen wir nun an, daß ein linearer Zusammenhang $Y = \vartheta^\top X + \varepsilon$ existiert in dem (sehr schwachen) Sinn, daß $E(\varepsilon|X) = 0$ gilt; anders gesagt, daß $E(Y|X) = \vartheta^\top X$ gilt. Dann gilt $PXY = PXX^\top\vartheta + PX\varepsilon$ und $PX\varepsilon = PE(X\varepsilon|X) = PXE(\varepsilon|X) = 0$, also $\vartheta = (PXX^\top)^{-1}PXY = \vartheta(P)$. Läßt sich dann der Kleinste-Quadrate-Schätzer verbessern? Die Modellannahme bedeutet, daß für alle (quadratintegrierbaren) k-dimensionalen Zufallsvektoren $W(X)$ gilt:

$$PW(X)(Y - \vartheta^\top X) = 0.$$

Wir erhalten also M-Schätzer als Lösungen $\hat\vartheta_W$ von $\mathbb{P}_n W(X)(Y - \vartheta^\top X) = 0$, also

$$\hat\vartheta_W = (\mathbb{P}_n W(X)X^\top)^{-1}\mathbb{P}_n W(X)Y = \Big(\sum_{i=1}^n W(X_i)X_i^\top\Big)^{-1}\sum_{i=1}^n W(X_i)Y_i.$$

Solche Schätzer heißen *gewichtete* Kleinste-Quadrate-Schätzer. Wie oben erhält man

$$\begin{aligned} n^{1/2}(\hat\vartheta_W - \vartheta) &= \Big(\frac{1}{n}\sum_{i=1}^n W(X_i)X_i^\top\Big)^{-1} n^{-1/2}\sum_{i=1}^n W(X_i)(Y_i - \vartheta^\top X_i) \\ &= n^{-1/2}\sum_{i=1}^n g_W(X_i, Y_i) + o_p(1) \end{aligned}$$

mit

$$g_W(X,Y) = (PW(X)X^\top)^{-1}W(X)(Y - \vartheta^\top X).$$

Also ist $\hat\vartheta_W$ asymptotisch normal mit Kovarianzmatrix

$$\begin{aligned}\Sigma(W) &= Pg_W g_W^\top \\ &= P(W(X)X^\top)^{-1}P(W(X)W(X)^\top\rho^2(X))P(XW(X)^\top)^{-1},\end{aligned}$$

wobei $\rho^2(X)$ die bedingte Varianz von ε gegeben X ist. Ähnlich wie im vorigen Beispiel rechnen wir nach, daß mit $W^*(X) = \rho^{-2}(X)X$ gilt:

$$\begin{aligned}&P(g_W - g_{W^*})g_{W^*}^\top \\ &= P\Big((PW(X)X^\top)^{-1}W(X) - (P\varrho^{-2}(X)XX^\top)^{-1}\varrho^{-2}(X)X\Big) \\ &\qquad \varepsilon^2\varrho^{-2}(X)X^\top(P\varrho^{-2}(X)XX^\top)^{-1} \\ &= (I_k - I_k)(P\varrho^{-2}(X)XX^\top)^{-1} = 0;\end{aligned}$$

hier bezeichnet I_k die k-dimensionale Einheitsmatrix. Also gilt wegen

$$\begin{aligned}g_W g_W^\top - g_{W^*}g_{W^*}^\top &= (g_W - g_{W^*})(g_W - g_{W^*})^\top \\ &\quad + (g_W - g_{W^*})g_{W^*}^\top + g_{W^*}(g_W - g_{W^*})^\top,\end{aligned}$$

daß

$$\Sigma(W) - \Sigma(W^*) = P(g_W - g_{W^*})(g_W - g_{W^*})^\top$$

positiv semidefinit ist. Also wird die Kovarianzmatrix $\Sigma(W)$ minimiert durch $W = W^*$. Das hängt vom unbekannten P ab. Wir müssen ρ durch einen geeigneten Schätzer $\hat\rho$ ersetzen. Solche Schätzer werden wir erst später kennenlernen. Wir erhalten

$$\hat\vartheta_* = \Big(\sum_{i=1}^n \hat\rho^{-2}(X_i)X_iX_i^\top\Big)^{-1}\sum_{i=1}^n \hat\rho^{-2}(X_i)X_iY_i.$$

Unter geeigneten Voraussetzungen hat dieser Schätzer die asymptotische Kovarianzmatrix $\Sigma(W^*) = (P\rho^{-2}(X)XX^\top)^{-1}$. Man kann zeigen, daß dieser Schätzer asymptotisch nicht verbesserbar ist.

Problem. Es gelte die Einschränkung $E(h_\vartheta(X,Y)|X) = 0$ mit unbekanntem d-dimensionalen Parameter ϑ. (Oben hatten wir $h_\vartheta(X,Y) = Y - \vartheta^\top X$.) Gesucht werden asymptotisch optimale Schätzer für ϑ (das ist i.w. bekannt) und für Pf. Zusatz: Was schätzen diese Schätzer, wenn die Einschränkung nicht gilt, und schätzen sie es optimal?

Bemerkung. Häufig wird im Regressionsmodell zusätzlich angenommen, daß X und ε unabhängig sind. Dann reduziert sich die Bedingung $E(\varepsilon|X) = 0$ auf $E\varepsilon = 0$. Die bedingte Varianz $\varrho^2(X) = E(\varepsilon^2|X)$ ist dann gleich der Varianz $E\varepsilon^2$. Die optimale Gewichtsfunktion ist also proportional zu X. Die Konstante hat keinen Einfluß auf den Schätzer. Also ist der (ungewichtete) Kleinste-Quadrate-Schätzer asymptotisch äquivalent zum optimalen gewichteten Kleinste-Quadrate-Schätzer. Allerdings lassen sich jetzt beide verbessern, indem man die Unabhängigkeit von ε und X ausnützt. Das behandeln wir in dieser Vorlesung nicht.

13 Ordnungsstatistiken und Stichprobenquantile

Seien $X_1, \ldots, X_n$ unabhängig mit stetiger Verteilungsfunktion F. Dann heißt $R_j = \sum_{i=1}^n \mathbf{1}(X_i \le X_j)$ der *Rang* von X_j. Die der Größe (dem Rang) nach geordneten Beobachtungen $X_{1:n} \le \cdots \le X_{n:n}$ heißen *Ordnungsstatistiken*. Das *p-Quantil* ist $\xi_p = F^{-1}(p) = \inf\{x : F(x) \ge p\}$. Die *empirische Verteilungsfunktion* ist

$$\mathbb{F}_n(t) = \mathbb{P}_n(-\infty, t] = \frac{1}{n}\sum_{i=1}^n \mathbf{1}(X_i \le t) = \frac{1}{n}|\{i : X_i \le t\}|.$$

Das *Stichproben-p-Quantil* ist $\hat{\xi}_p = \mathbb{F}_n^{-1}(p)$.

Satz 20 *Sei ξ_p der einzige Wert, der durch F auf p abgebildet wird. Sei $k = np + o(n)$. Dann gilt $X_{k:n} \to \xi_p$ in Wahrscheinlichkeit.*

Beweis. Sei $p_\varepsilon = F(\xi_p + \varepsilon)$. Dann gilt $p_\varepsilon > p$. Schreibe

$$\begin{aligned} P(X_{k:n} \le \xi_p + \varepsilon) &= P\Big(\sum_{i=1}^n \mathbf{1}(X_i \le \xi_p + \varepsilon) \ge k\Big) \\ &= P\Big(\frac{1}{n}\sum_{i=1}^n \mathbf{1}(X_i \le \xi_p + \varepsilon) - p_\varepsilon \ge \frac{k}{n} - p_\varepsilon\Big). \end{aligned}$$

Das konvergiert gegen 1 nach dem Gesetz der großen Zahl. Analog für $\xi_p - \varepsilon$ statt $\xi_p + \varepsilon$.

Satz 21 *Es habe F eine Dichte f, die in ξ_p stetig und positiv ist. Sei $k = np + o(n^{1/2})$. Dann gilt*

$$n^{1/2}(X_{k:n} - \xi_p) \Rightarrow N(0, p(1-p)/f^2(\xi_p)).$$

Beweis.

$$\begin{aligned} P(n^{1/2}(X_{k:n} - \xi_p) \le t) &= P(X_{k:n} \le \xi_p + n^{-1/2}t) \\ &= P\Big(\sum_{i=1}^{n} \mathbf{1}(X_i \le \xi_p + n^{-1/2}t) \ge k\Big) \\ &= P\Big(n^{-1/2}\sum_{i=1}^{n} Y_{ni} \ge u_n\Big) \end{aligned}$$

mit

$$\begin{aligned} Y_{ni} &= \mathbf{1}(X_i \le \xi_p + n^{-1/2}t) - F(\xi_p + n^{-1/2}t), \\ u_n &= n^{-1/2}(k - nF(\xi_p + n^{-1/2}t)). \end{aligned}$$

Die Zufallsvariable Y_{ni} nimmt den Wert $1 - F(\xi_p + n^{-1/2}t)$ mit Wahrscheinlichkeit $F(\xi_p + n^{-1/2}t)$ an, und den Wert $-F(\xi_p + n^{-1/2}t)$ mit Wahrscheinlichkeit $1-F(\xi_p+n^{-1/2}t)$. Es gilt $EY_{ni} = 0$ und $F(\xi_p+n^{-1/2}t) \to F(\xi_p) = p$, also $EY_{ni}^2 \to p(1-p)$ und mit einer Taylorentwicklung

$$u_n = n^{1/2}(p - F(\xi_p + n^{-1/2}t)) + o(1) \to -tf(\xi_p).$$

Nach einer Version des Zentralen Grenzwertsatzes für Dreiecksschemata gilt also

$$P\Big(n^{-1/2}\sum_{i=1}^{n} Y_{ni} \ge u_n\Big) \to 1 - \Phi(-tf(\xi_p)/(p(1-p))^{1/2}).$$

Hier ist Φ die Verteilungsfunktion der Standard-Normalverteilung $N(0,1)$.

14 Punktweise Konvergenz von Kernschätzern

Beispiel. Seien $X_1, \ldots, X_n$ unabhängig mit Dichte f. Wir wollen $f(x)$ schätzen. Sei $b > 0$. Ein erwartungstreuer Schätzer für $P[x - b/2, x + b/2]$ ist

$$\begin{aligned} \mathbb{P}_n[x - b/2, x + b/2] &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[x-b/2, x+b/2]}(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(-b/2 \le x - X_i \le b/2). \end{aligned}$$

Is f stetig in x, gilt andererseits für kleines b:

$$P[x - b/2, x + b/2] = \int_{x-b/2}^{x+b/2} f(t)\,dt \approx bf(x).$$

Wählen wir $b = b_n \to 0$ and $nb \to \infty$, so erwarten wir also, daß folgender Schätzer konsistent für $f(x)$ ist:

$$\hat{f}(x) = \frac{1}{nb} \sum_{i=1}^{n} \mathbf{1}_{[x-b/2, x+b/2]}(X_i) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{b} \mathbf{1}_{[-1/2,1/2]}\Big(\frac{x - X_i}{b}\Big).$$

Mit $K(t) = \mathbf{1}_{[-1/2,1/2]}(t)$ und $K_b(t) = K(t/b)/b$ läßt sich $\hat{f}$ schreiben als

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_b(x - X_i).$$

Das verallgemeinern wir im folgenden auf andere Dichten K.

Seien $X_1, \ldots, X_n$ unabhängig mit Dichte f. Ein *Kernschätzer* für $f(x)$ ist

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_b(x - X_i)$$

mit $K_b(t) = K(t/b)/b$, *Kern* K und Bandweite b. Konvergenzraten für $\hat{f}(x)$ ergeben sich aus der Chebyshev-Ungleichung

$$P(|\hat{f}(x) - f(x)| > a) \le a^{-2} E(\hat{f}(x) - f(x))^2.$$

Der *mittlere quadratische Fehler* (MSE) von $\hat{f}(x)$ läßt sich wie folgt zerlegen,

$$E(\hat{f}(x) - f(x))^2 = \operatorname{Var} \hat{f}(x) + (E\hat{f}(x) - f(x))^2,$$

die Summe aus Varianz und Quadrat des *Bias*. Für $r = 0, 1, 2, \ldots$ und $0 < \alpha \le 1$ sei $\operatorname{Lip}_{r,\alpha}(L)$ die Klasse der Funktionen, die beschränkt und r-mal differenzierbar sind und deren r-te Ableitungen in x *Lipschitz der Ordnung* α mit Konstante L sind:

$$|f^{(r)}(x) - f^{(r)}(y)| \le L|x - y|^{\alpha}.$$

Sei $\mathcal{K}_{r,\alpha}$ die Klasse der Funktionen, die beschränkt sind mit

$$\begin{aligned} \int K(t)\, dt &= 1, \\ \int t^j K(t)\, dt &= 0, \quad j = 1, \ldots, 1 \vee r, \\ \int |t^{r+\alpha} K(t)|\, dt &< \infty. \end{aligned}$$

Satz 22 *Ist $f \in \mathrm{Lip}_{r,\alpha}(L)$ und $K \in \mathcal{K}_{r,\alpha}$, so gilt für $b \downarrow 0$:*

$$\begin{aligned} E(\hat{f}(x) - f(x))^2 &\le \frac{1}{nb} f(x) \int K^2(t)\,dt \\ &\quad + b^{2(r+\alpha)} \Big(\frac{L}{r!} \int |t^{r+\alpha} K(t)|\,dt \Big)^2 + o\Big(\frac{1}{nb}\Big) \\ &= O\Big(\frac{1}{nb}\Big) + O\big(b^{2(r+\alpha)}\big). \end{aligned}$$

Also wird die optimale Rate $n^{-2(r+\alpha)/(2(r+\alpha)+1)}$ erzielt, wenn b proportional zu $n^{-1/(2(r+\alpha)+1)}$ ist.

Beweis. Schreibe Varianz und Bias als

$$\begin{aligned} \mathrm{Var}\,\hat{f}(x) &= \frac{1}{n} \mathrm{Var}\,K_b(x - X) \\ &= \frac{1}{n} \big(E K_b^2(x - X) - (E K_b(x - X))^2 \big) \\ &= \frac{1}{n} \Big(\int K_b^2(x - u) f(u)\,du - \Big(\int K_b(x - u) f(u)\,du \Big)^2 \Big); \\ E\hat{f}(x) - f(x) &= \int K_b(x - u)(f(u) - f(x))\,du. \end{aligned}$$

Für $m = 1, 2$ gilt

$$\begin{aligned} \int K_b^m(x - u) f(u)\,du &= b^{-m} \int K^m\Big(\frac{x - u}{b}\Big) f(u)\,du \\ &= b^{-m+1} \int K^m(t) f(x - bt)\,dt. \end{aligned}$$

Also

$$\mathrm{Var}\,\hat{f}(x) = \frac{1}{nb} \int K^2(t) f(x - bt)\,dt - \frac{1}{n} \Big(\int K(t) f(x - bt)\,dt \Big)^2.$$

Für $m = 1, 2$ gilt

$$\int |K^m(t)||f(x - bt) - f(x)|\,dt \le L b^\alpha \int |K^m(t)||t|^\alpha\,dt,$$

also mit Beschränktheit von K für die Varianz:

$$\mathrm{Var}\,\hat{f}(x) = \frac{1}{nb} f(x) \int K^2(t)\,dt + \frac{1}{nb} O(b^\alpha).$$

Für den Bias gilt für $r = 0$:

$$|E\hat{f}(x) - f(x)| = \Big| \int K(t)(f(x - bt) - f(x))\, dt \Big| \le b^\alpha L \int |K(t)t^\alpha|\, dt.$$

Für $r \ge 1$ und ein z zwischen x und $x - bt$ gilt

$$\begin{aligned} |E\hat{f}(x) - f(x)| &= \Big| \int K(t) \Big(\sum_{j=1}^{r} \frac{(-bt)^j}{j!} f^{(j)}(x) \\ &\qquad + \frac{(-bt)^r}{r!} (f^{(r)}(z) - f^{(r)}(x)) \Big)\, dt \Big| \\ &\le b^{r+\alpha} \frac{L}{r!} \int |K(t)t^{r+\alpha}|\, dt. \end{aligned}$$

Aus Satz 22 und der Chebyshev-Ungleichung ergibt sich insbesondere, daß $\hat{f}(x) - f(x) = O_p((nb)^{-1/2} + b^{r+\alpha})$.

15 Konvergenz von Kernschätzern in L_1

Der *integrierte mittlere quadratische Fehler* (MISE) von $\hat{f}$ ist

$$\int E(\hat{f}(x) - f(x))^2\, dx = E \int (\hat{f}(x) - f(x))^2\, dx.$$

Eine Schranke dafür läßt sich nicht einfach aus aus Satz 22 gewinnen, es sei denn, wir wüßten, daß die Dichte f auf einer beschränkten Menge lebt.

Sei λ das Lebesgue-Maß auf $\mathbb{R}$. Wir setzen $L_1 = L_1(\lambda)$. Die L_1-Norm einer Funktion f ist $\|f\|_1 = \int |f(x)|\, dx$. Wir benötigen einige Begriffe und Ergebnisse aus der reellen Analysis. Die *Translation* von f um y ist definiert durch $f_y(x) = f(x - y)$. Für $f \in L_1$ gilt $\|f_y\|_1 = \|f\|_1$.

Lemma 6 *(L_1-Stetigkeit der Translation) Für $f \in L_1$ ist $y \to f_y$ gleichmäßig L_1-stetig.*

Beweis. Sei $\varepsilon > 0$. Wähle g stetig mit Träger in $[-A, A]$ und $\|f - g\|_1 < \varepsilon$. Dann ist g gleichmäßig stetig. Also existiert $\delta > 0$ mit $|g(y) - g(z)| < \varepsilon/(3A)$ für $|y - z| < \delta$. Für $|y - z| < \delta$ gilt also $\|g_y - g_z\|_1 < (2A + \delta)\varepsilon/(3A) < \varepsilon$, also

$$\begin{aligned} \|f_y - f_z\|_1 &\le \|f_y - g_y\|_1 + \|g_y - g_z\|_1 + \|g_z - f_z\|_1 \\ &= \|f - g\|_1 + \|g_y - g_z\|_1 + \|g - f\|_1 < 3\varepsilon. \end{aligned}$$

Lemma 7 *Für $g, h \in L_1$ gilt*

$$\iint |g(x-bu) - g(x)||h(u)|\,du\,dx \to 0, \quad b \to 0.$$

Beweis. Es gilt $\|g_y - g\|_1 \le 2\|g\|_1$. Nach Lemma 6 ist $y \to g_y$ L_1-stetig. Nach dem Satz von der dominierten Konvergenz also

$$\iint |g(x-bu) - g(x)|\,dx|h(u)|\,du = \int \|g_{bu} - g\|_1 |h(u)|\,du \to 0, \quad b \to 0.$$

Eine Funktion $f : \mathbb{R} \to \mathbb{R}$ heißt *absolutstetig* auf dem endlichen Intervall $[a, b]$, wenn für alle $\varepsilon > 0$ ein $\delta > 0$ existiert, so daß für alle endlichen Familien von disjunkten Intervallen $(a_j, b_j] \subset (a, b]$ mit $\sum_j (b_j - a_j) < \delta$ gilt: $\sum_j |f(b_j) - f(a_j)| < \varepsilon$. Die Funktion heißt *absolutstetig*, wenn sie auf allen endlichen Intervallen absolutstetig ist.

Für eine Funktion $f : \mathbb{R} \to \mathbb{R}$ ist die *Totalvariation* auf einem Intervall $[a, b]$ definiert als

$$\sup_{(t_j)} \sum_j |f(t_j) - f(t_{j-1})|,$$

wobei sich das Supremum über alle endlichen Partitionen $a = t_0 < \cdots < t_k = b$ erstreckt. Eine auf $[a, b]$ absolutstetige Funktion hat endliche Totalvariation auf $[a, b]$.

Wir benötigen den Fundamentalsatz für Lebesgue-Integrale: Ist f absolutstetig auf $[a, b]$, so existiert f.s. die Ableitung f' und ist in L_1, und

$$f(x) - f(a) = \int_a^x f'(t)\,dt, \quad x \in [a, b].$$

Für $r = 0, 1, \ldots$ sei $\mathcal{L}_{r,0}$ die Klasse der Funktionen f, die $(r-1)$-mal differenzierbar sind mit $f^{(r-1)}$ absolutstetig und $f^{(r)} \in L_1$. Für $0 < \alpha \le 1$ sei $\mathcal{L}_{r,\alpha}$ die Klasse der Funktionen $f \in \mathcal{L}_{r,0}$, für die $y \to f_y^{(r)}$ L_1-*Lipschitz der Ordnung* α ist: Für ein $L > 0$ gilt

$$\|f_y^{(r)} - f^{(r)}\|_1 \le L|y|^\alpha, \quad y \in \mathbb{R}.$$

Die *Faltung* $f * g$ zweier Funktionen $f, g \in L_1$ ist definiert durch

$$f * g(x) = \int f(x-u)g(u)\,du.$$

Es gilt

$$\|f * g\|_1 \le \iint |f(x-u)g(u)|\,du\,dx = \|f\|_1 \|g\|_1.$$

Den Erwartungswert eines Kernschätzers können wir wie folgt schreiben:

$$\begin{aligned} E\hat{f}(x) &= EK_b(x-X) \\ &= \int K_b(x-y)f(y)\,dy = f * K_b(x) = \int f(x-bu)K(u)\,du. \end{aligned}$$

Es gilt $\|K_b\|_1 = \|K\|_1$, also

$$\|E\hat{f}\|_1 = \|f * K_b\|_1 \le \|f\|_1\|K_b\|_1 = \|K\|_1.$$

Satz 23 *Ist $f \in \mathcal{L}_{r,0}$ und $K \in \mathcal{K}_{r,0}$, so gilt*

$$\|f * K_b - f\|_1 = o(b^r).$$

Ist $0 < \alpha \le 1$ und $f \in \mathcal{L}_{r,\alpha}$, $K \in \mathcal{K}_{r,\alpha}$, so gilt

$$\|f * K_b - f\|_1 = O(b^{r+\alpha}).$$

Beweis. Durch Taylor-Entwicklung:

$$\begin{aligned} f(x-bu) - f(x) = &\sum_{j=1}^{r} \frac{(-bu)^j}{j!} f^{(j)}(x) \\ &+ \frac{(-bu)^r}{(r-1)!} \int_0^1 (1-t)^{r-1}(f^{(r)}(x-tbu) - f^{(r)}(x))\,dt. \end{aligned}$$

Mit $\int u^j K(u)\,du = 0$, $j = 1, \ldots, r$, gilt

$$\begin{aligned} &f * K_b(x) - f(x) \\ &= \frac{(-b)^r}{(r-1)!} \int_0^1 \int (1-t)^{r-1}(f^{(r)}(x-tbu) - f^{(r)}(x))u^r K(u)\,du\,dt. \end{aligned}$$

Für $\alpha = 0$ folgt aus Lemma 7, daß

$$\int\int |f^{(r)}(x-tbu) - f^{(r)}(x)||u^r K(u)|\,du\,dx \to 0, \quad 0 \le t \le 1.$$

Andererseits gilt

$$\begin{aligned} &\Big|\int\int (f^{(r)}(x-tbu) - f^{(r)}(x))u^r K(u)\,du\,dx\Big| \\ &\le \int\int (|f^{(r)}(x-tbu)| + |f^{(r)}(x)|)\,dx\,|u^r K(u)|\,du \\ &\le 2\int |f^{(r)}(x)|\,dx \int |u^r K(u)|\,du. \end{aligned}$$

Die Behauptung des Satzes für $\alpha = 0$ folgt also aus dem Satz von der dominierten Konvergenz. Für $\alpha > 0$ verwenden wir

$$\int |f^{(r)}(x - tbu) - f^{(r)}(x)|\, dx = \|f^{(r)}_{tbu} - f^{(r)}\|_1 \le Lb^\alpha |u|^\alpha, \quad u \in \mathbb{R}, |t| \le 1.$$

Satz 24 *Hat f ein endliches zweites Moment und ist $K \in \mathcal{K}_{1,1}$, so gilt*

$$E\|\hat{f} - f * K_b\|_1 = O((nb)^{-1/2}).$$

Beweis. Sei $V(x) = (1 + |x|)^2$. Mit der Schwarzschen Ungleichung gilt für jedes meßbare g:

$$\|g\|_1^2 = \Big(\int \frac{V^{1/2}|g|}{V^{1/2}} \Big)^2 \le \|V^{-1}\|_1 \|Vg^2\|_1,$$

und $\|V^{-1}\|_1$ ist endlich. Außerdem

$$EK_b^2(x - X) = \int K_b^2(x - y) f(y)\, dy = K_b^2 * f(x).$$

Es gilt also

$$\begin{aligned} E\|\hat{f} - f * K_b\|_1^2 &\le \|V^{-1}\|_1 E\|V(\hat{f} - f * K_b)^2\|_1 \\ &= \|V^{-1}\|_1 \|VE(\hat{f} - f * K_b)^2\|_1 \le \frac{1}{n}\|V^{-1}\|_1 \|V \cdot f * K_b^2\|_1. \end{aligned}$$

Es gilt $V(x + y) \le V(x)V(y)$ und

$$f * K_b^2(x) = \frac{1}{b^2} \int K^2\Big(\frac{x - y}{b}\Big) f(y)\, dy = \frac{1}{b} \int f(x - bu) K^2(u)\, du,$$

also

$$\begin{aligned} \|V \cdot f * K_b^2\|_1 &= \frac{1}{b} \iint V(x + bu) f(x) K^2(u)\, dx\, du \\ &\le \frac{1}{b} \int V(x) f(x)\, dx \int V(bu) K^2(u)\, du. \end{aligned}$$

Weil K beschränkt ist, ist das letzte Integral endlich. Mit der Schwarzschen Ungleichung ergibt sich also

$$\big(E\|\hat{f} - f * K_b\|_1\big)^2 \le E\|\hat{f} - f * K_b\|_1^2 = O((nb)^{-1}).$$

Der *erwartete L_1-Fehler* von $\hat{f}$ ist $E \int |\hat{f}(x) - f(x)|\, dx$. Es gilt mit der Dreiecksungleichung:

$$E \int |\hat{f}(x) - f(x)|\, dx = E\|\hat{f} - f\|_1 \le E\|\hat{f} - f * K_b\|_1 + \|f * K_b - f\|_1.$$

Ist $f \in \mathcal{L}_{r,\alpha}$ mit endlichem zweiten Moment und $K \in \mathcal{K}_{r,\alpha}$ mit $\alpha = 1$, falls $r = 1$, so gilt nach den Sätzen 23 und 24:

$$E \int |\hat{f}(x) - f(x)|\, dx = O(b^{r+\alpha} + (nb)^{-1/2}).$$

Insbesondere ergibt sich $\int |\hat{f}(x) - f(x)|\, dx = O(b^{r+\alpha} + (nb)^{-1/2})$ aus der Chebyshev-Ungleichung

$$P(\int |\hat{f}(x) - f(x)|\, dx > a) \le a^{-1} E \int |\hat{f}(x) - f(x)|\, dx.$$

Die optimale Bandweite und Konvergenzrate sind also dieselben wie für die Konvergenz der Wurzel des punktweisen mittleren quadratischen Fehlers, Satz 22.

Wir erwarten, daß $(nb)^{1/2}(\hat{f}(x) - f(x))$ asymptotisch normal mit Mittelwert 0 ist, solange b schneller als die optimale Bandweite gegen 0 geht, denn dann ist der Bias vernachlässigbar. Für die optimale Bandweite b wird $(nb)^{1/2}(\hat{f}(x) - f(x))$ unter geeigneten Annahmen immer noch asymptotisch normal sein, aber der Mittelwert der Grenzverteilung wird nicht 0 sein.

16 Nichtparametrische Regression und Nadaraya–Watson-Schätzer

Sei (X, Y) eine zweidimensionale Zufallsvariable. Die *Regressionsfunktion* von Y auf X ist der bedingte Erwartungswert

$$g(X) = E(Y|X).$$

Wir beobachten unabhängige Realisationen (X_i, Y_i), $i = 1, \ldots, n$. Der *Nadaraya–Watson-Schätzer* für $g(x)$ ist

$$\hat{g}(x) = \frac{\sum_{i=1}^n K_b(x - X_i) Y_i}{\sum_{i=1}^n K_b(x - X_i)}.$$

Hier ist K ein Kern und b eine Bandweite.

Sei $f(x, y)$ die Dichte von (X, Y), und $f_1(x) = \int f(x, y)\, dy$ die Dichte von X. Dann ist $f(x, y)/f_1(x)$ die bedingte Dichte von Y gegeben $X = x$, also $g(x) = \int y f(x, y)\, dy / f_1(x)$. Ist f_1 stetig, so konvergiert $\frac{1}{n} \sum_{i=1}^n K_b(x - X_i)$ gegen $f_1(x)$; siehe Kapitel 14. Für den Zähler von $\hat{g}(x)$ haben wir unter

geeigneten Annahmen

$$\begin{aligned} EK_b(x-X)Y &= \iint K_b(x-z)yf(z,y)\,dz\,dy \\ &= \iint K(u)f(x-bu,y)y\,du\,dy \\ &\to \int K(u)\,du \int yf(x,y)dy = \int yf(x,y)\,dy. \end{aligned}$$

Konvergenzraten erhält man wie bei Dichteschätzern.

17 Lokale polynomiale Glätter

Wir bleiben beim Schätzen der Regressionsfunktion g. Ist g ein (unbekanntes) Polynom mit (bekanntem) Grad r, so können wir g schätzen, indem wir die Beobachtungen (X_i, Y_i) durch ein Polynom $p(x) = \sum_{k=0}^r \vartheta_k x^k$ vom Grad r approximieren, das den mittleren quadratischen Fehler

$$\sum_{i=1}^n \Big(Y_i - \sum_{k=0}^r \vartheta_k X_i^k\Big)^2$$

minimiert. Durch Differentiation nach ϑ_j erhalten wir

$$\sum_{i=1}^n X_i^j \Big(Y_i - \sum_{k=0}^r \hat\vartheta_k X_i^k\Big) = 0, \quad j = 0,\dots,r,$$

also mit $Z_i = (X_i^0, \dots, X_i^r)^\top$:

$$\sum_{i=1}^n Z_i(Y_i - Z_i^\top \hat\vartheta) = 0$$

und deshalb

$$\hat\vartheta = \Big(\sum_{i=1}^n Z_i Z_i^\top\Big)^{-1} \sum_{i=1}^n Z_i Y_i$$

Der Schätzer $\hat g(x) = \sum_{k=0}^r \hat\vartheta_k x^k$ ist ein Kleinste-Quadrate-Schätzer; wir nennen ihn *polynomialen Glätter*.

Für das *Lageparameter-Modell* $Y = \vartheta + \varepsilon$ mit $E\varepsilon = 0$ ist $g(x) = \vartheta$ und $\hat\vartheta = \frac{1}{n}\sum_{i=1}^n Y_i$.

Für das *lineare Regressionsmodell* mit *Interzept* ϑ_0,

$$Y = \vartheta_0 + \vartheta_1 X + \varepsilon \quad \text{mit } E(\varepsilon|X) = 0,$$

ist $g(X) = \vartheta_0 + \vartheta_1 X$ und

$$ZZ^\top = \begin{pmatrix} 1 & X \\ X & X^2 \end{pmatrix}.$$

Für das *lineare Regressionsmodell ohne Interzept*,

$$Y = \vartheta X + \varepsilon \quad \text{mit } E(\varepsilon|X) = 0,$$

ist $g(X) = \vartheta X$. Wir können also den Index 0 weglassen und erhalten

$$\hat{\vartheta} = \Big(\sum_{i=1}^n X_i^2\Big)^{-1} \sum_{i=1}^n X_i Y_i.$$

Das hatten wir in Kapitel 12 schon allgemeiner hergeleitet, nämlich für $Y = \vartheta^\top X + \varepsilon$ mit Kovariablen*vektor* X.

Ist die Funktion g kein Polynom der Ordnung r, aber r-mal differenzierbar, so können wir g *lokal* durch ein Polynom approximieren und obige Methode lokal anwenden. Sei x fest. Eine Taylor-Entwicklung liefert für z nahe x:

$$g(z) \approx \sum_{k=0}^r \frac{g^{(k)}(x)}{k!}(z-x)^k.$$

Den Koeffizienten $\vartheta_0, \ldots, \vartheta_r$ entsprechen $g(x), g'(x), \ldots, g^{(r)}(x)/r!$. Wir minimieren den mittleren quadratischen Fehler

$$\sum_{i=1}^n \Big(Y_i - \sum_{k=0}^r \vartheta_k (X_i - x)^k\Big)^2 K_b(X_i - x).$$

Durch Differenzieren nach ϑ_j erhalten wir deshalb die Gleichung

$$\sum_{i=1}^n (X_i - x)^j \Big(Y_i - \sum_{k=0}^r \vartheta_k (X_i - x)^k\Big) K_b(X_i - x) = 0.$$

Sei $\hat{Q} = (\hat{Q}_{jk})_{j,k=0,\ldots,r}$ die Matrix mit Elementen

$$\hat{Q}_{jk} = \sum_{i=1}^n (X_i - x)^{j+k} K_b(X_i - x),$$

und sei $\hat{W} = (\hat{W}_0, \ldots, \hat{W}_r)^\top$ der Vektor mit Elementen

$$\hat{W}_j = \sum_{i=1}^n (X_i - x)^j K_b(X_i - x) Y_i.$$

Dann lassen sich die $r+1$ Gleichungen schreiben als $\hat{W} - \hat{Q}\vartheta = 0$. Wir erhalten also einen Schätzer für $g(x), g'(x), \ldots, g^{(r)}(x)$, den *lokalen polynomialen Glätter*, durch

$$\hat{\vartheta} = \Big(\hat{g}(x), \hat{g}'(x), \ldots, \frac{\hat{g}^{(r)}(x)}{r!}\Big)^\top = \hat{Q}^{-1}\hat{W}.$$

Für $r = 0$ ergibt sich wieder der Nadaraya–Watson-Schätzer

$$\hat{g}(x) = \frac{\hat{W}_0}{\hat{Q}_0} = \frac{\sum_{i=1}^n K_b(X_i - x)Y_i}{\sum_{i=1}^n K_b(X_i - x)}.$$

Dieser lokal konstante Glätter für eine Regressionsfunktion entspricht dem Kernschätzer $\hat{f}(x) = \frac{1}{n}\sum_{i=1}^n K_b(X_i - x)$ für eine Dichte.

18 Extreme Ordnungsstatistiken

Die Sätze 26 und 27 dieses Kapitels über extreme Ordnungsstatistiken werden mit Hilfe des folgenden Satzes 25 über das Verhalten extremer Ordnungsstatistiken bei einer Gleichverteilung bewiesen.

Satz 25 *Seien $X_1, \ldots, X_n$ unabhängig mit stetiger Verteilungsfunktion F. Dann gilt*

$$n(1 - F(X_{n:n})) \Rightarrow E_1.$$

Beweis. Es sind $F(X_1), \ldots, F(X_n)$ unabhängig und gleichverteilt auf $(0, 1)$. Für $t \geq 0$ gilt

$$\begin{aligned} P(n(1 - F(X_{n:n})) \leq t) &= P(F(X_{n:n}) \geq 1 - t/n) \\ &= 1 - P(F(X_{n:n}) < 1 - t/n) \\ &= 1 - (1 - t/n)^n \to 1 - e^{-t}. \end{aligned}$$

Definition. Der *linke* und *rechte Eckpunkt* einer Verteilungsfunktion F sind

$$x_\ell = \sup\{x : F(x) = 0\}, \quad x_r = \inf\{x : F(x) = 1\}.$$

Definition. Eine Verteilungsfunktion F mit endlichem rechten Eckpunkt x_r hat *terminalen Kontakt* der Ordnung m in x_r, wenn F links von x_r $(m+1)$-mal stetig differenzierbar ist mit $F(x_r) = 1$ und mit $F^{(j)}(x_r) = 0$, $j = 1, \ldots, m$, und $F^{(m+1)}(x_r) \neq 0$ für die linksseitigen Ableitungen.

Satz 26 *Seien $X_1, \dots, X_n$ unabhängig mit einer Verteilungsfunktion F, die terminalen Kontakt der Ordnung m in x_r hat. Dann existiert eine Folge b_n mit*

$$P\Big(\frac{X_{n:n} - x_r}{b_n} \le t\Big) \to \begin{cases} \exp(-(-t)^{m+1}), & t \le 0 \\ 1, & t > 0 \end{cases}$$

(Extremwertverteilung vom Typ 1).

Beweis. Für $s > 0$ gilt wegen $F^{(1)}(x_r) = \cdots = F^{(m)}(x_r) = 0$ die Taylorentwicklung

$$F(x_r - s) = F(x_r) + \frac{(-1)^{m+1}}{m!} s^{m+1} \int_0^1 (1-u)^m F^{(m+1)}(x_r - us)\, du.$$

Mit $s = x_r - X_{n:n}$ ergibt sich (mit gegen 1 strebender Wahrscheinlichkeit):

$$\begin{aligned} & n(1 - F(X_{n:n})) = n(F(x_r) - F(X_{n:n})) \\ & = \frac{(-1)^m}{m!} n(x_r - X_{n:n})^{m+1} \int_0^1 (1-u)^m F^{(m+1)}(x_r - u(x_r - X_{n:n}))\, du \\ & = \Big(\frac{x_r - X_{n:n}}{b_n}\Big)^{m+1} a_n \end{aligned}$$

mit

$$\begin{aligned} b_n &= \Big(\frac{(-1)^m (m+1)!}{n F^{(m+1)}(x_r)}\Big)^{1/(m+1)}, \\ a_n &= \frac{(m+1)\int_0^1 (1-u)^m F^{(m+1)}(x_r - u(x_r - X_{n:n}))\, du}{F^{(m+1)}(x_r)} = 1 + o_p(1). \end{aligned}$$

Also gilt mit Satz 25 für $t \le 0$:

$$\begin{aligned} P\Big(\frac{X_{n:n} - x_r}{b_n} \le t\Big) &= P\Big(\Big(\frac{x_r - X_{n:n}}{b_n}\Big)^{m+1} \ge (-t)^{m+1}\Big) \\ &= P(n(1 - F(X_{n:n})) \ge (-t)^{m+1}) + o(1) \\ &\to \exp(-(-t)^{m+1}). \end{aligned}$$

Definition. Eine Verteilungsfunktion F heißt vom *Cauchy-Typ* mit *Exponent* $k > 0$, wenn für ein $c > 0$ gilt:

$$x^k (1 - F(x)) \to c, \quad x \to \infty.$$

Satz 27 *Seien $X_1, \dots, X_n$ unabhängig mit stetiger Verteilungsfunktion F vom Cauchy-Typ mit Exponent k. Gelte $F(x^*_{rn}) = 1 - 1/n$. Dann gilt*

$$P\Big(\frac{X_{n:n}}{x^*_{rn}} \le t\Big) \to \begin{cases} \exp(-t^{-k}), & t > 0 \\ 0, & t \le 0 \end{cases}$$

(Extremwertverteilung vom Typ 2).

Beweis. Mit $1 - F(x^*_{rn}) = 1/n$ schreiben wir

$$n(1 - F(X_{n:n})) = \frac{1 - F(X_{n:n})}{1 - F(x^*_{rn})} = \Big(\frac{x^*_{rn}}{X_{n:n}}\Big)^k \frac{X^k_{n:n}(1 - F(X_{n:n}))}{x^{*k}_{rn}(1 - F(x^*_{rn}))}.$$

Nach Voraussetzung gilt

$$\frac{X^k_{n:n}(1 - F(X_{n:n}))}{x^{*k}_{rn}(1 - F(x^*_{rn}))} = 1 + o_p(1).$$

Also gilt mit Satz 25 für $t \ge 0$:

$$\begin{aligned} P\Big(\frac{X_{n:n}}{x^*_{rn}} \le t\Big) &= P\Big(\Big(\frac{x^*_{rn}}{X_{n:n}}\Big)^k \ge t^{-k}\Big) \\ &= P(n(1 - F(X_{n:n})) \ge t^{-k}) + o(1) \\ &\to \exp(-t^{-k}). \end{aligned}$$

19 Geglättete empirische Verteilungsfunktionen

Setzt man den Dichteschätzer in ein "glattes" Funktional der Dichte ein, so erhält man i.a. eine schnellere Rate; sie kann sogar die "parametrische" Rate $n^{-1/2}$ sein. Ähnliches gilt für Schätzer von Regressionsfunktionen. Die einfachsten Funktionale einer Dichte f sind *lineare* Funktionale

$$Eg(X) = \int g(x) f(x)\, dx$$

mit bekanntem g, zum Beispiel die Verteilungsfunktion

$$F(t) = \int \mathbf{1}_{(-\infty,t]}(x) f(x)\, dx.$$

Der übliche Schätzer für $F(t)$ ist die *empirische Verteilungsfunktion*

$$\mathbb{F}(t) = \mathbb{P}(-\infty, t] = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_i \le t).$$

Sei k ein Kern und b eine Bandweite. Der Dichteschätzer $\hat{f}$ von f ist

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} k_b(x - X_i).$$

Die Verteilungsfunktion zu $\hat{f}$ ist

$$\hat{\mathbb{F}}(t) = \int_{-\infty}^{t} \hat{f}(x)\,dx = \frac{1}{n}\sum_{i=1}^{n} K_b(t - X_i);$$

dabei ist $K(t) = \int_{-\infty}^{t} k(x)\,dx$ die Verteilungsfunktion zu k und $K_b(t) = K(t/b)$. Der Schätzer $\hat{\mathbb{F}}(t)$ unterscheidet sich von $\mathbb{F}(t)$ dadurch, daß der Indikator $\mathbf{1}(X_i \le t)$ durch eine geglättete Version $K_b(t - X_i)$ ersetzt ist. Wir nennen $\hat{\mathbb{F}}$ eine *geglättete* empirische Verteilungsfunktion. Es sei bemerkt, daß wir sie mit partieller Integration auch wie folgt schreiben können:

$$\hat{\mathbb{F}}(t) = \int K_b(t - x)\,d\mathbb{F}(x) = \int k_b(t - x)\mathbb{F}(x)\,dx = \int \mathbb{F}(t - bu)k(u)\,du.$$

Wir zeigen, daß $\hat{\mathbb{F}}(t)$ und $\mathbb{F}(t)$ asymptotisch äquivalent sind. Insbesondere hat $\hat{\mathbb{F}}(t)$ auch die Konvergenzrate $n^{-1/2}$ und ist asymptotisch normal mit Varianz

$$E(\mathbf{1}(X \le t) - F(t))^2 = F(t)(1 - F(t)).$$

Wie in Kapitel 15 bezeichne $\mathcal{L}_{1,1}$ die Klasse der Funktionen f, die absolut stetig sind und für die $y \mapsto f'_y$ L_1-Lipschitz ist. Wie in Kapitel 14 bezeichne $\mathcal{K}_{1,1}$ die Klasse der beschränkten und differenzierbaren Funktionen k mit $\int k(t)\,dt = 1$, $\int tk(t)\,dt = 0$ und $\int t^2|k(t)|\,dt < \infty$.

Satz 28 *Ist $f \in \mathcal{L}_{1,1}$, $k \in \mathcal{K}_{1,1}$ und $b = o(n^{-1/4})$, so gilt*

$$n^{1/2}(\hat{\mathbb{F}}(t) - \mathbb{F}(t)) = o_p(1).$$

Beweis. Zerlege $n^{1/2}(\hat{\mathbb{F}}(t) - F(t))$ in Varianz- und Bias-Term:

$$n^{1/2}(\hat{\mathbb{F}}(t) - F(t)) = n^{1/2}(\hat{\mathbb{F}}(t) - E\hat{\mathbb{F}}(t)) + n^{1/2}(E\hat{\mathbb{F}}(t) - F(t)).$$

Wir schreiben

$$E\hat{\mathbb{F}}(t) = \int_{-\infty}^{t}\int k_b(x - y)f(y)\,dy\,dx = \int_{-\infty}^{t}\int f(x - bu)k(u)\,du\,dx.$$

Mit einer Taylor-Entwicklung und $\int uk(u)\,du = 0$ ergibt sich für den Bias-Term:

$$E\hat{\mathbb{F}}(t) - F(t) = -b\int_{-\infty}^{t}\int\int_0^1 (f'(x-buv) - f'(x))dv\,uk(u)\,du\,dx,$$

also

$$\begin{aligned}|E\hat{\mathbb{F}}(t) - F(t)| &\le b\int\int_0^1 \|f'_{buv} - f'\|_1\,dv\,|uk(u)|\,du\\ &\le b^2L\int u^2|k(u)|\,du = O(b^2) = o(n^{-1/2}).\end{aligned}$$

Mit dem Satz von der dominierten Konvergenz gilt

$$E\big(n^{1/2}(\hat{\mathbb{F}}(t)-E\hat{\mathbb{F}}(t)-\mathbb{F}(t)+F(t)\big)^2 \le \int \big(K_b(t-x)-\mathbf{1}(x\le t)\big)^2 f(x)\,dx \to 0,$$

also für den Varianz-Term:

$$n^{1/2}(\hat{\mathbb{F}}(t) - E\hat{\mathbb{F}}(t)) = n^{1/2}(\mathbb{F}(t) - F(t)) + o_p(1),$$

und die Behauptung des Satzes folgt wegen

$$n^{1/2}(\hat{\mathbb{F}}(t) - \mathbb{F}(t)) = n^{1/2}(\hat{\mathbb{F}}(t) - E\hat{\mathbb{F}}(t) - \mathbb{F}(t) + F(t)) + n^{1/2}(E\hat{\mathbb{F}}(t) - F(t)).$$

Insbesondere sind der *empirische Prozeß* $n^{1/2}(\mathbb{F} - F)$ und der *geglättete empirische Prozeß* $n^{1/2}(\hat{\mathbb{F}} - F)$ in jedem t asymptotisch normal mit Varianz

$$E(\mathbf{1}(X \le t) - F(t))^2 = F(t)(1 - F(t)).$$

Wir zeigen, daß $n^{1/2}(\mathbb{F} - F)$ nicht nur punktweise, sondern auch als Zufallselement in L_1, ausgestattet mit der Borel-Algebra, asymptotisch normal ist. Um das auf $n^{1/2}(\hat{\mathbb{F}}-F)$ zu übertragen, müssen wir dann nur noch zeigen, daß die stochastische Approximation in Satz 28 nicht nur punktweise, sondern auch in L_1 gilt. Dazu verwenden wir ein Kriterium über die Kompaktheit von Mengen in L_1, Lemma 8. Den folgenden funktionalen zentralen Grenzwertsatz zitieren wir ohne Beweis.

Satz 29 *(Ledoux–Talagrand.) Seien $Z, Z_1, Z_2, \ldots$ unabhängige und identisch verteilte L_1-wertige Zufallselemente mit Erwartungswert 0. Dann konvergiert $n^{-1/2}\sum_{i=1}^n Z_i$ in L_1 in Verteilung gegen einen zentrierten Gaußschen Prozeß genau dann, wenn*

$$\begin{aligned}&t^2P(\|Z\|_1 > t) \to 0, \quad t\to\infty,\\ &\int (EZ^2(x))^{1/2}\,dx < \infty.\end{aligned}$$

Mit der Schwarzschen Ungleichung gilt für jede meßbare Funktion g und $\alpha > 1$:

$$\|g\|_1^2 = \Big(\int (1+|x|)^{-\alpha/2} (1+|x|)^{\alpha/2} |g(x)|\, dx \Big)^2 \le C_\alpha \int (1+|x|)^\alpha g^2(x)\, dx$$

mit $C_\alpha = \int (1+|x|)^{-\alpha}\, dx$.

Satz 30 *Hat F eine endliches Moment der Ordnung größer als 2, so konvergiert $n^{1/2}(\mathbb{F} - F)$ in L_1 in Verteilung gegen einen zentrierten Gaußschen Prozeß mit Kovarianzfunktion*

$$(x, y) \to F(x \wedge y) - F(x)F(y).$$

Beweis. Wir wenden Satz 29 auf $Z(x) = \mathbf{1}(X \le x) - F(x)$ an. Es gilt $EZ^2(x) = F(x)(1-F(x))$. Nach Voraussetzung hat F ein endliches Moment der Ordnung $1 + \alpha > 2$. Insbesondere gilt mit partieller Integration:

$$\int (1+|x|)^\alpha EZ^2(x)\, dx = \int (1+|x|)^\alpha F(x)(1-F(x))\, dx < \infty.$$

Also

$$\Big(\int (EZ^2(x))^{1/2}\, dx \Big)^2 \le C_\alpha \int (1+|x|)^\alpha EZ^2(x)\, dx < \infty,$$

$$E\|Z\|_1^2 \le C_\alpha \int (1+|x|)^\alpha EZ^2(x)\, dx < \infty,$$

und daher auch

$$P(\|Z\|_1 > t) \le t^{-2} E\big(\mathbf{1}(\|Z\|_1 > t)\|Z\|_1^2\big) = o(t^{-2}).$$

Ein Wahrscheinlichkeitsmaß P auf der Borel-Algebra $\mathcal{B}$ eines topologischen Raumes E heißt *von innen regulär*, wenn für jede Borelmenge B gilt:

$$PB = \sup\{PK : K \subset B,\ K \text{ kompakt}\}.$$

Ist E polnisch, so ist jedes Wahrscheinlichkeitsmaß von innen regulär.

Eine Familie $\mathcal{P}|\mathcal{B}$ von Wahrscheinlichkeitsmaßen heißt *straff*, wenn zu jedem $\varepsilon > 0$ ein kompaktes $K \subset E$ existiert, so daß

$$PK \ge 1 - \varepsilon, \quad P \in \mathcal{P}.$$

Ist P von innen regulär, so ist P straff. Insbesondere ist jedes Wahrscheinlichkeitsmaß P auf der Borel-Algebra eines polnischen Raumes straff. Gilt $P_n \Rightarrow P$, so ist $\{P_n : n \in \mathbb{N}\}$ ebenfalls straff. Die folgende Charakterisierung kompakter Mengen in L_1 beweisen wir nicht.

Lemma 8 *(Satz von Fréchet–Kolmogorov.) Eine abgeschlossene Teilmenge H von L_1 ist kompakt genau dann, wenn*

$$\begin{aligned} \sup_{h\in H} \|h\|_1 &< \infty, \\ \sup_{|t|<\delta} \sup_{h\in H} \|h_t - h\|_1 &\to 0, \quad \delta \downarrow 0, \\ \sup_{h\in H} \int_{|x|>c} |h(x)|\,dx &\to 0, \quad c \uparrow \infty. \end{aligned}$$

In Worten: Eine abgeschlossene Menge in L_1 ist kompakt genau dann, wenn sie beschränkt, gleichgradig stetig und gleichmäßig integrierbar ist.

Wir zeigen nun, daß die Behauptung von Lemma 7 gleichmäßig über kompakte Mengen gilt.

Lemma 9 *Sei $H \subset L_1$ kompakt. Es gelte $\int k(x)\,dx = 1$ und $b \to 0$. Dann gilt*

$$\sup_{h\in H} \|h * k_b - h\|_1 \to 0, \quad b \to 0.$$

Beweis. Es gilt

$$h * k_b(x) - h(x) = \int (h(x - bu) - h(x))k(u)\,du,$$

also

$$\sup_{h\in H} \|h * k_b - h\|_1 \le \int g_b(u)|k(u)|\,du$$

mit

$$g_b(u) = \sup_{h\in H} \int |h(x - bu) - h(x)|\,dx.$$

Aus der Gleichstetigkeit von H folgt, daß $g_b(u) \to 0$ für $b \to 0$. Aus der Beschränktheit von H folgt, daß $g_b(u) \le 2\sup_{h\in H} \|h\|_1 < \infty$. Die Behauptung folgt jetzt mit dem Satz von der dominierten Konvergenz.

Wie in Kapitel 15 bezeichne $\mathcal{L}_{2,0}$ die Klasse der Funktionen f, die differenzierbar sind mit f' absolut stetig und $f'' \in L_1$. Wie in Kapitel 14 bezeichne $\mathcal{K}_{2,0}$ die Klasse der Funktionen k, die beschränkt sind mit $\int k(t)\,dt = 1$ und $\int t^j k(t)\,dt = 0$ für $j = 1, 2$. Wir erhalten folgende L_1-Version von Satz 28.

Satz 31 *Ist $F \in \mathcal{L}_{2,0}$, $k \in \mathcal{K}_{2,0}$ und $b = O(n^{-1/4})$, so gilt*

$$\|\hat{\mathbb{F}} - \mathbb{F}\|_1 = o_p(n^{-1/2}).$$

Hat F eine endliches Moment der Ordnung größer als 2, so konvergiert $n^{1/2}(\hat{\mathbb{F}} - F)$ in Verteilung in L_1 gegen einen zentrierten Gaußschen Prozeß mit Kovarianzfunktion

$$(x, y) \to F(x \wedge y) - F(x)F(y).$$

Beweis. Nach Satz 30 ist $n^{1/2}(\mathbb{F} - F)$ straff in L_1. Für $\varepsilon > 0$ existiert also ein kompaktes $H \subset L_1$ mit $P(n^{1/2}(\mathbb{F} - F) \in H) \geq 1 - \varepsilon$ für $n \in \mathbb{N}$. Es gilt $\hat{\mathbb{F}} = \mathbb{F} * k_b$. Aus Lemma 9, angewandt für $h = n^{1/2}(\mathbb{F} - F)$, folgt also $\|\hat{\mathbb{F}} - F * k_b - \mathbb{F} + F\|_1 = o_p(n^{-1/2})$. Den Bias-Term behandeln wir ähnlich wie im Beweis von Satz 28. Durch Taylor-Entwicklung:

$$\begin{aligned} F(x - bu) - F(x) = -buf(x) + \frac{1}{2}b^2u^2f'(x) \\ + b^2u^2 \int_0^1 (1 - t)(f'(x - tbu) - f'(x))\, dt. \end{aligned}$$

Mit $\int uk(u)\, du = 0$ gilt daher

$$F * k_b(x) - F(x) = b^2 \int_0^1 (1 - t)(f'(x - tbu) - f'(x))u^2k(u)\, du\, dt.$$

Aus Lemma 7 folgt dann $\|F * k_b - F\|_1 = o(n^{-1/2})$. Die Konvergenz in Verteilung folgt jetzt aus Satz 30.

20 Faltungsschätzer

Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit Dichte f und Verteilungsfunktion F. Wir wollen die Dichte $g = f * f$ von $X_1 + X_2$ schätzen. Die Faltung $f * f$ ist ein "glattes" Funktional von f, das dem linearen Funktional $\int g(x)f(x)\, dx$ aus Kapitel 19 ähnelt. Einen Schätzer für $f * f$ kann man zum Beispiel verwenden, um zu testen, ob f normal ist. Denn dann müßte $f*f$ ebenfalls normal sein (mit dem doppelten Erwartungswert und der doppelten Varianz). Der entsprechend umstandardisierte Schätzer für $f * f$ müßte also nahe dem für f sein.

Ist $\hat f(x) = \frac{1}{n}\sum_{i=1}^n k_b(x - X_i)$ ein Kernschätzer für $f(x)$, so erhält man einen Schätzer für $f * f$ durch

$$\hat f * \hat f(x) = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n k_b * k_b(x - X_i - X_j),$$

eine *lokale von-Mises-Statistik*. Wir haben

$$\hat f * \hat f - f * f = 2f * (\hat f - f) + (\hat f - f) * (\hat f - f).$$

Gilt $\|\hat f - f\|_1 = O_p(a_n)$, so gilt also

$$\|\hat f * \hat f - f * f\|_1 \le 2\|f\|_1\|\hat f - f\|_1 + \|\hat f - f\|_1^2 = O_p(a_n).$$

Wir zeigen mit den Methoden von Kapitel 19, daß sogar die Rate $n^{-1/2}$ gilt. Es ist technisch bequemer, in $\hat f * \hat f$ die Summanden $i = j$ wegzulassen und die *lokale U-Statistik* als Schätzer zu nehmen:

$$\frac{2}{n(n-1)}\sum_{1\le i<j\le n} k_b * k_b(x - X_i - X_j).$$

Wir können auch $k_b * k_b$ durch einen allgemeinen Kern k_b ersetzen:

$$\hat g(x) = \frac{2}{n(n-1)}\sum_{1\le i<j\le n} k_b(x - X_i - X_j).$$

Diesen Schätzer könnte man auch als "gewöhnlichen" Kernschätzer ansehen, der auf $n(n-1)/2$ Beobachtungen $X_i + X_j$, $1 \le i < j \le n$, beruht.

Für $i \ne j$ haben wir

$$E(k_b(x - X_i - X_j)|X_i) = \int k_b(x - X_i - y)f(y)\,dy = f * k_b(x - X_i)$$

und

$$\begin{aligned} Ek_b(x - X_i - X_j) &= \iint k_b(x - y - z)f(y)\,dy\,f(z)\,dz \\ &= \int k_b(x - y)\int f(y - z)f(z)\,dz\,dy \\ &= k - b * (f * f) = g * k_b. \end{aligned}$$

Für $\hat g$ gilt also die *Hoeffding-Zerlegung*

$$\begin{aligned} &\hat g(x) = g * k_b(x) \\ &+ \frac{2}{n(n-1)}\sum_{i<j}\big(k_b(x - X_i - X_j) - f * k_b(x - X_i) - f * k_b(x - X_j) + g * k_b(x)\big) \\ &+ \frac{2}{n}\sum_{i=1}^n\big(f * k_b(x - X_i) - g * k_b(x)\big); \end{aligned}$$

anders geschrieben

$$\hat{g} = g * k_b + 2\mathbb{H} * k_b + \mathbb{U}$$

mit

$$\mathbb{H}(x) = \frac{1}{n}\sum_{i=1}^{n}(f(x - X_i) - Ef(x - X)) = \frac{1}{n}\sum_{i=1}^{n}(f(x - X_i) - g(x)),$$

$$\mathbb{U}(x) = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \big(k_b(x - X_i - X_j) - f * k_b(x - X_i) - f * k_b(x - X_j) + g * k_b(x)\big).$$

Lemma 10 *Für ein $\alpha > 1$ seien $\int |x|^\alpha f(x)\,dx$ und $\int (1 + |x|)^\alpha f^2(x)\,dx$ endlich. Dann konvergiert $n^{1/2}\mathbb{H}$ in Verteilung in L_1 gegen einen zentrierten Gaußschen Prozeß mit Kovarianzfunktion*

$$(x, y) \to \int f(x - z)f(y - z)f(z)\,dz - g(x)g(y).$$

Beweis. Wie im Beweis von Satz 30 haben wir mit $Z(x) = f(x-X)-g(x)$:

$$\Big(\int (EZ^2(x))^{1/2}\,dx\Big)^2 \le C_\alpha \int (1 + |x|)^\alpha EZ^2(x)\,dx,$$

$$E\|Z\|_1^2 \le C_\alpha \int (1 + |x|)^\alpha EZ^2(x)\,dx$$

und

$$P(\|Z\|_1 > t) \le t^{-2}E\big(\mathbf{1}(\|Z\|_1 > t)\|Z\|_1^2\big).$$

Es bleibt zu zeigen, daß $\int (1 + |x|)^\alpha EZ^2(x)\,dx$ endlich ist. Zunächst gilt

$$EZ^2(x) \le Ef^2(x - X) = f^2 * f(x).$$

Mit den Voraussetzungen an f und mit $(1 + |x|)^\alpha \le (1 + |x - y|)^\alpha(1 + |y|)^\alpha$ ergibt sich

$$\int (1 + |x|)^\alpha f^2 * f(x)\,dx \le \int (1 + |x|)^\alpha f^2(x)\,dx \int (1 + |x|)^\alpha f(x)\,dx < \infty.$$

Lemma 11 *Sei k eine beschränkte Dichte mit Mittelwert 0 und endlichem zweiten Moment. Für ein $\alpha > 1$ sei $\int |x|^\alpha f(x)\,dx$ endlich. Dann gilt*

$$\|\mathbb{U}\|_1 = O_p(n^{-1}b^{-1/2}).$$

Beweis. Ohne Einschränkung sei $\alpha \le 2$. Die Voraussetzungen an k implizieren, daß $\int (1+|x|)^\alpha k^2(x)\,dx$ endlich ist. Es gilt

$$E\mathbb{U}^2(x) \le \frac{2}{n(n-1)} E k_b^2(x - X_1 - X_2) \le \frac{2}{n(n-1)} k_b^2 * g(x).$$

Also

$$E\|\mathbb{U}\|_1^2 \le C_\alpha \int (1+|x|)^\alpha E\mathbb{U}^2(x)\,dx \le C_\alpha \frac{2}{n(n-1)} \int (1+|x|)^\alpha k_b^2 * g(x)\,dx.$$

Mit den Voraussetzungen an k ergibt sich

$$\begin{aligned}
&\int (1+|x|)^\alpha k_b^2 * g(x)\,dx \\
&\quad = \frac{1}{b^2} \iiint (1+|x|)^\alpha k^2\Big(\frac{x-t}{b}\Big) f(t-y) f(y)\,dt\,dy\,dx \\
&\quad = b^{-1} \iiint (1+|x+y+bz|)^\alpha f(x) f(y) k^2(z)\,dx\,dy\,dz \\
&\quad \le b^{-1} \Big(\int (1+|x|)^\alpha f(x)\,dx \Big)^2 \int (1+|x|)^\alpha k^2(x)\,dx = O(b^{-1}).
\end{aligned}$$

Also gilt $E\|\mathbb{U}\|_1^2 = O(n^{-2}b^{-1})$. Daraus folgt die Behauptung.

Satz 32 *Sei f wie in Lemma 10 und k wie in Lemma 11. Sei g L_1-Lipschitz der Ordnung $\gamma > 1/2$. Gelte $nb \to \infty$ und $nb^{2\gamma} \to 0$. Dann gilt*

$$\|\hat{g} - g - 2\mathbb{H}\|_1 = o_p(n^{-1/2}),$$

und $n^{1/2}(\hat{g}-g)$ konvergiert in Verteilung in L_1 gegen einen zentrierten Gaußschen Prozeß mit Kovarianzfunktion

$$(x,y) \to 4 \int f(x-z) f(y-z) f(z)\,dz - 4 g(x) g(y).$$

Beweis. Wir schreiben

$$\hat{g} - g = 2\mathbb{H} + g * k_b - g + 2(\mathbb{H} * k_b - \mathbb{H}) + \mathbb{U}.$$

Mit der Voraussetzung an g gilt

$$\|g * k_b - g\|_1 \le \iint |g(x-bu) - g(x)| dx\, k(u)\,du \le L b^\gamma = o_p(n^{-1/2}).$$

Nach Lemma 11 gilt $\|\mathbb{U}\|_1 = O_p(n^{-1}b^{-1/2})$. Aus Lemma 9, angewandt für $h = n^{1/2}\mathbb{H}$, folgt $\|\mathbb{H} * k_b - \mathbb{H}\|_1 = o_p(n^{-1/2})$. Also gilt die behauptete stochastische Entwicklung. Die behauptete Konvergenz in Verteilung folgt jetzt aus Lemma 10.

21 Rangtests

Im folgenden beschreiben wir einige Tests in nichtparametrischen Modellen. Resultate dazu beweisen wir nicht. Insbesondere testen wir die Lage einer unbekannten Verteilung, die Lagedifferenz der unbekannten Verteilung zweier unabhängiger Stichproben und die Abhängigkeit der Komponenten einer zweidimensionalen Zufallsvariablen. Eine Möglichkeit wäre, als Teststatistik den Schätzer eines geeigneten Funktionals zu wählen. Die Verteilung des Schätzers hängt aber im allgemeinen von der Verteilung der Beobachtungen ab; wir können also die kritische Grenze des Tests zu einem gegebenen Niveau nur asymptotisch bestimmen. Für Tests, die exakt das Niveau einhalten, brauchen wir Teststatistiken, deren Verteilung nicht von der Verteilung der Beobachtungen abhängt, d.h. *verteilungsfreie* Teststatistiken. In diesem Kapitel stellen wir verteilungsfreie Tests vor, die auf den "Rängen" der Beobachtungen basieren.

Testen des Medians

Definition. Ein *Median* m einer Zufallsvariablen X erfüllt $P(X \geq m) \geq 1/2$ und $P(X \leq m) \geq 1/2$.

Sei F eine unbekannte stetige Verteilungsfunktion mit (unbekanntem, aber) eindeutigem Median m. Seien $X_1, \ldots, X_n$ unabhängig mit Verteilungsfunktion F. Wir wollen die Hypothese $m = m_0$ testen.

Ein naheliegender Test beruht auf einem Schätzer für den Median. Sei $k = n/2 + o(n^{1/2})$. Dann gilt nach Satz 21 unter der Nullhypothese:

$$n^{1/2}(X_{k:n} - m_0) \Rightarrow N(0, 1/(4f^2(m_0))).$$

Ein kritischer Bereich für $m = m_0$ gegen $m > m_0$ zum asymptotischen Niveau α ist also

$$C_\alpha = \{n^{1/2}(X_{k:n} - m_0) > \xi_{1-\alpha}/(2\hat{f}(m_0))\};$$

dabei ist $\hat{f}(m_0)$ zum Beispiel ein Kernschätzer für $f(m_0)$ und $\xi_{1-\alpha}$ das $1-\alpha$-Quantil von $N(0,1)$. Da wir $f(m_0)$ nicht kennen, können wir keinen (nichttrivialen) auf $X_{k:n}$ beruhenden Test angeben, der exakt das Niveau α einhält.

Setze

$$R^+(X) = \begin{cases} 1 & X > m_0 \\ 0, & X < m_0 \end{cases}$$

und

$$T^+ = \sum_{i=1}^{n} R^+(X_i) = \#\{i = 1, \ldots, n : X_i > m_0\}.$$

Unter der Nullhypothese sind $R^+(X_1), \ldots, R^+(X_n)$ unabhängig verteilt nach $B_{1,1/2}$. Also ist T^+ verteilt nach $B_{n,1/2}$. Insbesondere hängt die Verteilung von T^+ nicht von der unbekannten Verteilung der Beobachtungen ab.

Der *Vorzeichentest* für $m = m_0$ gegen $m > m_0$ ist

$$\psi = \begin{cases} 1, & T^+ > c \\ a, & T^+ = c \\ 0, & T^+ < c, \end{cases}$$

wobei sich a und c aus $B_{n,1/2}(c, \infty) + aB_{n,1/2}\{c\} = \alpha$ ergeben. Unter einer Verteilung P mit Median $m > m_0$ gilt

$$P(R^+(X) = 1) = P(X > m_0) = p > 1/2,$$

also ist die Güte des Tests gleich $B_{n,p}(c, \infty) + aB_{n,p}\{c\}$, hängt also über p von P ab.

Testen des Medians bei Symmetrie

Definition. Der *Rang* von X_j in den Zufallsvariablen $X_1, \ldots, X_n$ ist

$$R_j = R(X_j) = \#\{i = 1, \ldots, n : X_i \leq X_j\}.$$

Setze $R = (R_1, \ldots, R_n)$. Sind $X_1, \ldots, X_n$ unabhängig mit stetiger Verteilungsfunktion F, so sind $X_1, \ldots, X_n$ f.s. verschieden. Insbesondere gilt $X_{1:n} < \cdots < X_{n:n}$ und $X_i = X_{R_i:n}$ f.s. Der Rangvektor R ist gleichverteilt über den Permutationen von $\{1, \ldots, n\}$. Sei nun zusätzlich bekannt, daß die zu F gehörende Verteilung symmetrisch um den Median m ist.

Nehmen wir zunächst an, daß die Verteilung ein endliches zweites Moment hat. Wegen der Symmetrie gilt $m = EX$. Wir können also m durch das Stichprobenmittel $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ schätzen. Nach dem zentralen Grenzwertsatz gilt unter der Nullhypothese

$$n^{1/2}(\bar{X} - m_0) \Rightarrow N(0, \sigma^2)$$

mit $\sigma^2 = E(X - EX)^2$. Ein kritischer Bereich für $m = m_0$ gegen $m > m_0$ zum asymptotischen Niveau α ist also

$$C_\alpha = \{n^{1/2}(\bar{X} - m_0) > \xi_{1-\alpha}\hat{\sigma}\};$$

dabei ist zum Beispiel $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$ die Stichprobenvarianz.

Um einen Test zum exakten Niveau α zu erhalten, müssen wir anders vorgehen. Setze

$$D_i = X_i - m_0.$$

Da F stetig ist, gilt $P(D_i = 0) = 0$ und $P(|D_i| = |D_j|) = 0$ für $i \neq j$. Wir dürfen also annehmen, daß $|D_1|, \ldots, |D_n|$ verschieden und ungleich 0 sind. Sei R_i^+ der Rang von $|D_i|$ in $|D_1|, \ldots, |D_n|$. Die *Vorzeichen-Rangstatistik* ist

$$T^+ = \sum_{D_i > 0} R_i^+.$$

Der *Vorzeichen-Rangtest* von Wilcoxon beruht auf dieser Teststatistik.

Zweistichproben-Lagetests

Sei F eine unbekannte stetige Verteilungsfunktion. Seien $X_1, \ldots, X_m$ und $Y_1, \ldots, Y_n$ unabhängig mit Verteilungsfunktionen F und $F(\cdot - \vartheta)$. Wir wollen die Hypothese $\vartheta = 0$ testen.

Es sind $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ f.s. verschieden. Sei $R(X_j)$ der Rang von X_j in $X_1, \ldots, X_m, Y_1, \ldots, Y_n$. Die *Rangsumme* von $X_1, \ldots, X_n$ ist

$$S_X = \sum_{i=1}^m R(X_i).$$

Der *Rangsummentest* von Wilcoxon beruht auf dieser Teststatistik. Statt der Rangsumme kann man auch die *Mann–Whitney-Statistik* verwenden,

$$\begin{aligned} W &= \sum_{i=1}^m \sum_{j=1}^n \mathbf{1}(X_i > Y_j) = \sum_{i=1}^m \Big(R(X_i) - \sum_{j=1}^m \mathbf{1}(X_i > X_j) \Big) \\ &= S_X - \frac{m(m+1)}{2}. \end{aligned}$$

Eine Verallgemeinerung sind die *linearen Rangstatistiken*

$$L = \sum_{i=1}^m a(R(X_i))$$

mit *Scores* $a(1) \leq \cdots \leq a(m+n)$.

Mehrstichproben-Lagetests

Sei F eine unbekannte stetige Verteilungsfunktion. Seien X_{ij}, $j = 1, \ldots, n_i$, $i = 1, \ldots, p$, unabhängig und haben $X_{i1}, \ldots, X_{in_i}$ die Verteilungsfunktion $F(\cdot - \vartheta_i)$. Wir wollen die Hypothese $\vartheta_1 = \ldots \vartheta_p = 0$ testen. Nach Voraussetzung sind die Beobachtungen f.s. verschieden. Sei $R(X_{ij})$ der Rang von X_{ij} in $X_{11}, \ldots, X_{1n_1}, \ldots, X_{p1}, \ldots, X_{pn_p}$. Wie beim Wilcoxon-Test setze

$$S_i = \sum_{j=1}^{n_i} R(X_{ij}), \quad i = 1, \ldots, p.$$

Es gilt mit $N = n_1 + \cdots + n_p$:

$$\sum_{i=1}^{p} n_i \frac{S_i}{n_i} = \sum_{i=1}^{p} S_i = \frac{N(N+1)}{2}.$$

Unter der Hypothese gilt also

$$E(S_1/n_1) = \cdots = E(S_p/n_p) = (N+1)/2.$$

Unter der Alternative gilt nicht überall Gleichheit. Die Teststatistik von *Kruskal–Wallis* ist

$$T = \frac{12}{N(N+1)} \sum_{i=1}^{p} n_i \Big(\frac{S_i}{n_i} - \frac{N+1}{2} \Big)^2.$$

Unter der Hypothese ist T asymptotisch verteilt nach χ^2_{p-1}.

Varianzanalyse

Sei F eine unbekannte stetige Verteilungsfunktion. Seien X_{ijk}, $k = 1, \ldots, n$, $i = 1, \ldots, p$, $j = 1, \ldots, q$, unabhängig, und haben $X_{ij1}, \ldots, X_{ijn}$ die Verteilungsfunktion $F(\cdot - \vartheta_{ij})$. Wir wollen die Hypothese testen, daß alle ϑ_{ij} gleich 0 sind. Nach Voraussetzung sind die Beobachtungen f.s. verschieden. Für die Spalten $j = 1, \ldots, q$ sei $R(X_{ijk})$ der Rang von X_{ijk} in der j-ten Spalte

$$X_{1j1}, \ldots, X_{1jn}, \ldots, X_{pj1}, \ldots, X_{pjn}.$$

Es gilt

$$\bar{R}_{\cdot\cdot\cdot} = \frac{1}{pqn} \sum_{j=1}^{q} \Big(\sum_{i,k} R(X_{ijk}) \Big) = \frac{1}{pqn} q \frac{pn(pn+1)}{2} = \frac{pn+1}{2}.$$

Für Zeile i setze

$$\bar{R}_{i..} = \frac{1}{qn} \sum_{j=1}^{q} \sum_{k=1}^{n} R(X_{ijk}).$$

Unter der Hypothese sind die $\bar{R}_{i..}$ ungefähr gleich, d.h. die $\bar{R}_{i..}$ sind alle ungefähr gleich $\bar{R}_{...}$. Unter der Hypothese gilt wie beim Mehrstichproben-Lagetest

$$E\bar{R}_{i..} = (pn+1)/2.$$

Die Teststatistik von *Friedman* ist

$$T = \frac{12q}{p(pn+1)} \sum_{i=1}^{p} \big(\bar{R}_{i..} - (pn+1)/2\big)^2.$$

Unter der Hypothese ist T asymptotisch verteilt nach χ^2_{p-1}.

Unabhängigkeitstests

Seien (X_i, Y_i), $i = 1, \ldots, n$, unabhängig und identisch verteilt mit endlicher Varianz. Der *Korrelationskoeffizient* von (X, Y) ist

$$\varrho = \frac{\mathrm{Cov}\,(X,Y)}{(\mathrm{Var}\,X\ \mathrm{Var}\,Y)^{1/2}}.$$

Wir wollen die Hypothese $\varrho = \varrho_0$ testen. Ein Schätzer für ϱ ist der *empirische Korrelationskoeffizient*

$$\hat{\varrho} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\big(\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2\big)^{1/2}}.$$

Ein kritischer Bereich für $\varrho = \varrho_0$ gegen $\varrho > \varrho_0$ zum asymptotischen Niveau α ist also

$$C_\alpha = \{n^{1/2}(\hat{\varrho} - \varrho_0) > \xi_{1-\alpha}\hat{c}\};$$

dabei ist $\hat{c}^2$ ein Schätzer für die asymptotische Varianz von $\hat{\varrho}$.

Eine verteilungsfreie Teststatistik erhält man, indem man die Korrelation der *Ränge* betrachtet. Sei $R(X_j)$ der Rang von X_j in $X_1, \ldots, X_n$ und $R(Y_j)$ der Rang von Y_j in $Y_1, \ldots, Y_n$. Es gilt

$$\frac{1}{n} \sum_{j=1}^{n} R(X_j) = \frac{1}{n} \sum_{j=1}^{n} R(Y_j) = \frac{n+1}{2}.$$

Der *Rangkorrelationskoeffizient von Spearman* ist

$$\hat{\varrho}_s = \frac{\sum(R(X_j) - (n+1)/2)(R(Y_j) - (n+1)/2)}{\left(\sum(R(X_j) - (n+1)/2)^2 \sum(R(Y_j) - (n+1)/2)^2\right)^{1/2}}.$$

Eine Verallgemeinerung des Rangkorrelationskoeffizienten sind Statistiken der Form

$$L = \sum_{j=1}^{n} a(R(X_j))b(R(Y_j))$$

mit Scores $a(1) \leq \cdots \leq a(n)$ und $b(1) \leq \cdots \leq b(n)$.

Ein Spezialfall ist der *Quadrantentest*, bei dem die Scores wie folgt gewählt werden:

$$a(j) = b(j) = \begin{cases} 0, & j \leq (n+1)/2) \\ 1, & j > (n+1)/2). \end{cases}$$

Die zugehörige Statistik L ist die Anzahl der Punkte (X_j, Y_j), die rechts oben von $\text{med}(X_1, \ldots, X_n), \text{med}(Y_1, \ldots, Y_n)$ liegen.

Ein weiterer Unabhängigkeitstest beruht auf dem *Rangkorrelationskoeffizienten von Kendall*,

$$\tau_K = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \text{sign}\,(R(X_i) - R(X_j))\text{sign}\,(R(Y_i) - R(Y_j)).$$

22 Schätzer in nichtparametrischen Modellen mit strukturellen Annahmen

Weiß man etwas über die Struktur der den Beobachtungen zugrundeliegenden Verteilung, so lassen sich häufig bessere als die empirischen Schätzer finden. Wir diskutieren insbesondere die Symmetrie der Verteilung um ein bekanntes oder unbekanntes Zentrum, die Unabhängigkeit der Komponenten mehrdimensionaler Beobachtungen und die Monotonie der Dichte.

Symmetrie

Seien $X_1, \ldots, X_n$ unabhängige Zufallsvariablen mit einer um 0 symmetrischen Verteilung P, also $PA = P(-A)$, $A \in \mathcal{B}$. Sei $h(X)$ quadratintegrierbar. Der übliche Schätzer für $Eh(X)$ ist der empirische Schätzer $H = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$. Da P symmetrisch um 0 ist, gilt $Eh(X) = Eh(-X)$. Also ist $H_- = \frac{1}{n}\sum_{i=1}^{n} h(-X_i)$ ebenfalls ein Schätzer für $Eh(X)$. Beide sind

asymptotisch normal mit Varianz $\operatorname{Var} h(X)$. Sie lassen sich konvex kombinieren. Der Schätzer $H_s = (H + H_-)/2$ ist asymptotisch normal mit Varianz

$$\frac{1}{2}\operatorname{Var} h(X) + \frac{1}{2}E(h(X) - Eh(X))(h(-X) - Eh(X)).$$

Nach der Schwarzschen Ungleichung gilt

$$|E(h(X) - Eh(X))(h(-X) - Eh(X))| \leq \operatorname{Var} h(X).$$

Also ist H_s nie schlechter als H oder H_-. Ist H schon symmetrisch, so gilt $H_- = H$, also $H_s = H$, und es gibt keine Verbesserung. Falls h sehr unsymmetrisch ist, wird H_s jedoch viel besser sein als H oder H_-. Im Extremfall einer antisymmetrischen Funktion, $h(-x) = -h(x)$, gilt $H_- = -H$, also $H_s = 0$. Allerdings weiß man dann, daß $Eh(X) = 0$ ist und braucht keinen Schätzer.

Symmetrie mit unbekanntem Zentrum

Seien $X_1, \ldots, X_n$ unabhängige Zufallsvariablen mit einer um ein unbekanntes Zentrum ϑ symmetrischen Verteilung P, also $P(\vartheta + A) = P(\vartheta - A)$, $A \in \mathcal{B}$. Dann ist $X = \vartheta + X - \vartheta$ verteilt wie $\vartheta - (X - \vartheta) = 2\vartheta - X$, also $Eh(X) = Eh(2\vartheta - X)$. Hat man einen Schätzer $\hat{\vartheta}$ für ϑ, zum Beispiel den Stichprobenmedian oder das Stichprobenmittel, so erhält man neben H einen neuen Schätzer für $Eh(X)$,

$$H_- = \frac{1}{n}\sum_{i=1}^{n} h(2\hat{\vartheta} - X_i).$$

Wir können jetzt nicht mehr sicher sein, daß eine Konvexkombination $aH + (1-a)H_-$ besser als H ist, denn das Schätzen von ϑ kann die Varianz von H_- erhöhen. Heuristisch gilt

$$\begin{aligned} H_- &= \frac{1}{n}\sum_{i=1}^{n} h(2\vartheta - X_i) + 2(\hat{\vartheta} - \vartheta)\frac{1}{n}\sum_{i=1}^{n} h'(2\vartheta - X_i) + o_p(n^{-1/2}) \\ &= \frac{1}{n}\sum_{i=1}^{n} h(2\vartheta - X_i) + 2(\hat{\vartheta} - \vartheta)Eh'(2\vartheta - X) + o_p(n^{-1/2}). \end{aligned}$$

Für $\hat{\vartheta} = \bar{X}$ also

$$H_- = \frac{1}{n}\sum_{i=1}^{n} \big(h(2\vartheta - X_i) + 2(X_i - \vartheta)Eh'(2\vartheta - X)\big) + o_p(n^{-1/2}).$$

Die asymptotische Varianz ist

$$\operatorname{Var} h(X) + 4E(X-\vartheta)^2(Eh'(2\vartheta - X))^2 + 4E(X-\vartheta)h(2\vartheta - X)Eh'(2\vartheta - X).$$

Unabhängigkeit

Seien (X_i, Y_i), $i = 1, \ldots, n$ unabhängige Zufallsvektoren. Ein Schätzer für den Erwartungswert $Eh(X, Y)$ ist der empirische Schätzer

$$H = \frac{1}{n}\sum_{i=1}^{n} h(X_i, Y_i) = \int h \, d\mathbb{E},$$

wobei $\mathbb{E}(s,t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \le s, Y_i \le t)$ die empirische Verteilungsfunktion der Beobachtungen ist. Es gilt

$$n^{1/2}(H - Eh(X,Y)) = n^{-1/2}\sum_{i=1}^{n}(h(X_i, Y_i) - Eh(X,Y)).$$

Also ist H asymptotisch normal mit Varianz $\operatorname{Var} h(X, Y)$.

Seien X und Y unabhängig mit Verteilungsfunktionen F und G. Dann ist $Eh(X,Y) = \iint h(x,y)\, dF(x)\, dG(y)$, und wir können $Eh(X,Y)$ mit der (verallgemeinerten) *von-Mises-Statistik*

$$H_M = \iint h(x,y)\, d\mathbb{F}(x)\, d\mathbb{G}(y) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} h(X_i, Y_j)$$

schätzen, wobei $\mathbb{F}(s) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \le s)$ und $\mathbb{G}(t) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}(Y_j \le t)$ die empirischen Verteilungsfunktionen zu $X_1, \ldots, X_n$ und $Y_1, \ldots, Y_n$ sind. (Bei den "verallgemeinerten" von-Mises- oder U-Statistiken hat man zwei (oder mehrere) Stichproben.) Um die asymptotische Verteilung zu finden, approximieren wir H_M durch eine lineare Statistik. Dazu verwenden wir die *Hoeffding-Zerlegung*

$$H_M - Eh(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(\bar{h}_{11}(X_i) + \bar{h}_{12}(Y_i)) + \mathbb{U}$$

mit

$$\mathbb{U} = \iint \bar{h}_2(x,y)\, d\mathbb{F}(x)\, d\mathbb{G}(y) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \bar{h}_2(X_i, Y_j).$$

Dabei bedeutet immer $\bar{k} = k - Ek(X,Y)$ und

$$\begin{aligned} h_{11}(X) &= E(h(X,Y)|X), \\ h_{12}(Y) &= E(h(X,Y)|Y), \\ h_2(X,Y) &= h(X,Y) - h_{11}(X) - h_{12}(Y). \end{aligned}$$

Es gilt

$$E\mathbb{U}^2 \leq \frac{1}{n^2} Eh^2(X,Y) = O(n^{-2}).$$

Die von-Mises-Statistik H_M heißt *ausgeartet* (unter der gegebenen Verteilung), wenn $E\bar{h}_{11}^2(X) = E\bar{h}_{12}^2(Y) = 0$. Ist H_M *nicht* ausgeartet, so erhalten wir die stochastische Entwicklung

$$n^{1/2}(H_M - Eh(X,Y)) = n^{-1/2} \sum_{i=1}^{n} (\bar{h}_{11}(X_i) + \bar{h}_{12}(Y_i)) + o_p(1).$$

Die Zufallsvariablen $\bar{h}_{11}(X)$ und $\bar{h}_{12}(Y)$ sind *orthogonal*,

$$E\bar{h}_{11}(X)\bar{h}_{12}(Y) = 0.$$

Also ist H_M asymptotisch normal mit Varianz

$$\operatorname{Var} h_{11}(X) + \operatorname{Var} h_{12}(Y) = E\bar{h}_{11}^2(X) + E\bar{h}_{12}^2(Y).$$

Die Hoeffding-Zerlegung ist ebenfalls orthogonal,

$$\begin{aligned} &E\bar{h}_2(X,Y)(\bar{h}_{11}(X) + \bar{h}_{12}(Y)) \\ &= E(\bar{h}(X,Y) - \bar{h}_{11}(X) - \bar{h}_{12}(Y))(\bar{h}_{11}(X) + \bar{h}_{12}(Y)) = 0. \end{aligned}$$

Insbesondere läßt sich die Varianz von H wie folgt zerlegen:

$$E\bar{h}^2(X,Y) = E\bar{h}_2^2(X,Y) + E\bar{h}_{11}^2(X) + E\bar{h}_{12}^2(Y).$$

Also hat H_M eine um $E\bar{h}_2^2(X,Y)$ kleinere Varianz als H.

Es gilt allgemeiner für Funktionen a und b mit $Ea(X) = Eb(Y) = 0$:

$$E(\bar{h}(X,Y) - \bar{h}_{11}(X) - \bar{h}_{12}(Y))(a(X) + b(Y)) = 0.$$

Also ist $\bar{h}_{11}(X) + \bar{h}_{12}(Y)$ die Projektion von $\bar{h}(X,Y)$ auf den Raum der Funktionen $a(X) + b(Y)$. Eine von-Mises-Statistik ist also genau dann ausgeartet, wenn $\bar{h}(X,Y)$ orthogonal zu diesem Raum ist. Für von-Mises-Statistiken höherer Ordnung, also für Funktionen h mit mehr als zwei Argumenten,

ergibt sich die Hoeffding-Zerlegung analog durch sukzessive Projektion auf Summen von Funktionen mit immer weniger Argumenten.

Monotone Dichte

Seien $X_1, \ldots, X_n$ unabhängige und positive Zufallsvariablen mit einer nichtwachsenden Dichte f. Die Dichte läßt sich mit einem Kernschätzer $\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} k_b(x - X_i)$ mit $k_b(x) = k(x/b)/b$, Kern k und Bandweite b schätzen. Dieser Schätzer ist aber i.a. selbst keine nichtwachsende Funktion und macht auch keinen Gebrauch von der Annahme an die Dichte. Obwohl das Modell unendlichdimensional ist, können wir hier den Maximum-Likelihood-Schätzer für f bestimmen, d.h. wir können die Likelihood-Funktion

$$f \to \prod_{i=1}^{n} f(X_i)$$

maximieren. Man kann sich überlegen, daß die Lösung eine linksstetige Treppenfunktion ist, die allenfalls in den Beobachtungen springt. Außerdem muß die Lösung auf $(-\infty, 0)$ und $(X_{n:n}, \infty)$ verschwinden. Sie ist insbesondere schon durch die Werte $f_i = \tilde{f}(X_{i:n})$ bestimmt. Das Maximierungsproblem reduziert sich also auf das endlichdimensionale Problem der Maximierung von

$$(f_1, \ldots, f_n) \to \prod_{i=1}^{n} f_i$$

unter den Nebenbedingungen

$$f_1 \geq \cdots \geq f_n \quad \text{und} \quad \sum_{i=1}^{n} (X_{i:n} - X_{i-1:n}) f_i = 1;$$

dabei ist $X_{0:n} = 0$ gewählt.

Man kann zeigen, daß sich die Lösung graphisch bestimmen läßt, indem man zunächst die kleinste konkave Majorante $\tilde{\mathbb{F}}$ der empirischen Verteilungsfunktion $\mathbb{F}$ bestimmt. Der gesuchte Maximum-Likelihood-Schätzer für die Dichte f ist die linksseitige Ableitung von $\tilde{\mathbb{F}}$.