Vorlesung Mathematische Statistik
WS 1991/92

Gerhard Keller

Mathematisches Institut
Universität Erlangen-Nürnberg

Dieses Vorlesungsskript stützt sich wesentlich auf das Buch „Vorlesungen zur Mathematischen Statistik“ von W.Winkler [11] und das Skriptum “Mathematische Statistik“ von U.Krengel [7].

Inhaltsverzeichnis

Kapitel 1

Das Bernoullische Versuchsschema

1.1 Grundbegriffe

Definition 1.1.1 *Ein Tripel $(M, \mathcal{M}, \mathcal{P})$ heißt ein* **statistischer Raum**, *falls $(M, \mathcal{M})$ ein meßbarer Raum und $\mathcal{P}$ eine nichtleere Familie von Wahrscheinlichkeitsverteilungen ist.*

Das Grundproblem der Statistik besteht darin, eine oder mehrere zufällige Realisierungen eines Experiments mit Werten in M zu beobachten und daraus auf die den Realisierungen zugrunde liegende Wahrscheinlichkeitsverteilung $P \in \mathcal{P}$ zu schließen.

Das einfachste Beispiel für einen statistischen Raum ist

$$M = \{0, 1\},\ \mathcal{M} = \text{Potenzmenge von } M,\ \mathcal{P} = \{P_p : p \in [0, 1]\},$$

wo P_p die Binomialverteilung zum Parameter p bezeichnet. Durch Produktbildung gelangen wir zum **endlichen Bernoullischen Versuchsschema** $(M, \mathcal{M}, \mathcal{P})^n = (M^n, \mathcal{M}^n, \mathcal{P}^n)$, wo $\mathcal{P}^n := \{P^n : P \in \mathcal{P}\}$ oder zum **unendlichen Bernoullischen Versuchsschema** $(M, \mathcal{M}, \mathcal{P})^{\mathbf{N}} = (M^{\mathbf{N}}, \mathcal{M}^{\mathbf{N}}, \mathcal{P}^{\mathbf{N}})$. Das zugehörige Experiment können wir interpretieren als Realisierung einer Folge $X_1, \ldots, X_n$ bzw. $X_1, X_2, X_3, \ldots$ von unabhängigen P_p-verteilten Zufallsvariablen. $(X_1, \ldots, X_n)$ wird als **n-fache Stichprobe** aus $(M, \mathcal{M}, \mathcal{P})$ oder auch als **Stichprobe vom Umfang n** bezeichnet. Man kann es aber auch als einfache Stichprobe aus $(M^n, \mathcal{M}^n, \mathcal{P}^n)$ interpretieren.

Definition 1.1.2

1. *Sei $(M, \mathcal{M}, \mathcal{P})$ ein statistischer Raum und $(D, \mathcal{D})$ ein meßbarer Raum. Eine meßbare Abbildung $S : (M, \mathcal{M}) \to (D, \mathcal{D})$ heißt* **Statistik**. *(In den meisten Fällen ist $(D, \mathcal{D}) = (\mathbf{R}, \mathcal{B})$ oder $= (\mathbf{R}^n, \mathcal{B}^n)$.)*

2. *Sei $\mathcal{P}_S = \{P \circ S^{-1} : P \in \mathcal{P}\}$. $(D, \mathcal{D}, \mathcal{P}_S)$ heißt der* **von S induzierte statistische Raum**.

Für das Bernoullische Versuchsschema (endlich oder unendlich) ist z.B. die relative Häufigkeit $H_n(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^n x_i$ eine $\mathbf{R}$-wertige Statistik. Der von ihr induzierte statistische Raum ist die Familie der Binomialverteilungen auf der Menge $\{\frac{k}{n} : k = 0. \ldots, n\}$.

1.2 Punktschätzung der unbekannten Wahrscheinlichkeit p

Im folgenden bezeichnen E_p bzw. V_p den Erwartungswert bzw. die Varianz unter P_p^n oder P_p^N, $X_{(n)} := (X_1, \ldots, X_n)$.

Wegen des Gesetzes der großen Zahl ist es intuitiv klar, daß $H_n(X_{(n)})$ ein geeigneter Schätzer für die unbekannte Wahrscheinlichkeit$p \in [0,1]$ sein sollte. H_n hat die beiden folgenden Eigenschaften:

- H_n ist **erwartungstreu**, d.h. $E_p H_n = p$ für alle $p \in [0,1]$.
- H_n ist **konsistent**, genauer: die Folge $(H_n)_{n\in N}$ ist konsistent, d.h. $H_n \to p$ P_p-stochastisch für $n \to \infty$. H_n ist sogar **stark konsistent**, da auch fast sichere Konvergenz vorliegt.
- H_n ist **gleichmäßig konsistent** bezüglich $p \in [0,1]$, d.h.

$$\lim_{n\to\infty} \sup_{p\in[0,1]} P_p^n\{|H_n - p| > \epsilon\} = 0 \quad \text{für alle } \epsilon > 0,$$

denn $P_p^n\{|H_n - p| > \epsilon\} \le \epsilon^{-2} V_p H_n = \epsilon^{-2}\frac{p(1-p)}{n} \le \epsilon^{-2}(4n)^{-1}$.

Einen ersten Eindruck von der Genauigkeit der Schätzung H_n für p erhalten wir aus $V_p H_n = \frac{p(1-p)}{n} \le \frac{1}{4n}$. Für p nahe bei 0.5 liegt also $V_p H_n$ nahe bei $\frac{1}{4n}$, für p nahe bei 0 oder 1 ist die Abschätzung aber zu grob, und man versucht, $n \cdot V_p H_n$ ebenfalls aus den Daten zu schätzen. Eine naheliegende Idee ist es, in der Formel für $n \cdot V_p H_n$ jedes p durch seinen Schätzer H_n zu ersetzen, wodurch man den Schätzer $S_n := H_n(1 - H_n)$ für $n \cdot V_p H_n = p(1-p)$ erhält. Mit H_n ist auch S_n konsistent. Da

$$\begin{aligned} E_p S_n &= E_p\left[\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n X_i(1-X_j)\right] \\ &= \frac{1}{n^2}\left(\sum_{i=1}^n E_p[X_i(1-X_i)] + \sum_{i\neq j} E_p[X_i(1-X_j)]\right) \\ &= \frac{(n-1)}{n}\,p(1-p), \end{aligned}$$

ist S_n nicht erwartungstreu, aber $\frac{n}{n-1}S_n$ ist ein konsistenter, erwartungstreuer Schätzer für $p(1-p)$.

Da

$$\sqrt{n}(H_n - p) \Longrightarrow \mathcal{N}(0, p(1-p)) \quad \text{(ZGS)},$$

geht

$$\sqrt{\frac{n}{S_n}}(H_n - p) \Longrightarrow \mathcal{N}(0,1),$$

d.h. für hinreichend große n ist

$$P_p^n\{|H_n - p| > \frac{\epsilon}{\sqrt{n}}\} \sim 2\Phi(\frac{-\epsilon}{\sqrt{S_n}}),$$

wo Φ die Verteilungsfunktion der Standardnormalverteilung bezeichnet. Dabei kann n als hinreichend groß angesehen werden, falls $\min\{np, n(1-p)\} \geq 5$. (Für genauere Approximationen der Binomialverteilung durch die Normalverteilung bei kleinem n siehe [6, S.86].) Für p nahe bei 0 oder 1 und moderate n ist diese Approximation also nicht brauchbar.

Darüberhinaus ist folgender Effekt zu beobachten: Bei festen n und k sieht $f_{k,n}(p) := P_p^n\{H_n = \frac{k}{n}\}$ wie in Abb.1.1 aus (dort für $k = 5$, $n = 25$). Man sieht, daß $f_{k,n}(p)$ ein

Abbildung 1.1:

Maximum bei $p = p_0 = \frac{k}{n}$ hat, aber nach $p = 0.5$ hin langsamer abfällt als zum Rand hin,so daß !

$$P_{p_0+\epsilon}^n\{H_n = \frac{k}{n}\} > P_{p_0-\epsilon}^n\{H_n = \frac{k}{n}\}, \quad \text{falls } p_0 < 0.5 \text{ und } \epsilon \text{ klein ist}$$

und umgekehrt für $p_0 > 0.5$. Diese Asymmetrie wird in den bisherigen Betrachtungen nicht berücksichtigt.

1.3 Konfidenzintervalle

Bisher haben wir einen einzigen Zahlenwert als Schätzer für den unbekannten Parameter p angegeben und versucht, dessen Genauigkeit durch S_n abzuschätzen. Nun wollen wir zu einem vorgegebenen $0 < \alpha < 1$ in Abhängigkeit von der beobachteten Häufigkeit H_n ein Intervall $J(H_n) = [p_u(H_n), p_o(H_n)]$ angeben, das folgende Eigenschaft hat:

- Für alle $p \in [0,1]$ ist $P_p\{p \in J(H_n)\} \geq 1 - \alpha$, d.h. $P_p\{p \notin J(H_n)\} \leq \alpha$.

Ein solches Intervall $J(H_n)$ heißt **Konfidenzintervall für p zum Konfidenzniveau** $1-\alpha$. Typische Werte sind $\alpha = 0.01$, $\alpha = 0.02$ oder $\alpha = 0.05$, bei besonderen Anforderungen an die Zuverlässigkeit der Schlußfolgerungen nimmt man aber auch z.B. $\alpha = 0.001$.

Wie oben bereits gesehen, gilt für $z > 0$ und hinreichend große n näherungsweise

$$P_p^n \left\{ |H_n - p| \leq z\sqrt{\frac{p(1-p)}{n}} \right\} = \Phi(z) - \Phi(-z) = 2\Phi(z) - 1.$$

Wählt man $z = z(\alpha)$ so, daß $\Phi(z) = 1 - \frac{\alpha}{2}$, so ist $2\Phi(z) - 1 = 1 - \alpha$, und $J(H_n)$ muß so gewählt werden, daß

$$|H_n - p| \leq z(\alpha)\sqrt{\frac{p(1-p)}{n}} \quad \text{für alle } p \in J(H_n).$$

(Ist $q = 1 - \frac{\alpha}{2}$, so ist $\Phi(z) = q$, und man bezeichnet $z = \lambda_q$ als das **Quantil der standardisierten Normalverteilung der Ordnung q**). Das führt zu

$$p_{o,u}(H_n) = \frac{n}{n + z(\alpha)^2} \left(H_n + \frac{z(\alpha)^2}{2n} \pm z(\alpha)\sqrt{\frac{H_n(1-H_n)}{n} + \frac{z(\alpha)^2}{4n^2}} \right),$$

so daß

$$p_{o,u}(H_n) = H_n \pm \frac{z(\alpha)\sqrt{S_n}}{\sqrt{n}} + \frac{z(\alpha)^2}{n}(0.5 - H_n) + O\left(\frac{1}{n^{3/2}}\right),$$

wobei die Konstante in O vom Wert von H_n abhängt. Die Intervallgrenzen sind also nicht symmetrisch bzgl. der Punktschätzung H_n. Da aus dieser Wahl von $p_{o,u}(H_n)$ folgt,daß

$$\lim_{n\to\infty} P_p^n\{p \in J(H_n)\} = 1 - \alpha,$$

bezeichnet man $J(H_n)$ auch als **asymptotisches Konfidenzintervall für p zum Konfidenzniveau** $1-\alpha$. Ein konkreter Zahlenwert: Für $\alpha = 0.02$ ist $z(\alpha) = 2.326$.

1.4 Prüfen von Hypothesen über die unbekannte Wahrscheinlichkeit p

In vielen Fällen stellt sich nicht das Problem, den Parameter p möglichst genau zu schätzen, sondern aufgrund der beobachteten Daten eine Hypothese über p, z.B. $p \leq p_0$, zu prüfen. Ein solches Problem tritt z.B. in der statistischen Qualitätskontrolle auf: Die Angabe eines Lieferanten, daß ein gewisser Lieferposten einen Ausschußsatz von höchstens 5% hat, soll überprüft werden. Als statistisches Modell dient das Bernoullische Versuchsschema, wobei $X_i = 1$ als „Teil i ist Ausschuß“ interpretiert wird. Die zufällig herausgegriffene Stichprobe werde mit $(X_1, \ldots, X_n)$ bezeichnet. Wir wollen hier annehmen (auch wenn das nicht ganz korrekt ist, siehe den nächsten Abschnitt), daß

die X_i u.i.v. sind. Geprüft wird die **Hypothese** $H_0 : p \leq 0.05$ (auch **Nullhypothese** genannt) gegen die **Alternative** $H_1 : p > 0.05$. Gesucht ist ein **Test** ϕ, d.h. eine Abbildung

$$\phi : \{0,1\}^n \to \{0,1\}$$

derart, daß bei der Interpretation

- $\phi(X_1, \ldots, X_n) = 0$ heißt Annahme von $H_o : p \leq 0.05$,
- $\phi(X_1, \ldots, X_n) = 1$ heißt Ablehnung von $H_o : p \leq 0.05$,

$P_p^n\{\phi = 1\}$ klein ist, wenn H_0 vorliegt, und auch $P_p^n\{\phi = 0\}$ möglichst klein ist, wenn H_1 vorliegt.

Sei $K_n = nH_n = \sum_{i=1}^n X_i$. Ein naheliegender Test für unser Problem ist

$$\phi(X_1, \ldots, X_n) = \begin{cases} 1 & \text{für } K_n \geq k_0 \\ 0 & \text{für } K_n < k_0 \end{cases},$$

wobei k_0 noch festzulegen ist. ϕ basiert also auf der Statistik K_n. Man bezeichnet $\{\phi = 1\}$ als den **kritischen Bereich** oder **Ablehnungsbereich** des Tests ϕ.

Die **Machtfunktion** von ϕ ist

$$m_\phi(p) = P_p^n\{\phi(X_1, \ldots, X_n) = 1\} = P_p^n\{K_n \geq k_0\}.$$

Da $m_\phi(p)$ die Wahrscheinlichkeit ist, H_0 abzulehnen, wird man bestrebt sein, $m_\phi(p)$ für $p \in H_0$ möglichst klein zu halten, damit ein solcher **Fehler 1. Art** unwahrscheinlich wird. Also wird man k_0 zu vorgegebenem $0 < \alpha < 1$ so wählen, daß

$$m_\phi(p) \leq \alpha \quad \text{für alle } p \in H_0. \tag{1.1}$$

Das ist sicher erfüllt, falls $k_0 = n + 1$, d.h. $\phi \equiv 0$, aber in diesem Fall würde ein **Fehler 2. Art**, nämlich H_0 zu akzeptieren obwohl $p \in H_1$, die Wahrscheinlichkeit $1 - m_\phi(p) = P_p^n\{\phi = 0\} = P_p^n\{K_n < n + 1\} = 1$ haben. Man wird also k_0 so wählen, daß der Fehler 2. Art unter der Nebenbedingung 1.1 minimiert wird.

Da $(p, k) \mapsto P_p^n\{K_n \geq k\}$ in p wächst und in k fällt, wählt man

$$k_0 = \min\left\{k : P_{p_0}^n\{K_n \geq k\} \leq \alpha\right\}.$$

Da die Variable k diskret ist, kann man Gleichheit i.a. nicht erreichen. In dem Beispiel der Gütekontrolle ergibt sich für $p_0 = 0.05$, $n = 25$ und $\alpha = 0.01$ der Wert $k_0 = 5$, d.h. ein Lieferposten wird zurückgewiesen, wenn in einer zufälligen Stichprobe vom Umfang 25 mehr als 4 Ausschußteile gefunden werden. Die maximale Wahrscheinlichkeit für einen Fehler 1. Art beträgt dabei $m_\phi(0.05) = 0.0072$. Die Wahrscheinlichkeit eines Fehlers 2. Art kann dagegen beliebig nahe bei $1 - 0.0072 = 0.9928$ liegen, wenn p sehr nah bei p_0 liegt. Da beide Fehler nicht gleichzeitig klein zu halten sind, ist diese asymmetrische Behandlung von H_0 und H_1 unumgänglich.

Abbildung 1.2:

Einen Test ϕ von H_0 gegen H_1, der (1.1) erfüllt, bezeichnet man als einen **Test zum Signifikanzniveau** α, kurz als α**-Test**.

Abb.1.2 zeigt die Machtfunktion in unserem Beispiel. Da die Teststatistik K_n asymptotisch normalverteilt ist ($n^{-1/2}(K_n - np) \Longrightarrow \mathcal{N}(0, p(1-p))$), kann die Machtfunktion für hinreichend große n annähernd mit der Normalverteilung ausgewertet werden:

$$m_\phi(p) = P_p^n\{K_n \geq k_0\} \sim 1 - \Phi\left(\frac{k_o - np}{\sqrt{np(1-p)}}\right).$$

Damit ergibt sich der kritische Wert k_0 aus $m_\phi(p) \leq \alpha$ zu

$$k_0 = \left[np_0 + \lambda_{1-\alpha}\sqrt{np_0(1-p_0)}\right] + 1.$$

1.5 Stichproben aus endlichen Grundgesamtheiten

Im Qualitätskontrollbeispiel des letzten Abschnitts wurde aus einer Warenlieferung vom Umfang N, die $A = Np$ Ausschußteile enthält, eine zufällige Stichprobe $(X_1, \ldots, X_n)$ herausgegriffen (ohne zurücklegen!). In diesem Modell sind die X_i nicht u.i.v., denn es ist zwar $P(X_1 = 1) = p$, aber $P(X_2 = 1 | X_1 = 1) = \frac{A-1}{N-1} \neq p$. $K_n = \sum_{i=1}^n X_i$ besitzt in der Tat eine **hypergeometrische** Verteilung $H(N, n, p)$:

$$P(K_n = k) = \frac{\binom{Np}{k}\binom{N-Np}{n-k}}{\binom{N}{n}}$$

für $\max\{0,\, n + Np - N\} \leq k \leq \min\{n,\, Np\}$ und 0 sonst.

Man beachte, daß $P(K_n = k) \to \binom{n}{k} p^k (1-p)^{n-k}$ mit $N \to \infty$.

Kapitel 2

Normalverteilte Beobachtungen

2.1 Aus der Normalverteilung hergeleitete Verteilungen

Sei $\mathbf{A}$ eine symmetrische, positiv definite $d \times d$-Matrix mit reellen Koeffizienten und Determinante ungleich 0. Dann gibt es eine Diagonalmatrix $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ und eine Orthogonalmatrix $\mathbf{B}$ derart, daß $\mathbf{A} = \mathbf{B^T \Lambda B}$ und $\lambda_1, \ldots \lambda_d > 0$ die Eigenwerte von $\mathbf{A}$ sind.

Definiere $f_{\mathbf{A}} : \mathbf{R}^d \to \mathbf{R}$ durch

$$f_{\mathbf{A}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d \cdot |\det(\mathbf{A})|^{-1}}} \exp(-\frac{1}{2}\,\mathbf{x^T A x}),$$

wo $\mathbf{x^T}$ den zu $\mathbf{x}$ transponierten Vektor bezeichnet. Dann ist $f_{\mathbf{A}} \geq 0$, und aus dem Integraltransformationssatz folgt

$$\begin{aligned} \int_{R^d} f_{\mathbf{A}}(\mathbf{x}) d\mathbf{x} &= \int_{R^d} f_{\mathbf{\Lambda}}(\mathbf{Bx}) d\mathbf{x} = \int_{R^d} f_{\mathbf{\Lambda}}(\mathbf{x}) d\mathbf{x} \\ &= \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi\lambda_i}} \int_{-\infty}^{+\infty} \exp(-\frac{x_i^2}{2\lambda_i})\, dx_i = 1, \end{aligned}$$

d.h. $f_{\mathbf{A}}$ ist eine Wahrscheinlichkeitsdichte auf $\mathbf{R}^d$.

Definition 2.1.1 *$f_{\mathbf{A}}$ heißt eine d-dimensionale Gaußsche Dichte oder Dichte einer d-dimensionalen Normalverteilung.*

Theorem 2.1.2 *Sei $\mathbf{X} = (X_1, \ldots, X_d)$ eine $\mathbf{R}^d$-wertige Zufallsvariable, deren Verteilung Dichte $f_{\mathbf{A}}$ hat. Bezeichne $\mathbf{V} = (v_{i,j}) = \mathbf{A}^{-1}$. Dann gilt:*

1. *X_i ist $\mathcal{N}(0, v_{i,i})$-verteilt für $i = 1, \ldots, d$.*
2. *$cov\,(X_i, X_j) = v_{i,j}$ für alle $i, j = 1, \ldots, d$.*
3. *Die X_i sind unabhängig genau dann, wenn $\mathbf{A}$ (und damit $\mathbf{V}$) Diagonalgestalt hat.*

Aufgrund dieser Eigenschaften sagt man, daß $\mathbf{X}$ *d-dimensional normalverteilt mit Erwartung* $\mathbf{0}$ *und Kovarianzmatrix* $\mathbf{V}$ *ist, symbolisch:* $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{V})$. *Ist* $\mu \in \mathbf{R}^n$, *so hat* $\mathbf{X} + \mu$ *eine* $\mathcal{N}(\mu, \mathbf{V})$*-Verteilung.*

Beweis: Übung

Sind insbesondere $X_1, \ldots, X_d$ unabhängig mit $X_i \sim \mathcal{N}(0, \sigma_i^2)$, so ist $(X_1, \ldots, X_d) \sim \mathcal{N}(\mathbf{0}, \mathbf{diag}(\sigma_\mathbf{1}^\mathbf{2}, \ldots, \sigma_\mathbf{d}^\mathbf{2}))$.

Theorem 2.1.3 *Ist* $\mathbf{X}$ *eine* $\mathbf{R}^n$*-wertige, nach* $\mathcal{N}(\mathbf{0}, \mathbf{V})$ *verteilte Zufallsvariable und ist* $\mathbf{D}$ *eine reelle* $m \times n$*-Matrix vom Rang* $m \leq n$, *so ist* $\mathbf{DX}$ *eine* $\mathbf{R}^m$*-wertige, nach* $\mathcal{N}(\mathbf{0}, \mathbf{DVD^T})$ *verteilte Zufallsvariable.*

Der Beweis wird mit charakteristischen Funktionen für $\mathbf{R}^d$-wertige Zufallsvariablen geführt, siehe z.B. [2, Section 29].

Theorem 2.1.4 (Mehrdimensionaler ZGS) *Seien* $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots$ *u.i.v.* $\mathbf{R}^d$*-wertige Zufallsvariablen mit Erwartungswert* μ *und endlicher Kovarianzmatrix* $\mathbf{V} = (v_{i,j})$, $v_{i,j} = \mathrm{cov}(\mathbf{X}_i^{(1)}, \mathbf{X}_j^{(1)})$. *Dann geht*

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} (\mathbf{X}^{(k)} - \mu) \Longrightarrow \mathcal{N}(\mathbf{0}, \mathbf{V}).$$

Beweis: Wir benutzen das **Cramér-Wold Kriterium** [2, Theorem 29.4]:

Für $\mathbf{R}^d$-wertige Zufallsvariablen $\mathbf{Y}^{(n)}$ und $\mathbf{Y}$ gilt:

$\mathbf{Y}^{(n)} \Rightarrow \mathbf{Y}$ genau dann, wenn $\sum_{i=1}^d t_i Y_i^{(n)} \Rightarrow \sum_{i=1}^d t_i Y_i$ für alle $\mathbf{t} \in \mathbf{R}^d$.

Sei also $\mathbf{t} = (t_1, \ldots, t_d)^T$ ein Spaltenvektor. Die $\mathbf{t}^T \mathbf{X}^{(k)}$ sind u.i.v. $\mathbf{R}$-wertige Zufallsvariablen mit

$$E[\mathbf{t}^T \mathbf{X}] = \mathbf{t}^T \mu$$

und

$$V(\mathbf{t}^T \mathbf{X}) = \sum_{i=1}^d \sum_{j=1}^d t_i t_j E[X_i X_j] - (\sum_{i=1}^d t_i E X_i)^2 = \sum_{i=1}^d \sum_{j=1}^d t_i t_j \mathrm{cov}(X_i, X_j) = \mathbf{t}^T \mathbf{V} \mathbf{t},$$

so daß

$$\mathbf{t}^T \left(\frac{1}{\sqrt{n}} \sum_{k=1}^{n} (\mathbf{X}^{(k)} - \mu) \right) \Longrightarrow \mathcal{N}(0, \mathbf{t}^T \mathbf{V} \mathbf{t}).$$

$\mathcal{N}(0, \mathbf{t}^T \mathbf{V} \mathbf{t})$ ist aber die Verteilung von $\mathbf{t}^T \mathbf{Y}$, wenn $\mathbf{Y}$ eine $\mathcal{N}(\mathbf{0}, \mathbf{V})$-verteilte Zufallsvariable ist (Satz 2.1.3), so daß der Satz aus dem Cramér-Wold Kriterium folgt. □

Eine direkte Folgerung aus dem Cramér-Wold Kriterium ist

Lemma 2.1.5 *Seien $\mathbf{Y}_n, \mathbf{Y}$ $\mathbf{R}^d$-wertige Zufallsvariablen mit $\mathbf{Y}_n \Longrightarrow \mathbf{Y}$. $(\mathbf{A}_n)_{n>0}$ sei eine Folge $\mathbf{R}^{d\times d}$-wertiger, d.h. Matrix-wertiger Zufallsvariablen, die gegen eine nichtzufällige $d \times d$-Matrix $\mathbf{A}$ konvergiert. Dann gilt $\mathbf{A}_n\mathbf{Y_n} \Longrightarrow \mathbf{AY}$.*

Beweis: Sei $\mathbf{t} \in \mathbf{R}^d$. Dann ist $\mathbf{t}^T(\mathbf{A}_n\mathbf{Y}_n) = (\mathbf{t}^T\mathbf{A})\mathbf{Y}_n + (\mathbf{t}^T(\mathbf{A}_n - \mathbf{A}))\mathbf{Y}_n$. Aus dem C-W Kriterium folgt $(\mathbf{t}^T\mathbf{A})\mathbf{Y}_n \Longrightarrow (\mathbf{t}^T\mathbf{A})\mathbf{Y} = \mathbf{t}^T(\mathbf{AY})$, und die Behauptung folgt aus

$$|(\mathbf{t}^T(\mathbf{A}_n - \mathbf{A}))\mathbf{Y}_n| \leq \|\mathbf{t}\|\,\|\mathbf{A}_n - \mathbf{A}\|\,\|\mathbf{Y_n}\| \to 0 \text{ in Wahrscheinlichkeit,}$$

denn $Y_n \Longrightarrow Y$ impliziert $\|Y - n\| \Longrightarrow \|Y\|$, so daß

$$\forall \epsilon > 0 \exists K > 0, n_0 \in \mathbf{N} \forall n \geq n_0 : P\{\|Y_n\| > K\} < \epsilon.$$

□

Definition 2.1.6 *1. Seien $X_1, \ldots, X_n$ unabhängige, $\mathcal{N}(\mu_i, 1)$ Zufallsvariablen. Die Verteilung von*

$$\sum_{i=1}^{n} X_i^2$$

heißt **χ^2-Verteilung mit n Freiheitsgraden und Nichtzentralitätsparameter** *$\delta = \sqrt{\sum_{i=1}^n \mu_i^2}$, kurz $\chi_n^2(\delta)$-Verteilung. Ist $\delta = 0$, d.h. $\mu_1 = \ldots = \mu_n = 0$, so heißt sie einfach χ^2-Verteilung mit n Freiheitsgraden, kurz χ_n^2-Verteilung.*

2. Seien X_1 und X_2 unabhängig, $X_1 \sim \mathcal{N}(\delta, 1)$, $X_2 \sim \chi_n^2$. Die Verteilung von

$$X_1 / \sqrt{\frac{X_2}{n}}$$

heißt **t-Verteilung mit n Freiheitsgraden und Nichtzentralitätsparameter** *δ, kurz $t_n(\delta)$-Verteilung. Ist $\delta = 0$, so heißt sie einfach t-Verteilung mit n Freiheitsgraden, kurz t_n-Verteilung.*

3. Seien X_1 und X_2 unabhängig, $X_1 \sim \chi_{n_1}^2(\delta)$, $X_2 \sim \chi_{n_2}^2$. Die Verteilung von

$$\frac{X_1}{n_1} / \frac{X_2}{n_2}$$

heißt **F-Verteilung mit (n_1, n_2) Freiheitsgraden und Nichtzentralitätsparameter** *δ, kurz $F_{n_1,n_2}(\delta)$-Verteilung. Ist $\delta = 0$, so heißt sie einfach F-Verteilung mit (n_1, n_2) Freiheitsgraden, kurz F_{n_1,n_2}-Verteilung.*

Theorem 2.1.7 *1. Die χ_n^2-Verteilung besitzt Momente jeder Ordnung. Ihr k-tes Moment lautet*

$$m_k = 2^k \frac{\Gamma(k + \frac{n}{2})}{\Gamma(\frac{n}{2})} = n \cdot (n+2) \cdot \ldots \cdot (n + 2k - 2).$$

Insbesondere ist die Erwartung $= n$ und die Varianz $= 2n$.

2. *Die t_n-Verteilung ist symmetrisch und besitzt Momente bis zur Ordnung $k = n-1$. Ihr k-tes Moment lautet*

$$m_k = \begin{cases} 0 & \text{für ungerades } k \\ n^{k/2} \frac{1 \cdot 3 \cdot \ldots \cdot (k-1)}{(n-2)(n-4)\ldots(n-k)} & \text{für gerades } k \end{cases}.$$

Insbesondere ist die Erwartung $= 0$ und die Varianz $= \frac{n}{n-2}$.

3. *Die F_{n_1,n_2}-Verteilung besitzt Momente der Ordnungen $k < \frac{n_1}{2}$. Ihr k-tes Moment lautet*

$$m_k = \left(\frac{n_2}{n_1}\right)^k \frac{\Gamma(\frac{n_1}{2}+k)\Gamma(\frac{n_2}{2}-k)}{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})} = \left(\frac{n_2}{n_1}\right)^k \frac{n_1 \cdot (n_1+2) \cdot \ldots \cdot (n_1+2k-2)}{(n_2-2k) \cdot (n_2-2k+2) \cdot \ldots \cdot (n_2-2)}.$$

Insbesondere ist der Erwartungswert $= \frac{n_2}{n_2-2}$.

Beweis: Übung.(Siehe z.B. [11].)

Seien nun $X_1, \ldots, X_n$ u.i.v. Wir bezeichnen das **empirische Mittel** der $X_1, \ldots, X_n$ mit

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

und ihre **empirische Varianz** mit

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

Beachte, daß $E\bar{X} = EX_1$ und

$$ES^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} EX_i^2 - \frac{2}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} E[X_i X_j] + \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} E[X_i X_j] \right) = V(X_1).$$

Sind die X_i normalverteilt, so gilt

Theorem 2.1.8 *Seien $X_1, \ldots, X_n$ u.i.v. nach $\mathcal{N}(\mu, \sigma^2)$. Dann sind $\bar{X}$ und S^2 unabhängig, $\bar{X}$ ist $\mathcal{N}(\mu, \frac{\sigma^2}{n})$-verteilt, und $\frac{n-1}{\sigma^2} S^2$ ist χ^2_{n-1}-verteilt, so daß $\sqrt{\frac{n}{S^2}} \bar{X}$ nach $t_{n-1}(\sqrt{\frac{n}{\sigma^2}}\mu)$ verteilt ist.*

Beweis:
Sei $\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(n)}$ das folgende Orthonormalsystem in $\mathbf{R}^n$:

$$\begin{aligned} \mathbf{a}^{(1)} &= \left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}} \right) \\ \mathbf{a}^{(k)} &= \left(-\frac{1}{\sqrt{k(k-1)}}, \ldots, -\frac{1}{\sqrt{k(k-1)}}, \sqrt{\frac{k-1}{k}}, 0, \ldots, 0 \right) \end{aligned}$$

für $k = 2, \ldots, n$, wobei $\mathbf{a}_{k+1}^{(k)}, \ldots, \mathbf{a}_n^{(k)} = 0$. Bezeichne $\mathbf{A}$ die Matrix mit den Zeilen $\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(n)}$ und $\mathbf{X} = (X_1, \ldots, X_n)^T$. Dann ist $\mathbf{Y} := \mathbf{AX} = \mathbf{A}(\mathbf{X} - \mu\mathbf{1}) + \mu\mathbf{A1}$ nach $\mathcal{N}(\mu\mathbf{A1}, \mathbf{A}\sigma^2\mathbf{I}\mathbf{A^T}) = \mathcal{N}(\mu\mathbf{A1}, \sigma^2\mathbf{I})$ verteilt (2.1.3), so daß die Y_i unabhängig normalverteilt sind. Da $Y_1 = (\mathbf{a}^{(1)})\mathbf{X} = \sqrt{n}\,\bar{X}$ und

$$\sum_{i=2}^{n} Y_i^2 = \mathbf{Y}^T\mathbf{Y} - Y_1^2 = \mathbf{X}^T\mathbf{X} - n\bar{X}^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 = (n-1)S^2,$$

sind $\bar{X}$ und S^2 unabhängig.

Aus $\mathbf{A1} = (\sqrt{n}, 0, \ldots, 0)^T$ folgt dann $\bar{X} = \frac{1}{\sqrt{n}}Y_1 \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$, und zusammen mit $\frac{n-1}{\sigma^2}S^2 = \sum_{i=1}^{n-1}(\frac{Y_i}{\sigma})^2$ auch $\frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}$. Schließlich ist

$$\sqrt{\frac{n}{S^2}}\bar{X} = \left(\sqrt{\frac{n}{\sigma^2}}\bar{X}\right) \Big/ \sqrt{\frac{\frac{n-1}{\sigma^2}S^2}{n-1}}$$

nach Definition $t_n(\sqrt{\frac{n}{\sigma^2}}\mu)$-verteilt. □

Ein analoges Ergebnis für d-dimensional normalverteilte ZV'en gilt ebenfalls [11, Satz 4.1].

Eine Konsequenz dieses Satzes und von Satz 2.1.7 ist, daß

$$E\bar{X} = \mu,\ V(\bar{X}) = \frac{\sigma^2}{n} \quad \text{und} \quad ES^2 = \sigma^2,\ V(S^2) = \frac{2\sigma^4}{n-1},$$

so daß $\bar{X}$ und S^2 erwartungstreue und konsistente Schätzer für μ bzw. σ^2 sind.

2.2 Tests bei Normalverteilung

Wir betrachten den statistischen Raum $(\mathbf{R}, \mathcal{B}, \mathcal{P})^n$, wo $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbf{R}, \sigma^2 \geq 0\}$. Dieser Raum kann z.B. n Messungen an Werkstücken mit Mittelwert μ und Streuung σ^2 repräsentieren. Unser Problem bestehe darin, die Hypothese $H_0 : \mu = \mu_0$ zu testen. (Wenn nichts weiteres festgelegt wird, heißt die Alternative $H_1 : \mu \neq \mu_0$.)

Eine denkbare Teststatistik für dieses Problem ist $\frac{\bar{X}-\mu_0}{\sigma}\sqrt{n}$. Da σ^2 aber ebenfalls unbekannt ist, muß man es durch die empirische Varianz S^2 ersetzen und kommt so zu der Teststatistik

$$T = \frac{\bar{X} - \mu_0}{S}\sqrt{n} = \frac{\bar{X} - \mu_0}{\sigma}\sqrt{n}\left(\sqrt{\frac{S^2}{\sigma^2}}\right)^{-1},$$

die unter $\mathcal{N}(\mu, \sigma^2)$ nach $t_{n-1}(\delta)$ verteilt ist mit $\delta = \frac{\sqrt{n}}{\sigma}(\mu - \mu_0)$. T induziert den statistischen Raum $(\mathbf{R}, \mathcal{B}, \mathcal{P}_T)$, wo $\mathcal{P}_T$ die Familie der nichtzentralen t-Verteilungen mit $n-1$ Freiheitsgraden ist. Ein auf T basierender Test ϕ ist dann eine meßbare Abbildung von $\mathbf{R}$ nach $\{0, 1\}$ wie in Abschnitt 1.4. Wir setzen $\phi(t) = I_K(t)$, wo $K = \{t \in \mathbf{R} : |t| \geq t_0\}$,

wobei noch t_0 in Abhängigkeit von einem angestrebten Signifikanzniveau α festgelegt werden muß.

Die Machtfunktion von ϕ ist

$$m_\phi(\delta) = P_{\delta,n}\{\phi(T) = 1\} = P_{\delta,n}\{|T| \geq t_0\},$$

wo $P_{\delta,n}$ die $t_{n-1}(\delta)$-Verteilung auf $(\mathbf{R}, \mathcal{B})$ bezeichne. Um das Signifikanzniveau, d.h. die maximale Wahrscheinlichkeit eines Fehlers 1. Art festzulegen, muß man nur die Verteilung von T unter H_0, d.h. t_{n-1} kennen, und da diese Verteilung stetig ist, kann man in

$$m_\phi(0) = P_{0,n}\{|T| \geq t_0\} \leq \alpha$$

durch Wahl von t_0 Gleichheit erzielen. Da die t_{n-1}-Verteilung symmetrisch ist, ist das äquivalent zu

$$P_{0,n}\{T > t_0\} = \frac{\alpha}{2}, \quad \text{d.h.} \quad P_{0,n}\{T \leq t_0\} = 1 - \frac{\alpha}{2}, \tag{2.1}$$

und t_0 ergibt sich als Quantil der Ordnung $1 - \frac{\alpha}{2}$ der t_{n-1}-Verteilung, kurz $t_0 = t_{n-1,1-\alpha/2}$.

Der oben beschriebene einfache t-Test dient also zum Prüfen des Erwartungswerts einer normalverteilten Grundgesamtheit mit unbekannter Varianz.

Soll die Hypothese $H_0 : \mu \leq \mu_0$ geprüft werden, so wird man den kritischen Bereich $K = \{t \in \mathbf{R} : t \geq t_0\}$ wählen, und gelangt wie in (2.1) zu $t_0 = t_{n-1,1-\alpha}$.

Zur Interpretation von $\phi = 0$ und $\phi = 1$: $\phi = 1$ heißt, daß $\mu \neq \mu_0$ mit Wahrscheinlichkeit $\geq 1 - \alpha$. $\phi = 0$ heißt aber **nicht**, daß $\mu = \mu_0$ mit großer Wahrscheinlichkeit, sondern nur, daß die Daten keine Evidenz gegen $\mu = \mu_0$ hergeben. Wenn es z.B. um die Frage geht, ob die Brote einer Großbäckerei, wie von der behauptet, mindestens $\mu_0 = 1000$g wiegen, so heißt $\phi = 1$, daß die Bäckerei mit Wahrscheinlichkeit mindestens $1 - \alpha$ systematisch zu kleine Brote backt, während $\phi = 0$ nur bedeutet, daß man die Bäckerei anhand der untersuchten Stichprobe zum Signifikanzniveau α nicht des Betrugs überführen kann.

Eine wichtige Variante des t-Tests ist der **doppelte t-Test zum Prüfen der Gleichheit der Erwartungswerte zweier normalverteilter Merkmale auf der Basis unabhängiger Stichproben**: Seien $(X_{i,1}, \ldots, X_{i,n_i})$ $(i = 1, 2)$ zwei unabhängige, nach $\mathcal{N}(\mu_i, \sigma_i^2)$ verteilte Stichproben. $\bar{X}_{i.}$ und S_i^2 mögen das empirische Mittel und die empirische Varianz der beiden Stichproben bezeichnen. Unter der zusätzlichen Voraussetzung $\sigma_1^2 = \sigma_2^2 = \sigma^2$ ergibt sich aus Satz 2.1.8, daß

$$\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{\sigma^2}$$

eine $\chi^2_{n_1+n_2-2}$-Verteilung hat und unabhängig von $(\bar{X}_{1.} - \bar{X}_{2.})$ ist (Übung). Die Statistik

$$\frac{\bar{X}_{1.} - \bar{X}_{2.}}{\sqrt{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}(n_1 + n_2 - 2)}$$

ist daher $t_{n_1+n_2-2}(\delta)$-verteilt mit $\delta = \sqrt{\frac{n_1 n_2}{n_1+n_2}}\frac{\mu_1-\mu_2}{\sigma}$ und zum Prüfen von $H_0 : \mu_1 = \mu_2$ geeignet. Die weitere Konstruktion des doppelten t-Tests erfolgt wie beim einfachen t-Test. Im Fall $\sigma_1^2 \neq \sigma_2^2$ ist das Problem wesentlich schwieriger (Behrens-Fischer Problem, siehe [1, Abschnitt 10.5].

Will man Hypothesen über die Streuung von normalverteilten Beobachtungen prüfen, z.B. ob in einem Fertigungsprozeß eine gewisse Fehlertoleranz nicht überschritten wird, geht man folgendermaßen vor: Ist $X_1, \ldots, X_n$ eine n-fache $\mathcal{N}(\mu, \sigma^2)$-verteilte Stichprobe mit unbekannten μ und σ^2 und will man $H_0 : \sigma^2 \leq \sigma_0^2$ testen, so eignet sich die unter ∂H_0 nach χ^2_{n-1} verteilte Statistik

$$T = \frac{n-1}{\sigma_0^2} S^2$$

als Teststatistik. Man wählt $\phi = I_K$ mit $K = [t_0, +\infty)$, und t_0 ergibt sich als Quantil $\chi^2_{n-1,1-\alpha}$ der Ordnung $1-\alpha$ der χ^2_{n-1}-Verteilung. Solche Tests heißen **χ^2-Streuungstests**.

Der doppelte t-Test gibt nur dann korrekte Resultate, wenn zumindest annähernd $\sigma_1^2 = \sigma_2^2$ gilt. (Er ist in der Praxis jedoch recht robust gegen Abweichungen von $\sigma_1^2 = \sigma_2^2$, vorausgesetzt, daß die Stichprobenumfänge nicht zu verschieden sind.) Einen Test von $H_0 : \sigma_1^2 = \sigma_2^2$ kann man in der Situation des doppelten t-Tests folgendermaßen konstruieren: Da die Statistiken $\frac{n_i-1}{\sigma_i^2} S_i^2$ für $i = 1, 2$ unabhängig und $\chi^2_{n_i-1}$-verteilt sind, besitzt S_1^2/S_2^2 unter H_0 eine F_{n_1-1,n_2-1}-Verteilung. Ein kritischer Bereich $K = (-\infty, t_u] \cup [t_o, +\infty)$ mit $t_u = F_{n_1-1,n_2-1,\alpha/2} < 1 < t_o = F_{n_1-1,n_2-1,1-\alpha/2}$ ergibt einen Test von H_0 zum Signifikanzniveau α, der Abweichungen nach oben und unten gleich gewichtet. ($F_{m,n,q}$ bezeichnet das Quantil der Ordnung q der $F_{m,n}$-Verteilung.)

2.3 Konfidenzintervalle

Ganz allgemein (nicht nur für normalverteilte Beobachtungen) besteht folgende Dualität zwischen Tests und Konfidenzbereichen: Den kritischen Bereich eines α-Tests kann man als Kompliment eines Konfidenzbereichs zum Konfidenzniveau $1-\alpha$ auffassen. Genauer:

Sei $\mathbf{X} = (X_1, \ldots, X_n)$ eine einfache Stichprobe aus dem statistischen Raum $(\mathbf{R}^n, \mathcal{B}^n, \mathcal{P})$, $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ eine beliebige Familie von Wahrscheinlichkeitsverteilungen auf $(\mathbf{R}^n, \mathcal{B}^n)$, und sei $\alpha > 0$. Zu jedem $\theta_0 \in \Theta$ sei ein α-Test ψ_{θ_0} für $H_0 : \theta = \theta_0$ konstruiert worden. Durch die Abbildung

$$\mathbf{x} \to J(\mathbf{x}) = \{\theta_0 \in \Theta : \psi_{\theta_0}(\mathbf{x}) = 0\}$$

wird jedem Beobachtungsvektor $\mathbf{x}$ eine Teilmenge $J(\mathbf{x})$ von Θ zugeordnet. Da

$$P_{\theta_0}\{\theta_0 \in J(\mathbf{X})\} = 1 - P_{\theta_0}\{\psi_{\theta_0}(\mathbf{X}) = 1\} \geq 1 - \alpha,$$

ist $J(\mathbf{x})$ ein Konfidenzbereich für θ_0 zum Konfidenzniveau $1-\alpha$.

Liegt umgekehrt durch $\mathbf{x} \to J(\mathbf{x})$ ein Konfidenzbereich für θ zum Konfidenzniveau $1-\alpha$ vor, so wird durch

$$\psi_{\theta_0}(\mathbf{x}) = I_{\Theta \setminus J(x)}(\theta_0)$$

ein α-Test für $H_0 : \theta = \theta_0$ definiert, da

$$P_{\theta_0}\{\psi_{\theta_0}(\mathbf{X}) = 1\} = P_{\theta_0}\{\theta_0 \notin J(\mathbf{X})\} \leq \alpha.$$

Ist $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbf{R}, \sigma^2 \geq 0\}$, so erhält man bei unbekanntem σ^2 Konfidenzintervalle für μ ausgehend vom t-Test für $H_0 : \mu = \mu_0$. Man wählt

$$\psi_{\mu_0}(\mathbf{X}) = I_{\left\{\left|\frac{\bar{X}-\mu_0}{S}\right|\sqrt{n} \geq t_0\right\}},$$

wo $t_0 = t_{n-1,1-\alpha/2}$, und erhält

$$\begin{aligned} J(\mathbf{X}) &= \{\mu_0 \in \mathbf{R} : \psi_{\mu_0}(\mathbf{X}) = 0\} \\ &= \left\{\mu_o \in \mathbf{R} : \left|\frac{\bar{X}-\mu_0}{S}\right|\sqrt{n} \leq t_0\right\} \\ &= \left(\bar{X} - t_0\frac{S}{\sqrt{n}}, \bar{X} + t_0\frac{S}{\sqrt{n}}\right) \end{aligned}$$

als Konfidenzintervall zum Konfidenzniveau $1-\alpha$.

2.4 Asymptotisch normalverteilte Statistiken

Sei $(\mathbf{R}, \mathcal{B}, \mathcal{P})$ ein statistischer Raum, $f : \mathbf{R} \to \mathbf{R}$ meßbar und quadratintegarierbar für alle $P \in \mathcal{P}$, und sei

$$\theta_f(P) := \int f(x)\, dP(x), \quad \sigma_f^2(P) := \int (f(x) - \theta_f(P))^2\, dP(x)$$

für alle $P \in \mathcal{P}$. Ist $X_1, \ldots, X_n$ eine n-fache Stichprobe aus $(\mathbf{R}, \mathcal{B}, \mathcal{P})$, so ist

$$\hat{f}_n := \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

ein erwartungstreuer, konsistenter Schätzer für $\theta_f(P)$, und aus dem ZGS folgt

$$\sqrt{\frac{n}{\sigma_f^2(P)}}(\hat{f}_n - \theta_f(P)) \Longrightarrow \mathcal{N}(0,1).$$

Da $\hat{S}_n^2 := \frac{1}{n-1}\sum_{i=1}^{n}(f(X_i) - \hat{f}_n)^2$ ein konsistenter Schätzer für $\sigma_f^2(P)$ ist, folgt (siehe Übung):

$$\sqrt{\frac{n}{\hat{S}_n^2}}(\hat{f}_n - \theta_f(P)) \Longrightarrow \mathcal{N}(0,1).$$

Für nicht so große n ist die Approximation von Verteilungen durch den ZGS oft nicht hinreichend exakt. In solchen Fällen kann man unter gewissen Voraussetzungen Korrekturterme zur Normalapproximation angeben, Stichwort: Edgeworth-Entwicklung.

Der folgende Satz erweist sich bei Untersuchungen zur asymptotischen Normalität oft als nützlich:

Theorem 2.4.1 *Sei $(Z_n)_{n>0}$ eine Folge $\mathbf{R}^d$-wertiger Zufallsvariablen mit folgenden Eigenschaften:*

1. *$Z_n \to z$ stochastisch für ein $z \in \mathbf{R}^d$.*
2. *$\sqrt{n}(Z_n - z) \Longrightarrow \mathcal{N}(\mathbf{0}, \mathbf{V})$.*

Ist dann $g : U \to \mathbf{R}^d$ stetig differenzierbar mit $\det(Dg(z)) \neq 0$, wo $U \subseteq \mathbf{R}^d$ eine Umgebung von z ist, so gilt:

1. *$g(Z_n) \to g(z)$ stochastisch und*
2. *$\sqrt{n}(g(Z_n) - g(z)) \Longrightarrow \mathcal{N}(\mathbf{0}, Dg(z)\mathbf{V}Dg(z)^T)$.*

(Das beinhaltet, daß $P\{Z_n \in U\} \to 1$.)

Beweis: Sei $t \in \mathbf{R}^d$. Nach dem Mittelwertsatz ist

$$t^T(g(Z_n) - g(z)) = t^T Dg(Z_n^*)(Z_n - z), \quad \text{falls } Z_n \in U, \tag{2.2}$$

für ein Z_n^*, das auf der Z_n und z verbindenden Strecke liegt. Daraus folgt $Z_n^* \to z$ stochastisch, so daß wegen der Stetigkeit von Dg

$$Dg(Z_n^*) \to Dg(z) \quad \text{und} \quad Dg(Z_n^*)(Dg(z))^{-1} \to \mathbf{1} \quad \text{stochastisch.}$$

Daraus und aus (2.2) folgt sofort $t^T g(Z_n) \to t^T g(z)$ stochastisch und außerdem mit Theorem 2.1.3 und Lemma 2.1.5

$$\begin{aligned} \sqrt{n} \cdot t^T(g(Z_n) - g(z)) &= \sqrt{n} \cdot \underbrace{t^T(Dg(Z_n^*)Dg(z)^{-1})}_{\to t^T} Dg(z)(Z_n - z) \\ &\Longrightarrow \mathcal{N}(0, t^T Dg(z)\mathbf{V}Dg(z)^T t) \sim t^T Y, \end{aligned}$$

wo Z eine $\mathcal{N}(\mathbf{0}, Dg(z)\mathbf{V}Dg(z)^T)$-verteilte Zufallsvariable ist. Daraus folgt die Behauptung mit Hilfe des Cramér-Wold Kriteriums. □

2.5 Ein einfaches lineares Regressionsmodell

Seien X und Y zwei (i.a. nicht unabhängige) Zufallsvariablen, z.B. die mittlere Januar- bzw. Februartemperatur eines Jahres in $^\circ C$ in Erlangen. Gesucht wird eine Borel-meßbare Funktion g, die aus Kenntnis des Wertes von X eine möglichst gute Vorhersage von Y liefert. Genauer: g soll so bestimt werden, daß $E[(Y - g(X))^2]$ minimiert wird. Die Funktion $g(X) = E[Y|X]$ leistet gerade das verlangte, denn für jede Borel-meßbare Funktion f gilt

$$\begin{aligned} E[(Y - E[Y|X] - f(X))^2|X] &= E[(Y - E[Y|X])^2|X] + f(X)^2 \\ &\geq E[(Y - E[X|Y])^2] \end{aligned} \tag{2.3}$$

mit Gleichheit genau dann, wenn $f = 0$ fast sicher. $g(X) = E[Y|X]$ heißt **Regressionsfunktion 1. Art**. Sie kann als Funktion von X eine sehr komplizierte Struktur haben, und daher schränkt man sich bei der suche nach Regressionsfunktionen häufig auf eine Teilklasse ein, innerhalb derer man versucht, $E[(Y - g(X))^2]$ zu minimieren. Man spricht dann von **Regression 2. Art**.

Eine wichtige solche Teilklasse ist die Klasse der **linearen Regressionsfunktionen**,

$$g(x) = \alpha + \beta x. \tag{2.4}$$

Die Minimierung von $E[(Y-g(X))^2]$ in dieser Klasse bedeutet, $E[(Y-\alpha-\beta X)^2]$ durch Wahl von α und β zu minimieren. Sei dazu $\mu_X = E[X]$, $\mu_Y = E[Y]$ und $\sigma_X^2 = V(X)$.

$$\begin{aligned} \frac{\partial}{\partial \alpha} E[(Y - \alpha - \beta X)^2] &= -2(\mu_Y - \alpha - \beta \mu_X) = 0 \\ \frac{\partial}{\partial \beta} E[(Y - \alpha - \beta X)^2] &= -2(\mathrm{cov}(X,Y) + \mu_X \mu_Y - \alpha \mu_X - \beta(\sigma_X^2 + \mu_X^2)) = 0 \end{aligned}$$

führt auf

$$\alpha = \mu_Y - \beta \mu_X \quad \text{und} \quad \beta = \frac{\mathrm{cov}(X,Y)}{\sigma_X^2}, \tag{2.5}$$

und da

$$\frac{\partial^2}{\partial(\alpha,\beta)^2} E[(Y - \alpha - \beta X)^2] = \begin{pmatrix} 2 & 2\mu_X \\ 2\mu_X & 2(\sigma_X^2 + \mu_X^2) \end{pmatrix}$$

positiv definit ist, erhalten wir

Lemma 2.5.1

$$g(X) = \mu_Y + \frac{\mathrm{cov}(X,Y)}{\sigma_X^2}(X - \mu_X) \tag{2.6}$$

minimiert den mittleren quadratischen Fehler unter allen linearen Regressionsfunktionen.

Für später halten wir fest:

Lemma 2.5.2 *Sei $g(X)$ gegeben durch (2.6).*

1. $E[Y - g(X)] = 0$.
2. *X und $(Y - g(X))$ sind unkorreliert, $g(X)$ und $(Y - g(X))$ sind unkorreliert.*
3. $$E[(Y - g(X))^2] = \sigma_Y^2 - \frac{\mathrm{cov}(X,Y)^2}{\sigma_X^2}.$$

Beweis:

1. Folgt sofort aus (2.6).
2. $$\begin{aligned} E[X(Y - g(X))] &= E[X(Y - \mu_Y)] - \frac{\mathrm{cov}(X,Y)}{\sigma_X^2} E[X(X - \mu_X)] \\ &= \mathrm{cov}(X,Y) - \mathrm{cov}(X,Y) = 0 \end{aligned}$$
3. $$\begin{aligned} & E[(Y - g(X))^2] \\ = \; & E[(Y - \mu_Y)^2] - 2E[(Y - \mu_Y)\frac{\mathrm{cov}(X,Y)}{\sigma_X^2}(X - \mu_X)] + E[(\frac{\mathrm{cov}(X,Y)}{\sigma_X^2}(X - \mu_X))^2] \\ = \; & \sigma_Y^2 - 2\frac{\mathrm{cov}(X,Y)^2}{\sigma_X^2} + \frac{\mathrm{cov}(X,Y)^2}{\sigma_X^2} \\ = \; & \sigma_Y^2 - \frac{\mathrm{cov}(X,Y)^2}{\sigma_X^2} \end{aligned}$$

□

Das folgende Lemma besagt, daß die gemäß (2.6) ermittelte lineare Regression im Falle normalverteilter Zufallsvariablen bereits eine Regression 1. Art ist:

Lemma 2.5.3 *Hat (X,Y) eine 2-dimensionale Normalverteilung, so ist*

$$E[Y|X] = \mu_Y + \frac{\mathrm{cov}(X,Y)}{\sigma_X^2}(X - \mu_X).$$

Beweis: Da

$$(X - \mu_X, Y - g(X)) = (X - \mu_X, Y - \mu_Y - \frac{\mathrm{cov}(X,Y)}{\sigma_X^2}(X - \mu_X)),$$

folgt aus Theorem 2.1.3, daß $(X - \mu_X, Y - g(X))$ normalverteilt ist, und daher aus Lemma 2.5.2 und Theorem 2.1.2, daß $X - \mu_X$ und damit auch X unabhängig von $Y - g(X)$ ist. Ist nun f eine beliebige Borel-meßbare Funktion, für die $E[f(X)^2] < \infty$ ist, so folgt:

$$\begin{aligned} & E[(Y - f(X))^2] \\ = \; & E[(Y - g(X) + g(X) - f(X))^2] \\ = \; & E[(Y - g(X))^2] + 2E[Y - g(X)]E[g(X) - f(X)] + E[(g(X) - f(X))^2] \\ \geq \; & E[(Y - g(X))^2], \end{aligned}$$

d.h. $g(X)$ ist eine Regressionsfunktion 1. Art und $g(X) = E[Y|X]$, siehe (2.3). □

Die Aufgabe des Statistikers bei der linearen Regression besteht nun darin, aus der Kenntnis von n Beobachtungspaaren $(X_1, Y_1), \ldots, (X_n, Y_n)$ die Adäquatheit der linearen Hypothese (2.4) zu beurteilen und, falls sie vernünftig erscheint, die Regressionskonstanten α und β gemäß (2.5) zu schätzen. Sind (X_1, Y_i), $i = 1, \ldots, 89$ z.B. die mittleren Januar- und Februartemperaturen der Jahre 1901 bis 1989 und wird aufgrund dieser Daten das lineare Modell (2.4) nicht verworfen, so kann Ende Januar 1990 die mittlere Februartemperatur Y_{90} mit Hilfe der geschätzten α und β aus der Kenntnis von X_{90} prognostiziert werden.

Ein heuristisch naheliegendes Verfahren, Schätzer $\hat{\alpha}$ und $\hat{\beta}$ für α und β zu konstruieren, besteht darin, die Größen μ_X, μ_Y, σ_X^2 und $\mathrm{cov}(X, Y)$ in (2.5) durch die entsprechenden Stichprobenmittel $\bar{X}$ und $\bar{Y}$ bzw. die Stichprobenvarianz S_X^2 bzw. die Stichprobenkovarianz

$$C_{X,Y} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

zu ersetzen. Das führt auf die Schätzer

$$\hat{\beta} = \frac{C_{X,Y}}{S_X^2} \quad \text{und} \quad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}. \tag{2.7}$$

Die Konsistenz dieser Schätzer folgt leicht aus der Konsistenz von $\bar{X}$, $\bar{Y}$, S_X^2 und $C_{X,Y}$. Sind die Beobachtungsvektoren normalverteilt, so können auch exakte Verteilungsaussagen für $\hat{\alpha}$ und $\hat{\beta}$ gemacht werden, auf die wir an dieser Stelle jedoch nicht eingehen. Auf die asymptotischen Verteilungseigenschaften kommen wir noch zurück.

Um festzustellen, ob überhaupt ein linearer Zusammenhang zwischen X und Y besteht, wählt man als Testgröße

$$\rho^2 := \frac{V(g(X))}{V(Y)}.$$

Da $Y - g(X)$ und $g(X)$ unkorreliert sind (Lemma 2.5.2), ist

$$V(Y) = V(Y - g(X) + g(X)) = V(Y - g(X)) + V(g(X)),$$

so daß $1 - \rho^2 = V(Y - g(X))/V(Y)$, d.h. $1 - \rho^2$ mißt die Variabilität in Y, die nicht durch einen linearen Zusammenhang mit der Zufallsvariablen X erklärt werden kann. Da die Anteile in der Zerlegung

$$1 - \rho^2 = \frac{V(Y - E[Y|X])}{V(Y)} + \frac{V(E[Y|X] - g(X))}{V(Y)}$$

unbekannt sind, bleibt aber offen, ob ein großer Wert von $1 - \rho^2$ daher rührt, daß Y von X nahezu unabhängig ist (großer 1. Summand), oder ob nur die Hypothese einer rein linearen Abhängigkeit nicht gerechtfertigt ist (großer 2. Summand). Im normalverteilten Fall, wo ja $g(X) = E[Y|X]$, mißt ρ^2 also die Variabilität in Y, die durch die lineare Abhängigkeit von der Zufallsvariablen X erklärt werden kann.

Weiter ist

$$\rho^2 = (\sigma_Y)^{-2} \cdot \frac{\mathrm{cov}(X,Y)^2}{\sigma_X^4} \cdot \sigma_X^2 = \frac{\mathrm{cov}(X,Y)^2}{\sigma_X^2 \sigma_Y^2},$$

d.h. ρ ist gerade der Korrelationskoefizient von X und Y. Daher kann ρ^2 aus den Daten konsistent geschätzt werden, nämlich durch

$$r^2 = \frac{C_{X,Y}^2}{S_X^2 S_Y^2}.$$

Um zu testen, ob überhaupt ein linearer Zusammenhang zwischen X und Y besteht, ist $H_0 : \rho^2 = 0$ gegen $H_1 : \rho^2 > 0$ zu testen. Durch eine längere Rechnung (siehe [11, S.79-81]) kann man zeigen, daß bei normalverteilten (X, Y) unter H_0, d.h. $\rho = 0$, gilt:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad \text{ist } t_{n-2}\text{-verteilt.}$$

Also läßt sich auf Basis der Statistik t ein Test von H_0 gegen H_1 konstruieren.

Ist $\rho \neq 0$, so läßt sich zeigen, daß bei Existenz vierter Momente für X und Y gilt $\sqrt{n}(r-\rho) \Longrightarrow \mathcal{N}(0, \gamma^2)$ für ein $\gamma > 0$. (Übung). Seien nun (X, Y) wieder normalverteilt. Dann kann man zeigen, daß $\gamma = 1 - \rho^2$, und aus Satz 2.4.1 folgt nun leicht

$$Z_n := \sqrt{n}\left(\frac{1}{2}\ln\frac{1+r}{1-r} - \frac{1}{2}\ln\frac{1+\rho}{1-\rho}\right) \Longrightarrow \mathcal{N}(0,1).$$

Die obige Situation kann durch den statistischen Raum$(\mathbf{R}^2, \mathcal{B}^2, \mathcal{P})^n$ beschrieben werden, wo $\mathcal{P}$ z.B. die Familie aller Verteilungen auf $\mathbf{R}^2$ oder, was in der Praxis oft angenommen wird, die Familie aller 2-dimensionalen Normalverteilungen ist.

Dieses Modell, bei dem sowohl die X_i als auch die Y_i Zufallsvariablen sind, wird auch als **Modell II der Regressionsanalyse** bezeichnet. Bei ihm besteht eine wechselseitige stochastische Abhängigkeit zwischen den X_i und den Y_i, denn durch Korrelationen

wird nie ein kausaler Zusammenhang ausgedrückt! Die unsymmetrische Behandlung von X und Y beruht nur auf dem Ziel, Y aus der Kenntnis von X vorherzusagen. Sie kann in der Praxis dadurch erzwungen werden, daß die X-Werte dem Experimentator regelmäßig früher bekannt werden als die Y-Werte (Beispiel Januar- und Februartemperaturen). Trotzdem bleibt die gegenseitige stochastische Abhängigkeit bestehen.

Sind die X_i dagegen kontrollierbare Einflußgrößen (z.B. die Temperatur bei Messungen an chemischen Reaktionsabläufen), so wird man ein anderes Modell wählen, das solche einseitigen stochastischen Abhängigkeiten beschreibt. Man spricht dann vom **Modell I der Regressionsanalyse**.

Werden die X_i durch Vektoren im $\mathbf{R}^p$ ersetzt, d.h. untersucht man die gleichzeitige Abhängigkeit von Y von p Größen, so spricht man von **multipler Regression**. Schränkt man sich wieder auf die lineare Regresion ein, so erhält man Schätzer für die Regresionskonstanten ähnlich wie im Fall der einfachen linearen Regression, wobei die Varianzmatrix von X und die Kovarianzmatrix von (X, Y) jetzt die Rolle von σ_X^2 bzw. $\mathrm{cov}(X, Y)$ spielen.

Anders als in der Regressionsanalyse, wo man den Einfluß einer kontinuierlich variierenden (kontrollierbaren oder zufälligen) Größe X auf eine zufällige Größe Y untersucht, interessiert man sich in der **Varianzanalyse** für den Einfluß einer (oder mehrerer) qualitativer Größen X auf die zufällige Größe Y. Wir werden die Varianzanalyse (hoffentlich) am Ende des Semesters im Lichte der dann bereitgestellten allgemeinen Schätz- und Testtheorie behandeln.

Kapitel 3

Nichtparametrische Verfahren

3.1 Empirische Verteilungsfunktionen

Bisher konnten unsere Familien $\mathcal{P}$ von Wahrscheinlichkeitsverteilungen immer durch eine endliche Anzahl von reellen Parametern parametrisiert werden. Dazu benötigen wir aber Vorausinformationen über die möglichen zugrunde liegenden Verteilungen. Da das nicht immer der Fall ist, möchte man manchmal für $\mathcal{P}$ die Familie aller Wahrscheinlichkeitsverteilungen auf $\mathbf{R}$ oder zumindest aller Wahrscheinlichkeitsverteilungen mit stetiger Verteilungsfunktion zulassen. Arbeitet man unter einer so allgemeinen Annahme, spricht man von **nichtparametrischer Statistik**. Die Aufgabe besteht dabei, grob gesprochen, nicht darin, einen oder mehrere reelle Parameter zu schätzen oder zu testen, sondern die zugrunde liegende Verteilungsfunktion zu schätzen bzw. Eigenschaften dieser Verteilungsfunktion zu testen.

So wie das GGZ und der ZGS die theoretische Grundlage für die Konstruktion von Schätzern und Tests im parametrischen Fall bilden, sind nun die folgenden Begriffsbildungen und Sätze von grundlegender Bedeutung:

Definition 3.1.1 *Sei* $(X_1, \ldots, X_n)$ *eine Stichprobe. Als* **empirische Verteilungsfunktion von** $(X_1, \ldots, X_n)$ *bezeichnen wir*

$$\hat{F}_n(z) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, z]}(X_i).$$

F_n ist die Verteilungsfunktion der zufälligen Wahrscheinlichkeitsverteilung $\frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$. Sind die X_i u.i.v. mit Verteilungsfunktion F, so folgt aus dem starken GGZ:

$$\lim_{n \to \infty} \hat{F}_n(z) = E[I_{(-\infty, z]}(X_1)] = P\{X_1 \leq z\} = F(z) \quad \text{fast sicher.} \tag{3.1}$$

Es gilt in der Tat

Theorem 3.1.2 (Satz von Glivenko-Cantelli) *Seien $(X_1, \ldots, X_n)$ u.i.v. mit Verteilungsfunktion F. Dann konvergiert*

$$D_n := \sup_{x \in R} |\hat{F}_n(x) - F(x)|$$

fast sicher gegen 0.

Beweis: Wir nutzen die Monotonie von F und $\hat{F}_n$ aus: Für $N \in \mathbf{N}$ und $k = 1, \ldots, N$ sei

$$x_{N,k} := \inf\{x \in \mathbf{R} : \; F(x) \geq \frac{k}{N}\},$$

also das **Quantil der Ordnung** $\frac{k}{N}$ von F. Beachte: Ist F bei x unstetig, so können mehrere $x_{N,k}$ mit x übereinstimmen. Setzt man noch $x_{N,0} = -\infty$, $x_{N,N+1} = +\infty$, so bilden die (eventuell auch leeren) Intervalle $I_{N,k} := [x_{N,k}, x_{N,k+1})$ $(k = 0, \ldots, N)$ eine Überdeckung von $\mathbf{R}$, und für $z \in I_{N,k}$ ist ist wegen der Monotonie und Rechtsstetigkeit von F $|F(z) - F(x_{N,k})| \leq \frac{1}{N}$.

Für $z \in I_{N,k}$ ist

$$\begin{aligned}
|\hat{F}_n(z) - F(z)| &\leq |\hat{F}_n(z) - \hat{F}_n(x_{N,k})| + |\hat{F}_n(x_{N,k}) - F(x_{N,k})| + |F(x_{N,k}) - F(z)| \\
&\leq |\hat{F}_n(x_{N,k+1}-) - \hat{F}_n(x_{N,k})| + |\hat{F}_n(x_{N,k}) - F(x_{N,k})| + \frac{1}{N} \\
&\leq |\hat{F}_n(x_{N,k+1}-) - F(x_{N,k+1}-)| + |F(X_{N,k+1}-) - F(X_{N,k})| \\
&\quad +2|\hat{F}_n(x_{N,k}) - F(x_{N,k})| + \frac{1}{N},
\end{aligned}$$

und aus (3.1) folgt

$$\limsup_{n \to \infty} \sup_{z \in R} |\hat{F}_n(z) - F(z)| \leq \frac{2}{N} \quad \text{fast sicher,}$$

da $\lim_{n\to\infty} \hat{F}_n(x_{N,k+1}-) = P\{X_1 < x_{N,k+1}\} = F(x_{N,k+1}-)$. Für $N \to \infty$ folgt daraus der Satz. □

So wie der Satz von Glivenko-Cantelli eine Verschärfung des starken GGZ darstellt, gibt es auch eine Aussage zur Verteilungskonvergenz von D_n, die wir hier ohne Beweis angeben:

Theorem 3.1.3 (Satz von Kolmogorov-Smirnov) *Seien $X_1, \ldots, X_n$ u.i.v. mit* **stetiger** *Verteilungsfunktion F, D_n wie in (3.1.2) und $D_n^+ := \sup_{x \in R}(\hat{F}_n(x) - F(x))$. Dann gilt:*

1.

$$\lim_{n \to \infty} P\{\sqrt{n}\, D_n^+ \leq x\} = \begin{cases} 1 - e^{-2x^2} & \text{für } x > 0 \\ 0 & \text{sonst} \end{cases}$$

2.

$$\lim_{n\to\infty} P\{\sqrt{n}\,D_n \le x\} = K(x) := \begin{cases} \sum_{k=-\infty}^{+\infty}(-1)^k e^{-2k^2x^2} & \text{für } x > 0 \\ 0 & \text{sonst} \end{cases}$$

K heißt die Verteilungsfunktion der Kolmogorov-Verteilung.

Für den Vergleich zweier Stichproben ist der folgende Satz von Interesse:

Theorem 3.1.4 *Seien $X_1, \ldots, X_{n_1}$ und $Y_1, \ldots, Y_{n_2}$ u.i.v. Zufallsvariablen mit stetiger Verteilungsfunktion F, und seien $\hat{F}_{n_1}$ und $\hat{G}_{n_2}$ die empirischen Verteilungsfunktionen zu $X_1, \ldots, X_{n_1}$ bzw. $Y_1, \ldots, Y_{n_2}$. Setze*

$$D_{n_1,n_2} := \sup_{x\in R} |\hat{F}_{n_1}(x) - \hat{G}_{n_2}(x)|, \quad D^+_{n_1,n_2} := \sup_{x\in R}(\hat{F}_{n_1}(x) - \hat{G}_{n_2}(x)).$$

Dann haben die Statistiken

$$\sqrt{\frac{n_1 n_2}{n_1+n_2}} D_{n_1,n_2} \quad \text{bzw.} \quad \sqrt{\frac{n_1 n_2}{n_1+n_2}} D^+_{n_1,n_2} \quad \text{für } \sqrt{\frac{n_1 n_2}{n_1+n_2}} \to \infty$$

asymptotisch die gleiche Verteilung wie $\sqrt{n}D_n$ bzw. $\sqrt{n}D_n^+$ für $n \to \infty$.

Einen Beweis dieses Satzes findet man z.B. in [8, Kapitel VIII]. Wir geben hier nur den Beweis für die D^+-Statistik im Spezialfall $n_1 = n_2 = n$ an:

Wir ordnen die $2n$ Werte $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ der Größe nach und bezeichnen Sie in dieser Reihenfolge mit $Z_1 < Z_2 < \ldots < Z_{2n}$. Da F stetig ist, sind die Werte fast sicher tatsächlich alle verschieden. Für $k = 1, \ldots, n$ sei

$$\xi_k = \begin{cases} 1 & \text{falls } Z_k \text{ eines der } X_k \text{ ist,} \\ -1 & \text{sonst,} \end{cases}$$

und $S_k = \xi_1 + \ldots + \xi_k$. Man rechnet leicht nach, daß

$$n \cdot D^+_{n,n} = \sup_{x\in R} n(\hat{F}_n(x) - \hat{G}_n(x)) = \max_{1\le k\le 2n} S_k.$$

Da unter den ξ_k genau n 1-en und n (-1)-en sind, gibt es $\binom{2n}{n}$ verschiedene Arten, die Werte der ξ_k zu arrangieren, und da die Z_k u.i.v. sind, hat jedes dieser Arrangements die Wahrscheinlichkeit $\binom{2n}{n}^{-1}$.

Sei $r \in \mathbf{N}$ und $\max_k S_k \ge r$. Setze $l := \min\{k : S_k = r\}$. Für $k = 1, \ldots, 2n$ sei

$$\bar{\xi}_k = \begin{cases} \xi_k, & \text{falls } k \le l, \\ -\xi_k, & \text{falls } k > l. \end{cases}$$

$(\bar{\xi}_k)_k$ ist eine Folge von $(n+r)$ 1-en und $(n-r)$ (-1)-en, da $\sum_{k=1}^{2n} \bar{\xi}_k = 2\sum_{k=1}^{l} \xi_k - \sum_{k=1}^{2n} \xi_k = 2r$. Durch die Zuordnung $(\xi_k)_k \mapsto (\bar{\xi}_k)_k$ wird eine bijektive Zuordnung

Abbildung 3.1:

zwischen den $\{-1,1\}$-Folgen mit Summe 0 und Partialsummenmaximum $\geq r$ und den $\{-1,1\}$-Folgen mit Summe $2r$ hergestellt, so daß es genau $\binom{2n}{n-r}$ Folgen dieser Art gibt (siehe Abb.3.1). Also ist

$$P\{nD_{n,n}^+ \geq r\} = P\{\max_{1\leq k\leq 2n} S_k \geq r\} = \binom{2n}{n-r} / \binom{2n}{n}, \tag{3.2}$$

womit wir nebenbei die exakte Verteilung von $nD_{n,n}^+$ bestimmt haben.

Wir erhalten somit:

$$\begin{aligned}
& \lim_{n\to\infty} P\{\sqrt{\frac{n}{2}} D_{n,n}^+ \geq z\} \\
= & \lim_{n\to\infty} P\{n \cdot D_{n,n}^+ \geq r\}, \text{ wo } r = \min\{k \in \mathbf{Z} : k \geq z\sqrt{2n}\} \approx z\sqrt{2n} \\
= & \lim_{n\to\infty} \binom{2n}{n-r} / \binom{2n}{n} \\
= & \lim_{n\to\infty} \frac{(2n)!\, n!\, n!}{(n-r)!\,(n+r)!\,(2n)!} \\
= & \lim_{n\to\infty} \frac{(n/e)^{2n} 2\pi n}{((n-r)/e)^{n-r}((n+r)/e)^{n+r} 2\pi\sqrt{n^2-r^2}} \quad \text{(Stirling Formel)} \\
= & \lim_{n\to\infty} \left(\frac{n-r}{n+r}\right)^r \left(\frac{n^2}{n^2-r^2}\right)^{n+1/2} \\
= & \lim_{n\to\infty} \left(1 - \frac{2r}{r^2/2z^2 + r}\right)^r \left(1 + \frac{2nz^2}{n^2 - 2nz^2}\right)^n \\
= & e^{-4z^2} e^{2z^2} \\
= & e^{-2z^2}
\end{aligned}$$

□

In (3.2) hatten wir bereits gesehen, daß die exakte Verteilung von $D_{n,n}^+$ nicht von der zugrunde liegenden stetigen Verteilungsfunktion F abhängt. Das motiviert

Definition 3.1.5 *Sei $(M, \mathcal{M}, \mathcal{P})$ ein statistischer Raum, T eine Statistik auf $(M, \mathcal{M}, \mathcal{P})$. T heißt* **verteilungsfrei auf** *$(M, \mathcal{M}, \mathcal{P})$, falls $P \circ T^{-1}$ für alle $P \in \mathcal{P}$ dieselbe Wahrscheinlichkeitsverteilung ist.*

Allgemein gilt

Theorem 3.1.6 *Die Statistiken D_n, D_n^+, D_{n_1,n_2}, D_{n_1,n_2}^+ sind verteilungsfrei auf $(\mathbf{R}, \mathcal{B}, \mathcal{P})^n$ bzw. $(\mathbf{R}, \mathcal{B}, \mathcal{P})^{n_1+n_2}$, wo $\mathcal{P}$ die Familie aller stetigen Wahrscheinlichkeitsverteilungen auf $(\mathbf{R}, \mathcal{B})$ ist.*

Beweis: Sei $\phi : (0,1) \to \mathbf{R}$ die **Quantilfunktion** zu F,

$$\phi(u) = \inf\{x \in \mathbf{R} : u \le F(x)\}.$$

Es ist

$$\phi(u) \le x \quad \Longleftrightarrow \quad u \le F(x), \tag{3.3}$$

also $F(\phi(u)) \ge u$ und $\phi(F(x)) \le x$, und aus der Stetigkeit von F folgt sogar

$$F(\phi(u)) = \inf\{F(x) : x \in \mathbf{R}, u \le F(x)\} = u.$$

Seien nun $U_1, \ldots, U_n$ unabhängige, auf $(0,1)$ gleichverteilte Zufallsvariablen. Wegen (3.3) ist $P\{\phi(U_i) \le x\} = P\{U_i \le F(x)\} = F(x)$, so daß die Zufallsvariablen $\phi(U_i)$ unabhängig und nach F verteilt sind, also o.B.d.A. $X_i = \phi(U_i)$. Wieder wegen (3.3) folgt

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n I_{\{\phi(U_i)\le x\}} = \frac{1}{n}\sum_{i=1}^n I_{\{U_i \le F(x)\}} = \hat{H}_n(F(x)),$$

wo $\hat{H}_n$ die empirische Verteilungsfunktion von $U_1, \ldots, U_n$ ist. Also:

$$D_n = \sup_{x\in R} |\hat{F}_n(x) - F(x)| = \sup_{x\in R} |\hat{H}_n(F(x)) - F(x)| = \sup_{u\in(0,1)} |\hat{H}_n(u) - u|.$$

Daher hat D_n unter jedem stetigen F die gleiche Verteilung wie unter der Gleichverteilung auf $(0,1)$. Für die anderen Statistiken argumentiert man ähnlich. □

Dieser Satz ermöglicht es, die Verteilungen der D-Statistiken zu tabellieren und auch bei kleinen Stichprobenumfängen Tests auf ihnen zu basieren.

3.2 Geordnete Stichproben

Wenn wir ein n-tupel $(x_1, \ldots, x_n)$ der Größe nach anordnen, so bezeichnen wir das Ergebnis mit

$$g_{(n)}(x_1, \ldots, x_n) = (x_n^{(1)} \leq x_n^{(2)} \leq \ldots \leq x_n^{(n)}). \tag{3.4}$$

$\mathrm{Rg}(x_i)$, die **Rangzahl** von x_i, bezeichne die Position von x_i in (3.4), und wir setzen

$$r_{(n)}(x_1, \ldots, x_n) = (\mathrm{Rg}(x_1), \ldots, \mathrm{Rg}(x_n)). \tag{3.5}$$

Bei Gleichheit von x_i und x_j legen wir (willkürlich) fest, daß $\mathrm{Rg}(x_i) < \mathrm{Rg}(x_j)$, falls $i < j$. In jedem Fall ist $r_{(n)}(x_1, \ldots, x_n) \in \Pi_n$, der Menge aller Permutationen von $\{1, \ldots, n\}$.

Seien nun $X_1, \ldots, X_n$ u.i.v. Zufallsvariablen mit stetiger Verteilungsfunktion F. Aus der Stetigkeit von F folgt, daß die X_i f.s. paarweise verschieden sind. Daher sind

$$R_{(n)} = r_{(n)}(X_1, \ldots, X_n)$$

und

$$X_{(n)}^g = g_{(n)}(X_1, \ldots, X_n)$$

fast sicher ohne Willkür definiert. $X_{(n)}^g$ heißt **Ordnungsstatistik**, $R_{(n)}$ **Rangstatistik** von $(X_1, \ldots, X_n)$. Der folgende Zusammenhang mit der empirischen Verteilungsfunktion ist evident:

$$\hat{F}_n(X_n^{(k)}) = \frac{k}{n}$$

Theorem 3.2.1 *Seien $X_1, \ldots, X_n$ u.i.v. Zufallsvariablen mit stetiger Verteilungsfunktion F. Dann sind die wie oben definierten Zufallsvariableen $X_{(n)}^g$ und $R_{(n)}$ unabhängig voneinander. Es gilt:*

1. $$P\{R_{(n)} = \pi\} = \frac{1}{n!} \text{ für alle } \pi \in \Pi_n$$

   und

   $$P\{X_{(n)}^g \in A\} = n!\, P\{(X_1, \ldots, X_n) \in A \cap V_n\} \text{ für alle } A \in \mathcal{B}^n,$$

   wo $V_n = \{(x_1, \ldots, x_n) \in \mathbf{R}^n : x_1 < \ldots < x_n\}$.

2. *Für die von $g_{(n)}$ erzeugte σ-Algebra $\sigma(g_{(n)})$ gilt:*

   $$\sigma(g_{(n)}) = \{A \in \mathcal{B}^n : \tilde{\pi}(A) = A \ \forall \pi \in \Pi_n\},$$

   $\tilde{\pi}(x_1, \ldots, x_n) = (x_{\pi(1)}, \ldots, x_{\pi(n)})$.

Beweis:

1. Seien $\pi, \psi \in \Pi$, $\psi = \pi^{-1}$. $R_{(n)} = \pi$ ist äquivalent zu $X^g_{(n)} = \tilde{\psi}(X_1, \ldots, X_n) = (X_{\psi(1)}, \ldots, X_{\psi(n)})$. Da die Verteilung von $(X_1, \ldots, X_n)$ unter Koordinatenpermutation invariant ist, folgt:

$$\begin{aligned} & P\{R_{(n)} = \pi,\, X^g_{(n)} \in A\} \\ = \;& P\{r_{(n)}(X_1, \ldots, X_n) = \pi,\, \tilde{\psi}(X_1, \ldots, X_n) \in A\} \\ = \;& P\{r_{(n)}(\tilde{\psi}(X_1, \ldots, X_n) = (1, \ldots, n)),\, \tilde{\psi}(X_1, \ldots, X_n) \in A\} \qquad (3.6) \\ = \;& P\{r_{(n)}(X_1, \ldots, X_n) = (1, \ldots, n),\, (X_1, \ldots, X_n) \in A\} \\ = \;& P\{(X_1, \ldots, X_n) \in A \cap V_n\}. \end{aligned}$$

   Insbesondere hängt $P\{R_{(n)} = \pi,\, X^g_{(n)} \in A\}$ nicht von $\pi \in \Pi_n$ ab, so daß

$$P\{R_{(n)} = \pi,\, X^g_{(n)} \in A\} = (\operatorname{card} \Pi_n)^{-1} P\{X^g_{(n)} \in A\} = \frac{1}{n!} P\{X^g_{(n)} \in A\} \qquad (3.7)$$

   für alle $\pi \in \Pi_n$. Für $A = \mathbf{R}^n$ ergibt sich $P\{R_{(n)} = \pi\} = \frac{1}{n!}$. Zusammen mit (3.7) folgt daraus die Unabhängigkeit von $X^g_{(n)}$ und $R_{(n)}$, und durch Summation über π erhält man aus (3.6) $P\{X^g_{(n)} \in A\} = n! \cdot P\{(X_1, \ldots, X_n) \in A \cap V_n\}$.

2. Da $g_{(n)} = g_{(n)} \circ \pi$ für alle $\pi \in \Pi_n$, ist die „$\subseteq$"-Inklusion klar. Sei umgekehrt $\tilde{\pi}(A) = A$ für alle $\pi \in \Pi_n$. Dann ist $g_{(n)}(A) \subseteq A$ und daher

$$A \subseteq g^{-1}_{(n)}(g_{(n)}(A)) \subseteq g^{-1}_{(n)}(A) = \bigcup_{\pi \in \Pi_n} \pi(A) = A.$$

□

Bemerkung 3.2.2 Da wir im Beweis des letzten Satzes nicht wirklich die Unabhängigkeit der X_i sondern nur die Invarianz ihrer gemeinsamen Verteilung unter Koordinatenpermutationen benutzt haben, bleibt der Satz unter dieser schwächeren Annahme der **Vertauschbarkeit** der Verteilungen richtig.

Der Zusammenhang zwischen der Rangstatistik und der empirischen Verteilungsfunktion kann benutzt werden, um die Verteilungen der Ranggrößen $X^{(k)}_n$ zu bestimmen:

$$\begin{aligned} P\{X^{(k)}_n \le x\} &= P\{\hat{F}_n(x) \ge \frac{k}{n}\} = \sum_{j=k}^{n} \binom{n}{j} F(x)^j (1 - F(x))^{n-j} \qquad (3.8) \\ &= k \binom{n}{k} \int_0^{F(x)} t^{k-1} (1-t)^{n-k}\, dt \qquad (3.9) \end{aligned}$$

Die letzte Identität erhält man durch mehrfache partielle Integration. !

Für $0 < q < 1$ bezeichnet man die Ranggröße $X^{(k)}_n$ mit $k = [nq]+1$ als das **Stichprobenquantil der Ordnung** q. Der folgende Satz besagt, daß die Stichprobenquantile eine konsistente Folge von Schätzern für die Quantile der zugrunde liegenden Verteilungsfunktion bilden.

Theorem 3.2.3 *Seien $X_1, X_2, \ldots$ u.i.v. Zufallsvariablen mit Verteilungsfunktion F, und sei $0 < q < 1$. Wenn die Gleichung $F(x) = q$ genau eine Lösung $x = \lambda_q$ besitzt, so gilt*

$$X_n^{([nq]+1)} \to \lambda_q \quad \textit{stochastisch.}$$

Beweis: Da λ_q die einzige Lösung von $F(x) = q$ ist, ist für $\epsilon > 0$

$$\lim_{n\to\infty} \left(F(\lambda_q + \epsilon) - \frac{[nq]+1}{n} \right) = F(\lambda_q + \epsilon) - q =: \delta(\epsilon) > 0.$$

Für hinreichend große n ist daher $\frac{[nq]+1}{n} < F(\lambda_q + \epsilon) - \frac{\delta(\epsilon)}{2}$, und es folgt:

$$\begin{aligned} \limsup_{n\to\infty} P\{X_n^{([nq]+1)} > \lambda_q + \epsilon\} &= \limsup_{n\to\infty} P\{\hat{F}_n(\lambda_q + \epsilon) < \frac{[nq]+1}{n}\} \\ &\leq \limsup_{n\to\infty} P\{\hat{F}_n(\lambda_q + \epsilon) < F(\lambda_q + \epsilon) - \frac{\delta(\epsilon)}{2}\} \\ &= 0 \end{aligned}$$

nach dem Satz von Glivenko-Cantelli. Analog zeigt man

$$\limsup_{n\to\infty} P\{X_n^{([nq]+1)} < \lambda_q - \epsilon\} = 0.$$

□

Hat F eine Dichte, die bei λ_q positiv und stetig ist, so gilt auch ein ZGS für die Konvergenz der Stichprobenquantile, siehe Satz 3.7 in [11] und die dort angegebene Literatur.

3.3 Der Wilcoxon Test

Einer der bekanntesten auf Rängen basierenden Tests ist der **Wilcoxon 2-Stichproben Test**: Seien $X_1, \ldots, X_n$ und $Y_1, \ldots, Y_m$ unabhängige Stichproben mit stetigen Verteilungsfunktionn F bzw. G. Getestet werden soll $H_0 : F = G$ gegen $H_1 : F < G$. Der zur Modellierung benutzte statistische Raum ist

$$(M, \mathcal{M}, \mathcal{P}) = (\mathbf{R}^{n+m}, \mathcal{B}^{n+m}, \{F^n \times G^m : F, G \text{ stetige Verteilungsfunktionen}\}).$$

Als **Wilcoxon Statistik** bezeichnet man

$$W_{n,m} = \sum_{i=1}^{n} R_i,$$

wo R_i der Rang von X_i in der kombinierten Stichprobe $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ ist. Ist $F < G$, so erwartet man, daß die X_i tendenziell niedrigere Ränge haben als die Y_j. Das wird im folgenden Satz präzisiert:

Theorem 3.3.1 *1.* $W_{n,m} = \sum_{i=1}^{n} \sum_{j=1}^{m} I_{\{X_i \geq Y_j\}} + \frac{n(n+1)}{2}$

2. Sei $\theta(F,G) := \int G(x)\, dF(x) = P\{Y_1 \leq X_1\}$. *Dann ist*

$$E[W_{n,m}] = nm\,\theta(F,G) + \frac{n(n+1)}{2},$$

und es gibt Konstanten a, b, c *derart, daß*

$$V(W_{n,m}) = nm(a(n-1) + b(m-1) + c).$$

3. Unter $H_0 : F = G$ *ist* $\theta(F,G) = \frac{1}{2}$, *also* $E[W_{n,m}] = \frac{1}{2}n(n+m+1)$, *und* $a = b = \frac{1}{12}$, $c = \frac{1}{4}$, *also* $V(W_{n,m}) = \frac{1}{12}nm(n+m+1)$.

4. Gehen $n, m(n) \to \infty$ *derart, daß* $n/(n+m(n)) \to \alpha$ *für ein* $0 < \alpha < 1$, *so ist* $\lim_{n\to\infty} \frac{1}{nm(n)} W_{n,m(n)} = \theta(F,G) + \frac{\alpha}{2(1-\alpha)}$ *fast sicher.*

5. Mit $m, n \to \infty$ *wie vorher gilt*

$$\sqrt{n+m}\left(\frac{1}{nm}W_{n,m} - \theta(F,G) - \frac{n+1}{2m}\right) \Longrightarrow \mathcal{N}(0,\sigma^2),$$

wo $\sigma^2 = \alpha^{-1}V(G(X_1)) + (1-\alpha)^{-1}V(F(Y_1))$. *Ist* $F = G$, *so ist* $\sigma^2 = \frac{1}{12\alpha(1-\alpha)}$.

Bemerkung 3.3.2 Da aus $F < G$ folgt $\theta = P\{Y_1 \leq X_1\} > \frac{1}{2}$, kann die Statistik $\frac{1}{nm}W_{n,m}$ zum Testen von $H_0 : F = G$ gegen $H_1 : F < G$ benutzt werden. Da die Rangstatistik unter H_0 verteilungsfrei ist, gilt das gleiche für $W_{n,m}$. Das erlaubt es, die Verteilung von $(W_{n,m} - E[W_{n,m}])/\sqrt{V(W_{n,m})}$ für kleine Stichprobenumfänge exakt zu tabellieren. Tafeln dafür findet man z.B. in [9], wo auch viele andere nichtparametrische Statistiken unter Anwendungsgesichtspunkten beschrieben sind.

Beweis des Satzes:

1.

$$W = \sum_{i=1}^{n} R_i = \sum_{i=1}^{n}\sum_{j=1}^{n} I_{\{X_j \leq X_i\}} + \sum_{i=1}^{n}\sum_{j=1}^{m} I_{\{Y_j \leq X_i\}} = \frac{n(n+1)}{2} + W',$$

wo $W' := \sum_{i=1}^{n}\sum_{j=1}^{m} I_{\{Y_j \leq X_i\}}$.

2. Da

$$P\{Y_j \leq X_i\} = E[P(Y_j \leq X_i | X_i)] = E[G(X_i)] = \int G(x)\, dF(x) = \theta, \qquad (3.10)$$

ist $E[W] = E[W'] + \frac{n(n+1)}{2} = nm\theta + \frac{n(n+1)}{2}$. Weiter ist

$$\begin{aligned} V(W) &= V(W') \qquad (3.11)\\ &= \sum_{i,j=1}^{n}\sum_{k,l=1}^{m} P\{Y_k \le X_i,\ Y_l \le X_j\} - n^2m^2\theta^2 \\ &= \sum_{i\neq j}\sum_{k\neq l} P\{Y_k \le X_i\}P\{Y_l \le X_j\} + \sum_{i\neq j}\sum_{k=1}^{m} P\{X_i \ge Y_k,\ X_j \ge Y_k\} \\ &\quad + \sum_{i=1}^{n}\sum_{k\neq l} P\{Y_k \le X_i,\ Y_l \le X_i\} + \sum_{i=1}^{n}\sum_{k=1}^{m} P\{Y_k \le X_i\} - n^2m^2\theta^2 \\ &= n(n-1)m(m-1)\theta^2 + nm(n-1)\int (1-F(y))^2\, dG(y) \\ &\quad + nm(m-1)\int G(x)^2\, dF(x) + nm\theta - n^2m^2\theta^2 \\ &= nm\left[(n-1)\left(1 - 2\int F(y)\, dG(y) + \int F(y)^2\, dG(y) - \theta^2\right)\right. \\ &\quad \left. + (m-1)\left(\int G(x)^2\, dF(x) - \theta^2\right) + \theta - \theta^2\right]. \qquad (3.12) \end{aligned}$$

3. Sei $F = G$. Da F stetig ist, ist für alle $k \in \mathbf{N}$:

$$\int F(x)^k\, dF(x) = \frac{1}{k+1}(F(+\infty)^{k+1} - F(-\infty)^{k+1}) = \frac{1}{k+1}.$$

   Insbesondere ist $\theta = \frac{1}{2}$ und $\int F(x)^2\, dF(x) = \frac{1}{3}$. Der Rest folgt nun durch Einsetzen in (3.11).

4. Aus (3.11) folgt sofort, daß $V(\frac{1}{nm}W_{n,m}) = O(\frac{1}{n+m})$, so daß $\frac{1}{nm}W_{n,m}$ stochastisch gegen $\lim_{n,m\to\infty} E[\frac{1}{nm}W_{n,m}] = \theta + \frac{\alpha}{2(1-\alpha)}$ konvergiert.

   Zum Beweis des starken GGZ und des ZGS benutzen wir die **Projektionsmethode**: Beachte zunächst, daß

$$\begin{aligned} U_j &:= E[W'|X_j] - nm\theta = \sum_{i=1}^{n}\sum_{k=1}^{m} P(Y_k \le X_i|X_j) - nm\theta \\ &= -m\theta + mG(X_j) \quad \text{und} \\ V_l &:= E[W'|Y_l] - nm\theta = \sum_{i=1}^{n}\sum_{k=1}^{m} P(Y_k \le X_i|Y_l) - nm\theta \\ &= -n\theta + n(1-F(Y_l)). \end{aligned}$$

   Beachte $E[U_j] = E[V_l] = 0$. Aus dem starken GGZ für u.i.v Zufallsvariablen folgt daher, falls $\frac{n}{n+m} \to \alpha$:

$$\lim_{n\to\infty} \frac{1}{nm}(\sum_{j=1}^{n} U_j + \sum_{l=1}^{m} V_l)$$

$$
\begin{aligned}
&= \lim_{n\to\infty}\left(\frac{1}{n}\sum_{j=1}^{n}(-\theta+G(X_j))+\frac{1}{m}\sum_{l=1}^{m}(-\theta+(1-F(Y_l)))\right)\\
&= 0. \qquad (3.13)
\end{aligned}
$$

Aus dem ZGS für Schemata unabhängiger Zufallsvariablen folgt

$$
\begin{aligned}
&\frac{\sqrt{n+m}}{nm}(\sum_{j=1}^{n}U_j+\sum_{l=1}^{m}V_l)\\
= \;&\frac{1}{\sqrt{n+m}}\left(\alpha^{-1}\sum_{j=1}^{n}(G(X_j)-\theta)+(1-\alpha)^{-1}\sum_{l=1}^{m}(1-F(Y_l)-\theta)\right)\cdot(1+o(1))\\
\Longrightarrow \;&\mathcal{N}(0,\sigma^2), \qquad (3.14)
\end{aligned}
$$

wo

$$
\begin{aligned}
\sigma^2 &= \lim_{n\to\infty}\frac{\alpha^{-2}nV(G(X_1))+(1-\alpha)^{-2}mV(G(Y_1))}{n+m}\\
&= \alpha^{-1}V(G(X_1))+(1-\alpha)^{-1}V(F(Y_1)).
\end{aligned}
$$

Im Fall $F=G$ ist

$$
V(G(X_1))=V(F(Y_1))=\int F(x)^2\,dF(x)-(\int F(x)\,dF(x))^2=\frac{1}{12},
$$

so daß $\sigma^2=\frac{1}{12\alpha(1-\alpha)}$.
Zu zeigen bleibt, daß

$$
\Delta_{n,m}:=\frac{1}{nm}(W'_{n,m}-\sum_{j=1}^{n}U_j-\sum_{l=1}^{m}V_l-nm\theta)
$$

in einem geeigneten Sinn gegen 0 konvergiert: Da $E[U_j]=E[V_l]=E[W']-nm\theta=0$, ist

$$
\begin{aligned}
E[\Delta] &= 0,\\
V(\Delta) &= V(\frac{1}{nm}\sum_{j=1}^{n}\sum_{l=1}^{m}(I_{\{Y_l\le X_j\}}-G(X_j)+\theta-(1-F(Y_l))+\theta-\theta))\\
&= \frac{1}{n^2m^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{l=1}^{m}E[(I_{\{Y_k\le X_i\}}-G(X_i)-(1-F(Y_k))+\theta)\\
&\quad (I_{\{Y_l\le X_j\}}-G(X_j)-(1-F(Y_l))+\theta)]\\
&= \frac{1}{n^2m^2}\sum_{i=1}^{n}\sum_{k=1}^{m}E[(I_{\{Y_k\le X_i\}}-G(X_i)-F(Y_k)+\theta)^2]\\
&=: \frac{\gamma}{nm}=\frac{\gamma}{\alpha(1-\alpha)(n+m)^2}
\end{aligned}
$$

mit einer von F und G abhängigen Konstanten $\gamma < \infty$, da für $i \neq j$ oder $k \neq l$ die Erwartung eines Produkts bzgl. X_i bzw. Y_k schon identisch 0 ist. Aus dem Borel Cantelli Lemma folgt, daß $\lim_{n\to\infty} \Delta_{n,m(n)} = 0$ fast sicher, also wegen (3.13) $\lim_{n\to\infty} \frac{1}{nm(n)} W'_{n,m(n)} = \theta$ f.s.

5. Ebenso ist $V(\sqrt{n+m}\Delta_{n,m}) = \frac{\gamma}{\alpha(1-\alpha)(n+m)}$, so daß $\sqrt{n+m}\Delta_{n,m} \to 0$ stochastisch. Aus (3.14) folgt daher $\sqrt{n+m}(\frac{1}{nm} W'_{n,m} - \theta) \Longrightarrow \mathcal{N}(0, \sigma^2)$.

□

Kapitel 4

Grundbegriffe der Mathematischen Statistik

4.1 Dominierbare statistische Räume, Exponentialräume

Ist $(M, \mathcal{M}, \mathcal{P})$ ein statistischer Raum, so heißt eine Menge $A \in \mathcal{M}$ **$\mathcal{P}$-Nullmenge**, falls $P(A) = 0$ für alle $P \in \mathcal{P}$. Eine meßbare $\mathbf{R}^m$-wertige Funktion f heißt **$\mathcal{P}$-integrierbar**, falls $E_P|f| < \infty$ für alle $P \in \mathcal{P}$, u.s.w.

Definition 4.1.1 *1. Ein statistischer Raum $(M, \mathcal{M}, \mathcal{P})$ heißt* **dominierbar**, *falls es ein σ-endliches Maß μ auf $(M, \mathcal{M})$ gibt, so daß $P << \mu$ für alle $P \in \mathcal{P}$, kurz: $\mathcal{P} << \mu$. μ heißt dominierendes Maß.*

2. $\mathcal{P}$ heißt äquivalent zu μ, kurz $\mathcal{P} \equiv \mu$, falls $\mathcal{P} << \mu$ und falls jede $\mathcal{P}$-Nullmenge eine μ-Nullmenge ist.

3. $\mathcal{P}_1$ heißt äquivalent zu $\mathcal{P}_2$, kurz $\mathcal{P}_1 \equiv \mathcal{P}_2$, falls jede $\mathcal{P}_1$-Nullmenge auch eine $\mathcal{P}_2$-Nullmenge ist und umgekehrt.

Bemerkung 4.1.2 Das Maß μ in dieser Definition kann o.B.d.A. als Wahrscheinlichkeitsmaß angenommen werden, denn aus der σ-Endlichkeit von μ folgt die Existenz eines strikt positiven $f \in L^1_\mu$.

Theorem 4.1.3 *Sei $(M, \mathcal{M}, \mathcal{P})$ ein statistischer Raum. Äquivalent sind:*

1. $(M, \mathcal{M}, \mathcal{P})$ ist dominierbar.

2. Es gibt eine abzählbare Teilklasse $\mathcal{P}_0 \subseteq \mathcal{P}$, die zu $\mathcal{P}$ äquivalent ist.

3. Es gibt $P_i \in \mathcal{P}$, $c_i > 0$ $(i \in \mathbf{N})$ derart, daß $\sum_{i=1}^{\infty} c_i P_i =: \nu$ ein zu $\mathcal{P}$ äquivalentes Wahrscheinlichkeitsmaß ist.

Beweis: Die Implikationen $2 \Rightarrow 3$ und $3 \Rightarrow 1$ sind klar. Wir zeigen $1 \Rightarrow 2$: O.B.d.A. sei $\mathcal{P}$ durch ein Wahrscheinlichkeitsmaß μ dominiert. Sei

$$\Gamma := \{\mathcal{F} \subseteq \mathcal{P} : \ \mathcal{F} \text{ höchstens abzählbar}\}.$$

Für $\mathcal{F} \in \Gamma$ setze

$$\alpha(\mathcal{F}) := \sup\{\mu(A) : \ P(A) = 0 \ \forall P \in \mathcal{F}\}.$$

Ein einfaches Ausschöpfungsargument zeigt: Es gibt ein $A_{\mathcal{F}} \in \mathcal{M}$ derart, daß $\alpha(\mathcal{F}) = \mu(A_{\mathcal{F}})$ und $P(A_{\mathcal{F}}) = 0$ für alle $P \in \mathcal{F}$. Setze

$$\alpha := \inf\{\alpha(\mathcal{F}) : \ \mathcal{F} \in \Gamma\}.$$

Ein ebenso einfaches Ausschöpfungsargument zeigt: Es gibt ein $\mathcal{P}_0 \in \Gamma$ mit $\alpha = \alpha(\mathcal{P}_0)$.

Sei nun $A \in \mathcal{M}$ beliebig mit $P(A) = 0$ für alle $P \in \mathcal{P}_0$. Zu zeigen ist: $P(A) = 0$ für alle $P \in \mathcal{P}$. Sei dazu $P' \in \mathcal{P}$ beliebig, $\mathcal{F} := \mathcal{P}_0 \cup \{P'\}$. Da $P(A \cup A_{\mathcal{F}}) \leq P(A) + P(A_{\mathcal{F}}) = 0$ für alle $P \in \mathcal{P}_0$, ist

$$\mu(A \cup A_{\mathcal{F}}) \leq \alpha(\mathcal{P}_0) = \alpha \leq \alpha(\mathcal{F}) = \mu(A_{\mathcal{F}}),$$

also $\mu(A \setminus A_{\mathcal{F}}) = 0$, so daß aus $P' << \mu$ und $P' \in \mathcal{F}$ folgt:

$$P'(A) \leq P'(A \setminus A_{\mathcal{F}}) + P'(A_{\mathcal{F}}) = 0.$$

□

Eine wichtige Klasse dominierbarer statistischer Räume bilden die Exponentialräume:

Definition 4.1.4 *Ein dominierbarer statistischer Raum $(M, \mathcal{M}, \mathcal{P})$, $\mathcal{P} = \{P_\vartheta : \ \vartheta \in \Theta\}$, ist ein ($k$-parametriger)* **Exponentialraum**, *falls die zugehörigen Dichten die Form*

$$f_\vartheta(x) := \frac{dP_\vartheta}{d\mu}(x) = C(\vartheta)h(x)\exp\left(\sum_{j=1}^{k} c_j(\vartheta)h_j(x)\right)$$

haben, wo $\{1, c_1(\vartheta), \ldots, c_k(\vartheta)\}$ ein linear unabhängiges System reeller Funktionen auf Θ und $\{1, h_1(x), \ldots, h_k(x)\}$ ein auf dem Komplement jeder μ-Nullmenge unabhängiges System reeller, meßbarer Funktionen ist. $\mathcal{P}$ heißt dann auch **Exponentialfamilie**.

Durch Übergang zum dominierenden Maß $d\mu^*(x) = h(x)d\mu(x)$ kann man o.B.d.A. annehmen, daß $h = 1$.

Bemerkung 4.1.5

1. Ist $\mathcal{P}$ eine Exponentialfamilie, so gilt $P \equiv P'$ für alle $P, P' \in \mathcal{P}$.
2. Würde man keine lineare Unabhängigkeit fordern, so könnte eine k-parametrige Exponentialfamilie auch mit weniger als k Parametern dargestellt werden.

Beispiel 4.1.6 Exponentialfamilien sind:

1. Binomialverteilungen: $\mathcal{P} = \{b(p,n) : \ 0 < p < 1\}$, n fest, $\mu =$ Zählmaß auf $\mathbf{Z}$, $\vartheta = p$.

$$\begin{aligned} f_\vartheta(k) &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \underbrace{(1-p)^n}_{C(\vartheta)} \cdot \underbrace{\binom{n}{k}}_{h(k)} \cdot \exp(\underbrace{(\log \frac{p}{1-p})}_{c_1(\vartheta)} \cdot \underbrace{k}_{h_1(k)}) \end{aligned}$$

2. Normalverteilungen mit festem σ^2: $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2) : \ \mu \in \mathbf{R}, \sigma^2 \text{ fest}\}$, dominierendes Maß = Lebesgue-Maß auf $\mathbf{R}$.

$$\begin{aligned} f_\mu(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) \\ &= \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{\mu^2}{2\sigma^2})}_{C(\mu)} \cdot \underbrace{\exp(-\frac{x^2}{2\sigma^2})}_{h(x)} \cdot \exp(\underbrace{\frac{\mu}{\sigma^2}}_{c_1(\mu)} \cdot \underbrace{x}_{h_1(x)}) \end{aligned}$$

3. Normalverteilungen: $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2) : \ \mu \in \mathbf{R}, \sigma^2 > 0\}$, dominierendes Maß = Lebesgue-Maß auf $\mathbf{R}$.

$$\begin{aligned} f_{\mu,\sigma^2}(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) \\ &= \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{\mu^2}{2\sigma^2})}_{C(\mu)} \cdot \exp(\underbrace{\frac{\mu}{\sigma^2}}_{c_1(\mu,\sigma^2)} \cdot \underbrace{x}_{h_1(x)} + \underbrace{\frac{-1}{2\sigma^2}}_{c_2(\mu,\sigma^2)} \cdot \underbrace{x^2}_{h_2(x)}) \end{aligned}$$

4. Normalverteilungen mit festem μ (Übung) !
5. d-dimensionale Normalverteilungen (Übung) !
6. Poissonverteilungen (Übung) !

Bemerkung 4.1.7 1. Kann ein Experiment durch den Exponentialraum $(M, \mathcal{M}, \mathcal{P})$ beschrieben werden, so kann seine n-fache unabhängige Wiederholung durch den durch μ^n dominierten Exponentialraum $(M, \mathcal{M}, \mathcal{P})^n$ beschrieben werden, wobei

$$\frac{dP_\vartheta^n}{d\mu^n}(x_1, \ldots, x_n) = C(\vartheta)^n \left(\prod_{i=1}^n h(x_i) \right) \cdot \exp\left(\sum_{j=1}^k c_j(\vartheta) (\sum_{i=1}^n h_j(x_i)) \right).$$

2. Durch den Übergang von ϑ zu $\xi(\vartheta) := (c_1(\vartheta), \ldots, c_k(\vartheta))$ erhält man einen neuen Parameterraum $\subseteq \mathbf{R}^k$. Dieser kann zum **natürlichen Parameterraum**

$$\Xi = \left\{ \xi \in \mathbf{R}^k : \ 0 < \int_M \exp\left(\sum_{j=1}^k \xi_j h_j(x) \right) d\mu^*(x) < \infty \right\}$$

erweitert werden. Zur Vereinfachung der Schreibweise nennt man die neuen Parameter wieder $\vartheta \in \Theta$ (und manchmal das Maß μ^* wieder μ).

Mit diesen Konventionen haben wir

Lemma 4.1.8 *Sei $\mathcal{P} = \{P_\vartheta : \vartheta \in \Theta\}$ eine k-parametrige Exponentialfamilie mit natürlichem Parameterraum $\Theta \subseteq \mathbf{R}^k$.*

1. *Θ ist konvex.*

2. *Sei $\Theta_C := \{z \in \mathbf{C}^k : \ \Re z \in \Theta\}$, wo $\Re z = (\Re z_1, \ldots, \Re z_k)$, $G \subseteq \Theta_C$ offen, und sei $\varphi : M \to \mathbf{R}$ $\mathcal{M}$-meßbar und für alle $P_{\Re z}$ mit $z \in G$ integrierbar. Dann ist*

$$g(z) := \int_M \varphi(x) \exp\left(\sum_{j=1}^k z_j h_j(x) \right) d\mu^*(x)$$

eine auf G holomorphe Funktion in dem Sinn, daß sie in jeder Variablen z_j separat holomorph ist.

3. *Sei $\Theta \subseteq \mathbf{R}^k$ offen, $G = \Theta_C$, $(M, \mathcal{M}) = (\mathbf{R}^k, \mathcal{B}^k)$ und $h_j(x) = x_j$ für alle j. Ist $g(z) = 0$ für alle $z \in \Theta$, so ist $\varphi = 0$ μ^*-f.s.*

4. *Für $j = 1, \ldots, k$ ist*

$$\frac{\partial}{\partial z_j} g(z) = \int_M \varphi(x) h_j(x) \exp\left(\sum_{j=1}^k z_j h_j(x) \right) d\mu^*(x)$$

auf G.

5. *Für $j = 1, \ldots, k$ ist*

$$\frac{\partial}{\partial \vartheta_j} \log(C(\vartheta)) = -\int_M h_j(x) dP_\vartheta(x) = -E_\vartheta[h_j].$$

6. *Für $i, j = 1, \ldots, k$ ist*

$$\frac{\partial^2}{\partial \vartheta_i \partial \vartheta_j} \log(C(\vartheta)) = E_\vartheta[h_i] E_\vartheta[h_j] - E_\vartheta[h_i h_j] = -\mathrm{cov}_\vartheta(h_i, h_j)$$

eine negativ definite Matrix.

Beweis:

1. Folgt aus der Konvexität der e-Funktion.

2. Folgt aus Aufgabe 11.

3. Durch Übergang zu einem anderen Referenzmaß können wir zunächst annehmen, daß $0 \in \Theta$, denn ist $b \in \Theta$, so gilt

$$f_\vartheta(x) = C(\vartheta) \exp\left(\sum_{j=1}^k b_j x_j\right) \exp\left(\sum_{j=1}^k (z_j - b_j) x_j\right).$$

   Schreibe $\varphi = \varphi^+ - \varphi^-$, und bilde g^+, g^- wie in 2. mit $\varphi = \varphi^+$ bzw. $\varphi = \varphi^-$. Dann ist

$$g^+(z) = g^-(z) \quad \text{für alle } z \in \Theta. \tag{4.1}$$

   Aus $g^+(0) = g^-(0)$ folgt $\int \varphi^+ d\mu^* = \int \varphi^- d\mu^*$. Haben diese Integrale den gemeinsamen Wert 0, so folgt $\varphi = 0$ μ^*-f.s., und wir sind fertig. Sonst können wir durch Normalisierung von φ mit einer Konstanten annehmen, daß $\varphi^+ d\mu^*$ und $\varphi^- d\mu^*$ Wahrscheinlichkeitsmaße auf $\mathbf{R}^k$ sind.

   Durch k-fache Anwendung des Identitätssatzes für holomorphe Funktionen in einer Veränderlichen auf die Identität (4.1) folgt, daß $g^+(z) = g^-(z)$ für alle $z \in \Theta_C$, insbesondere also für alle rein imaginären $z = (it_1, \ldots, it_k)$, da $0 \in \Theta$. Also haben $\varphi^+ d\mu^*$ und $\varphi^- d\mu^*$ identische charakteristische Funktionen, und aus dem Eindeutigkeitssatz für charakteristische Funktionen (k-dimensionale Version) folgt $\varphi^+ d\mu^* = \varphi^- d\mu^*$, d.h. $\varphi^+ = \varphi^-$ μ^*-f.s., so daß $\varphi = 0$ μ^*-f.s.

4. Wie bei 2.

5. Da

$$C(\vartheta)^{-1} = \int_M \exp\left(\sum_{j=1}^k \vartheta_j h_j(x)\right) d\mu^*(x),$$

   folgt aus 4., daß

$$\begin{aligned} \frac{\partial}{\partial \vartheta_j} \log(C(\vartheta)) &= -C(\vartheta) \cdot \frac{\partial}{\partial \vartheta_j} \frac{1}{C(\vartheta)} \\ &= -C(\vartheta) \int_M h_j(x) dP_\vartheta(x). \end{aligned}$$

6. Aus 4. und 5. folgt

$$\begin{aligned} \frac{\partial^2}{\partial \vartheta_i \partial \vartheta_j} \log(C(\vartheta)) &= -\frac{\partial}{\partial \vartheta_i} C(\vartheta) \cdot \frac{1}{C(\vartheta)} E_\vartheta[h_j] - C(\vartheta) \cdot \frac{\partial}{\partial \vartheta_i} \left(\frac{1}{C(\vartheta)} E_\vartheta[h_j]\right) \\ &= E_\vartheta[h_i] E_\vartheta[h_j] - E_\vartheta[h_i h_j] = -\mathrm{cov}_\vartheta(h_i, h_j). \end{aligned}$$

Diese Matrix der 2. Ableitungen ist negativ definit, da das System $\{1, h_1, \ldots, h_k\} \subseteq L^2_{P_\vartheta}$ linear unabhängig ist, denn für $\lambda \in \mathbf{R}^k$ ist

$$-\sum_{i=1}^{k}\sum_{j=1}^{k}\lambda_i \mathrm{cov}_\vartheta(h_i,h_j)\lambda_j = -\mathrm{cov}_\vartheta(\sum_{i=1}^{k}\lambda_i h_i, \sum_{j=1}^{k}\lambda_j h_j) = -V_\vartheta(\sum_{i=1}^{k}\lambda_i h_i) \leq 0$$

mit Gleichheit genau dann, wenn $\sum_{i=1}^{k}\lambda_i h_i$ P_ϑ-f.s. konstant ist. □

4.2 Grundbegriffe der statistischen Entscheidungstheorie

Die statistische Entscheidungstheorie hat sich als wichtiges Hilfsmittel erwiesen, um bei gegebener Problemstellung die Eignung verschiedener Test- oder Schätzverfahren zu beurteilen. Das benutzte Modell sieht folgendermaßen aus:

$(M, \mathcal{M}, \{P_\vartheta : \vartheta \in \Theta\})$ sei ein statistischer Raum. Der Statistiker hat aufgrund einer Beobachtung $x \in M$ (i.a. wird das ein Beobachtungsvektor sein) eine Entscheidung $e = \delta(x) \in E$ zu treffen, wo E die Menge aller möglichen Entscheidungen ist. Man nimmt an, daß auf E eine σ-Algebra $\mathcal{E}$ erklärt ist und daß die **Entscheidungsfunktion** $\delta : M \to E$ $\mathcal{M}$-$\mathcal{E}$-meßbar ist.

Die Konsequenz einer Entscheidung (es kann ja eine „gute“ oder eine „schlechte“ Entscheidung sein) wird durch eine **Verlustfunktion** $L : \Theta \times E \to \mathbf{R}$ modelliert, die so interpretiert wird, daß, wenn die beobachtete Situation durch P_ϑ beschrieben wird und sich der Statistiker für e entscheidet, ein Verlust von $L(\vartheta, e)$ entsteht.

Um die Lösung von Optimierungsproblemen des Typs

> „Bei gegebenem $(M, \mathcal{M}, \mathcal{P})$ und gegebener Verlustfunktion L minimiere den Verlust $L(\vartheta, \delta(x))$ (in einem später zu präzisierenden Sinn) durch Wahl der Entscheidungsfunktion δ“

in möglichst vielen Fällen zu ermöglichen, wäre es wünschenswert, mit einem konvexen Raum von Entscheidungsfunktionen zu arbeiten. Daher erweitert man den Begriff der Entscheidungsfunktion zum Begriff der randomisierten Entscheidungsfunktion:

Definition 4.2.1 *Sei $(M, \mathcal{M}, \mathcal{P})$ ein statistischer Raum, $(E, \mathcal{E})$ ein meßbarer Raum. Ein stochastischer Kern $\delta(x, A)$ von $(M, \mathcal{M})$ nach $(E, \mathcal{E})$ heißt* **randomisierte Entscheidungsfunktion**. *Ist $L : \Theta \times E \to \mathbf{R}$ eine Verlustfunktion und δ eine randomisierte Entscheidungsfunktion, so wird mit*

$$R(\vartheta, \delta) := \int \left(\int L(\vartheta, e)\delta(x, de) \right) dP_\vartheta(x)$$

die zugehörige **Risikofunktion** *bezeichnet, falls dieses Integral wohldefiniert ist. Bezeichnet Δ die Menge der randomisierten Entscheidungsfunktionen mit wohldefinierter Risikofunktion, so ist $R : \Theta \times \Delta \to \mathbf{R}$.*

Nichtrandomisierte Entscheidungsfunktionen kann man als spezielle randomisierte Entscheidungsfunktionen interpretieren, indem man der nichtrandomisierten Entscheidungsfunktion $\delta(x)$ den Kern $\delta(x, A) = I_A(\delta(x))$ zuordnet. Daher unterscheiden wir notationsmäßig nicht zwischen diesen beiden Klassen. Ist $\delta \in \Delta$ nichtrandomisiert, so ist $R(\vartheta, \delta) = \int L(\vartheta, \delta(x)) dP_\vartheta(x)$.

Eine Entscheidung gemäß einer randomisierten Entscheidungsfunktion wird herbeigeführt, indem man die Entscheidung mit einem von den Beobachtungen unabhängigen Zufallsmechanismus mit Verteilung $\delta(x, .)$ „auswürfelt". Es widerspricht sicher dem „gesunden Menschenverstand", eine möglichst optimale Entscheidung dem Zufall zu überlassen, und in den meisten Fällen wird sich auch eine nichtrandomisierte Entscheidungsfunktion als am besten geeignet herausstellen, aber in manchen Fällen muß man randomisieren, um z.B. Tests mit exaktem vorgegebenen Niveau konstruieren zu können.

Beispiel 4.2.2 Gegeben sei ein statistischer Raum $(M, \mathcal{M}, \{P_\vartheta : \vartheta \in \mathbf{R}\})$.

1. Test von $H_0 : \vartheta \leq \vartheta_0$ gegen $H_1 : \vartheta > \vartheta_0$:

   $E = \{0, 1\}$, $\delta(x) = \phi(x)$, so daß $\delta(X) = 0$ Annahme von H_0 heißt. Eine naheliegende Verlustfunktion wäre

   $$L(\vartheta, 1) = \begin{cases} a & \text{für } \vartheta \leq \vartheta_0 \\ 0 & \text{für } \vartheta > \vartheta_0 \end{cases},$$
   $$L(\vartheta, 0) = \begin{cases} 0 & \text{für } \vartheta \leq \vartheta_0 \\ b & \text{für } \vartheta > \vartheta_0 \end{cases},$$

   wobei a und b positive reelle Zahlen sind. Die zugehörige Risikofunktion ist

   $$R(\vartheta, \phi) = \begin{cases} a \cdot P_\vartheta\{\phi(X) = 1\} & \text{für } \vartheta \leq \vartheta_0 \\ b \cdot P_\vartheta\{\phi(X) = 0\} & \text{für } \vartheta > \vartheta_0 \end{cases}.$$

   Ein Fehler 1.Art wird also mit der „Strafe" a belegt, ein Fehler 2.Art mit b.

2. Punktschätzung des unbekannten Parameters ϑ:

   $E = \Theta = \mathbf{R}$, $\delta(X) = T(X)$, wo $T(X)$ eine Schätzstatistik für ϑ ist. Als Verlustfunktion wählt man oft die **quadratische Verlustfunktion**

   $$L(\vartheta, e) = a(\vartheta - e)^2,$$

   wo $a > 0$. Dann ist

   $$R(\vartheta, T) = a \int_M (\vartheta - T(x))^2 dP_\vartheta(x).$$

3. Konfidenzschätzung des unbekannten Parameters ϑ:

$E = \{(a,b) : \ a,b \in \mathbf{R}, a < b\}$, $\delta(X) = J(X)$, wo $J(X)$ ein Konfidenzschätzer für ϑ ist.Als Verlustfunktion kann man z.B. wählen

$$L(\vartheta, J) = \begin{cases} 0 & \text{für } \vartheta \in J \\ a & \text{für } \vartheta \notin J \end{cases},$$

wo $a > 0$. Als Risikofunktion erhält man:

$$R(\vartheta, J) = a \cdot P_\vartheta\{\vartheta \notin J(X)\}.$$

Die Güte einer Entscheidungsfunktion wird nun allein nach ihrer Risikofunktion beurteilt. Da aber für $\delta_1, \delta_2 \in \Delta$ nur in den seltensten Fällen $R(\vartheta, \delta_1) \leq R(\vartheta, \delta_2)$ für alle $\vartheta \in \Theta$ gilt (wir sagen dann: δ_1 ist **mindestens so gut** wie δ_2), ist der Vergleich zweier Risikofunktionen nicht ganz unproblematisch. Wir erwähnen die folgenden Kriterien:

- **Zulässigkeit:** Ist δ_1 mindestens so gut wie δ_2 und gibt es mindestens ein $\vartheta \in \Theta$, für das $R(\vartheta, \delta_1) < R(\vartheta, \delta_2)$, so heißt δ_1 **besser** als δ_2. $\delta_0 \in \Delta$ heißt **zulässig**, falls es kein besseres $\delta \in \Delta$ gibt.

- **Minimax-Prinzip:** δ_1 wird δ_2 vorgezogen, falls

$$\sup_{\vartheta \in \Theta} R(\vartheta, \delta_1) < \sup_{\vartheta \in \Theta} R(\vartheta, \delta_2).$$

  $\delta_0 \in \Delta$ heißt **Minimax- Entscheidungsfunktion**, falls

$$\sup_{\vartheta \in \Theta} R(\vartheta, \delta_0) = \inf_{\delta \in \Delta} \sup_{\vartheta \in \Theta} R(\vartheta, \delta).$$

  Man versucht also, das Risiko im schlimmsten Fall möglichst gering zu halten.

- **Einschränkung auf Teilklassen von Entscheidungsfunktionen:** Um die Klasse der in Betracht zu ziehenden Entscheidungsfunktionen zu verkleinern, kann man sich z.B. auf **erwartungstreue** Entscheidungsfunktionen einschränken. Das sind solche, für die

$$R(\vartheta, \delta) = \int \left(\int L(\vartheta, e) \delta(x, de) \right) dP_\vartheta(x) \leq \int \left(\int L(\vartheta', e) \delta(x, de) \right) dP_\vartheta(x)$$

  für alle $\vartheta, \vartheta' \in \Theta$.

  Die Einschränkung auf erwartungstreue Entscheidungsfunktionen schließt bei der Suche nach **gleichmäßig besten** Entscheidungsfunktionen δ (das sind solche, für die δ mindestens so gut wie jedes andere zur Konkurrenz zugelassene δ') z.B. die folgenden pathologischen Konkurrenten aus, die die Existenz einer gleichmäßig besten Entscheidungsfunktion verhindern würden: Gibt es zu jedem $\vartheta \in \Theta$ ein $e_\vartheta \in E$ mit $L(\vartheta, e_\vartheta) = 0$, so betrachte die Entscheidungsfunktionen $\delta_\vartheta(x) := e_\vartheta$. Dann ist $R(\vartheta, \delta_\vartheta) = 0$ für alle $\vartheta \in \Theta$, und ein gleichmäßig bestes δ müßte $R(\vartheta, \delta) = 0$ für alle $\vartheta \in \Theta$ haben, eine unsinnige Forderung.

- **Bayes Prinzip:** Man versieht Θ mit einer σ-Algebra $\mathcal{G}$ und einer Wahrscheinlichkeitsverteilung π (der **a-priori-Verteilung**) und ordnet jedem $\delta \in \Delta$ das **Bayes Risiko**

$$r_\pi(\delta) = \int R(\vartheta, \delta)\ d\pi(\vartheta)$$

zu. Ein $\delta_0 \in \Delta$, das dieses Risiko minimiert, heißt **Bayes Entscheidungsfunktion**.

Wie in [10, Abschnitt II.7.7] ausgeführt, kann man unter entsprechenden Meßbarkeitsannahmen $P(\vartheta, .) := P_\vartheta(.)$ als stochastischen Kern (reguläre bedingte Wahrscheinlichkeitsverteilung) von $(\Theta, \mathcal{G})$ nach $(M, \mathcal{M})$ auffassen. Sei $Q := \pi \times P$ und $\bar{P}(A) := Q(\Theta \times A)$ für $A \in \mathcal{M}$. Ist nun z.B. $\Theta = \mathbf{R}^d$ und $M = \mathbf{R}^n$, so existieren die regulären bedingten Wahrscheinlichkeitsverteilungen Q_x auf $(\Theta, \mathcal{G})$, und man erhält

$$\begin{aligned} r_\pi(\delta) &= \int_\Theta \int_M \left(\int_E L(\vartheta, e)\delta(x, de) \right) dP_\vartheta(x) d\pi(\vartheta) \\ &= \int_{\Theta \times M} \left(\int_E L(\vartheta, e)\delta(x, de) \right)\ dQ(\vartheta, x) \\ &= \int_M \left(\int_\Theta (\int_E L(\vartheta, e)\delta(x, de))\ dQ_x(\vartheta) \right) d\bar{P}(x) \\ &= \int_M \left(\int_E (\int_\Theta L(\vartheta, e)\ dQ_x(\vartheta))\delta(x, de) \right) d\bar{P}(x). \end{aligned}$$

Daraus erhält man eine Bayes Entscheidungsfunktion $\delta(x, .)$, indem man für jedes $x \in M$

$$\int_E (\int_\Theta L(\vartheta, e)\ dQ_x(\vartheta))\delta(x, de)$$

durch Wahl von $\delta(x, .)$ minimiert.

Ist $\Theta = E = \mathbf{R}$ und $L(\vartheta, e) = (\vartheta - e)^2$, so wird $\int_\Theta L(\vartheta, e)\ dQ_x(\vartheta)$ durch Wahl von $e_x = \int_R \vartheta\ dQ_x(\vartheta)$ eindeutig minimiert, und man kommt zu der nichtrandomisierten Entscheidungsfunktion $\delta(x) = e_x$.

Insbesondere muß zur Bestimmung einer Bayes Entscheidungsfunktion die Risikofunktion nicht explizit berechnet werden, denn in das zu minimierende Integral gehen nur die Verlustfunktion und die bedingte Verteilung $dQ_x(\vartheta)$ ein. Q_x heißt die **a-posteriori-Verteilung** nach Beobachtung von x.

Man kann Bayes'sche Statistik folgendermaßen interpretieren: Das Vorwissen des Experimentators wird durch die a-priori-Verteilung π repräsentiert, und nach der Beobachtung von x wird sein nun modifiziertes Wissen durch die a-posteriori-Verteilung Q_x beschrieben. In Abschnitt 5.3 werden wir darauf noch einmal zurückkommen.

4.3 „Frequentismus“ versus „Bayesianismus“

Die letzte Bemerkung des vorherigen Abschnitts beleuchtet den wesentlichen Unterschied zwischen sogenannten Frequentisten, die sich beim Vergleich von Risikofunktionen vornehmlich an den ersten drei Kriterien orientieren, und Bayesianern, die nach Bayes-Regeln suchen. (Die Formel für die a-posteriori-Verteilung wurde ursprünglich um 1750 von dem Reverend Thomas Bayes, einem anglikanischen Geistlichen, entdeckt.)

Der **Frequentist** stützt sich auf Vergleichskriterien für Risikofunktionen, die die explizite Bestimmung der Risikofunktion voraussetzen. Das beinhaltet eine Integration mit dP_ϑ, also eine Mittelung der Verlustfunktion über alle potentiell möglichen Beobachtungen. Die Interpretation eines Wertes $R(\vartheta, \delta)$ der Risikofunktion lautet dann: Wenn 1000 Experimentatoren unabhängig voneinander das gleiche Experiment durchführen (d.h. mit voneinander unabhängig erhobenen Daten) und die Entscheidungsfunktion δ benutzen, so ist ihr mittlerer Verlust bei vorliegen von ϑ ungefähr $R(\vartheta, \delta)$. Die „Minimierung“ des Risikos nach einem der ersten drei beschriebenen Kriterien bedeutet also, den mittleren Verlust über alle potentiell möglichen Versuchsausgänge zu „minimieren“.

Der **Bayesianer** dagegen verbietet sich die Vorstellung, daß ein Experiment viele Male unabhängig wiederholt werden oder von vielen Experimentatoren gleichzeitig unabhängig voneinander durchgeführt werden kann, da das in vielen Fällen völlig unrealistisch ist. Diese Vorstellung mag zwar bei Experimenten in den Natur- und Ingenieurwisenschaften und bei kontrollierten klinischen Studien der Realität nahekommen, bei sozialwissenschaftlichen Erhebungen oder medizinischen Feldstudien muß man aber von einem einzigen, nicht wiederholbaren Experiment ausgehen. Man denke nur an eine Fragestellung wie die Untersuchung des Zusammenhangs zwischen Einkommen und Wohnungsgröße bei Familien in der Bundesrepublik. Durch das Arbeiten mit dem Bayes-Risiko vermeidet der Bayesianer den Rückgriff auf zwar potentiell beobachtbare aber nicht tatsächlich beobachtete Daten bei der Auswahl seiner Entscheidungsfunktion, denn er muß $\int_\Theta L(\vartheta, e) dQ_X(\vartheta)$ nur für das tatsächlich beobachtete X auswerten.

Das folgende Beispiel findet man in dem sehr lesenswerten Artikel “Controversies in the Foundations of Statistics“ von B.Efron [4]:

Wir gehen von der Vorstellung aus, daß jeder Mensch einen gewissen I.Q. Wert μ hat (was immer mit dieser Maßzahl auch tatsächlich gemessen wird). Der I.Q. ist eine positive relle Zahl, und ist idealerweise so skaliert, daß gerade die Hälfte der beschriebenen Population einen I.Q. ≤ 100 hat.

Nimmt eine Person mit I.Q. μ an einem I.Q. Test teil, so ist das Ergebnis X ein Wert, der mehr oder weniger nahe bei dem wahren I.Q. μ liegt. In unserem Modell sei X nach $\mathcal{N}(\mu, \sigma^2)$ verteilt, wo wir σ^2 als einen festen, nicht von der speziellen Person abhängenden Parameter ansehen. Der Frequentist wird nicht zögern, den Wert X als Schätzer für den wahren I.Q. zu verwenden. X ist erwartungstreu (d.h. $E_\mu[X] = \mu$), und wir werden im nächsten Kapitel sehen, daß er unter allen erwartungstreuen Schätzern gleichmäßig kleinstes Risiko bei quadratischer Verlustfunktion hat.

Der Bayesianer bezieht in seine Schätzung dagegen das Wissen um die Verteilung der I.Q.'s in der Gesamtbevölkerung mit ein. Aus umfangreichen Untersuchungen sei ihm bekannt, daß die I.Q.'s μ der Gesamtbevölkerung ungefähr nach $\mathcal{N}(m, s^2)$ verteilt sind. (In den U.S.A. hat man etwa $m = 100$ und $s = 15$.) Die a-posteriori-Verteilung Q_X von μ errechnet man leicht als

$$\mathcal{N}(m + C(X - m), D), \quad \text{wo } C = \frac{1/\sigma^2}{1/s^2 + 1/\sigma^2} \text{ und } D = \frac{1}{1/s^2 + 1/\sigma^2}.$$

Bei $X = 140$ und $\sigma = 7.5$ erhält man z.B. $Q_X = \mathcal{N}(132, (6.7)^2)$, und der Bayesianer wird μ als Erwartungswert von Q_X schätzen, in unserem Beispiel also als 132. Dieser Wert liegt deutlich unter $X = 140$, da die Wahrscheinlichkeit, eine Testperson mit I.Q. 120 zu erwischen, die zufällig einen Testwert von 140 produziert, viel größer ist als die, eine Testperson mit I.Q. 160 zu erwischen, die zufällig einen Testwert von 140 produziert. Dabei ist die individuelle Wahrscheinlichkeit für einen Testwert von 140 bei beiden Personen die gleiche, nur gibt es viel weniger Personen mit I.Q. 160 als mit I.Q. 120.

Problematisch am Bayesschen Ansatz ist die Willkür, die in der Wahl der a-priori-Verteilung liegt. Ist der zugrunde liegende Parameterraum ein endliches Intervall, zum Beispiel $(0, 1)$ im Falle der Bernoulli Verteilungen, so ist es nur natürlich, völliges Nichtwissen über den Parameter durch die Wahl der Gleichverteilung als a-priori-Verteilung auszudrücken. Variiert aber, wie in unserem Beispiel, der unbekannte Parameter z.B. über $\mathbf{R}$, so gibt es keine geeignete „uniforme" a-priori-Verteilung. Man benötigt also ein Vorwissen über die Verteilung von μ, was hier durch vorherige umfangreiche Studien erworben wurde, i.a. aber nicht zur Verfügung steht.

Der Vorteil des Bayesschen gegenüber dem frequentistischen Ansatz liegt in der größeren logischen Konsistenz. Als Beispiel für die Schwierigkeiten, in die ein Frequentist (nicht aber ein Bayesianer) geraten kann, stehe die folgende Fortsetzung des I.Q. Beispiels, ebenfalls aus dem Artikel von Efron:

Nachdem der Frequentist seine Schätzung $X = 140$ für den unbekannten I.Q. der Testperson abgegeben hat, wird ihm von dem Institut, das mit der Durchführung der I.Q. Tests betraut ist, folgendes mitgeteilt:

> „An dem Tag, als wir Ihnen das Testergebnis von $X = 140$ mitteilten, funktionierte unsere Maschine zur automatischen Auswertung der Testbögen nicht korrekt. Alle Testergebnisse mit Werten ≤ 100 wurden als $= 100$ übermittelt. Für Testergebnisse über 100 traten jedoch keine Übermittlungsfehler auf."

Man beachte, daß dem Frequentisten der korrekte Wert übermittelt wurde, so daß sich seine Datenlage durch diese Nachricht nicht ändert. Trotzdem hat er jetzt das folgende Problem:

Bezeichne $P_\mu = \mathcal{N}(\mu, \sigma^2)$. Der dem Statistiker übermittelte Wert ist eine Realisierung der Zufallsvariablen

$$Y = \max\{X, 100\} \geq X,$$

und da $P_\mu(Y > X) = P_\mu(X > 100) > 0$ für alle $\mu \in \mathbf{R}$, folgt

$$E_\mu[Y] > E_\mu[X] = \mu.$$

Y ist also kein erwartungstreuer Schätzer für μ, und damit kann man aus frequentistischer Sicht nicht mehr sicher sein, einen in einem wohldefinierten Rahmen optimalen Schätzer gewählt zu haben. Der Frequentist müßte nach der obigen Mitteilung also seinen Schätzer um einen Term modifizieren, der die Verzerrung des Schätzers nach oben aufhebt, falls das überhaupt möglich ist.

Der Bayesianer hat dagegen keine Probleme. Seine Schätzung beruht nur auf der a-posteriori-Verteilung von μ, gegeben die Beobachtung, d.h. auf $\pi_X(B) = Q(B \times M|X)$ bzw. $\pi_Y(B) = Q(B \times M|Y)$ $(B \in \mathcal{G})$, wo $Q = \pi \times P$ wie im vorangegangenen Abschnitt eine Wahrscheinlichkeitsverteilung auf $(\Theta \times M, \mathcal{G} \times \mathcal{M})$ ist. Wir werden zeigen, daß für alle $B \in \mathcal{G}$ gilt $\pi_X(B) = \pi_Y(B)$ fast sicher auf $\{X = Y > 100\}$. Also muß der Bayesianer seine Schätzung aufgrund der obigen Nachricht nicht ändern. Sie ist noch genauso gut begründet wie vorher, denn sie hängt nur von den tatsächlich beobachteten Daten ab.

Beweis von "$\pi_X = \pi_Y$ fast sicher auf $\{X = Y > 100\}$": Sei $f = I_{B \times M}$. Dann ist

$$\pi_X(B) = E_Q[f|X] \quad \text{und} \quad \pi_Y(B) = E_Q[f|Y] \; Q\text{-f.s..}$$

Ist $A \in \sigma(X)$, $A \subseteq \{X > 100\} = \{Y > 100\}$, so ist $A \in \sigma(Y)$, so daß

$$\int_A \pi_Y(B)\, dQ = \int_A f\, dQ = \int_A \pi_X(B)\, dQ.$$

Da $\pi_Y(B)$ und $\pi_X(B)$ $\sigma(X)$-meßbar sind und da $A \in \sigma(X)$ mit $A \subseteq \{X > 100\} = \{Y > 100\}$ beliebig war, folgt $\pi_Y(B) = \pi_X(B)$ Q-f.s. auf $\{X = Y > 100\}$.

4.4 Suffizienz

In den vorangegangenen Kapiteln haben wir gesehen, daß für die Konstruktion vieler Tests und Schätzer nicht die explizite Kenntnis der gesamten Stichprobe erforderlich ist, sondern daß sie oft nur die Kenntnis des Wertes einer auf den Daten basierenden Statistik erfordern. Es stellt sich die Frage, wie weit man die in den Daten enthaltene Information reduzieren darf, ohne wesentliches aufzugeben. Das führt auf den Begriff der Suffizienz, dessen Grundidee bis auf R.A. Fisher (1925) zurückgeht.

Definition 4.4.1 *Sei $(M, \mathcal{M}, \mathcal{P})$ ein statistischer Raum.*

1. *Eine Teil-σ-Algebra $\mathcal{F} \subseteq \mathcal{M}$ heißt* **suffizient**, *wenn es für jedes $A \in \mathcal{M}$ eine $\mathcal{F}$-meßbare Funktion $p_{\mathcal{F}}(x, A)$ gibt, die für alle $P \in \mathcal{P}$ eine Version der bedingten Wahrscheinlichkeit $P(A|\mathcal{F})(x)$ darstellt.*

2. *Eine Statistik $T : (M, \mathcal{M}) \to (D, \mathcal{D})$ heißt* **suffizient**, *falls $T^{-1}(\mathcal{D})$ suffizient ist.*

Suffizienz ist innerhalb eines Modells $(M, \mathcal{M}, \mathcal{P})$ definiert, d.h. bei einem inadäquaten Modell kann durch Übergang von $X_1, \ldots, X_n$ zu $T(X_1, \ldots, X_n)$ relevante Information verloren gehen. Z.B. kann in einer Stichprobe $X_1, \ldots, X_n$ noch eventuell ein „Ausreißer" erkannt werden, in $T(X_1, \ldots, X_n) = \bar{X}$ aber i.a. nicht mehr.

Definition 4.4.2 *Sei $(M, \mathcal{M}, \mathcal{P})$ ein statistischer Raum. Zwei auf $(M, \mathcal{M})$ definierte Statistiken S und T heißen äquivalent, falls sie die gleiche σ-Algebra erzeugen.*

Bemerkung 4.4.3

1. Sind die Statistiken S und T äquivalent, so ist eine suffizient genau dann, wenn die andere suffizient ist.
2. Zum Nachweis der Äquivalenz von S und T ist die folgende triviale Bemerkung oftmals nützlich: $\sigma(S) \subseteq \sigma(T)$, falls es eine meßbare Abbildung Ψ gibt, so daß $S = \Psi \circ T$.

Bemerkung 4.4.4 Ist $(M, \mathcal{M}) = (\mathbf{R}, \mathcal{B})$, so können die $p_{\mathcal{F}}(x, A)$ so gewählt werden, daß $p_{\mathcal{F}}(x, .)$ für jedes $x \in M$ ein Wahrscheinlichkeitsmaß auf $(M, \mathcal{M})$ ist. Wir schreiben dann $P_{\mathcal{F}}(x, A)$ statt $p_{\mathcal{F}}(x, A)$ und betrachten $P_{\mathcal{F}}(x, A)$ als stochastischen Kern von $(M, \mathcal{F})$ nach $(M, \mathcal{M})$.

Das zeigt man ähnlich wie in [10, Abschnitt II.7.7] oder in [2, Theorem 33.3], wo der Fall einer einelementigen Familie $\mathcal{P}$ behandelt wird. Die notwendigen Modifikationen des Beweises sind z.B. im Beweis zu Satz 3.15 in [12] beschrieben.

Diese Bemerkung bleibt richtig, wenn $(M, \mathcal{M})$ **Borelsch** ist, d.h. wenn es eine bijektive, bimeßbare Abbildung von $(M, \mathcal{M})$ auf eine Borelsche Teilmenge von $\mathbf{R}$ (versehen mit der von $(\mathbf{R}, \mathcal{B})$ induzierten σ-Algebra) gibt. Z.B. ist der $\mathbf{R}^n$, $n \in \mathbf{N}$, Borelsch, siehe die Folgerung zu [10, Satz II.7.5].

Der folgende Satz zeigt, daß unter gewissen Meßbarkeitsannahmen durch Reduktion der Daten mit einer suffizienten Statistik kein Verlust an Entscheidungsfähigkeit im Sinne der Risikotheorie auftritt.

Theorem 4.4.5 (Satz von Bahadur) *Sei $(M, \mathcal{M}, \{P_\vartheta : \vartheta \in \Theta\})$ ein statistischer Raum, $(M, \mathcal{M})$ Borelsch und $\mathcal{F} \subseteq \mathcal{M}$ suffizient. Ist $\delta \in \Delta$ eine randomisierte Entscheidungsfunktion, so hat die $\mathcal{F}$-meßbare randomisierte Entscheidungsfunktion*

$$\begin{aligned} \delta^*(x_0, D) &:= (P_{\mathcal{F}} \times \delta)(x_0, M \times D), \quad \textit{also} \\ \delta^*(x_0, D) &= \int_M \delta(x, D)\ P_{\mathcal{F}}(x_0, dx) \quad (x_0 \in M, D \in \mathcal{E}) \end{aligned}$$

bei gegebener Verlustfunktion L eine Risikofunktion, die mit der von δ identisch ist.

Ist δ erwartungstreu, so auch δ^*.

Bemerkung 4.4.6 Der Satz bleibt richtig, wenn statt $(M, \mathcal{M})$ der Entscheidungsraum $(E, \mathcal{E})$ Borelsch ist. Man definiert dann

$$\delta^*(x_0, D) := E_\vartheta[\delta(., D)|\mathcal{F}](x_0)$$

und zeigt, daß dafür eine von ϑ unabhängige Festsetzung existiert, die in D σ-additiv ist. Eine Beweisskizze findet man in [7, S.97 ff.]. Ist $(M, \mathcal{M})$ Borelsch, so stimmt diese Definition von δ^* mit der im Satz von Bahadur gegebenen überein.

Beweis des Satzes von Bahadur: Als Produkt zweier stochastischer Kerne ist $P_{\mathcal{F}} \times \delta$ ein stochastischer Kern, und als dessen Projektion auf eine Koordinate ist auch δ^* ein stochastischer Kern, also eine randomisierte Entscheidungsfunktion.

Sei

$$\begin{aligned} \pi_2 : M \times M &\to M, & (x,y) &\mapsto y, \\ \pi_{1,3} : M \times M \times E &\to M \times E, & (x,y,e) &\mapsto (x,e) \end{aligned}$$

und $L(\vartheta, e)$ die gegebene Verlustfunktion. Beachte, daß

$$\begin{aligned} (P_\vartheta \times P_{\mathcal{F}} \times \delta) \circ \pi_{1,3}^{-1} &= P_\vartheta \times \delta^* \quad \text{und} \\ (P_\vartheta \times P_{\mathcal{F}}) \circ \pi_2^{-1} &= P_\vartheta. \end{aligned}$$

Letzteres folgt aus

$$(P_\vartheta \times P_{\mathcal{F}}) \circ \pi_2^{-1}(A) = \int_M P_{\mathcal{F}}(x, A) dP_\vartheta(x) = \int_M P_\vartheta(A|\mathcal{F})(x) dP_\vartheta(x) = P_\vartheta(A)$$

für alle $A \in \mathcal{M}$. Damit erhält man

$$\begin{aligned} & \int_M \left(\int_E L(\vartheta', e) \delta^*(x, de) \right) dP_\vartheta(x) \\ = & \int_{M \times E} L(\vartheta', e) d(P_\vartheta \times \delta^*)(x, e) \\ = & \int_{M \times M \times E} L(\vartheta', e) d(P_\vartheta \times P_{\mathcal{F}} \times \delta)(x, y, e) \\ = & \int_{M \times M} \left(\int_E L(\vartheta', e) \delta(y, de) \right) d(P_\vartheta \times P_{\mathcal{F}})(x, y) \\ = & \int_M \left(\int_E L(\vartheta', e) \delta(y, de) \right) dP_\vartheta(y). \end{aligned}$$

Für $\vartheta = \vartheta'$ ist das gerade die Gleichheit der Risikofunktionen, und aus der Identität für beliebige ϑ, ϑ' folgt, daß δ^* erwartungstreu ist, genau dann wenn δ es ist. □

Ein wichtiges Hilfsmittel zur Beantwortung der Frage, ob eine gegebene Statistik suffizient ist, wird durch das sogenannte **Faktorisierungskriterium** von Neyman und Fisher bereitgestellt. Zu seiner Vorbereitung beweisen wir zunächst

Lemma 4.4.7 *Seien $S : (M, \mathcal{M}) \to (C, \mathcal{C})$ und $T : (M, \mathcal{M}) \to (D, \mathcal{D})$ Statistiken und μ ein Wahrscheinlichkeitsmaßauf $(M, \mathcal{M})$. Für alle $c \in C$ sei $\{c\} \in \mathcal{C}$. Genau dann ist $\sigma(S) \subseteq \sigma(T) \bmod \mu$, wenn es eine meßbare Abbildung $\Psi : D \to C$ gibt, so daß $S = \Psi \circ T$ μ-f.s.*

Beweis: „$\Leftarrow$“: $\sigma(S) = S^{-1}(\mathcal{C}) = T^{-1}(\Psi^{-1}\mathcal{C}) \bmod \mu$ und $T^{-1}(\Psi^{-1}\mathcal{C}) \subseteq T^{-1}(\mathcal{D}) = \sigma(T)$. „$\Rightarrow$“: Für $c \in C$ sei $\mathcal{U}_c = \{U \in \mathcal{D} : T^{-1}U = S^{-1}\{c\} \bmod \mu\}$. Da $S^{-1}\{c\} \in \sigma(S) \subseteq \sigma(T) \bmod \mu$, ist $\mathcal{U}_c \neq \emptyset$ für alle $c \in C$. Durch ein Ausschöpfungsargument findet man eine Menge $U_c \in \mathcal{U}_c$ mit $\mu(T^{-1}U_c) = \sup\{\mu(T^{-1}U) : U \in \mathcal{U}_c\}$.

Theorem 4.4.8 (Halmos/Savage) *Sei $(M, \mathcal{M}, \mathcal{P})$ durch ein zu $\mathcal{P}$ äquivalentes Wahrscheinlichkeitsmaß $\nu = \sum_{i=1}^{\infty} c_i P_i$ wie in Theorem (4.1.3) dominiert. Dann gilt:*

1. *Eine Teil-σ-Algebra $\mathcal{F} \subseteq \mathcal{M}$ ist suffizient genau dann, wenn es für alle $P \in \mathcal{P}$ $\mathcal{F}$-meßbare Dichten f_P gibt, so daß $\frac{dP}{d\nu} = f_P$ fast sicher.*

2. *Eine Statistik $T : (M, \mathcal{M}) \to (D, \mathcal{D})$ ist suffizient genau dann, wenn es für alle $P \in \mathcal{P}$ eine meßbare Abbildung $g_P : D \to \mathbf{R}$ gibt, so daß $\frac{dP}{d\nu} = g_P \circ T$ fast sicher.*

Beweis: Aussage 2 folgt aus Aussage 1 mit dem Faktorisierungslemma III/14 aus MS II. Wir beweisen nur Aussage 1:
„$\Rightarrow$“
Sei $\nu = \sum_{i=1}^{\infty} c_i P_i \equiv \mathcal{P}$. Für $A \in \mathcal{M}$ und $F \in \mathcal{F}$ gilt

$$\nu(A \cap F) = \sum_{i=1}^{\infty} c_i P_i(A \cap F) = \sum_{i=1}^{\infty} c_i \int_F p_{\mathcal{F}}(x, A) dP_i(x) = \int_F p_{\mathcal{F}}(x, A) d\nu(x),$$

d.h. $p_{\mathcal{F}}(., A)$ ist auch eine Version von $\nu(A|\mathcal{F})$.
Setze

$$f_P := \frac{dP_{|\mathcal{F}}}{d\nu_{|\mathcal{F}}} \quad (\mathcal{F}\text{-meßbar}).$$

Wir zeigen, daß $f_P = dP/d\nu$ fast sicher: Für $A \in \mathcal{M}$ ist

$$\begin{aligned} P(A) &= \int_M P(A|\mathcal{F}) dP = \int_M p_{\mathcal{F}}(x, A) dP(x) \\ &= \int_M p_{\mathcal{F}}(x, A) dP_{|\mathcal{F}}(x) = \int_M \nu(A|\mathcal{F}) dP_{|\mathcal{F}} \\ &= \int_M \nu(A|\mathcal{F}) f_P d\nu_{|\mathcal{F}} = \int_M \nu(A|\mathcal{F}) f_P d\nu \\ &= \int_A f_P d\nu. \end{aligned}$$

„$\Leftarrow$“
Sei $f_P = dP/d\nu$ $\mathcal{F}$-meßbar, und für $A \in \mathcal{M}$ sei $p_{\mathcal{F}}(x, A)$ eine Festlegung von $\nu(A|\mathcal{F})(x)$. Zu zeigen: Für jedes $P \in \mathcal{P}$ ist $p_{\mathcal{F}}(x, A)$ auch eine Version von $P(A|\mathcal{F})(x)$. Für $F \in \mathcal{F}$ und $P \in \mathcal{P}$ gilt:

$$\begin{aligned} \int_F p_{\mathcal{F}}(x, A) dP(x) &= \int_F \nu(A|\mathcal{F}) f_P d\nu = \int E_\nu[I_F \cdot I_A \cdot f_P|\mathcal{F}] d\nu \\ &= \int_{A \cap F} f_P d\nu = P(A \cap F). \end{aligned}$$

□

Theorem 4.4.9 (Neymansches Faktorisierungskriterium) *Sei $(M, \mathcal{M}, \mathcal{P})$ durch μ dominiert. Dann gilt:*

1. *$\mathcal{F} \subseteq \mathcal{M}$ ist suffizient genau dann, wenn es Versionen der $dP/d\mu(x)$ von der Form $f_P(x)h(x)$ gibt, wobei f_P $\mathcal{F}$-meßbar und h $\mathcal{M}$-meßbar (und von P unabhängig) ist.*

2. *Eine Statistik $T : (M, \mathcal{M}) \to (D, \mathcal{D})$ ist suffizient genau dann, wenn es Versionen der $dP/d\mu(x)$ von der Form $g_P(T(x)) \cdot h(x)$ gibt, wobei $g_P : D \to \mathbf{R}$ $\mathcal{D}$-meßbar und h $\mathcal{M}$-meßbar (und von P unabhängig) ist.*

Beweis: „$\Rightarrow$“
Sei ν wie im Satz von Halmos/Savage, $h := d\nu/d\mu$ (vergl. Theorem 4.1.3). Dann ist

$$\frac{dP}{d\mu} = \frac{dP}{d\nu} \cdot h,$$

und die Behauptung folgt aus dem Satz von Halmos/Savage.
„$\Leftarrow$“

Wir zeigen nur Aussage 1. Die Aussage 2. folgt daraus wie im Beweis des Satzes von Halmos/Savage.

Sei $dP/d\mu = f_P \cdot h$ für alle $P \in \mathcal{P}$ und $\nu = \sum_{i=1}^{\infty} c_i P_i \equiv \mathcal{P}$ wie im Satz von Halmos/Savage. Setze $f_\nu := \sum_{i=1}^{\infty} c_i f_{P_i}$. Da

$$P_i\{f_\nu = 0\} \leq P_i\{f_{P_i} = 0\} = \int_{\{f_{P_i}=0\}} f_{P_i} h \, d\mu = 0 \quad \forall i,$$

folgt $\nu\{f_\nu = 0\} = 0$. Daher ist für alle $A \in M$ und $P \in \mathcal{P}$:

$$\int_A \frac{f_P}{f_\nu} d\nu = \sum_{i=1}^{\infty} c_i \int_A \frac{f_P}{f_\nu} dP_i = \sum_{i=1}^{\infty} c_i \int_A \frac{f_P}{f_\nu} f_{P_i} h \, d\mu = \int_A \frac{f_P}{f_\nu} f_\nu h \, d\mu = P(A),$$

also

$$\frac{dP}{d\nu} = \frac{f_P}{\sum_{i=1}^{\infty} c_i f_{P_i}}$$

ν-fast sicher, und der letzte Ausdruck ist nach Voraussetzung $\mathcal{F}$-meßbar. Die Suffizienz folgt nun aus dem Satz von Halmos/Savage. □

Korollar 4.4.10 *Ist $\{P_\vartheta : \vartheta \in \Theta\}$ eine exponentielle Familie mit Dichten*

$$f_\vartheta(x) = C(\vartheta) \exp\left(\sum_{j=1}^{k} c_j(\vartheta) h_j(x)\right) \cdot h(x),$$

so ist die Statistik $T = (h_1, \ldots, h_k)$ suffizient.

Beispiel 4.4.11 (vergl Beispiel 4.1.6.)

1. Für die Familie $\mathcal{P} = \{b(p,n) : \ 0 < p < 1\}$ ist $K_n =$„Anzahl der 1-en“ suffizient.

2. Für die Familie $\mathcal{P} = \{(\mathcal{N}(\mu,\sigma^2))^{\times n} : \mu \in \mathbf{R}, \sigma^2 > 0\}$ ist $(\bar{X}, S_X^2)$ suffizient, denn

$$\begin{aligned} f_{\mu,\sigma^2}(x_1,\ldots,x_n) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp(-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp(-\frac{n\mu}{2\sigma^2}) \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^n x_i\right), \end{aligned}$$

   woraus die Suffizienz von $T(X_1,\ldots,X_n) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ folgt. Da diese Statistik zu $(\bar{X}, S_X^2)$ äquivalent ist, folgt die Behauptung aus Bemerkung 4.4.3.

Theorem 4.4.12 *Ist $\{P_\vartheta : \vartheta \in \Theta\}$ eine vertauschbare Familie von Wahrscheinlichkeitsverteilungen auf $\mathbf{R}^n$, so ist die Statistik $g_{(n)}(x_1,\ldots,x_n) = (x_n^{(1)},\ldots,x_n^{(n)})$ suffizient.*

Beweis: Nach Theorem 3.2.1 ist $\mathcal{F} := \sigma(g_{(n)}) = \{B \in \mathcal{B}^n : \ \tilde{\pi}(B) = B \ \forall \pi \in \Pi_n\}$. Setze

$$p_{\mathcal{F}}(x,A) = \frac{1}{n!}\sum_{\pi\in\Pi_n} I_{\tilde{\pi}(A)}(x) = \frac{1}{n!}\sum_{\pi\in\Pi_n} I_A(\tilde{\pi}(x)).$$

Da $p_{\mathcal{F}}(\tilde{\pi}(x),A) = p_{\mathcal{F}}(x,A)$ für alle $\pi \in \Pi_n$, ist $p_{\mathcal{F}}(.,A)$ $\mathcal{F}$-meßbar.

Seien nun $F \in \mathcal{F}$ und $\vartheta \in \Theta$ beliebig. Es gilt:

$$\begin{aligned} \int_F p_{\mathcal{F}}(x,A)dP_\vartheta(x) &= \frac{1}{n!}\sum_{\pi\in\Pi_n}\int_F I_{\tilde{\pi}(A)}dP_\vartheta \\ &= \frac{1}{n!}\sum_{\pi\in\Pi_n} P_\vartheta(\tilde{\pi}(F\cap A)), \quad \text{da } \tilde{\pi}(F) = F, \\ &= \frac{1}{n!}\sum_{\pi\in\Pi_n} P_\vartheta(F\cap A), \quad \text{da } P_\vartheta\circ\tilde{\pi} = P_\vartheta, \\ &= P_\vartheta(F\cap A). \end{aligned}$$

□

Wir betrachten nun die Statistik

$$U(X_1,\ldots,X_n) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2, \ldots, \sum_{i=1}^n X_i^n).$$

Theorem 4.4.13 *U ist äquivalent zu $g_{(n)}$, insbesondere also auch suffizient für jede Familie $\{P_\vartheta : \vartheta \in \Theta\}$ wie im vorhergehenden Theorem.*

Beweis: Wir betrachten zunächst die Statistik

$$S(x_1, \ldots, x_n) = (s_1(x_1, \ldots, x_n), \ldots, s_n(x_1, \ldots, x_n)),$$

wo die s_j die elementarsymmetrischen Polynome in $x_1, \ldots, x_n$ sind,

$$s_1 = \sum_i x_i,\ s_2 = \sum_{i<j} x_i x_j, \ldots, s_n = x_1 \cdot \ldots \cdot x_n.$$

Die $s_j(x_1, \ldots, x_n)$ bestimmen sich eindeutig aus

$$f(x) := \prod_{i=1}^{n}(x - x_i) = \sum_{j=0}^{n}(-1)^j s_j x^{n-j}, \tag{4.2}$$

wenn man $s_0 = 1$ setzt. Bezeichne $u_j(x_1, \ldots, x_n) = \sum_{i=1}^n x_i^j$. Wir werden die folgende Beziehung zwischen den s_j und u_j zeigen (Newtonsche Identitäten):

$$u_k - s_1 u_{k-1} + s_2 u_{k-2} - + \ldots + (-1)^{k-1} s_{k-1} u_1 + (-1)^k k s_k = 0 \tag{4.3}$$

für $k = 1, \ldots, n$. Aus ihnen ergibt sich leicht die Äquivalenz von U und S. Wir zeigen (4.3): Aus (4.2) folgt einerseits

$$x\frac{f'}{f} = x\sum_{i=1}^{n}\frac{1}{x - x_i} = \sum_{i=1}^{n}\frac{1}{1 - \frac{x_i}{x}} = \sum_{i=1}^{n}\sum_{j=0}^{\infty} x_i^j x^{-j} = \sum_{j=0}^{\infty} u_j x^{-j},$$

wo $u_0 = n$ gesetzt ist, und andererseits

$$xf' = \sum_{i=0}^{n}(-1)^i s_i\,(n-i)x^{(n-i)}.$$

Durch Vergleich der Koeffizienten von x^{n-i} folgt für $i = 1, \ldots, n$:

$$(-1)^i s_i\,(n-i) = (-1)^i n s_i + (-1)^{i-1} u_1 s_{i-1} + \ldots + u_i s_0,$$

also gerade (4.3).

Nun zeigen wir die Äquivalenz von S und $g_{(n)}$: Ist $S(x) = z$, so folgt aus der Eindeutigkeit der Linearfaktorzerlegung in (4.2), daß

$$S^{-1}\{z\} = \{\tilde{\pi}(x) : \pi \in \Pi_n\}.$$

Sei $V = \{x \in \mathbf{R}^n : x_1 \le x_2 \le \ldots \le x_n\}$. Dann ist $|V \cap S^{-1}(z)| \le 1$, also ist

$$S_{|V} : V \to S(V) = S(\mathbf{R}^n)$$

injektiv (und natürlich auch stetig). $\Psi = (S_{|V})^{-1}$ existiert also auf $S(\mathbf{R}^n)$ und ist meßbar, denn die Borel Mengen in V werden von den kompakten Teilmengen von V

erzeugt und $\Psi^{-1}(K) = S(K)$ ist kompakt für jedes kompakte $K \subseteq V$. Die Äquivalenz von S und $g_{(n)}$ folgt schließlich aus

$$S \circ g_{(n)} = S \quad \text{und} \quad g_{(n)} = \Psi \circ S.$$

□

Wir beschließen diesen Abschnitt mit zwei Lemmata:

Lemma 4.4.14 *Sei $(M, \mathcal{M}, \mathcal{P})$ ein statistischer Raum, $\mathcal{F} \subseteq \mathcal{M}$ suffizient. Dann existiert zu jedem $\mathcal{P}$-integrierbarem $g : M \to \mathbf{R}$ ein $\mathcal{F}$-meßbares $f : M \to \mathbf{R}$ derart, daß*

$$f = E_P[g|\mathcal{F}] \quad \mathcal{P}\text{-f.s. für alle } P \in \mathcal{P}.$$

Beweis: Für $g = I_A$, $A \in \mathcal{M}$, ist das die Definition der Suffizienz. Für Elementarfunktionen g folgt das Lemma daraus sofort wegen der Linearität der bedingten Erwartung.

Sei nun $g \geq 0$ meßbar. Dann gibt es Elementarfunktionen $0 \leq g_n$ mit $g_n \nearrow g$. Für jedes n sei f_n eine $\mathcal{F}$-meßbare gemeinsame Version der $E_P[g_n|\mathcal{F}]$. Dann gilt $f_1 \leq f_2 \leq f_3 \leq \ldots$ $\mathcal{P}$-f.s., und es bleibt zu zeigen, daß die $\mathcal{F}$-meßbare Funktion $f = \sup_n f_n$ eine gemeinsame Version der $E_P[g|\mathcal{F}]$ ist. Seien dazu $P \in \mathcal{P}$ und $F \in \mathcal{F}$ beliebig.

$$\int_F f\, dP = \sup_n \int_F f_n\, dP = \sup_n \int_F E_P[g_n|\mathcal{F}] dP = \sup_n \int_F g_n\, dP = \int_F g\, dP.$$

□

Lemma 4.4.15 *Seien $(M_i, \mathcal{M}_i, \mathcal{P}_i)$ statistische Räume mit suffizienten Statistiken $T_i : (M_i, \mathcal{M}_i) \to (D_i, \mathcal{D}_i)$ $(i = 1, 2)$. Setze $M = M_1 \times M_2$, $\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2$, $\mathcal{P} = \{P_1 \times P_2 : P_i \in \mathcal{P}_i\}$.*

Dann ist $T = T_1 \times T_2 : (M, \mathcal{M}) \to (D_1 \times D_2, \mathcal{D}_1 \times \mathcal{D}_2)$ suffizient für $(M, \mathcal{M}, \mathcal{P})$, und für alle $A_1 \in \mathcal{M}_1$, $A_2 \in \mathcal{M}_2$ ist $p_{\mathcal{F}_1 \times \mathcal{F}_2}((x_1, x_2), A_1 \times A_2) = p_{\mathcal{F}_1}(x_1, A_1) \cdot p_{\mathcal{F}_2}(x_2, A_2)$, wo $\mathcal{F}_i = T_i^{-1}(\mathcal{D}_i)$.

Beweis: Sei $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2$ und sei $\mathcal{G}$ die Menge aller $A \in \mathcal{M}$, für die es eine von $P \in \mathcal{P}$ unabhängige Version $p_{\mathcal{F}}((x_1, x_2), A)$ von $P(A|\mathcal{F})(x_1, x_2)$ gibt.

Seien $A_1 \in \mathcal{M}_1$, $A_2 \in \mathcal{M}_2$. Dann ist für alle $P_1 \in \mathcal{P}_1$, $P_2 \in \mathcal{P}_2$ und $F_1 \in \mathcal{F}_1$, $F_2 \in \mathcal{F}_2$

$$\begin{aligned}
& \int_{F_1 \times F_2} p_{\mathcal{F}_1}(x_1, A_1) p_{\mathcal{F}_2}(x_2, A_2) d(P_1 \times P_2)(x_1, x_2) \\
= & \int_{F_1} P_1(A_1|\mathcal{F}_1) dP_1 \int_{F_2} P_2(A_2|\mathcal{F}_2) dP_2 \\
= & P_1(F_1 \times A_1) P_2(F_2 \times A_2) \\
= & (P_1 \times P_2)((F_1 \times F_2) \cap (A_1 \times A_2)),
\end{aligned}$$

d.h. $A = A_1 \times A_2 \in \mathcal{G}$.

Um zu zeigen, daß $\mathcal{M} \subseteq \mathcal{G}$, reicht es wegen der $\cap$-Stabilität des $\mathcal{M}$-Erzeugers $\{A_1 \times A_2 : A_i \in \mathcal{M}_i\}$ zu zeigen, daß $\mathcal{G}$ ein Dynkin-System ist. Das folgt sofort aus

$$P(M|\mathcal{F}) = 1 \quad P - \text{f.s. für alle } P \in \mathcal{P},$$

$$\begin{aligned} & P(A \setminus B|\mathcal{F})(x) = P(A|\mathcal{F})(x) - P(B|\mathcal{F})(x) \\ = \; & p_{\mathcal{F}}(x, A) - p_{\mathcal{F}}(x, B) \quad P\text{-f.s.} \qquad\qquad (A, B \in \mathcal{M}, B \subseteq A) \end{aligned}$$

und

$$P(\cup_{i=1}^{\infty} B_i|\mathcal{F})(x) = \sum_{i=1}^{\infty} P(B_i|\mathcal{F})(x) = \sum_{i=1}^{\infty} p_{\mathcal{F}}(x, B_i) \; P\text{-f.s.}$$
$$(B_i \in \mathcal{M} \text{ paarweise disjunkt}).$$

□

4.5 Vollständigkeit

Das Ziel der Datenreduktion mittels einer Statistik ist, die Daten so weit wie möglich zu komprimieren, ohne dabei Entscheidendes zu verlieren. Letzteres wird durch die Suffizienz garantiert, ersteres durch die Vollständigkeit:

Definition 4.5.1 *Sei $(M, \mathcal{M}, \mathcal{P})$ ein statistischer Raum.*

1. *Sei $\mathcal{F}$ eine Teil-σ-Algebra von $\mathcal{M}$. $\mathcal{P}$ heißt* **vollständig** *bzgl. $\mathcal{F}$, falls für jede $\mathcal{F}$-meßbare Funktion f gilt:*

$$E_P[f] = 0 \text{ für alle } P \in \mathcal{P} \Rightarrow f = 0 \; \mathcal{P}\text{-f.s.}.$$

2. *Eine Statistik T auf $(M, \mathcal{M}, \mathcal{P})$ mit Werten in $(D, \mathcal{D})$ heißt* **vollständig**, *wenn die Familie $\{P \circ T^{-1} : P \in \mathcal{P}\}$ vollständig bzgl. $\mathcal{D}$ ist.*

Bemerkung 4.5.2 1. Äquivalent sind:

   (a) T ist vollständig.

   (b) $\mathcal{P}$ ist vollständig bzgl. $\sigma(T) = T^{-1}(\mathcal{D})$.

   (c) Für jede $\mathcal{D}$-meßbare Funktion $f : D \to \mathbf{R}$ gilt:

$$E_P[f \circ T] = 0 \text{ für alle } P \in \mathcal{P} \Rightarrow f \circ T = 0 \; \mathcal{P}\text{-f.s.}$$

Das folgt, da für jedes $\mathcal{D}$-meßbare f gilt

$$\begin{array}{rcl} E_{P\circ T^{-1}}[f] = 0 & \Longleftrightarrow & E_P[f\circ T] = 0 \quad \text{und} \\ f = 0\ \mathcal{P}\circ T^{-1}\text{-f.s.} & \Longleftrightarrow & f\circ T = 0\ \mathcal{P}\text{-f.s.} \end{array}$$

und da sich jedes $\sigma(T)$-meßbare g als $g = f\circ T$ darstellen läßt.

2. Ist von zwei äquivalenten Statistiken eine vollständig, so auch die andere.

Sei $(M,\mathcal{M},\mathcal{P})$ ein statistischer Raum. Eine Teil-σ-Algebra $\mathcal{F}\subseteq\mathcal{M}$ heißt minimal-suffizient, falls für jede weitere suffiziente Teil-σ-Algebra $\mathcal{G}\subseteq\mathcal{M}$ gilt, daß $\mathcal{F}\subseteq\mathcal{G}$ mod $\mathcal{P}$, d.h. $\forall F\in\mathcal{F}\exists G\in\mathcal{G}: F = G \bmod \mathcal{P}$.

Theorem 4.5.3 *Sei $(M,\mathcal{M},\mathcal{P})$ ein statistischer Raum. Ist $\mathcal{P}$ vollständig bzgl. der suffizienten Teil-σ-Algebra $\mathcal{F}$, so ist $\mathcal{F}$ minimal-suffizient.*

Beweis: Sei $F\in\mathcal{F}$, $g(x) := p_{\mathcal{G}}(x,F)$. O.B.d.A. $g\le 1$. Sei $f(x)$ eine $\mathcal{F}$-meßbare gemeinsame Version der $E_P[g|\mathcal{F}]$, vergleiche Lemma 4.4.14. O.B.d.A. ist $0\le f\le 1$.

Insbesondere $E_P[f - I_F] = 0$ für alle $P\in\mathcal{P}$, so daß aus der Vollständigkeit von $\mathcal{P}$ bzgl. $\mathcal{F}$ folgt: $f = I_F$ $\mathcal{P}$-f.s. Nun ist für beliebiges $P\in\mathcal{P}$

$$\begin{array}{rl} & E_P[f^2 - g^2] = E_P[I_F^2 - P(F|\mathcal{G})^2] = E_P[(I_F - P(F|\mathcal{G}))^2] \\ = & E_P[(f-g)^2] = E_P[(g - E_P[g|\mathcal{F}])^2] = E_P[g^2 - (E_P[g|\mathcal{F}])^2] \\ = & E_P[g^2 - f^2] \\ = & -E_P[f^2 - g^2]. \end{array}$$

Daraus folgt $E_P[(f-g)^2] = 0$, also $g = f = I_F$ P-f.s., und da g $\mathcal{G}$-meßbar ist, folgt die Behauptung. □

Bemerkung 4.5.4 Es ist nicht schwer, zu zeigen, daß zu jedem dominierbaren statistischen Raum $(M,\mathcal{M},\mathcal{P})$ eine minimalsuffiziente σ-Algebra exisitert, siehe [7, Satz 11.8]. Ist $\mathcal{P}$, versehen mit der Metrik der Totalvariation ein separabler metrischer Raum, so existiert auch eine „minimalsuffiziente“ Statistik, d.h. eine Statistik derart, daß sich jede suffiziente Statistik mod $\mathcal{P}$ als Funktion dieser Statistik schreiben läßt.

Theorem 4.5.5 *Sei $(M,\mathcal{M},\{P_\vartheta : \vartheta\in\Theta\})$ ein Exponentialraum mit Dichten*

$$f_\vartheta(x) = C(\vartheta)h(x)\exp\left(\sum_{j=1}^{k} c_j(\vartheta)h_j(x)\right).$$

Enthält die Menge $\Gamma := \{(c_1(\vartheta),\ldots,c_k(\vartheta)) : \vartheta\in\Theta\}$ ein nicht ausgeartetes k-dimensionales Rechteck, so ist $T(x) = (h_1(x),\ldots,h_k(x))$ eine vollständige Statistik.

Beweis: Durch Übergang zu natürlichen Parametern und durch Übergang von μ zu $\bar{h}\mu$ können wir annehmen, daß $h \equiv 1$ und $c_j(\vartheta) = \vartheta_j$.

Sei nun $f \circ T$ $\mathcal{P}$-integrierbar, $\int f \circ T \, dP_\vartheta = 0$ für alle $\vartheta \in \Theta$. Zu zeigen $f \circ T = 0$ $\mathcal{P}$-f.s.

Aus $0 = \int f \circ T \, dP_\vartheta = \int f(Tx) \exp(\sum_{j=1}^k \vartheta_j h_j(x)) \, d\mu(x)$ folgt:

$$\int f(t) \exp\left(\sum_{j=1}^k \vartheta_j t_j\right) d(\mu \circ T^{-1})(t) = 0.$$

Aus Lemma 4.1.8 folgt nun $f = 0$ $\mu \circ T^{-1}$-f.s. und damit $f \circ T = 0$ μ-f.s. Wegen $P << \mu$ für alle $P \in \mathcal{P}$ folgt daraus die Behauptung. □

Theorem 4.5.6 *Die Statistiken*

$$g_{(n)}(x_1, \ldots, x_n) = (x_n^{(1)}, \ldots, x_n^{(n)})$$

und

$$U(x_1, \ldots, x_n) = (\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2, \ldots, \sum_{i=1}^n x_i^n)$$

sind vollständig auf $(\mathbf{R}^n, \mathcal{B}^n, \mathcal{P}^n)$*, falls* $\mathcal{P}$ *mit einem Maß* ν *auch alle zu* ν *absolut stetigen Wahrscheinlichkeitsverteilungen enthält. Insbesondere kann* $\mathcal{P}$ *die Familie aller Verteilungen oder aller stetigen Verteilungen auf* $\mathbf{R}$ *sein.*

Beweis: Wegen der Äquivalenz von U und $g_{(n)}$ (Theorem 4.4.13) reicht es, die Vollständigkeit von U nachzuweisen:

Sei $\nu \in \mathcal{P}$ beliebig. Sei $U_j(x) = \sum_{i=1}^n x_i^j$. $\mathcal{P}^n$ enthält die n-parametrige Exponentialfamilie $\mathcal{P}_\nu^* := \{P_\vartheta : \vartheta \in \mathbf{R}^n\}$ mit den Dichten

$$\begin{aligned} f_\vartheta(x) &= C(\vartheta) \exp\left(-\sum_{i=1}^n x_i^{2n}\right) \exp\left(\sum_{j=1}^n \vartheta_j U_j(x)\right) \\ &= \prod_{i=1}^n \left(C(\vartheta)^{\frac{1}{n}} \cdot \exp(-x_i^{2n}) \cdot \exp(\sum_{j=1}^n \vartheta_j x_i^j)\right) \end{aligned}$$

bzgl. ν^n. Also ist U vollständig bzgl. $\mathcal{P}_\nu^*$, d.h. aus $E_P[f \circ U] = 0$ $(P \in \mathcal{P}_\nu^*)$ folgt $f = 0$ $\mathcal{P}_\nu^*$-f.s. und damit ν^n-f.s. . Da $\nu^n \in \mathcal{P}^n$ beliebig und $\mathcal{P}_\nu^* \subseteq \mathcal{P}^n$ für jede Wahl von ν, folgt aus $E_P[f \circ U] = 0$ $(P \in \mathcal{P}^n)$ sofort $f = 0$ $\mathcal{P}^n$-f.s. , d.h. U ist vollständig für $\mathcal{P}^n$. □

Für den Zweistichprobenfall erhält man

Theorem 4.5.7 *Die Statistiken*

$$\tilde{g}(x_1, \ldots, x_n, y_1, \ldots, y_m) = (g_{(n)}(x_1, \ldots, x_n), g_{(m)}(y_1, \ldots, y_m))$$

und

$$\tilde{U}(x_1, \ldots, x_n, y_1, \ldots, y_m) = (U(x_1, \ldots, x_n), U(y_1, \ldots, y_m))$$

auf $(\mathbf{R}^{n+m}, \mathcal{B}^{n+m}, \mathcal{P})$ *mit* $\mathcal{P} = \{P_1^n \times P_2^m : P_1, P_2$ *(stetige) Verteilungen auf* $\mathbf{R}\}$ *sind äquivalent. Beide sind suffizient und vollständig.*

Beweis: Die Äquivalenz folgt sofort aus der von $g_{(n)}$ und U. Die Suffizienz folgt dann aus Theorem 4.4.12 und Lemma 4.4.15. Die Vollständigkeit von $\tilde{U}$ beweist man wie im Einstichprobenfall: $\mathcal{P}$ enthält die $(n+m)$-parametrige Exponentialfamilie mit den Dichten

$$f_\vartheta(x, y) = C(\vartheta) \exp\left(-\sum_{i=1}^{n} x_i^{2n} - \sum_{i=1}^{m} y_i^2 m\right) \exp\left(\sum_{j=1}^{n} \vartheta_j \sum_{i=1}^{n} x_i^j + \sum_{j=1}^{m} \vartheta_{n+j} \sum_{i=1}^{m} y_i^j\right).$$

□

Kapitel 5

Theorie der Punktschätzungen

5.1 Erwartungstreue Punktschätzungen mit gleichmäßig kleinstem Risiko

Sei $(\mathbf{R}, \mathcal{B}, \{P_\vartheta : \vartheta \in \Theta\})$ ein statistischer Raum. Bezeichnung: $\vartheta = (\vartheta_1, \ldots, \vartheta_s)^* \in \mathbf{R}^s$ (* bezeichnet die Transposition). Eine n-fache Stichprobe werde mit $X = (X_1, \ldots, X_n)^*$ bezeichnet, also als Spaltenvektor notiert. Der dem n-fachen Experiment zugeordnete statistische Raum sei $(M, \mathcal{M}, \mathcal{P}) := (\mathbf{R}, \mathcal{B}, \{P_\vartheta : \vartheta \in \Theta\})^n$.

Weiter sei $g : \Theta \to \mathbf{R}^k$ eine Funktion, die zu schätzende Kenngröße der Verteilung P_ϑ. Die statistische Aufgabenstellung bestehe darin, auf der Basis des Beobachtungsvektors X eine Punktschätzung für $g(\vartheta) = (g_1(\vartheta), \ldots, g_k(\vartheta))$ anzugeben, d.h. wir suchen geeignete Statistiken $T : M = \mathbf{R}^n \to \mathbf{R}^k$. Im Spezialfall $\Theta \subseteq \mathbf{R}^k$, $g(\vartheta) = \vartheta$, ist also der unbekannte Parameter ϑ selbst zu schätzen.

Ziel dieses Abschnitts ist es, Schätzer zu finden, die unter allen erwartungstreuen Schätzern gleichmäßig kleinste Risikofunktion haben.

Definition 5.1.1 *1.* $b(\vartheta, T) := E_\vartheta[T] - g(\vartheta)$ *heißt* **Bias** *oder* **Verzerrung** *des Schätzers T für $g(\vartheta)$.*

2. Ist $b(\vartheta, T) = 0$ für alle $\vartheta \in \Theta$, so heißt T **erwartungstreu** *(engl.: unbiased). Erwartungstreue ist also bzgl. der Familie $\{P_\vartheta : \vartheta \in \Theta\}$ und der Funktion g definiert.*

Um die Theorie der Punktschätzer in einen entscheidungstheoretischen Rahmen zu stellen, faßt man sie als nichtrandomisierte Entscheidungsfunktionen $\delta(x) = T(x)$ mit Werten in $\mathbf{R}^k$ auf und betrachtet eine quadratische Verlustfunktion

$$L(\vartheta, t) = (g(\vartheta) - t)^* \mathbf{A} (g(\vartheta) - t),$$

wo $\mathbf{A}$ eine positiv definite $k \times k$-Matrix ist. Dann ist nämlich

$$\begin{aligned} & E_\vartheta[L(\vartheta', T)] \\ = \; & E_\vartheta[(T - g(\vartheta'))^* \mathbf{A}(T - g(\vartheta'))] \\ = \; & E_\vartheta[(T - E_\vartheta[T])^* \mathbf{A}(T - E_\vartheta[T])] \\ & + (E_\vartheta[T] - g(\vartheta'))^* \mathbf{A}(E_\vartheta[T] - g(\vartheta')) \\ =: \; & V_\vartheta^A(T) + L(\vartheta', E_\vartheta[T]). \end{aligned} \tag{5.1}$$

Man sieht nun leicht, daß erwartungstreue Schätzer erwartungstreu im Sinne der Entscheidungstheorie sind: Erwartungstreue von T im Sinne der Entscheidungstheorie heißt

$$E_\vartheta[L(\vartheta', T)] - E_\vartheta[L(\vartheta, T)] = L(\vartheta', E_\vartheta[T]) - L(\vartheta, E_\vartheta[T]) \geq 0 \quad \forall \vartheta, \vartheta' \in \Theta.$$

Das ist sicher (und i.a. nur dann) erfüllt, falls $L(\vartheta, E_\vartheta[T]) = 0$, d.h. $E_\vartheta[T] = g(\vartheta)$.

Für $\vartheta = \vartheta'$ führt (5.1) auf

$$R(\vartheta, T) = \underbrace{V_\vartheta^A(T)}_{\geq 0} + \underbrace{b(\vartheta, T)^* \mathbf{A} b(\vartheta, T)}_{\geq 0}.$$

Diese Formel legt es nahe, sich bei der Suche nach Schätzern mit kleiner Risikofunktion auf erwartungstreue Schätzer einzuschränken.

Beispiel 5.1.2 Sei $\Theta = \{\vartheta = (\mu, \sigma^2) : \; \mu \in \mathbf{R}, \sigma^2 > 0\}$, $P_\vartheta = \mathcal{N}(\mu, \sigma^2)$, $g(\vartheta) = \mu$, $L(\vartheta, t) = (\mu - t)^2$, $T_1(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^n x_i$ und $T_2(x_1, \ldots, x_n) = a \cdot T_1(x_1, \ldots, x_n)$ für ein $0 \leq a < 1$.

Es ist

$$\begin{aligned} V_\vartheta(T_1) = \tfrac{\sigma^2}{n}, \; b(\vartheta, T_1) = 0, \; R(\vartheta, T_1) = \tfrac{\sigma^2}{n} \\ V_\vartheta(T_2) = a^2 \tfrac{\sigma^2}{n}, \; b(\vartheta, T_2) = (a-1)\mu, \; R(\vartheta, T_2) = a^2 \tfrac{\sigma^2}{n} + (1-a)^2 \mu^2. \end{aligned}$$

In einer Umgebung von $\mu = 0$ ist daher $R(\vartheta, T_2) < R(\vartheta, T_1)$, obwohl (oder gerade weil) T_2 nicht erwartungstreu ist. Für große μ^2 ist dagegen $R(\vartheta, T_2) > R(\vartheta, T_1)$.

Theorem 5.1.3 (Rao-Blackwell) *Sei $(M, \mathcal{M}, \{P_\vartheta : \vartheta \in \Theta\})$ ein statistischer Raum. Ist T ein erwartungstreuer Punktschätzer für $g(\vartheta)$ und S irgendeine suffiziente Statistik, so ist auch jede $\sigma(S)$-meßbare gemeinsame Version $\tilde{T} = E[T|S]$ der bedingten Erwartungen $E_\vartheta[T|\sigma(S)]$ (siehe Lemma 4.4.14) ein erwartungstreuer Punktschätzer für $g(\vartheta)$, und es gilt*

$$R(\vartheta, \tilde{T}) \leq R(\vartheta, T) \quad \forall \vartheta \in \Theta$$

mit Gleichheit für alle $\vartheta \in \Theta$ genau dann, wenn T bzgl. $\sigma(S)$ mod $\mathcal{P}$ meßbar ist, d.h. wenn $T = \tilde{T}$ $\mathcal{P}$-f.s.

Beweis: $\tilde{T}$ ist erwartungstreu, da $E_\vartheta[\tilde{T}] = E_\vartheta[T] = g(\vartheta)$. Außerdem ist

$$\begin{aligned} R(\vartheta, T) &= E_\vartheta[(T - g(\vartheta))^* \mathbf{A} (T - g(\vartheta))] \\ &= \underbrace{E_\vartheta[(T - \tilde{T})^* \mathbf{A} (T - \tilde{T})]}_{\geq 0} \\ &\quad -2 \underbrace{E_\vartheta[(\tilde{T} - g(\vartheta))^* \mathbf{A} (T - E_\vartheta[T|S])]}_{=0} + R(\vartheta, \tilde{T}) \end{aligned}$$

mit Gleichheit für alle $\vartheta \in \Theta$ genau dann, wenn $T = \tilde{T}$ $\mathcal{P}$-f.s. □

Theorem 5.1.4 (Lehmann-Scheffé) *Sei $(M, \mathcal{M}, \{P_\vartheta : \vartheta \in \Theta\})$ ein statistischer Raum. Ist T ein erwartungstreuer Punktschätzer für $g(\vartheta)$ und S eine vollständige und suffiziente Statistik, so ist $\tilde{T} = E[T|S]$ ein erwartungstreuer Punktschätzer für $g(\vartheta)$ mit gleichmäßig kleinstem Risiko (unter allen erwartungstreuen Punktschätzern für $g(\vartheta)$) und als solcher $\mathcal{P}$-f.s. eindeutig bestimmt.*

Man nennt $\tilde{T}$ einen **UMV Schätzer** *(uniformly minimal variance).*

Beweis: Aus dem Satz von Rao-Blackwell folgt, daß $\tilde{T}$ ein erwartungstreuer, $\sigma(S)$-meßbarer Schätzer für $g(\vartheta)$ ist, und da S vollständig ist, ist $\tilde{T}$ als solcher $\mathcal{P}$-f.s. eindeutig bestimmt. Aus dem Satz von Rao-Blackwell folgt nun auch, daß die erwartungstreuen UMV Schätzer für $g(\vartheta)$ gerade die erwartungstreuen $\sigma(S)$-meßbaren Schätzer für $g(\vartheta)$ sind. □

Beispiel 5.1.5 (Wilcoxon 2-Stichproben Statistik) Der der Wilcoxon 2-Stichproben Statistik

$$W_{n,m} = \sum_{i=1}^{n} R_i = \underbrace{\sum_{i=1}^{n} \sum_{j=1}^{m} I_{\{X_i \geq Y_j\}}}_{=:W'_{n,m}} + \frac{n(n+1)}{2}$$

zugrunde liegende statistische Raum ist

$$(M, \mathcal{M}, \mathcal{P}) = (\mathbf{R}^{n+m}, \mathcal{B}^{n+m}, \{F^n \times G^m : F, G \text{ stetige Verteilungsfunktionen}\}).$$

$\frac{1}{nm} W'_{n,m}$ ist ein erwartungstreuer Schätzer für

$$\theta(F, G) = \int G(x)\, dF(x) = P\{Y_1 \leq X_1\}.$$

Da $W'_{n,m}(x_1, \ldots, x_n, y_1, \ldots, y_m) = W'_{n,m}(g_{(n)}(x_1, \ldots, x_n), g_{(m)}(y_1, \ldots, y_m))$, ist $W'_{n,m}$ Funktion einer vollständigen und suffizienten Statistik, siehe Theorem 4.5.7. Also ist $W'_{n,m}$ eindeutig bestimmter UMV Schätzer für die Testgröße θ.

Beispiel 5.1.6 (U-Statistiken) Mit dem gleichen statistischen Raum läßt sich das vorhergehende Beispiel folgendermaßen verallgemeinern (die Verteilungen müssen dabei nicht stetig sein): Sei $\psi : \mathbf{R}^{r+s} \to \mathbf{R}$ meßbar und bzgl. $dF(x_1) \dots dF(x_r) dG(y_1) \dots dG(y_s)$ integrierbar. Setze

$$\theta(\psi; F, G) := \int \psi(x_1, \dots, x_r, y_1, \dots, y_s) dF(x_1) \dots dF(x_r) dG(y_1) \dots dG(y_r).$$

Für $n \geq r$ und $m \geq s$ ist

$$T := \psi(X_1, \dots, X_r, Y_1, \dots, Y_s)$$

ein erwartungstreuer Schätzer für θ. Wegen Theorem 4.5.7 ist

$$\tilde{T} := E[T | g_{(n)}(X_1, \dots, X_n), g_{(m)}(Y_1, \dots, Y_m)]$$

ein UMV Schätzer für θ. Man rechnet mit Hilfe von Theorem 3.2.1 leicht nach, daß

$$\begin{aligned} \tilde{T} &= \frac{1}{n!m!} \sum_{\pi \in \Pi_n, \sigma \in \Pi_m} \psi(X_{\pi(1)}, \dots, X_{\pi(r)}, Y_{\sigma(1)}, \dots, Y_{\sigma(s)}) \\ &= \binom{n}{r}^{-1} \binom{m}{s}^{-1} \sum_{\substack{1 \leq i_1 < \dots < i_r \leq n \\ 1 \leq j_1 < \dots < j_s \leq m}} \psi_{\text{symm}}(X_{i_1}, \dots, X_{i_r}, Y_{j_1}, \dots, Y_{j_s}), \end{aligned}$$

wo ψ_{symm} die Symmetrisierung von ψ in $x_1, \dots, x_r$ und $y_1, \dots, y_s$ bezeichnet.

Solche Statistiken heißen **U-Statistiken**. Ihre Konsistenz und asymptotische Normalität wird wie bei der Wilcoxon Statistik mit der Projektionsmethode bewiesen.

Ein einfaches Beispiel für den Fall $r = 1$, $s = 0$ (Einstichprobenfall) ist $\psi(x) = I_{\{x \in I\}}$, wo I ein Intervall in $\mathbf{R}$ ist. In diesem Fall ist $\theta(\psi, F) = \int_I dF = P\{X \in I\}$ und $\tilde{T} = \frac{1}{n} \sum_{i=1}^n I_{\{X_i \in I\}}$ ist ein UMV Schätzer für θ.

Beispiel 5.1.7 Wir betrachten den Poissonraum $(\mathbf{N}, \wp(\mathbf{N}), \mathcal{P} = \{P_\lambda : \lambda > 0\})^n$, wo $P_\lambda(k) = e^{-\lambda} \frac{\lambda^k}{k!}$ $(k \in \mathbf{N})$. Zu schätzen ist $g(\lambda) = P_\lambda(k_0)$ für ein $k_0 \in \mathbf{N}$. $T = I_{\{X_1 = k_0\}}$ ist ein erwartungstreuer Schätzer für θ. Der zugehörige U-Statistik Schätzer ist aber nicht mehr UMV, da wir nun mit einer viel kleineren Familie $\mathcal{P}$ arbeiten.

Da $\mathcal{P}^n$ eine Exponentialfamilie mit Dichten

$$f_\lambda^n(x_1, \dots, x_n) = e^{-n\lambda} \frac{1}{x_1! \cdots x_n!} e^{(x_1 + \dots + x_n) \log \lambda}$$

zum Zählmaß auf $\mathbf{N}$ ist, ist $\bar{X}$ eine suffiziente und vollständige Statistik. Also erhält man einen UMV Schätzer für θ als $\tilde{T} = E[T | \bar{X}]$. Explizit errechnet man für $k \geq k_0$

$$\begin{aligned} \tilde{T}_{|n\bar{X} = k} &= E[T | n\bar{X} = k] \\ &= P(X_1 = k_0 | X_1 + \dots + X_n = k) \end{aligned}$$

$$
\begin{aligned}
&= \frac{P(X_1 = k_0,\, X_2 + \ldots + X_n = k - k_0)}{P(X_1 + \ldots + X_n = k)} \\
&= \frac{e^{-\lambda}\lambda^{k_0} \cdot e^{-(n-1)\lambda}(n-1)^{k-k_0}\lambda^{k-k_0} \cdot k!}{k_0! \cdot (k-k_0)! \cdot e^{-n\lambda} n^k \lambda^k} \\
&= \binom{k}{k_0} (\frac{n-1}{n})^k (n-1)^{-k_0},
\end{aligned}
$$

während $\tilde{T}_{|n\bar{X}=k} = 0$ für $k < k_0$. Ist $k = nx$, x fest, so strebt dieser Ausdruck für $n \to \infty$ gegen $P_x(k_0)$, wie zu erwarten war.

Ist $(M, \mathcal{M}, \{P_\vartheta : \vartheta \in \Theta\})$ mit $\Theta \subseteq \mathbf{R}^k$ ein durch μ dominierter statistischer Raum, so gibt es unter gewissen Regularitätsannahmen an die **Likelihood Funktion**

$$\vartheta \mapsto f_\vartheta(x), \text{ wo } f_\vartheta = \frac{dP_\vartheta}{d\mu},$$

für erwartungstreue Punktschätzer von $g(\vartheta)$ eine durch g und die Familie $\{P_\vartheta : \vartheta \in \Theta\}$ bestimmte untere Schranke für die Risikofunktion. Wir betrachten hier nur den Fall, wo g Werte in $\mathbf{R}$ annimmt und $L(\vartheta, t) = (g(\vartheta) - t)^2$ ist.

Theorem 5.1.8 (Informationsungleichung, auch Cramér-Rao Ungleichung) *Sei $(M, \mathcal{M}, \{P_\vartheta : \vartheta \in \Theta\})$, $\Theta \subseteq \mathbf{R}^s$ offen, ein durch μ dominierter statistischer Raum, f_ϑ wie oben, X eine Stichprobe. $S : M \to \mathbf{R}$ sei ein erwartungstreuer Punktschätzer für $g : \Theta \to \mathbf{R}$, $S \in L^2_{P_\vartheta}$ für alle $\vartheta \in \Theta$. Wir setzen voraus*

1. *$\vartheta \mapsto \log f_\vartheta(X)$ ist für alle Werte von X auf Θ differenzierbar (insbesondere also $f_\vartheta(X) \neq 0$), und für alle $\vartheta \in \Theta$ ist*

$$\operatorname{grad} \log f_\vartheta(X) = \frac{1}{f_\vartheta(X)} \operatorname{grad} f_\vartheta(X) \in L^2_{P_\vartheta} \text{ und}$$

$$E_\vartheta[\operatorname{grad} \log f_\vartheta(X)] = E_\mu[\operatorname{grad} f_\vartheta(X)] \stackrel{!}{=} \operatorname{grad} E_\mu[f_\vartheta(X)] = 0.$$

2.

$$E_\mu[(\operatorname{grad} f_\vartheta(X))S] = \operatorname{grad} E_\mu[f_\vartheta(X)S] \quad \forall \vartheta \in \Theta.$$

3. *Die Kovarianzmatrix $I(\vartheta)$ von $\operatorname{grad} \log f_\vartheta(X)$, gegeben durch*

$$(I(\vartheta))_{i,j} = E_\vartheta[\frac{\partial}{\partial \vartheta_i} \log f_\vartheta \cdot \frac{\partial}{\partial \vartheta_j} \log f_\vartheta],$$

ist invertierbar.

Dann ist

$$R(\vartheta, S) = E_\vartheta[(S - g(\vartheta))^2] \geq (\operatorname{grad} g(\vartheta))^* I(\vartheta)^{-1} \operatorname{grad} g(\vartheta).$$

Beweis: Da

$$\begin{aligned} \mathrm{grad}E_\vartheta[S] &= \mathrm{grad}E_\mu[f_\vartheta(X)S] = E_\mu[(\mathrm{grad}f_\vartheta(X))S] \\ &= E_\vartheta[\mathrm{grad}(\log f_\vartheta(X))\,(S-g(\vartheta))], \end{aligned}$$

gilt für $u = \mathrm{grad}E_\vartheta[S]$ und $v = I(\vartheta)^{-1}u$

$$\begin{aligned} v^*u &= E_\vartheta[(v^*\mathrm{grad}\log f_\vartheta(X))\cdot(S-g(\vartheta))] \\ &\leq \left(E_\vartheta[(v^*\mathrm{grad}\log f_\vartheta(X))^2]E_\vartheta[(S-g(\vartheta))^2]\right)^{\frac{1}{2}}, \end{aligned}$$

woraus folgt

$$E_\vartheta[(S-g(\vartheta))^2] \geq \frac{(v^*u)^2}{E_\vartheta[(v^*\mathrm{grad}\log f_\vartheta(X))^2]}.$$

Da

$$\begin{aligned} & E_\vartheta[(v^*\mathrm{grad}\log f_\vartheta(X))^2] \\ = \; & E_\vartheta[v^*\mathrm{grad}\log f_\vartheta(X)(\mathrm{grad}\log f_\vartheta(X))^*v] \\ = \; & v^*E_\vartheta[\mathrm{grad}\log f_\vartheta(X)(\mathrm{grad}\log f_\vartheta(X))^*]v \\ = \; & v^*I(\vartheta)I(\vartheta)^{-1}u \\ = \; & v^*u, \end{aligned}$$

folgt

$$E_\vartheta[(S-g(\vartheta))^2] \geq v^*u = u^*I(\vartheta)^{-1}u,$$

da $I(\vartheta)$ symmetrisch ist. □

Definition 5.1.9 *Unter den Annahmen von Theorem 5.1.8 heißt S ein* **effizienter** *Schätzer für $g(\vartheta)$, falls $R(\vartheta,S) = (\mathrm{grad}g(\vartheta))^*I(\vartheta)^{-1}\mathrm{grad}g(\vartheta)$.*

Bemerkung 5.1.10

1. Ein erwartungstreuer, effizienter Schätzer existiert nicht in jedem Fall, siehe Beispiel 5.1.12.

2. $I(\vartheta)$ wird als **Fisher Informationsmatrix** bezeichnet. Macht man außer 1. in Theorem 5.1.8 noch die Annahmen

$$\vartheta \mapsto f_\vartheta(X) \text{ ist zweimal differenzierbar und}$$

$$E_\mu\left[\frac{\partial^2}{\partial\vartheta_i\partial\vartheta_j}f_\vartheta(X)\right] = \frac{\partial^2}{\partial\vartheta_i\partial\vartheta_j}E_\mu[f_\vartheta(X)] = 0 \quad \forall 1 \leq i,j \leq k,$$

so ist

$$I(\vartheta)_{i,j} = -E_\vartheta\left[\frac{\partial^2}{\partial\vartheta_i\partial\vartheta_j}\log f_\vartheta(X)\right], \tag{5.2}$$

denn

$$\begin{aligned}
&\frac{\partial^2}{\partial\vartheta_i\partial\vartheta_j}\log f_\vartheta(X) = \frac{\partial}{\partial\vartheta_i}\left(\frac{\partial}{\partial\vartheta_j}f_\vartheta(X)\cdot\frac{1}{f_\vartheta(X)}\right)\\
= &\left(\frac{\partial^2}{\partial\vartheta_i\partial\vartheta_j}f_\vartheta(X)\right)\frac{1}{f_\vartheta(X)} - \frac{\partial f_\vartheta(X)}{\partial\vartheta_i}\frac{\partial f_\vartheta(X)}{\partial\vartheta_j}\frac{1}{f_\vartheta(X)^2}\\
= &\left(\frac{\partial^2}{\partial\vartheta_i\partial\vartheta_j}f_\vartheta(X)\right)\frac{1}{f_\vartheta(X)} - \frac{\partial}{\partial\vartheta_i}\log f_\vartheta\cdot\frac{\partial}{\partial\vartheta_j}\log f_\vartheta
\end{aligned}$$

und

$$E_\vartheta\left[\frac{1}{f_\vartheta(X)}\frac{\partial^2}{\partial\vartheta_i\partial\vartheta_j}f_\vartheta(X)\right] = E_\mu\left[\frac{\partial^2}{\partial\vartheta_i\partial\vartheta_j}f_\vartheta(X)\right] = 0.$$

3. Die Bezeichnung „Informationsmatrix“ ist auf den Zusammenhang mit der **Kullback Information** von P_ϑ gegenüber $P_{\vartheta'}$ zurückzuführen:

$$\begin{aligned}
K(P_\vartheta, P_{\vartheta'}) &:= \int \log\frac{dP_\vartheta}{dP_{\vartheta'}}\,dP_\vartheta\\
&= E_\vartheta[\log\frac{f_\vartheta(X)}{f_{\vartheta'}(X)}] \quad \text{(wo } X \text{ eine Stichprobe aus } M \text{ ist).}
\end{aligned}$$

$K(P_\vartheta, P_{\vartheta'})$ ist eine Art Abstandsmaß zwischen P_ϑ und $P_{\vartheta'}$, denn aus der Jensen Ungleichung folgt:

$$K(P_\vartheta, P_{\vartheta'}) = \int(-\log\frac{dP_{\vartheta'}}{dP_\vartheta})\,dP_\vartheta \geq -\log(\int\frac{dP_{\vartheta'}}{dP_\vartheta}\,dP_\vartheta) = 0$$

mit Gleichheit genau dann, wenn $\frac{dP_{\vartheta'}}{dP_\vartheta} = 1$ P_ϑ-f.s., d.h. wenn $P_\vartheta = P_{\vartheta'}$.

Unter geeigneten Regularitätsannahmen besitzt nun $K(P_\vartheta, P_{\vartheta'})$ die folgende Approximation 2.Ordnung in $(\vartheta - \vartheta')$:

$$\begin{aligned}
&K(P_\vartheta, P_{\vartheta'})\\
= &-E_\vartheta[\log f_{\vartheta'}(X) - \log f_\vartheta(X)]\\
= &-\underbrace{(E_\vartheta[\operatorname{grad}\log f_\vartheta(X)])^*}_{=0}(\vartheta' - \vartheta) + \frac{1}{2}(\vartheta' - \vartheta)^* I(\vartheta)(\vartheta - \vartheta') + o\left(\|\vartheta - \vartheta'\|^2\right)
\end{aligned}$$

4. Hat der dominierbare statistische Raum $(M, \mathcal{M}, \{P_\vartheta : \vartheta \in \Theta\})$ die Informationsmatrix $I(\vartheta)$, so hat der statistische Raum $(M, \mathcal{M}, \{P_\vartheta : \vartheta \in \Theta\})^n$ die Informationsmatrix $I^{(n)}(\vartheta) = n \cdot I(\vartheta)$. Unter den Regularitätsannahmen in 2 folgt das sofort aus (5.2), und auch unter den Annahmen von Theorem 5.1.8 ist das leicht zu zeigen.

!

Bemerkung 5.1.11 Ist $\{P_\vartheta : \vartheta \in \Theta\}$ eine k-parametrige Exponentialfamilie mit natürlichem Parameterraum,

$$f_\vartheta(x) = C(\vartheta)h(x)\exp\left(\sum_{i=1}^{k}\vartheta_i h_i(x)\right),$$

so ist $h_j(X)$ ein effizienter, erwartungstreuer Schätzer für $g(\vartheta) = -\frac{\partial}{\partial\vartheta_j}\log C(\vartheta) = E_\vartheta[h_j(X)]$, denn

$$\begin{aligned} I(\vartheta)_{i,j} &= E_\vartheta[\frac{\partial}{\partial\vartheta_i}\log f_\vartheta(X)\frac{\partial}{\partial\vartheta_j}\log f_\vartheta(X)] \\ &= E_\vartheta[(-E_\vartheta[h_i(X)] + h_i(X))(-E_\vartheta[h_j(X)] + h_j(X))] \\ &= \mathrm{cov}_\vartheta(h_i(X), h_j(X)) = -\frac{\partial^2}{\partial\vartheta_i\partial\vartheta_j}\log C(\vartheta) \\ &= (\text{grad } g(\vartheta))_i, \end{aligned}$$

d.h. $(\text{grad } g_\vartheta)^*$ kann als j-te Zeile von $I(\vartheta)$ gedeutet werden so daß

$$(\text{grad } g(\vartheta))^* I(\vartheta)^{-1}(\text{grad } g(\vartheta)) = e_j^* \cdot \text{grad } g(\vartheta) = V_\vartheta(h_j(X)).$$

Im Kontrast dazu steht das folgende Beispiel.

Beispiel 5.1.12 Sei $X_1, \ldots, X_n$ eine n-fache Stichprobe aus $(\mathbf{R}, \mathcal{B}, \mathcal{P})$, $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbf{R}, \sigma^2 > 0\})$, S^2 die empirische Varianz der X_i. Wir untersuchen die Statistiken cS^2 $(c > 0)$.

Beachte zunächst, daß $\frac{n-1}{\sigma^2}S^2$ nach χ^2_{n-1} verteilt ist, also

$$E[S^2] = \sigma^2\frac{n-1}{n-1} = \sigma^2 \text{ und } E[S^4] = \sigma^4\frac{(n-1)(n+1)}{(n-1)^2} = \sigma^4\frac{n+1}{n-1}.$$

Daraus folgt

$$\begin{aligned} R((\mu,\sigma^2), cS^2) &= E_{(\mu,\sigma^2)}[(cS^2 - \sigma^2)^2] \\ &= c^2 E_{(\mu,\sigma^2)}[S^4] - 2c\sigma^2 E_{(\mu,\sigma^2)}[S^2] + \sigma^4 \\ &= \sigma^4(c^2\frac{n+1}{n-1} - 2c + 1). \end{aligned}$$

Durch differenzieren nach c erhält man, daß $R((\mu,\sigma^2), cS^2)$ für $c = c_{min} = \frac{n-1}{n+1}$ minimal wird und dort den Wert $\frac{2\sigma^4}{n+1}$ annimmt. Dagegen ist $R((\mu,\sigma^2), S^2) = \sigma^4\frac{2}{n-1} > \frac{2\sigma^4}{n+1}$.

S^2 ist also kein zulässiger Schätzer für σ^2, obwohl er der eindeutig bestimmte erwartungstreue Schätzer mit kleinster Varianz ist (folgt aus der Suffizienz und Vollständigkeit von $(\bar{X}, S^2)$).

Um zu zeigen, daß S^2 nicht effizient ist, bestimmen wir zunächst die Informationsmatrix von $(\mathbf{R}, \mathcal{B}, \mathcal{P})$: Mit $\vartheta = (\mu, \sigma^2)$,

$$\begin{aligned} U_1 &:= \frac{\partial}{\partial \mu} \log f_\vartheta(x) = \frac{x-\mu}{\sigma^2} \\ U_2 &:= \frac{\partial}{\partial \sigma^2} \log f_\vartheta(x) = -\frac{1}{2\sigma^2} + \frac{(x-\mu)^2}{2\sigma^4} \end{aligned}$$

ist

$$I(\vartheta) = \begin{pmatrix} V_\vartheta(U_1) & \mathrm{cov}_\vartheta(U_1, U_2) \\ \mathrm{cov}_\vartheta(U_1, U_2) & V_\vartheta(U_2) \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix},$$

also

$$I(\vartheta)^{-1} = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}.$$

Da für die Informationsmatrix $I^{(n)}(\vartheta)$ von $(M, \mathcal{M}, \{P_\vartheta : \vartheta \in \Theta\})^n$ gilt: $I^{(n)}(\vartheta) = n \cdot I(\vartheta)$, folgt für $g(\vartheta) = \sigma^2$

$$(\mathrm{grad} g(\vartheta))^* (I^{(n)}(\vartheta))^{-1} \mathrm{grad} g(\vartheta) = \frac{2\sigma^4}{n} < \frac{2\sigma^4}{n-1} = R(\vartheta, S^2).$$

S^2 ist also nicht effizient, und da S^2 unter allen erwartungstreuen Schätzern für σ^2 die kleinste Varianz hat, gibt es keine erwartungstreuen, effizienten Schätzer für σ^2.

Schließlich bemerken wir noch, daß $R(\vartheta, c_{min} S^2) = \sigma^4 \frac{2}{n+1} < \sigma^4 \frac{2}{n}$. Beachte, daß $c_{min} S^2$ kein erwartungstreuer Schätzer für σ^2 ist!

Definition 5.1.13 *Sei $(M, \mathcal{M}, \{P_\vartheta : \vartheta \in \Theta\})$ ein dominierbarer statistischer Raum, für den die Fisher Informationsmatrix $I(\vartheta)$ wohldefiniert ist. $g : \Theta \to \mathbf{R}$ sei eine auf $\Theta \subseteq \mathbf{R}^s$ differenzierbare Funktion.*

Eine konsistente Folge $(S_n)_{n>0}$ quadratintegrierbarer Schätzer für $g(\vartheta)$ heißt **asymptotisch effizient**, *falls*

$$\begin{aligned} & \lim_{n\to\infty} \frac{V_\vartheta(S_n)}{(\mathrm{grad} g(\vartheta))^* (I^{(n)}(\vartheta))^{-1} \mathrm{grad} g(\vartheta)} \\ = & \lim_{n\to\infty} \frac{n V_\vartheta(S_n)}{(\mathrm{grad} g(\vartheta))^* I(\vartheta)^{-1} \mathrm{grad} g(\vartheta)} \\ = & \ 1 \quad \textit{für alle } \vartheta \in \Theta. \end{aligned}$$

Gelegentlich wird statt dessen verlangt, daß

$$\sqrt{n}(S_n - g(\vartheta)) \Longrightarrow \mathcal{N}(0, (\mathrm{grad} g(\vartheta))^* I(\vartheta)^{-1} \mathrm{grad} g(\vartheta)) \quad \textit{unter } P_\vartheta \textit{ für alle } \vartheta \in \Theta.$$

Bemerkung 5.1.14 1. S^2 aus Beispiel 5.1.12 ist also asymptotisch effizient.

2. Beachte, daß in der Definition von asymptotischer Effizienz nicht die Erwartungstreue der S_n sondern nur ihre Konsistenz vorausgesetzt wurde. $c_{min}S^2$ aus Beispiel 5.1.12 ist also auch asymptotisch effizient, wobei für jedes n sogar $V_{\mu,\sigma^2}(c_{min}S^2) < \sigma^4\frac{2}{n}$ = asymptotische Varianz. Es gibt sogar Beispiele für konsistente, asymptotisch normale Schätzer S_n, für die

$$\lim_{n\to\infty} \frac{V_\vartheta(S_n)}{(\text{grad } g(\vartheta))^*(I^{(n)}(\vartheta))^{-1}\text{grad } g(\vartheta)} \leq 1 \quad \forall \vartheta \in \Theta$$

und strikter Ungleichung für mindestens ein $\vartheta_0 \in \Theta$. Solche Folgen von Schätzern heißen **supereffizient**.

5.2 Maximum Likelihood Schätzungen

Ein recht plausibles und in der Praxis einfach zu handhabendes Verfahren zur Konstruktion von Schätzern ist die **Maximum Likelihood Methode**, die wir im folgenden vorstellen wollen. Dazu sei immer $(M, \mathcal{M}, \mathcal{P} = \{P_\vartheta : \vartheta \in \Theta\})$ ein durch μ dominierter statistischer Raum und

$$f_\vartheta := \frac{dP_\vartheta}{d\mu} \quad \forall \vartheta \in \Theta.$$

Definition 5.2.1 *1. Eine (lokale)* **Maximum Likelihood Approximation** *der Beobachtung X gegeben $\{P_\vartheta : \vartheta \in \Theta\}$ ist jedes $\hat{\vartheta} \in \Theta$, das die* **Likelihood Funktion** $\vartheta \mapsto f_\vartheta(X)$ *oder, was damit gleichbedeutend ist, die* **Log-Likelihood Funktion** $\vartheta \mapsto \log f_\vartheta(X)$ *(lokal) maximiert.*

2. Sei $g : \Theta \to \mathbf{R}^d$ eine Abbildung. Eine (lokale) **Maximum Likelihood Schätzung** *von $g(\vartheta)$ auf der Basis der Beobachtung X ist*

$$\hat{g}(X) = g(\hat{\vartheta}(X)),$$

wo $\hat{\vartheta}(X)$ eine (lokale) Maximum Likelihood Approximation von X ist.

Bemerkung 5.2.2 1. Die Dichte einer n-fachen Stichprobe $X = (X_1, \ldots, X_n)$ bzgl. μ^n ist $f_\vartheta^n(x_1, \ldots, x_n) := \prod_{i=1}^n f_\vartheta(x_i)$. Daher gilt: Eine (lokale) Maximum Likelihood Approximation von $X = (X_1, \ldots, X_n)$ (gegeben $\{P_\vartheta^n : \vartheta \in \Theta\}$) ist jedes $\hat{\vartheta} \in \Theta$, das $f_\vartheta^n(X)$ oder, was damit gleichbedeutend ist, $\sum_{i=1}^n \log f_\vartheta(X_i)$ (lokal) maximiert.

2. Ist $\Theta \subseteq \mathbf{R}^s$, so kann unter geeigneten Differenzierbarkeitsannahmen an $\vartheta \mapsto f_\vartheta(x)$ eine lokale Maximum Likelihood Approximation als Lösung der **Maximum Likelihood Gleichungen**

$$\frac{\partial}{\partial\vartheta_j} f_\vartheta(X) = 0 \text{ (bzw. } \frac{\partial}{\partial\vartheta_j}\log f_\vartheta(X) = 0) \quad \forall j = 1, \ldots, s$$

bestimmt werden.

3. Eine sehr lesenswerte Darstellung der Grundlagen der Maximum Likelihood Methode findet man in [5].

Theorem 5.2.3 *Sei $\mathcal{P}$ eine k-parametrige Exponentialfamilie. Dann ist $\hat{\vartheta}$ eine lokale Maximum Likelihood-Approximation von $(X_1, \ldots, X_n)$ genau dann, wenn es das System der* **Maximum Likelihood Gleichungen**

$$\int_M h_j(x)\, dP_{\vartheta}(x) = \frac{1}{n}\sum_{i=1}^{n} h_j(X_i) \quad \left(= \int_M h_j(x) d\hat{F}_n(x)\right) \quad (j = 1, \ldots, k)$$

löst, wo $\hat{F}_n$ die empirische Verteilungsfunktion von $X_1, \ldots, X_n$ ist.

Beweis: Wir können o.B.d.A. annehmen, daß $\mathcal{P}$ natürliche Parameter hat. Dann gilt

$$\Psi(\vartheta) := \sum_{i=1}^{n} \log f_{\vartheta}(X_i) = \sum_{i=1}^{n} \left(\log C(\vartheta) + \log h(X_i) + \sum_{j=1}^{k} \vartheta_j h_j(X_i) \right),$$

und um ein Maximum dieser Funktion von ϑ zu bestimmen, lösen wir die Gleichungen

$$\frac{\partial}{\partial \vartheta_j}\Psi(\vartheta) = 0 \quad (j = 1, \ldots, k),$$

wo

$$\begin{aligned} \frac{\partial}{\partial \vartheta_j}\Psi(\vartheta) &= n \frac{\partial}{\partial \vartheta_j} \log C(\vartheta) + \sum_{i=1}^{n} h_j(X_i) \\ &= n \left(- \int_M h_j(x)\, dP_{\vartheta}(x) + \frac{1}{n}\sum_{i=1}^{n} h_j(X_i) \right) \quad (j = 1, \ldots, k) \end{aligned} \tag{5.3}$$

nach Lemma 4.1.8. Die 2. Ableitungen ergeben sich durch Ableiten des ersten Terms in (5.3) nach ϑ_l als

$$\frac{\partial^2}{\partial \vartheta_l \partial \vartheta_j}\Psi(\vartheta) = -n \frac{\partial}{\partial \vartheta_l} \int_M h_j(x)\, dP_{\vartheta}(x) = -n \mathrm{cov}_{\vartheta}(h_j, h_l).$$

Das folgt aus Lemma 4.1.8, in dem auch gezeigt wurde, daß diese Matrix negativ definit ist, so daß tatsächlich ein Maximum vorliegt.

□

Beispiel 5.2.4 Die folgenden Exponentialfamilien wurden in Beispiel 4.1.6 untersucht:

1. Bernoulli Verteilungen: $\mathcal{P} = \{P_p : \ 0 < p < 1\}$, $\mu =$ Zählmaß auf $\{0, 1\}$.

$$f_p(k) = \underbrace{(1-p)}_{C(p)} \cdot \exp(\underbrace{(\log \frac{p}{1-p})}_{c_1(\vartheta)} \cdot \underbrace{k}_{h_1(k)}).$$

Die Maximum Likelihood Gleichung lautet also

$$E_p[h_1] = p = \frac{1}{n}\sum_{i=1}^{n} h_1(X_i) = \bar{X},$$

d.h. die Maximum Likelihood Methode liefert gerade den erwartungstreuen Schätzer mit gleichmäßig kleinster Varianz.

2. Normalverteilungen: $\mathcal{P} = \{P_\vartheta = \mathcal{N}(\mu, \sigma^2) : \ \mu \in \mathbf{R}, \sigma^2 > 0\}$, $\vartheta = (\mu, \sigma^2)$.

$$f_\vartheta(x) \ = \ \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{\mu^2}{2\sigma^2})}_{C(\vartheta)} \cdot \exp(\underbrace{\frac{\mu}{\sigma^2}}_{c_1(\vartheta)} \cdot \underbrace{x}_{h_1(x)} + \underbrace{\frac{-1}{2\sigma^2}}_{c_2(\vartheta)} \cdot \underbrace{x^2}_{h_2(x)}).$$

Die Maximum Likelihood Gleichungen lauten also

$$E_\vartheta[h_1] = \mu = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X},$$
$$E_\vartheta[h_2] = \mu^2 + \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \frac{n-1}{n}S^2 + \bar{X}^2,$$

und wir erhalten den Schätzer $\hat{\vartheta} = (\bar{X}, \frac{n-1}{n}S^2)$ für $\vartheta = (\mu, \sigma^2)$. (Man sieht, daß $\hat{\sigma}^2 = 0$ im Fall $n = 1$, also $\hat{\vartheta} \notin \Theta$. In diesem Fall existiert der Maximum Likelihood Schätzer also nicht.)

Dieser Schätzer ist nicht erwartungstreu, da $E_\vartheta[S^2] = \sigma^2$, und in der Tat sind nach der Maximum Likelihood Methode gewonnene Schätzer i.a. nicht erwartungstreu. Man sieht aber, daß der Schätzer konsistent und asymptotisch erwartungstreu ist.

In Beispiel 5.1.12 haben wir gesehen, daß

$$R(\vartheta, cS^2) = \sigma^4(c^2\frac{n+1}{n-1} - 2c + 1).$$

Insbesondere ist $R(\vartheta, c_{min}S^2) = \sigma^4\frac{2}{n+1}$, $R(\vartheta, \frac{n-1}{n}S^2) = \sigma^4(\frac{2}{n} - \frac{1}{n^2})$, $R(\vartheta, S^2) = \sigma^4\frac{2}{n-1}$, so daß der Maximum Likelihood Schätzer $\frac{n-1}{n}S^2$ zwar nicht zulässig aber besser als S^2 ist.

3. Lineare Regression: Das in Kapitel 2 besprochene Modell der linearen Regression legte den statistischen Raum $(\mathbf{R}^2, \mathcal{B}^2, \mathcal{P})^n$ mit

$$\mathcal{P} = \{\mathcal{N}(\mu, V) : \ \mu \in \mathbf{R}^2, V = \begin{pmatrix} \sigma_X^2 & \mathrm{cov}(X,Y) \\ \mathrm{cov}(X,Y) & \sigma_Y^2 \end{pmatrix} \text{ Kovarianzmatrix}\}$$

zugrunde. Sind $(X_1, Y_1), \ldots, (X_n, Y_n)$ u.i.v. Beobachtungspaare, so ergibt sich als Maximum Likelihood Approximation $\hat{\mu}_X = \bar{X}$, $\hat{\mu}_Y = \bar{Y}$, $\hat{\sigma}_X^2 = \frac{n-1}{n}S_X^2$, $\hat{\sigma}_Y^2 = \frac{n-1}{n}S_Y^2$, $\hat{\mathrm{cov}}(X,Y) = \frac{n-1}{n}C_{X,Y}$. (Übung) !

Daraus erhält man wegen $\beta = \frac{\mathrm{cov}(X,Y)}{\sigma_X^2}$ und $\alpha = \mu_Y - \beta\mu_X$ die Maximum Likelihood Schätzer

$$\hat{\beta} = \frac{C_{X,Y}}{S_X^2} \text{ und } \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X},$$

d.h. gerade die in Kapitel 2 heuristisch gewonnenen Schätzer.

Maximum Likelihood Schätzer haben oft günstige asymptotische Eigenschaften wie Konsistenz und asymptotische Normalität. Der Beweis solcher Eigenschaften fällt um so leichter, je „regulärer" die Abbildung $\vartheta \mapsto f_\vartheta(x)$ ist. Wir beschränken uns hier auf Ergebnisse für Exponentialfamilien. Einen allgemeineren Satz findet man in [11, Abschnitt 6.3].

Theorem 5.2.5 *Sei $(M, \mathcal{M}, \mathcal{P})$ ein statistischer Raum, $\mathcal{P} = \{P_\vartheta : \vartheta \in \Theta\}$ eine k-parametrige Exponentialfamilie, $X_1, X_2, \ldots$ eine unendliche Stichprobe aus $(M, \mathcal{M}, \mathcal{P})$. Außerdem sei $\Theta \subseteq \mathbf{R}^k$ und die Abbildung $\Gamma : \Theta \to \Gamma(\vartheta) \subseteq \mathbf{R}^k, \vartheta \mapsto (c_1(\vartheta), \ldots, c_k(\vartheta))^*$ ein Diffeomorphismus. Dann geht die Wahrscheinlichkeit, daß die Maximum Likelihood Gleichungen*

$$\int_M h_j(x)\, dP_\vartheta(x) = \frac{1}{n}\sum_{i=1}^n h_j(X_i) = \int_M h_j(x) d\hat{F}_n(x) \quad (j = 1, \ldots, k)$$

eine eindeutige Lösung $\hat{\vartheta}_n$ in Θ haben, für $n \to \infty$ gegen 1, und es gilt:

1. *$\hat{\vartheta}_n$ ist eine konsistente Folge von Schätzern für ϑ,*
2. $$\sqrt{n}(\hat{\vartheta}_n - \vartheta) \Longrightarrow \mathcal{N}(\mathbf{0}, (D\Gamma)^{-1}_{|\vartheta} V_\vartheta^{-1} ((D\Gamma)^{-1}_{|\vartheta})^*),$$

   wo $v_{j,l|\vartheta} = \mathrm{cov}_\vartheta(h_j, h_l) = -\frac{\partial^2}{\partial c_j \partial c_l} \log(C(\vartheta))$.

3. *Definiere $\Psi : \Theta \to \mathbf{R}^k$ durch $\Psi(\vartheta) = (E_\vartheta[h_1], \ldots, E_\vartheta[h_k])^*$. Ist $\Psi : \Theta \to \Psi(\Theta) \subseteq \mathbf{R}^k$ invertierbar und haben die Maximum Likelihood Gleichungen bei gegebenem n fast sicher eine Lösung $\hat{\vartheta}_n$, so ist $\hat{\vartheta}_n$ suffizient und vollständig für $(M, \mathcal{M}, \mathcal{P})^n$.*

Beweis: Wegen Theorem 2.4.1 reicht es, den Fall zu betrachten, wo Θ der natürliche Parameterraum und daher $\Gamma = \mathrm{Id}$ ist.

Da $\Psi(\vartheta) = (-\frac{\partial}{\partial\vartheta_1} \log C(\vartheta), \ldots, -\frac{\partial}{\partial\vartheta_k} \log C(\vartheta))^*$, folgt aus Lemma 4.1.8, daß

$$D\Psi_{j,l}(\vartheta) = -\frac{\partial^2}{\partial\vartheta_j \partial\vartheta_l} \log(C(\vartheta)) = \mathrm{cov}_\vartheta(h_j, h_l) = V_\vartheta$$

positiv definit ist, so daß es zu festem $\vartheta \in \Theta$ eine Umgebung U von $\Psi(\vartheta)$ gibt, auf der Ψ^{-1} existiert und differenzierbar ist.

Sei $H_j^{(n)} = \frac{1}{n}\sum_{i=1}^n h_j(X_i)$ $(j = 1, \ldots, k)$, $H^{(n)} = (H_1^{(n)}, \ldots, H_k^{(n)})^*$. Die Maximum Likelihood Gleichungen schreiben sich damit als

$$\Psi(\hat{\vartheta}_n) = H^{(n)}, \tag{5.4}$$

und die Suffizienz und Vollständigkeit von $\hat{\vartheta}_n$ für $(M, \mathcal{M}, \mathcal{P})^n$ folgt sofort aus der von $H^{(n)}$ (beachte Bemerkung 4.1.7, Bemerkung 4.4.3, Korollar 4.4.10 und Theorem 4.5.5). Da $H^{(n)} \to \Psi(\vartheta)$ stochastisch, folgt aus der Invertierbarkeit von Ψ auf U, daß (5.4) fast sicher für große n eindeutig lösbar ist,

$$\hat{\vartheta}_n = \Psi^{-1}(H^{(n)}).$$

Da außerdem

$$\sqrt{n} \cdot (H^{(n)} - \Psi(\vartheta)) \Longrightarrow \mathcal{N}(\mathbf{0}, V_\vartheta),$$

folgt aus Theorem 2.4.1, daß

$$\begin{array}{rcl} \hat{\vartheta}_n & \to & \vartheta \text{ stochastisch und} \\ \sqrt{n}(\hat{\vartheta}_n - \vartheta) & \Longrightarrow & \mathcal{N}(\mathbf{0}, (D\Psi_{|\vartheta})^{-1} V_\vartheta ((D\Psi_{|\vartheta})^{-1})^*) \\ & = & \mathcal{N}(\mathbf{0}, (V_\vartheta^{-1}). \end{array}$$

□

Bemerkung 5.2.6 Die Übertragbarkeit vieler asymptotischer Ergebnisse für Maximum Likelihood Schätzer von Exponentialräumen auf allgemeinere dominierbare statistische Räume rührt daher, daß sich viele Familien von Dichten lokal (d.h. in einer Umgebung von $\vartheta = \vartheta_0$) durch Exponentialfamilien approximieren lassen.

Bemerkung 5.2.7 Die Maximum Likelihood Gleichungen für eine k-parametrige Exponentialfamilie können wir auch folgendermaßen aufschreiben:

$$\int_M h_j(x)(dP_\vartheta(x) - d\hat{F}_n(x)) = 0 \ (j = 1, \ldots, k).$$

Das legt die folgende Interpretation nahe: Gegeben $(X_1, \ldots, X_n)$, d.h. $\hat{F}_n$, wird P_ϑ in der k-dimensionalen Mannigfaltigkeit $\{P_\vartheta : \vartheta \in \Theta\}$ so bestimmt, daß die Differenz $(P_\vartheta - \hat{F}_n)$ in dem k-codimensionalen linearen Teilraum aller signierten Maße liegt, die „orthogonal“ zu $h_1, \ldots, h_k$ sind.

Bemerkung 5.2.8 Den günstigen asymptotischen Eigenschaften von Maximum Likelihood Schätzern stehen Nachteile bei endlichen Stichproben gegenüber. Wir haben bereits gesehen (Beispiel 2), daß ein Maximum Likelihood Schätzer weder erwartungstreu noch zulässig zu sein braucht und eventuell gar nicht existiert.

5.3 Bayes Schätzungen

Sei nun $(M, \mathcal{M}, \mathcal{P} = \{P_\vartheta : \vartheta \in \Theta\})$ ein statistischer Raum, $\mathcal{G}$ eine σ-Algebra auf Θ und π eine a-priori-Verteilung, siehe das Ende von Abschnitt 4.2. Wie in [10, Abschnitt II.7.7] ausgeführt, kann man unter entsprechenden Meßbarkeitsannahmen $P(\vartheta, .) := P_\vartheta(.)$ als stochastischen Kern (reguläre bedingte Wahrscheinlichkeitsverteilung) von $(\Theta, \mathcal{G})$ nach $(M, \mathcal{M})$ auffassen. Sei $Q := \pi \times P$ und $\bar{P}(A) := Q(\Theta \times A)$ für $A \in \mathcal{M}$. Ist nun z.B. Θ ein vollständiger separabler metrischer Raum, so existieren die regulären bedingten Wahrscheinlichkeitsverteilungen Q_x auf $(\Theta, \mathcal{G})$.

Definition 5.3.1 *Sei $g : \Theta \to \mathbf{R}$ meßbar. Die* **Bayes Schätzung** *von $g(\vartheta)$ auf der Basis der Beobachtung X und unter der a-priori-Verteilung π ist*

$$g_\pi(X) = \int_\Theta g(\vartheta)\, dQ_X(\vartheta).$$

Ähnlich wie am Ende von Abschnitt 4.2 zeigt man:

Theorem 5.3.2 *$g_\pi : M \to \mathbf{R}$ ist meßbar, und der Schätzer g_π für $g(\vartheta)$ minimiert das Bayes-Risiko bei quadratischer Verlustfunktion. Das heißt: Ist $T : M \to \mathbf{R}$ meßbar,*

$$r_\pi(T) := \int_\Theta \int_M (g(\vartheta) - T(x))^2\, dP_\vartheta(x)\, d\pi(\vartheta),$$

so ist $r_\pi(g_\pi) \le r_\pi(T)$ mit Gleichheit genau dann, wenn $T(x) = g_\pi(x)$ für $\bar{P}$-fast alle $x \in M$.

Beweis: Übung!

Bemerkung 5.3.3 Sei nun $\{P_\vartheta : \vartheta \in \Theta\}$ durch μ dominiert, $f_\vartheta := dP_\vartheta/d\mu$. Dann wird ein konzeptioneller Zusammenhang zwischen Maximum Likelihood Schätzungen und Bayes Schätzungen deutlich: Bei beiden beruht die Schätzung nur auf der Likelihood Funktion $\vartheta \mapsto f_\vartheta(X)$ für die tatsächliche Beobachtung X. Im einen Fall wird sie maximiert, im anderen Fall bestimmt sich der Schätzer durch

$$g_\pi(X) = \frac{\int g(\vartheta) f_\vartheta(X)\, d\pi(\vartheta)}{\int f_\vartheta(X)\, d\pi(\vartheta)}.$$

In dieser Tatsache liegt auch der Grund, weshalb beide nicht so recht in die Systematik der "klassischen" Statistik passen: Risikotheoretisch sind sie i.a. nicht optimal. Daher sind beide Verfahren, so bedeutsam sie für die Praxis in verschiedenen Zweigen der Angewandten Statistik auch immer waren, von den Mathematischen Statistikern lange Zeit vernachlässigt worden.

Kapitel 6

Elemente der Testtheorie

6.1 Grundbegriffe

Sei $(M, \mathcal{M}, \{P_\vartheta : \vartheta \in \Theta\})$ ein statistischer Raum, $\Theta_0, \Theta_1 \subseteq \Theta$ und $\Theta_0 \cap \Theta_1 = \emptyset$. (Oftmals ist $\Theta_1 = \Theta \setminus \Theta_0$.) Ein Testproblem besteht darin, auf der Grundlage einer Beobachtung $X \in M$ über die Annahme der Nullhypothese $H_0 : \vartheta \in \Theta_0$ zu entscheiden, wenn als Alternative nur die Gegenhypothese $H_1 : \vartheta \in \Theta_1$ möglich ist. Die Entscheidung wird mit einer (randomisierten) Entscheidungsfunktion δ auf $(M, \mathcal{M})$ mit Werten in $\{0, 1\}$ herbeigeführt, wobei die Entscheidung für 0 Annahme von H_0 bedeutet.

δ kann eindeutig charakterisiert werden durch

$$\phi : (M, \mathcal{M}) \to ([0, 1], \mathcal{B}), \quad \phi(x) = \delta(x, \{1\}),$$

und umgekehrt definiert jede meßbare Funktion $\phi : M \to [0, 1]$ auf diese Weise eine randomisierte Entscheidungsfunktion. ϕ wird als **Test** bezeichnet, und D sei die Menge aller Tests von H_0 gegen H_1. Ist δ nichtrandomisiert, so nimmt $\phi(x)$ nur die Werte 0 oder 1 an, und ϕ ist ein Test im Sinne von Kapitel 1.

$$m_\phi : \Theta \to [0, 1], \quad m_\phi(\vartheta) := E_\vartheta[\phi]$$

heißt die **Machtfunktion** von ϕ. Für ein nichtrandomisiertes ϕ ist $m_\phi(\vartheta) = P_\vartheta\{\phi = 1\}$ und stimmt daher mit der früher gegebenen Definition überein.

$$D(\alpha) := \{\phi \in D : m_\phi(\vartheta) \leq \alpha \, \forall \vartheta \in \Theta_0\}$$

ist die Menge der **Tests zum Niveau** α, kurz α**-Tests**. Die Größe

$$\sup_{\vartheta \in \Theta_0} m_\phi(\vartheta) \leq \alpha$$

heißt **Umfang** von ϕ.

Unter allen α-Tests interessieren uns diejenigen am meisten, für die $E_\vartheta[1 - \phi] = 1 - m_\phi(\vartheta)$ für $\vartheta \in \Theta_1$ (Fehler 2. Art) möglichst klein ist. Diese Fehler werden durch die **Macht** von ϕ,

$$M(\phi) := \inf_{\vartheta \in \Theta_1} m_\phi(\vartheta)$$

kontrolliert.

Wie schon in Beispiel 4.2.2 gezeigt, kann man zwei Tests ϕ_1 und ϕ_2 im Rahmen der Risikotheorie folgendermaßen vergleichen: Definiere eine Verlustfunktion $L(\vartheta, e)$ durch

$$\begin{aligned} L(\vartheta,1) &= \begin{cases} a & \text{für } \vartheta \in \Theta_0 \\ 0 & \text{für } \vartheta \in \Theta_1 \end{cases}, \\ L(\vartheta,0) &= \begin{cases} 0 & \text{für } \vartheta \in \Theta_0 \\ b & \text{für } \vartheta \in \Theta_1 \end{cases}, \end{aligned}$$

wobei a und b positive reelle Zahlen sind. Ein Fehler 1.Art wird also mit der „Strafe“ a belegt, ein Fehler 2.Art mit b. Für $\vartheta, \vartheta' \in \Theta$ ist

$$\begin{aligned} & \int \left(\int L(\vartheta', e) \delta_\phi(x, de) \right) dP_\vartheta(x) \\ = & \int \left(L(\vartheta',1)\phi(x) + L(\vartheta',0)(1-\phi(x)) \right) dP_\vartheta(x) \\ = & \; L(\vartheta',1) m_\phi(\vartheta) + L(\vartheta',0)(1 - m_\phi(\vartheta)) \\ = & \begin{cases} a \cdot m_\phi(\vartheta) & \text{für } \vartheta' \in \Theta_0 \\ b \cdot (1 - m_\phi(\vartheta)) & \text{für } \vartheta' \in \Theta_1 \end{cases}. \end{aligned} \tag{6.1}$$

Für $\vartheta = \vartheta'$ ist das gerade die Risikofunktion. Insbesondere gilt für $\vartheta \in \Theta_1$:

$$R(\vartheta, \phi_1) \leq R(\vartheta, \phi_2) \qquad \Longleftrightarrow \qquad m_{\phi_1}(\vartheta) \geq m_{\phi_2}(\vartheta),$$

und der risikotheoretische Vergleich zweier α-Tests hat sich nur auf die Machtfunktion zu stützen.

Sind $\phi_1, \phi_2 \in D(\alpha)$, so heißt ϕ_1 **schärfer** (oder **mächtiger**) als ϕ_2, falls $m_{\phi_1}(\vartheta) \geq m_{\phi_2}(\vartheta)$ für alle $\vartheta \in \Theta_1$.

Ein Test ϕ heißt **unverfälscht**, wenn $\sup_{\vartheta \in \Theta_0} m_\phi(\vartheta) \leq \inf_{\vartheta \in \Theta_1} m_\phi(\vartheta)$.

Lemma 6.1.1 *Ein Test ϕ ist unverfälscht genau dann, wenn er (im Sinne der Entscheidungstheorie) erwartungstreu ist.*

Beweis: Erwartungstreue (im Sinne der Entscheidungstheorie) ist wegen (6.1) äquivalent zu

$$a \cdot m_\phi(\vartheta) \leq b \cdot (1 - m_\phi(\vartheta)) \quad \forall \vartheta \in \Theta_0$$

und

$$b \cdot (1 - m_\phi(\vartheta)) \leq a \cdot m_\phi(\vartheta) \quad \forall \vartheta \in \Theta_1,$$

und das wiederum ist äquivalent zu

$$\sup_{\vartheta \in \Theta_0} m_\phi(\vartheta) \leq \frac{b}{a+b} \leq \inf_{\vartheta \in \Theta_1} m_\phi(\vartheta).$$

□

Lemma 6.1.2 *Ist $\phi \in D(\alpha)$ zulässig in $D(\alpha)$, so auch in D.*

Beweis: Sei ϕ zulässig in $D(\alpha)$. Angenommen, es gibt ein $\phi' \in D$, das besser als ϕ ist. Insbesondere ist dann

$$a \cdot m_{\phi'}(\vartheta) = R(\vartheta, \phi') \leq R(\vartheta, \phi) = a \cdot m_\phi(\vartheta) \leq a \cdot \alpha \quad \forall \vartheta \in \Theta_0,$$

so daß $\phi' \in D(\alpha)$. Aus der Zulässigkeit von ϕ in $D(\alpha)$ folgt nun, daß ϕ' nicht besser als ϕ sein kann. Also gibt es in D keinen besseren Test als ϕ, d.h. ϕ ist zulässig in D. □

6.2 Einfache Hypothesen, Neyman-Pearson Tests

Wir beschränken uns in diesem Abschnitt auf den Fall einfacher Hypothesen, d.h. $\Theta_0 = \{\vartheta_0\}$ und $\Theta_1 = \{\vartheta_1\}$. Wir können dann annehmen, daß $\mathcal{P} = \{P_{\vartheta_0}, P_{\vartheta_1}\}$ durch ein μ dominiert ist (z.B. durch $\mu = P_{\vartheta_0} + P_{\vartheta_1}$). Sei

$$f_0 = \frac{dP_{\vartheta_0}}{d\mu}, \quad f_1 = \frac{dP_{\vartheta_1}}{d\mu}.$$

Definition 6.2.1 *Ein Test ϕ heißt* **Neyman-Pearson Test** *(NP Test), wenn es ein $c \in [0, +\infty]$ gibt, so daß*

$$\phi(x) = \begin{cases} 1 & falls \quad f_1(x) > c f_0(x) \\ 0 & falls \quad f_1(x) < c f_0(x) \\ \in [0,1] & falls \quad f_1(x) = c f_0(x) \end{cases}$$

(Beachte die Konvention $0 \cdot \infty = 0$.) Für NP Tests ϕ^*, ϕ' benutzen wir entsprechend c^*, c'.

Theorem 6.2.2 (Neyman-Pearson Lemma) *Für das Testen einer einfachen Hypothese $H_0 : \vartheta = \vartheta_0$ gegen eine einfache Alternative $H_1 : \vartheta = \vartheta_1$ gilt:*

1. *Zu jedem $0 \leq \alpha \leq 1$ existiert ein NP Test ϕ^* mit (Umfang=) $m_{\phi^*}(\vartheta_0) = \alpha$ und $\phi^* = \gamma^* = \text{const}$ auf $\{x : f_1(x) = c^* f_0(x)\}$.*
2. *Ein NP Test ϕ^* ist am schärfsten unter allen ϕ mit $m_\phi(\vartheta_0) \leq m_{\phi^*}(\vartheta_0)$.*
3. *Ist ϕ am schärfsten unter allen α-Tests, so ist ϕ μ-f.s. ein NP Test. Ist $m_\phi(\vartheta_0) < \alpha$, so ist $M(\phi) = m_\phi(\vartheta_1) = 1$.*
4. *Jeder zulässige Test ist μ-f.s. ein NP Test.*
5. *Sei ϕ^* ein NP Test. ϕ^* ist zulässig $\iff$*
   (a) $c^* \neq 0$ oder

(b) $c^* = 0$ *und* $\phi^*(x) = 0$ P_{ϑ_0}*-f.s. auf* $\{f_1 = 0\}$.

Beweis: Wir schreiben P_0 für P_{ϑ_0} und P_1 für P_{ϑ_1}.

1. Ist $\alpha = 0$, so erhalten wir für $c^* = \infty$ und $\gamma^* = 0$ einen NP Test ϕ^* mit

$$m_{\phi^*}(\vartheta_0) = \int_{\{f_1 > \infty \cdot f_0\}} f_0 \, d\mu \leq \int_{\{f_0 = 0\}} f_0 \, d\mu = 0 = \alpha.$$

Betrachte nun den Fall $\alpha > 0$. Für $c \in \mathbf{R}$ setze

$$\rho(c) := P_0\{x : f_1(x) > c f_0(x)\}.$$

- $\rho : \mathbf{R} \to \mathbf{R}$, $\rho(c)$ fällt in c,
- $\rho(c - 0) \geq \rho(c)$,
- $\rho(c + 0) = \rho(c)$, da $\rho(c) = P_0\{f_1 > c f_0\} = P_0(\cup_n \{f_1 > (c + \frac{1}{n}) f_0\}) = \lim_{n \to \infty} \rho(c + \frac{1}{n})$,
- $$\rho(-\epsilon) = \int_{\{f_1 > -\epsilon f_0\}} f_0 \, d\mu = \int_{\{f_1 \geq 0\}} f_0 \, d\mu = 1 \quad \forall \epsilon > 0,$$

  also $\rho(0 - 0) = 1$, und
- $$\lim_{c \to +\infty} \rho(c) = \int_{\{f_1 > \infty \cdot f_0\}} f_0 \, d\mu = 0.$$

Setze nun $c^* := \inf\{c \geq 0 : \rho(c) \leq \alpha\}$. Da $\alpha > 0$, ist $0 \leq c^* < \infty$. Es gilt

$$\rho(c^*) = \rho(c^* + 0) \leq \alpha \leq \rho(c^* - 0).$$

Ist $\rho(c^* - 0) = \rho(c^*)$, also $P_0\{f_1 = c f_0\} = 0$, so setze $\gamma^* = 0$. Dann ist $m_{\phi^*}(\vartheta_0) = P_0\{f_1 > c^* f_0\} = \rho(c^*) = \alpha$.

Ist $\rho(c^* - 0) > \rho(c^*)$, so setze $\gamma^* = \frac{\alpha - \rho(c^*)}{\rho(c^* - 0) - \rho(c^*)}$. Dann ist ebenfalls

$$m_{\phi^*}(\vartheta_0) = P_0\{f_1 > c^* f_0\} + \gamma^* P_0\{f_1 = c^* f_0\} = \rho(c^*) + \gamma^*(\rho(c^* - 0) - \rho(c^*)) = \alpha.$$

2. Sei ϕ^* ein NP Test, $\phi \in D$ mit $m_\phi(\vartheta_0) \leq m_{\phi^*}(\vartheta_0)$. Dann gilt

$$\begin{aligned} m_{\phi^*}(\vartheta_1) - m_\phi(\vartheta_1) &= \int (\phi^* - \phi) f_1 d\mu \\ &= \underbrace{\int (\phi^* - \phi)(f_1 - c^* f_0) d\mu}_{=:A} + \underbrace{c^* \int (\phi^* - \phi) f_0 d\mu}_{=:B}. \end{aligned} \quad (6.2)$$

Da $B = c^*(m_{\phi^*}(\vartheta_0) - m_\phi(\vartheta_0)) \geq 0$ nach Voraussetzung, bleibt zu zeigen, daß $A \geq 0$:

- Auf $\{\phi^* - \phi > 0\}$ ist $\phi^* > 0$, also $f_1 \geq c^* f_0$.
- Auf $\{\phi^* - \phi < 0\}$ ist $\phi^* < 1$, also $f_1 \leq c^* f_0$.

Zusammen folgt:

$$(\phi^* - \phi)(f_1 - c^* f_0) \geq 0, \tag{6.3}$$

also $A \geq 0$.

3. Sei ϕ ein schärfster α-Test und ϕ^* der in 1 konstruierte NP Test. Da ϕ^* als NP Test auch am schärfsten ist, ist $m_\phi(\vartheta_1) = m_{\phi^*}(\vartheta_1)$, und da $m_\phi(\vartheta_0) \leq \alpha = m_{\phi^*}(\vartheta_0)$, folgt wie in (6.2), daß

$$0 = m_{\phi^*}(\vartheta_1) - m_\phi(\vartheta_1) = A + B \quad \text{mit } A, B \geq 0,$$

also $A = B = 0$.

Sei $S = \{\phi \neq \phi^*\} \cap \{f_1 - c^* f_0 \neq 0\}$. Wegen (6.3) ist $(\phi^* - \phi)(f_1 - c^* f_0) > 0$ auf S, so daß $\mu(S) = 0$ wegen $A = 0$. Also ist $\{\phi \neq \phi^*\} \subseteq \{f_1 = c^* f_0\}$ μ-f.s., so daß ϕ μ-f.s. ein NP Test ist.

Ist $m_\phi(\vartheta_0) < \alpha$, so folgt aus

$$0 = B = c^*(m_{\phi^*}(\vartheta_0) - m_\phi(\vartheta_0)) = c^*(\alpha - m_\phi(\vartheta_0)),$$

daß $c^* = 0$. Insbesondere ist $\phi^* = 1$ auf $\{f_1 > 0\}$, also $m_{\phi^*}(\vartheta_1) = \int \phi^* f_1 \, d\mu = 1$. Da ϕ schärfster α-Test ist, muß auch $m_\phi(\vartheta_1) = 1$ sein.

4. Sei ϕ zulässig, $\alpha := m_\phi(\vartheta_0)$. Da ϕ zulässig ist, ist ϕ am schärfsten zum Niveau α, also nach dem eben gezeigten μ-f.s. ein NP Test.

5. Sei nun ϕ^* ein NP Test.

   $\Rightarrow$: Ist ϕ^* zulässig und $c^* = 0$, so betrachte $\phi = I_{\{f_1 > 0\}}$. Dann ist

   $$m_\phi(\vartheta_0) = E_{\vartheta_0}[\phi] \leq E_{\vartheta_0}[\phi^*] = m_{\phi^*}(\vartheta_0)$$

   und

   $$m_\phi(\vartheta_1) = \int \phi f_1 \, d\mu = 1 \geq m_{\phi^*}(\vartheta_1),$$

   und wegen der Zulässigkeit von ϕ^* herrscht in beiden Ungleichungen Gleichheit, so daß aus $\phi \leq \phi^*$ insbesondere folgt: $\phi = \phi^*$ P_{ϑ_0}-f.s., d.h. $\phi^*(x) = 0$ P_{ϑ_0}-f.s. auf $\{f_1 = 0\}$.

   $\Leftarrow$: Setze $\alpha := m_{\phi^*}(\vartheta_0)$. Wegen Aussage 2 ist ϕ^* am schärfsten zum Niveau α. Sei ϕ ein Test mit

   $$m_\phi(\vartheta_0) \leq m_{\phi^*}(\vartheta_0) = \alpha \quad \text{und} \quad m_\phi(\vartheta_1) \geq m_{\phi^*}(\vartheta_1).$$

   Zu zeigen: In beiden Ungleichungen herrscht Gleichheit.

Da ϕ^* am schärfsten zum Niveau α ist, ist auch ϕ am schärfsten zum Niveau α, und es gilt $m_{\phi^*}(\vartheta_1) = m_\phi(\vartheta_1)$. Annahme:

$$m_\phi(\vartheta_0) < m_{\phi^*}(\vartheta_0) = \alpha. \tag{6.4}$$

Dann folgt aus 3

$$m_{\phi^*}(\vartheta_1) = m_\phi(\vartheta_1) = 1,$$

also $\phi = \phi^* = 1$ μ-f.s. auf $\{f_1 > 0\}$. Sowohl aus (5a) als auch aus (5b) folgt andererseits, daß $\phi^* = 0$ μ-f.s. auf $\{f_1 = 0, f_0 > 0\}$. Also ist $\phi^* \leq \phi$ μ-f.s. auf $\{f_0 > 0\}$ im Widerspruch zu (6.4).

□

Korollar 6.2.3 *Jeder NP Test ist unverfälscht, d.h.* $m_{\phi^*}(\vartheta_0) \leq m_{\phi^*}(\vartheta_1)$. *Ist* $P_{\vartheta_0} \neq P_{\vartheta_1}$ *und* $0 < m_{\phi^*}(\vartheta_0) < 1$, *so gilt sogar* $m_{\phi^*}(\vartheta_0) < m_{\phi^*}(\vartheta_1)$.

Beweis: Sei ϕ^* ein NP Test mit $m_{\phi^*}(\vartheta_0) = \alpha$, und betrachte den Test $\phi \equiv \alpha$. Da ϕ^* am schärfsten in $D(\alpha)$ ist, folgt: $m_{\phi^*}(\vartheta_0) = \alpha = m_\phi(\vartheta_1) \leq m_{\phi^*}(\vartheta_1)$.

Wäre $m_{\phi^*}(\vartheta_0) = m_{\phi^*}(\vartheta_1)$, also $m_\phi(\vartheta_1) = m_{\phi^*}(\vartheta_1) = \alpha$, so wäre $m_\phi = m_{\phi^*}$, und da ϕ^* als NP Test am schärfsten in $D(\alpha)$ ist, wäre auch ϕ am schärfsten in $D(\alpha)$ und insbesondere μ-f.s. ein NP Test. Wegen $0 < \alpha < 1$ würde folgen: $f_1 = cf_0$ μ-f.s., also $P_{\vartheta_0} = P_{\vartheta_1}$, ein Widerspruch. □

6.3 Tests bei isotonen Dichtequotienten

Will man sich bei der Gewinnung optimaler Tests von der Beschränkung auf einfache Hypothesen lösen, so muß man speziellere Annahmen über den zugrundeliegenden statistischen Raum machen. Eine Möglichkeit bietet

Definition 6.3.1 *Sei* $(M, \mathcal{M}, \{P_\vartheta : \vartheta \in \Theta\})$ *ein dominierter statistischer Raum mit Dichten* f_ϑ, $P_\vartheta \neq P_{\vartheta'} \forall \vartheta \neq \vartheta'$.

$\{P_\vartheta : \vartheta \in \Theta\}$ *(oder* $(M, \mathcal{M}, \{P_\vartheta : \vartheta \in \Theta\})$*) hat* **monotonen Dichtequotienten** *(oder* **monotonen Likelihood Quotienten***), wenn eine Zufallsvariable* $T : M \to \mathbf{R}$ *und für alle* ϑ, ϑ' *eine monotone Abbildung* $h_{\vartheta,\vartheta'} : \mathbf{R} \to [0, +\infty]$ *existieren, so daß*

$$\frac{f_{\vartheta'}(x)}{f_\vartheta(x)} = h_{\vartheta,\vartheta'}(T(x)) \quad (P_\vartheta + P_{\vartheta'}) - f.s.$$

Ist Θ *totalgeordnet und ist für* $\vartheta < \vartheta'$ *die Funktion* $h_{\vartheta,\vartheta'}$ *monoton steigend, so hat* $\{P_\vartheta : \vartheta \in \Theta\}$ *einen* **isotonen** *Dichtequotienten.*

Beachte: Die $h_{\vartheta,\vartheta'}$ *werden nicht als strikt monoton vorausgesetzt.*

Beispiel 6.3.2 Wir betrachten den statistischen Raum $(\mathbf{R}, \mathcal{B}, \mathcal{P})^n$ mit $\mathcal{P} = \{\mathcal{N}(\mu, 1) : \mu \in \mathbf{R}\}$.

Für $\mu < \mu'$ ist

$$\frac{f_{\mu'}}{f_\mu} = \exp(-\frac{n}{2}(\mu'^2 - \mu^2)) \exp((\mu' - \mu) \sum_{i=1}^{n} x_i),$$

also $T(x) = \sum_{i=1}^n x_i$.

Allgemeiner gilt:

Bemerkung 6.3.3 Ist $(M, \mathcal{M}, \{P_\vartheta : \vartheta \in \Theta\})$ ein 1-parametriger Exponentialraum mit $f_\vartheta = C(\vartheta)h(x)\exp(c_1(\vartheta)h_1(x))$, ist Θ totalgeordnet und ist c_1 isoton in ϑ, so hat $(M, \mathcal{M}, \{P_\vartheta : \vartheta \in \Theta\})$ einen isotonen Dichtequotienten, denn für $\vartheta < \vartheta'$ ist

$$\frac{f_{\vartheta'}(x)}{f_\vartheta(x)} = \frac{C(\vartheta')}{C(\vartheta)} \exp((c_1(\vartheta') - c_1(\vartheta))h_1(x)).$$

Theorem 6.3.4 *Sei $(M, \mathcal{M}, \{P_\vartheta : \vartheta \in \Theta\})$ ein statistischer Raum mit isotonem Dichtequotienten, $\vartheta_0 \in \Theta$. Dann gilt:*

1. *Zu jedem $\alpha \in [0, 1]$ existiert ein schärfster α-Test von $H_0 : \vartheta \leq \vartheta_0$ gegen $H_1 : \vartheta > \vartheta_0$, der die folgende Form hat:*

$$\phi(x) = \begin{cases} 1 & falls \quad T(x) > \tau \\ 0 & falls \quad T(x) < \tau \\ \gamma & falls \quad T(x) = \tau \end{cases},$$

   wo sich $\tau \in [-\infty, \infty]$ und $\gamma \in [0, 1]$ (nicht notwendig eindeutig) aus $m_\phi(\vartheta_0) = \alpha$ bestimmen lassen.

2. *Der in 1. für ϑ_0 bestimmte Test ϕ ist auch für alle $\vartheta_0' \in \Theta$ am schärfsten für $H_0' : \vartheta \leq \vartheta_0'$ gegen $H_1' : \vartheta > \vartheta_0'$ zum Niveau $\alpha' = m_\phi(\vartheta_0')$.*

3. *$m_\phi(\vartheta)$ ist strikt monoton wachsend auf $\{\vartheta \in \Theta : 0 < m_\phi(\vartheta) < 1\}$.*

4. *Sei $\vartheta < \vartheta_0$. ϕ minimiert $m_\psi(\vartheta)$ unter allen Tests ψ mit $m_\psi(\vartheta_0) = \alpha$.*

Zum Beweis benötigen wir

Lemma 6.3.5 *$(M, \mathcal{M}, \{P_\vartheta : \vartheta \in \Theta\})$ habe isotonen Dichtequotienten (in T) wie in Definition 6.3.1, und $g : \mathbf{R} \to \mathbf{R}$ sei eine beschränkte isotone Funktion. Dann ist $\vartheta \mapsto E_\vartheta[g \circ T]$ isoton.*

Beweis: Sei $\vartheta < \vartheta'$,

$$A := \{h_{\vartheta,\vartheta'} < 1\}, \quad B := \{h_{\vartheta,\vartheta'} \geq 1\}.$$

Die Isotonie von $h_{\vartheta,\vartheta'}$ impliziert $a := \sup A \leq b := \inf B$, und aus der Isotonie von g folgt $g(a) \leq g(b)$.

Sei $Q_\vartheta := P_\vartheta \circ T^{-1}$. Beachte zunächst, daß

$$\begin{aligned} & \int_A (h_{\vartheta,\vartheta'} - 1) dQ_\vartheta + \int_B (h_{\vartheta,\vartheta'} - 1) dQ_\vartheta \\ = & \int_M (h_{\vartheta,\vartheta'} \circ T - 1) f_\vartheta d\mu = \int_M (f_{\vartheta'} - f_\vartheta) d\mu \\ = & \ 1 - 1 = 0. \end{aligned}$$

Da $(h_{\vartheta,\vartheta'} - 1) < 0$ auf A und $(h_{\vartheta,\vartheta'} - 1) \geq 0$ auf B, folgt nun

$$\begin{aligned} & E_{\vartheta'}[g \circ T] - E_\vartheta[g \circ T] \\ = & \int_M g \circ T (f_{\vartheta'} - f_\vartheta) d\mu \\ = & \int_M g \circ T (h_{\vartheta,\vartheta'} \circ T - 1) f_\vartheta d\mu \\ = & \int_R g (h_{\vartheta,\vartheta'} - 1) dQ_\vartheta \\ \geq & \ g(a) \int_A (h_{\vartheta,\vartheta'} - 1) dQ_\vartheta + g(b) \int_B (h_{\vartheta,\vartheta'} - 1) dQ_\vartheta \\ = & \ (g(b) - g(a)) \int_B (h_{\vartheta,\vartheta'} - 1) dQ_\vartheta \\ \geq & \ 0. \end{aligned}$$

□

Beweis zu Theorem 6.3.4:

1. Sei zunächst $\vartheta_1 > \vartheta_0$ fest. Setze

$$\tau := \inf\{t \in \mathbf{R} : P_{\vartheta_0}\{T > t\} \leq \alpha\}.$$

Wegen der üblichen Monotonie- und einseitigen Stetigkeitseigenschaften von Verteilungsfunktionen ist

$$P_{\vartheta_0}\{T > \tau\} \leq \alpha \leq P_{\vartheta_0}\{T \geq \tau\}.$$

Setze $c := h_{\vartheta_0,\vartheta_1}(\tau)$ und

$$\phi(x) := \begin{cases} 1 & falls \quad T(x) > \tau \\ 0 & falls \quad T(x) < \tau \\ \gamma & falls \quad T(x) = \tau \end{cases},$$

wo

$$\gamma := \begin{cases} \frac{\alpha - P_{\vartheta_0}\{T>\tau\}}{P_{\vartheta_0}\{T=\tau\}} & falls \quad P_{\vartheta_0}\{T=\tau\} \neq 0 \\ 0 & sonst \end{cases} .$$

Es gilt

$$f_1(x) > cf_0(x) \Rightarrow h_{\vartheta_0,\vartheta_1}(T(x)) > c \Rightarrow T(x) > \tau \Rightarrow \phi(x) = 1,$$
$$f_1(x) < cf_0(x) \Rightarrow h_{\vartheta_0,\vartheta_1}(T(x)) < c \Rightarrow T(x) < \tau \Rightarrow \phi(x) = 0,$$

so daß ϕ ein NP Test ist. Aus der Definition von ϕ folgt sofort $m_\phi(\vartheta_0) = \alpha$, und wegen des NP Lemmas ist ϕ am schärfsten zum Niveau α.

Setze nun

$$g(t) := \begin{cases} 1 & falls \quad t > \tau \\ \gamma & falls \quad t = \tau \\ 0 & falls \quad t < \tau \end{cases} .$$

Dann ist $\phi(x) = g(T(x))$, und aus dem vorangegangenen Lemma folgt

$$E_\vartheta[\phi] \leq E_{\vartheta_0}[\phi] = \alpha \quad \forall \vartheta \leq \vartheta_0.$$

Also ist ϕ auch schärfster α-Test von H_0 gegen $\{\vartheta_1\}$.

Da die Festsetzung von τ und damit die von ϕ unabhängig von ϑ_1 erfolgte, ist ϕ auch schärfster α-Test von H_0 gegen $H_1 : \vartheta > \vartheta_0$.

2. Für $\vartheta_0' < \vartheta_1'$ gilt:

$$\phi(x) = g(T(x)) = \begin{cases} 1, \text{ falls } h_{\vartheta_0',\vartheta_1'}(T(x)) > h_{\vartheta_0',\vartheta_1'}(\tau) =: c' \\ 0, \text{ falls } h_{\vartheta_0',\vartheta_1'}(T(x)) < h_{\vartheta_0',\vartheta_1'}(\tau) \end{cases} .$$

Also ist ϕ ein NP Test von $\{\vartheta_0'\}$ gegen $\{\vartheta_1'\}$ zum Niveau $\alpha' = m_\phi(\vartheta_0')$. Wie oben folgt, daß ϕ auch schärfster α'-Test von H_0' gegen H_1' ist.

3. folgt nun sofort aus Korollar 6.2.3.

4. Sei $\vartheta < \vartheta_0$. Setze $c' := 1/h_{\vartheta,\vartheta_0}(\tau)$. Dann ist

$$\phi'(x) := \begin{cases} 1 & falls \quad f_\vartheta(x) > c' f_{\vartheta_0}(x) \\ 0 & falls \quad f_\vartheta(x) < c' f_{\vartheta_0}(x) \\ 1 - \phi(x) & falls \quad f_\vartheta(x) = c' f_{\vartheta_0}(x) \end{cases}$$

ein NP Test von $\{\vartheta_0\}$ gegen $\{\vartheta\}$. Da

$$T(x) > \tau \Rightarrow h_{\vartheta,\vartheta_0}(T(x)) \geq 1/c' \Rightarrow f_\vartheta(x) \leq c' f_{\vartheta_0}(x),$$
$$T(x) < \tau \Rightarrow h_{\vartheta,\vartheta_0}(T(x)) \leq 1/c' \Rightarrow f_\vartheta(x) \geq c' f_{\vartheta_0}(x),$$
$$T(x) = \tau \Rightarrow h_{\vartheta,\vartheta_0}(T(x)) = 1/c' \Rightarrow \phi'(x) = 1 - \phi(x),$$

ist $\phi' = 1 - \phi$, also $m_{\phi'}(\vartheta_0) = 1 - m_\phi(\vartheta_0) = 1 - \alpha$.

Ist nun ψ ein weiterer Test mit $m_\psi(\vartheta_0) = \alpha$, so ist $1 - \psi$ ein Test von $\{\vartheta_0\}$ gegen $\{\vartheta\}$ zum Niveau $1 - \alpha$, und da $\phi' = 1 - \phi$ als NP Test am schärfsten zu diesem Niveau ist, ist $m_{1-\psi}(\vartheta) \leq m_{1-\phi}(\vartheta)$, also $m_\psi(\vartheta) \geq m_\phi(\vartheta)$.

□

Fortsetzung von Beispiel 6.3.2: Gesucht ist ein schärfster Test von $H_0 : \mu \leq \mu_0$ gegen $H_1 : \mu > \mu_0$. Da $T = n\bar{X}$ stetig verteilt ist, hat ϕ f.s. die Struktur

$$\phi(X_1, \ldots, X_n) = \begin{cases} 1, & \text{falls } n\bar{X} > c \\ 0, & \text{falls } n\bar{X} < c \end{cases} .$$

Zu gegebenem α ist τ nun so zu bestimmen, daß

$$m_\phi(\mu_0) = P_{\mu_0}\{\phi = 1\} = P_{\mu_0}\{n\bar{X} > \tau\} = \alpha,$$

was wegen der Normalität von $n\bar{X}$ immer möglich ist. Es ergibt sich $\tau = \sqrt{n}u_{1-\alpha} + n\mu_0$, wo $u_{1-\alpha}$ das $(1-\alpha)$-Quantil der Standardnormalverteilung bezeichnet.

Sei nun der Einfachheit halber $\mu_0 = 0$. ϕ ist (nach Konstruktion) auch schärfster α-Test von $\{\mu = 0\}$ gegen $\{\mu > 0\}$, und aus Symmetriegründen ist $\phi'(x) = \phi(-x)$ schärfster α-Test von $\{\mu \geq 0\}$ bzw. $\{\mu = 0\}$ gegen $\{\mu < 0\}$.

Gäbe es auch einen schärfsten α-Test ψ von $\{\mu = 0\}$ gegen $\{\mu \neq 0\}$, so wäre er auch schärfster Test von $\{\mu = 0\}$ gegen $\{\mu = \mu_-\}$ und von $\{\mu = 0\}$ gegen $\{\mu = \mu_+\}$, $\mu_- < 0 < \mu_+$. Als solcher hätte er einen strikten Ablehnungsbereich, der wegen des NP Lemmas f.s. die Form $\{T > c_+\}$ und auch $\{T < c_-\}$ haben muß. Für normalverteiltes T können diese beiden Bereiche aber nicht fast sicher übereinstimmen.

6.4 Ungünstigste a priori-Verteilungen

Eine weitere Möglichkeit, ein Testproblem mit zusammengesetztem H_0 auf eines mit einfachem H_0 (bei einfachem H_1) zu reduzieren, besteht in der Suche nach ungünstigsten a priori-Verteilungen.

Sei $\Theta_0 \subseteq \Theta$ beliebig, $\Theta_1 = \{\vartheta_1\}$. Wir nehmen an, daß Θ_0 mit einer σ-Algebra $\mathcal{T}_0$ ausgerüstet ist und daß $P_\vartheta(A)$ als stochastischer Kern von $(\Theta_0, \mathcal{T}_0)$ nach $(M, \mathcal{M})$ aufgefaßt werden kann. Für ein Wahrscheinlichkeitsmaß λ auf $(\Theta_0, \mathcal{T}_0)$ und für $A \in \mathcal{M}$ sei

$$P_\lambda(A) := \int_{\Theta_0} P_\vartheta(A) d\lambda(\vartheta).$$

Mit $\phi_{\lambda,\alpha}$ werde ein schärfster α-Test von $P = P_\lambda$ gegen $P = P_{\vartheta_1}$ bezeichnet, der wegen des NP Lemmas existiert.

Definition 6.4.1 *λ heißt eine* **ungünstigste a priori-Verteilung** *zum Niveau α bzgl. Θ_0 und $\Theta_1 = \{\vartheta_1\}$, wenn für jede andere Wahrscheinlichkeitsverteilung ν auf $(\Theta_0, \mathcal{T}_0)$ gilt:*

$$m_{\phi_{\nu,\alpha}}(\vartheta_1) \geq m_{\phi_{\lambda,\alpha}}(\vartheta_1).$$

(Ein solches λ braucht aber weder zu existieren noch eindeutig zu sein.)

Theorem 6.4.2 *Sei λ eine Wahrscheinlichkeitsverteilung auf $(\Theta_0, \mathcal{T}_0)$. Ist $\phi_{\lambda,\alpha}$ ein α-Test von Θ_0 gegen $\Theta_1 = \{\vartheta_1\}$, so gilt:*

1. *$\phi_{\lambda,\alpha}$ ist schärfster α-Test von Θ_0 gegen Θ_1.*
2. *Sei $\{P_\vartheta : \vartheta \in \Theta_0 \cup \{\vartheta_1\}\}$ durch μ dominiert. Ist $\phi_{\lambda,\alpha}$ als schärfster α-Test von $P = P_\lambda$ gegen $P = P_{\vartheta_1}$ μ-f.s. eindeutig bestimmt, so auch als schärfster α-Test von Θ_0 gegen Θ_1.*
3. *λ ist ungünstigste a priori Verteilung zum Niveau α*

Beweis:

1. Sei $\phi' \in D(\alpha)$. Zu zeigen:
$$m_{\phi'}(\vartheta_1) \leq m_{\phi_{\lambda,\alpha}}(\vartheta_1). \tag{6.5}$$
Es gilt
$$\int_M \phi' \, dP_\lambda = \int_{\Theta_0} (\int_M \phi'(x) dP_\vartheta(x)) d\lambda(\vartheta) \leq \int_{\Theta_0} \alpha \, d\lambda(\vartheta) = \alpha.$$
Also ist ϕ' Test von $P = P_\lambda$ gegen $P = P_{\vartheta_1}$ zum Niveau α, und da $\phi_{\lambda,\alpha}$ schärfster α-Test für diese Situation ist, folgt (6.5).
2. Ist nun ϕ' auch ein schärfster α-Test von Θ_0 gegen Θ_1, so gilt Gleichheit in (6.5), und da ϕ' auch α-Test von $P = P_\lambda$ gegen $P = P_{\vartheta_1}$ ist, ist $\phi' = \phi_{\lambda,\alpha}$ μ-f.s.
3. Sei ν eine andere Wahrscheinlichkeitsverteilung auf $(\Theta_0, \mathcal{T}_0)$. Dann ist $\phi_{\lambda,\alpha}$ Test zum Niveau α von $P = P_\nu$ gegen $P = P_{\vartheta_1}$, denn
$$\int_M \phi_{\lambda,\alpha} \, dP_\nu = \int_{\Theta_0} (\int_M \phi_{\lambda,\alpha}(x) dP_\vartheta(x)) d\nu(\vartheta) \leq \int_{\Theta_0} \alpha \, d\nu(\vartheta) = \alpha.$$
Da $\phi_{\nu,\alpha}$ schärfster α-Test von $P = P_\nu$ gegen $P = P_{\vartheta_1}$ ist, folgt
$$m_{\phi_{\lambda,\alpha}}(\vartheta_1) \leq m_{\phi_{\nu,\alpha}}(\vartheta_1).$$

□

Beispiel 6.4.3 [Zeichentest] (Siehe Krengel [7, S.69])

Wir betrachten den statistischen Raum $(\mathbf{R}, \mathcal{B}, \mathcal{P})^n$, wo $\mathcal{P}$ die Familie aller Wahrscheinlichkeitsverteilungen auf $\mathbf{R}$ ist. Formal: $\Theta = \mathcal{P}$ und $\mathcal{P}^n = \{P^n : P \in \Theta\}$. Um den vorherigen Satz anwenden zu können, versehen wir $\mathcal{P}$ mit der Potenzmenge von $\mathcal{P}$ als σ-Algebra.

Sei $u \in \mathbf{R}$, $U = (-\infty, u]$, $p_0 \in (0,1)$,
$$\Theta_0 = \{P \in \mathcal{P} : P(U) \geq p_0\}, \; \Theta_1 = \mathcal{P} \setminus \Theta_0.$$

Für $u = 0$ und $p_0 = \frac{1}{2}$ erhält man den klassischen (Vor-)**Zeichentest**.

Gesucht ist ein schärfster α-Test von Θ_0 gegen Θ_1. Zunächst ordnet man jedem $P \in \mathcal{P}$ ein Tripel (p, P^+, P^-) zu, wobei

$$p = P(U),\ P^+ = P(.|U^C),\ P^- = P(.|U).$$

Ist $p = 0$ ($p = 1$), so kann man P^- (P^+) beliebig definieren. Dann ist $P = pP^- + (1 - p)P^+$.

Sei nun zunächst $P_1 = (p_1, P_1^+, P_1^-)$ mit $p_1 < p_0$ eine einfache Alternative. Die zu P_1 ähnlichste unter allen Verteilungen in Θ_0 ist (p_0, P_1^+, P_1^-), so daß wir als ungünstigste a priori Verteilung λ die Einheitsmasse auf diesem Element von $\mathcal{P}$ erwarten, also

$$P_\lambda = (p_0, P_1^+, P_1^-).$$

Da $P_1 << P_\lambda$ und

$$\frac{dP_1}{dP_\lambda} = \frac{p_1}{p_0} I_U + \frac{1 - p_1}{1 - p_0} I_{U^C},$$

also

$$\frac{dP_1^n}{dP_\lambda^n}(x) = \left(\frac{p_1}{p_0}\right)^{m(x)} \left(\frac{1 - p_1}{1 - p_0}\right)^{n - m(x)},$$

wo $m(x) = \sum_{i=1}^n I_U(x_i)$. Da $p_1 < p_0$, hängt der Dichtequotient antiton von $m(x)$ ab, so daß ein schärfster Test von $P = P_\lambda$ gegen $P = P_1$ nach dem NP Lemma die Form

$$\phi_{\lambda,\alpha}(x) = \begin{cases} 1 & falls \quad m(x) < c \\ 0 & falls \quad m(x) > c \\ \gamma & falls \quad m(x) = c \end{cases},$$

wo γ und c so zu bestimmen sind, daß

$$m_{\phi_{\lambda,\alpha}}(P_\lambda) = P_\lambda\{m < c\} + \gamma P_\lambda\{m = c\} = \alpha.$$

Sind die $X_1, \ldots, X_n$ nach $P = (p, P^+, P^-)$ verteilt, so ist m nach $B(n, p)$ verteilt (unabhängig von P^+ und P^-). Aus Lemma 6.3.5 folgt: $m_{\phi_{\lambda,\alpha}}(P) = m_{\phi_{\lambda,\alpha}}(p(P))$ ist fallende, nur von $p(P)$ abhängige Funktion. Also ist

$$m_{\phi_{\lambda,\alpha}}(p, P^+, P^-) \leq m_{\phi_{\lambda,\alpha}}(p_0, P^+, P^-) = m_{\phi_{\lambda,\alpha}}(p_0, P_1^+, P_1^-) = m_{\phi_{\lambda,\alpha}}(P_\lambda) = \alpha,$$

so daß $\phi_{\lambda,\alpha}$ ein α-Test von Θ_0 gegen $P = P_1$ ist. Aus dem vorherigen Satz folgt nun: $\phi_{\lambda,\alpha}$ ist schärfster α-Test von Θ_0 gegen $P = P_1$.

Da die Festlegung von c und γ (und damit von $\phi_{\lambda,\alpha}$) unabhängig von der speziellen Alternative $P = P_1$ erfolgte, ist $\phi_{\lambda,\alpha}$ sogar schärfster α-Test von Θ_0 gegen Θ_1.

Beispiel 6.4.4 [Anwendungen auf Tests für Normalverteilungen] (Siehe Krengel [7, S.71]) Wir betrachten den statistischen Raum $(\mathbf{R}, \mathcal{B}, \mathcal{P})^n$, wo $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2) : \mu \in$

$\mathbf{R}, \sigma^2 > 0\}$. Die Dichte von $\mathcal{N}(\mu, \sigma^2)^n$ bezgl. des n-dimensionalen Lebesguemaßes ist

$$\begin{aligned} & f_{\mu,\sigma^2}(x_1, \ldots, x_n) \\ = \; & (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right) \\ = \; & (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \bar{x})^2\right) \exp\left(-\frac{1}{2\sigma^2} n(\bar{x} - \mu)^2\right). \end{aligned}$$

Sie hängt insbesondere außer von den Parametern nur von $\bar{x}$ und $u(x) := \sum_{i=1}^n (x_i - \bar{x})^2$ ab.

Sei nun $\Theta = \{\vartheta = (\mu, \sigma^2) : \mu \in \mathbf{R}, \sigma^2 > 0\}$. Ein plausibler Ansatz zur Konstruktion von Tests von Θ_0 gegen Θ_1 ist die **Maximum-Likelihood-Methode**: Bilde

$$q(x) := \frac{\sup_{\vartheta \in \Theta_0} f_\vartheta(x)}{\sup_{\vartheta \in \Theta_1} f_\vartheta(x)},$$

und setze in Verallgemeinerung der NP Methode

$$\phi(x) = \begin{cases} 1, \text{ falls } q(x) \le c \\ 0, \text{ falls } q(x) > c \end{cases}.$$

Sind Θ_0 und Θ_1 einfach, so ist ϕ ein NP Test. Aber schon, wenn $\Theta_0 \cup \Theta_1$ aus drei Parametern besteht, läßt sich diese Methode nicht mehr allgemein entscheidungstheoretisch begründen. Für viele der praktisch vorkommenden Verteilungsannahmen ergibt die ML Methode aber trotzdem einen Test, der sich auch vom Standpunkt der Risikotheorie als optimal erweist.

1. Betrachte $\Theta_0 = \{\vartheta \in \Theta : \sigma^2 \le \sigma_0^2\}$, $\Theta_1 = \Theta \setminus \Theta_0 = \{\vartheta \in \Theta : \sigma^2 > \sigma_0^2\}$.

   Die Suprema im Zähler und auch im Nenner von $q(x)$ werden sicher nur für $\mu = \bar{x}$ erreicht. Da

   $$\sigma^2 \mapsto (2\pi\sigma^2)^{-n/2} \exp(-u(x)/2\sigma^2)$$

   für $\sigma^2 < n^{-1}u(x)$ steigt und für $\sigma^2 > n^{-1}u(x)$ fällt, ist

   $$q(x) = \begin{cases} \left(\frac{u(x)}{n\sigma_0^2}\right)^{n/2} \exp\left(\frac{n}{2} - \frac{u(x)}{2\sigma_0^2}\right) & \text{für} \quad \sigma_0^2 \le n^{-1}u(x) \\ \left(\frac{u(x)}{n\sigma_0^2}\right)^{-n/2} \exp\left(-\frac{n}{2} + \frac{u(x)}{2\sigma_0^2}\right) & \text{für} \quad \sigma_0^2 \ge n^{-1}u(x) \end{cases},$$

   und da $q(x)$ als Funktion von $u(x)$ streng monoton fällt, hat $\phi(x)$ die Form

   $$\phi(x) = \begin{cases} 1, \text{ falls } u(x) \ge c \\ 0, \text{ falls } u(x) < c \end{cases}.$$

   ϕ ist also ein χ^2-Streuungstest, vergl. Kapitel 2.2. Wir zeigen durch Wahl einer ungünstigsten a priori-Verteilung: ϕ ist schärfster Test von $\{\sigma^2 \le \sigma_0^2\}$ gegen $\{\sigma^2 > \sigma_0^2\}$:

Sei zunächst $\vartheta_1 = (\mu_1, \sigma_1^2)$ mit $\sigma_1^2 > \sigma_0^2$ fest gewählt. Da $\mathcal{N}(\mu_1, \sigma_1^2)$ „flacher“ als die $\mathcal{N}(\mu, \sigma^2)$ mit $\sigma^2 \leq \sigma_0^2$ ist, mitteln wir über $\{\sigma^2 \leq \sigma_0^2\}$ so, daß eine ähnlich flache Verteilung wie $\mathcal{N}(\mu_1, \sigma_1^2)$ entsteht. Dazu sollte λ auf $\{\sigma^2 = \sigma_0^2\}$ konzentriert sein, und man zeigt, daß

$$\lambda = \mathcal{N}(\mu_1, \frac{\sigma_1^2 - \sigma_0^2}{n}) \quad \text{als Verteilung auf } \{\sigma^2 = \sigma_0^2\}$$

! auf die Dichte

$$f_\lambda(x_1, \ldots, x_n) = \text{const} \cdot \exp(-\frac{u(x)}{2\sigma_0^2}) \exp(-\frac{n(\bar{x} - \mu_1)^2}{2\sigma_1^2})$$

führt:

$$\begin{aligned} & f_\lambda(x_1, \ldots, x_n) \\ = \; & \text{const} \exp\left(-\frac{u(x)}{2\sigma_0^2}\right) \int \exp\left(-\frac{n}{2\sigma_0^2}(\bar{x} - \mu)^2\right) \exp\left(-\frac{n}{2(\sigma_1^2 - \sigma_0^2)}(\mu - \mu_1)^2\right) d\mu \end{aligned}$$

mit einer Konstanten, die nur von σ_0^2 und σ_1^2 abhängt. Das Integral läßt sich weiter auswerten zu

$$\begin{aligned} & \int \ldots d\mu \\ = \; & \int \exp\left(-\frac{n}{s} \left[\mu^2 \underbrace{\left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma_1^2 - \sigma_0^2} \right)}_{=:\frac{1}{\tau}} \right.\right. \\ & \qquad \left.\left. -2\mu \left(\frac{\bar{x}}{\sigma_0^2} + \frac{\mu_1}{\sigma_1^2 - \sigma_0^2} \right) + \frac{\bar{x}^2}{\sigma_0^2} + \frac{\mu_1^2}{\sigma_1^2 - \sigma_0^2} \right] \right) d\mu \\ = \; & \int \exp\left(-\frac{n}{2} \left[\frac{1}{\tau} \left(\mu - \tau(\frac{\bar{x}}{\sigma_0^2} + \frac{\mu_1}{\sigma_1^2 - \sigma_0^2}) \right)^2 \right.\right. \\ & \qquad \left.\left. -\tau \left(\frac{\bar{x}}{\sigma_0^2} + \frac{\mu_1}{\sigma_1^2 - \sigma_0^2} \right)^2 + \frac{\bar{x}}{\sigma_0^2} + \frac{\mu_1^2}{\sigma_1^2 - \sigma_0^2} \right] \right) d\mu \\ = \; & \text{const} \cdot \exp\left(-\frac{n}{2} \left[\bar{x}^2 \left(\frac{1}{\sigma_0^2} - \tau \frac{1}{\sigma_0^4} \right) \right.\right. \\ & \qquad \left.\left. -2\tau \frac{\bar{x}}{\sigma_0^2} \frac{\mu_1}{\sigma_1^2 - \sigma_0^2} + \mu_1^2 \left(\frac{1}{\sigma_1^2 - \sigma_0^2} - \frac{\tau}{(\sigma_1^2 - \sigma_0^2)^2} \right) \right] \right) d\mu \\ = \; & \text{const} \cdot \exp\left(-\frac{n}{2\sigma_1^2} (\bar{x} - \mu)^2 \right). \end{aligned}$$

Also ist

$$\frac{f_{\mu_1, \sigma_1^2}(x)}{f_\lambda(x)} = \text{const} \cdot \exp\left((\frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_1^2}) u(x) \right).$$

Da der Dichtequotient isoton in $u(x)$ ist, hat $\phi_{\lambda,\alpha}$ die Form

$$\phi_{\lambda,\alpha}(x) = \begin{cases} 1, & \text{falls} \quad u(x) \geq \tilde{c} \\ 0, & \text{falls} \quad u(x) < \tilde{c} \end{cases} .$$

Da, wie wir weiter unten zeigen, $u(X)$ unter P_λ und $\mathcal{N}(\mu, \sigma_0^2)$, μ beliebig, die gleiche Verteilung besitzt und die f_{μ,σ^2} bei festem μ einen isotonen Dichtequotienten als Funktion von $u(x)$ haben, ist für $\sigma^2 \leq \sigma_0^2$

$$E_{\mu,\sigma^2}[\phi_{\lambda,\alpha}] \leq E_{\mu,\sigma_0^2}[\phi_{\lambda,\alpha}] = E_{P_\lambda}[\phi_{\lambda,\alpha}] = \alpha,$$

also $\phi_{\lambda,\alpha}$ ein α-Test von $\Theta_0 = \{\sigma^2 \leq \sigma_0^2\}$ gegen $\{(\mu_1, \sigma_1^2)\}$. Nach Satz 6.4.1 ist $\phi_{\lambda,\alpha}$ sogar ein schärfster Test für diese Situation. Da die Wahl von c schließlich unabhängig von (μ_1, σ_1^2) erfolgte, ist $\phi_{\lambda,\alpha}$ sogar ein schärfster α-Test von Θ_0 gegen Θ_1.

Bleibt zu zeigen, daß $u(X)$ unter P_λ und $\mathcal{N}(\mu, \sigma_0^2)$, μ beliebig, die gleiche Verteilung besitzt:

$$\begin{aligned} & P_\lambda\{u(X) \leq z\} \\ = & \int I_{\{u(x)\leq z\}} f_\lambda(x)\, dx \\ = & \int I_{\{u(x)\leq z\}} \underbrace{(2\pi\sigma_0^2)^{-\frac{n}{2}} \exp\left(-\frac{u(X)}{2\sigma_0^2} - \frac{n(\bar{x}-\mu_1)^2}{2\sigma_0^2}\right)}_{=f_{\mu_1,\sigma_0^2}(x)} \\ & \qquad \underbrace{\cdot \text{const} \cdot \exp\left(-\frac{n}{2}(\bar{x}-\mu_1)^2\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right)\right)}_{=:\psi(\bar{x})}\, d\mu \\ = & E_{\mu_1,\sigma_0^2}\left[I_{\{u(X)\leq z\}}\right] \cdot E_{\mu_1,\sigma_0^2}\left[\psi(\bar{X})\right]. \end{aligned}$$

Angewandt auf $z = +\infty$ ergibt das $E_{\mu_1,\sigma_0^2}[\psi(\bar{X})] = 1$, so daß $P_\lambda\{u(X) \leq z\} = P_{\mu_1,\sigma_0^2}\{u(X) \leq z\}$.

2. Betrachtet man umgekehrt $\Theta_0' = \{\sigma^2 \geq \sigma_0^2\}$ und $\Theta_1' = \{\sigma^2 < \sigma_0^2\}$, so ergibt sich nach der ML Methode $q'(x) = 1/q(x)$ und daher

$$\phi'(x) = \begin{cases} 1, \text{ falls } u(x) \leq c' \\ 0, \text{ falls } u(x) > c' \end{cases} .$$

ϕ' ist jedoch nicht schärfster Test von Θ_0' gegen Θ_1', sondern nur am schärfsten unter allen unverfälschten (d.h. erwartungstreuen) Tests.

Wir zeigen zunächst die zweite Aussage. Sei ψ ein unverfälschter Test von Θ_0' gegen Θ_1'. Wegen der Stetigkeit von $m_\psi(\mu, \sigma^2)$ in σ^2, ist $m_\psi(\mu, \sigma_0^2) = \text{const} = \alpha$ für alle $\mu \in \mathbf{R}$. Zu zeigen ist daher:

Sei $\mu \in \mathbf{R}$ und $\sigma^2 < \sigma_0^2$. Dann maximiert ϕ' $m_\psi(\mu, \sigma^2)$ unter allen Tests ψ mit $m_\psi(\mu, \sigma_0^2) = \alpha$.

Sei dazu ψ ein Test mit $m_\psi(\mu, \sigma_0^2) = \alpha$. Dann ist

$$m_{1-\psi}(\mu, \sigma_0^2) = 1 - \alpha = 1 - m_{\phi'}(\mu, \sigma_0^2) = m_\phi(\mu, \sigma_0^2).$$

Kann man nun Theorem 6.3.4 anwenden, so folgt wegen $\sigma^2 < \sigma_0^2$ aus der letxten Aussage dieses Theorems, daß $m_\phi(\mu, \sigma^2) \leq m_{1-\psi}(\mu, \sigma^2)$ und damit $m_{\phi'}(\mu, \sigma^2) \geq m_\psi(\mu, \sigma^2)$. Um Theorem 6.3.4 anwenden zu können, ist noch zu zeigen, daß $\mathcal{N}(\mu, \sigma^2)$ bei festem μ einen isotonen Dichtequotienten in σ^2 hat. Betrachte dazu $\sigma_1^2 < \sigma_0^2$. Dann ist

$$\frac{f_{\mu,\sigma_0^2}}{f_{\mu,\sigma_1^2}}(x) = \left(\frac{\sigma_0^2}{\sigma_1^2}\right)^{-n/2} \exp\left(\underbrace{\left(\frac{1}{2\sigma_1^2} - \frac{1}{2\sigma_0^2}\right)}_{>0} \sum_{i=1}^n (x_i - \mu)^2 \right) \tag{6.6}$$

streng monoton wachsend in $\sum_{i=1}^n (x_i - \mu)^2$.

Nun wird gezeigt, daß kein schärfster Test von Θ_0' gegen Θ_1' existiert: Betrachte (μ_1, σ_1^2) mit $\sigma_1^2 < \sigma_0^2$. Die zugehörige Verteilung kann man in Θ_0' am besten durch $\mathcal{N}(\mu_1, \sigma_0^2)$ approximieren, und wir wählen als Ansatz für eine ungünstigste a priori Verteilung λ =Punktmasse auf (μ_1, σ_0^2). Dann ergibt sich

$$\frac{f_{\mu_1,\sigma_1^2}}{f_\lambda}(x) = \frac{f_{\mu,\sigma_1^2}}{f_{\mu,\sigma_0^2}}(x),$$

und aus (6.6) folgt, daß

$$\phi_{\lambda,\alpha}(x) = \begin{cases} 1, & \text{falls} \quad \sum_{i=1}^n (x_i - \mu_1)^2 \leq c_0 \\ 0, & \text{falls} \quad \sum_{i=1}^n (x_i - \mu_1)^2 > c_0 \end{cases},$$

wo sich c_0 wieder aus dem Niveau α bestimmt. $\phi_{\lambda,\alpha}$ ist als schärfster α-Test von P_λ gegen (μ_1, σ_1^2) Lebesgue-fast sicher eindeutig bestimmt. Das folgt aus dem NP Lemma, denn ein schärfster α-Test hat fast sicher NP Struktur, und da die Teststatistik stetig verteilt ist, nimmt er nur die Werte 1 und 0 an. Dann ist aber c_0 durch α eindeutig bestimmt.

Wir zeigen nun, daß $\phi_{\lambda,\alpha}$ ein α-Test von Θ_0' gegen (μ_1, σ_1^2) ist. Für $\sigma^2 \geq \sigma_0^2$ und $\mu \in \mathbf{R}$ gilt

$$\begin{aligned} & m_{\phi_{\lambda,\alpha}}(\mu, \sigma^2) = P_{\mu,\sigma^2}\{\sum_{i=1}^n (X_i - \mu_1)^2 \leq c_0\} \leq P_{\mu_1,\sigma^2}\{\sum_{i=1}^n (X_i - \mu_1)^2 \leq c_0\} \\ = \; & P_{0,\sigma^2}\{\sum_{i=1}^n X_i^2 \leq c_0\} \leq P_{0,\sigma_0^2}\{\sum_{i=1}^n X_i^2 \leq c_0\} \\ = \; & P_{\mu_1,\sigma_0^2}\{\sum_{i=1}^n (X_i - \mu_1)^2 \leq c_0\} = \alpha. \end{aligned}$$

Also ist $\phi_{\lambda,\alpha}$ nach Theorem 6.4.1 schärfster α-Test von Θ_0' gegen $\{(\mu_1, \sigma_1^2)\}$ und als solcher fast sicher eindeutig bestimmt. Wegen der Abhängigkeit des Tests von μ_1 kann kein schärfster α-Test von Θ_0' gegen Θ_1' existieren, denn ein solcher wäre auch am schärfsten gegen $\{(\mu_1, \sigma_1^2)\}$ für alle μ_1.

3. Für $\Theta_0'' = \{\mu \leq 0\}$ und $\Theta_1'' = \{\mu > 0\}$ führt die ML Methode auf den t-Test (Kapitel 2). Im nächsten Abschnitt werden wir zeigen, daß der t-Test zwar nicht am schärfsten, aber am schärfsten unter allen unverfälschten (d.h. erwartungstreuen) Tests ist. !

6.5 Optimalität des t-Tests als unverfälschter Test

Betrachte den statistischen Raum $(\mathbf{R}, \mathcal{B}, \mathcal{P})^n$, $n \geq 2$, wo

$$\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbf{R}, \sigma^2 > 0\}.$$

Ziel dieses Abschnitts ist es, folgenden Satz zu beweisen:

Theorem 6.5.1 *Der t-Test ist der schärfste unverfälschte Test für das Testproblem $H_0 : \mu \leq 0$ gegen $H_1 : \mu > 0$.*

Die dabei benutzte Beweistechnik kann auch bei Tests einseitiger Hypothesen über einen eindimensionalen Parameter in allgemeineren Exponentialräumen angewandt werden, siehe [12, Kapitel 4].

Bezeichne im folgenden $\vartheta = (\mu, \sigma^2)$, $\Theta = \mathbf{R} \times \mathbf{R}_+$, $P_\vartheta = \mathcal{N}(\mu, \sigma^2)$, $\Theta_0 = \{(\mu, \sigma^2) \in \Theta : \mu \leq 0\}$, $\Theta_1 = \{(\mu, \sigma^2) \in \Theta : \mu > 0\}$ und $\Gamma = \{(\mu, \sigma^2) \in \Theta : \mu = 0\}$

Für späteren Gebrauch notieren wir noch einmal die Dichte von $\mathcal{N}(\mu, \sigma^2)^n$ bezgl. des n-dimensionalen Lebesguemaßes:

$$\begin{aligned}
& f_{\mu,\sigma^2}(x_1, \ldots, x_n) \\
= \; & (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right) \\
= \; & (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{n\mu^2}{2\sigma^2}\right) \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2\right) \exp\left(\frac{\mu}{\sigma^2}\sum_{i=1}^n x_i\right) \\
= \; & (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \bar{x})^2\right) \exp\left(-\frac{1}{2\sigma^2} n(\bar{x} - \mu)^2\right).
\end{aligned}$$

Lemma 6.5.2 *Sei ϕ ein unverfälschter Test von Θ_0 gegen Θ_1 mit Umfang α. Dann ist $m_\phi(\vartheta) = \alpha$ für alle $\vartheta \in \Gamma$. Man sagt: ϕ ist α-**ähnlich** (oder auch nur ähnlich) auf Γ, dem gemeinsamen Rand von Θ_0 und Θ_1.*

Beweis: Sei $(0, \sigma^2) \in \Gamma \subseteq \Theta_0$. Aus der Unverfälschtheit und der Stetigkeit von $\mu \mapsto m_\phi(\mu, \sigma^2)$ folgt:

$$\alpha \geq m_\phi(0, \sigma^2) = \lim_{\mu \to 0, \mu > 0} m_\phi(\mu, \sigma^2) \geq \sup_{\vartheta \in \Theta_0} m_\phi(\vartheta) = \alpha.$$

□

Sei $V = \sum_{i=1}^n X_i^2$, und bezeichne mit

$$\hat{D}_\alpha := \{\phi = g(\bar{X}, V) : \phi \text{ auf } \Gamma \text{ } \alpha\text{-ähnlich}\}.$$

Lemma 6.5.3 *Sei ϕ ein t-Test vom Umfang α.*

1. $\phi \in \hat{D}_\alpha$.
2. $m_\phi(\mu, \sigma^2) \leq m_\phi(0, \sigma^2) = \alpha$ *für alle* $(\mu, \sigma^2) \in \Theta_0$.
3. *Ist ϕ schärfster Test aus $\hat{D}_\alpha$ von Γ gegen Θ_1, so ist ϕ schärfster, unverfälschter α-Test von Θ_0 gegen Θ_1.*

Beweis:

1. Als t-Test hat ϕ die Form

$$\phi(X_1, \ldots, X_n) = I_{\{\bar{X}/\sqrt{S^2} \geq c\}} = I_{\{\frac{\bar{X}}{\sigma}/\sqrt{\frac{S^2}{\sigma^2}} \geq c\}}.$$

   Da $\frac{\bar{X}}{\sigma}$ und $\sqrt{\frac{S^2}{\sigma^2}}$ unabhängig sind und ihre Verteilungen nicht von σ^2 abhängen, hängt auch die Verteilung von $\phi(X_1, \ldots, X_n)$ nicht von σ^2 ab, insbesondere ist ϕ auf Γ ähnlich.

2. Sei $(\mu, \sigma^2) \in \Theta_0$, also $\mu \leq 0$. Da $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$, hat die Verteilung von $\bar{X}$ bei festem σ^2 isotonen Dichtequotienten in μ und $\bar{x}$. (Beispiel 6.3.2). Also folgt aus Lemma 6.3.5, daß

$$\begin{aligned}
& m_\phi(\mu, \sigma^2) \\
= \; & E_{\mu,\sigma^2}[E_{\mu,\sigma^2}[I_{\{\bar{X} \geq c\sqrt{S^2}\}} | S^2]] \\
\leq \; & E_{\mu,\sigma^2}[E_{0,\sigma^2}[I_{\{\bar{X} \geq c\sqrt{S^2}\}} | S^2]] \\
= \; & E_{0,\sigma^2}[E_{0,\sigma^2}[\phi | S^2]], \quad \text{da die Verteilung von } S^2 \text{ nicht von } \mu \text{ abhängt,} \\
= \; & m_\phi(0, \sigma^2)
\end{aligned}$$

   Also ist

$$\alpha = \sup_{\mu \leq 0, \sigma^2 > 0} m_\phi(\mu, \sigma^2) \leq \sup_{\sigma^2 > 0} m_\phi(0, \sigma^2) \leq \alpha,$$

   und da ϕ nach Teil 1. des Beweises auf Γ ähnlich ist, ist $m_\phi(0, \sigma^2) = \alpha$ für alle $\sigma^2 > 0$, und ϕ ist auf Γ αähnlich. Wegen der speziellen Gestalt des t-Tests ist $\phi \in \hat{D}_\alpha$, was auch Aussage 1 beweist.

3. Sei nun ϕ' ein unverfälschter α-Test von Θ_0 gegen Θ_1. Wir nehmen zunächst an, daß ϕ' Umfang α hat. Dann ist ϕ' auf Γ α-ähnlich (Lemma 6.5.2). Da $(\bar{X}, V)$ eine suffiziente Statistik für unser Testproblem ist, ist $\tilde{\phi} := E[\phi'|\bar{X}, V]$ wohldefiniert (unabhängig von ϑ), und es ist

$$m_{\tilde{\phi'}}(\vartheta) = E_\vartheta[E[\phi'|\bar{X}, V]] = E_\vartheta[\phi'] = m_{\phi'}(\vartheta).$$

Insbesondere ist auch $\tilde{\phi}$ auf Γ α-ähnlich und somit $\tilde{\phi} \in \hat{D}_\alpha$. Damit folgt aus der Prämisse von Aussage 3 des Lemmas, daß

$$m_\phi(\vartheta) \geq m_{\tilde{\phi}}(\vartheta) = m_{\phi'}(\vartheta) \quad \text{für alle } \vartheta \in \Theta_1.$$

Hat ϕ' Umfang $\alpha' < \alpha$, so gibt es einen Test $\phi'' \geq \phi'$ mit Umfang α, und für $\vartheta \in \Theta_1$ ist $m_\phi(\vartheta) \geq m_{\phi''}(\vartheta) \geq m_{\phi'}(\vartheta)$. Also ist ϕ schärfster unverfälschter α-Test von Θ_0 gegen Θ_1. □

Beweis des Theorems:
Wegen Lemma 6.5.3 reicht es zu zeigen:

> Der t-Test ϕ vom Umfang α ist am schärfsten unter allen Tests $\phi' \in \hat{D}_\alpha$ von Γ gegen Θ_1. (6.7)

Für $v \in \mathbf{R}$ sei Q^v_ϑ eine Version der bedingten Verteilung von $\bar{X}$ unter P_ϑ gegeben $V = v$, $Q^v_\vartheta(A) = P_\vartheta(\bar{X} \in A|V = v)$.

In unserem Fall läßt sich Q^v_ϑ leicht explizit bestimmen: Die gemeinsame Dichte $h(u, w)$ von $\bar{X}$ und $(n-1)S^2$ ist wegen der Unabhängigkeit von $\bar{X}$ und S^2 das Produkt einer normalen Dichte mit einer „gestreckten“ χ^2-Dichte (siehe [12, S.82]).

$$h_{\mu,\sigma^2}(u, w) = \frac{1}{\sqrt{2\pi\frac{\sigma^2}{n}}} \exp\left(-n\frac{(u-\mu)^2}{2\sigma^2}\right) \frac{1}{(2\sigma^2)^{(n-1)/2}\Gamma(\frac{n-1}{2})} w^{\frac{n-3}{2}} e^{-\frac{w}{2\sigma^2}} I_{\{w>0\}}.$$

Daher haben $\bar{X}$ und $V = (n-1)S^2 + n\bar{X}^2$ die gemeinsame Dichte

$$\begin{aligned} h'(u, v) &= h(u, v - nu^2) \\ &= C(n, \mu, \sigma^2) e^{-\frac{nu^2}{2\sigma^2}} e^{\frac{n\mu}{\sigma^2}u} (v - nu^2)^{\frac{n-3}{2}} e^{-\frac{v}{2\sigma^2}} e^{\frac{nu^2}{2\sigma^2}} I_{\{v>nu^2\}} \\ &= C(n, \mu, \sigma^2) e^{-\frac{v}{2\sigma^2}} v^{\frac{n-3}{2}} e^{\frac{n\mu}{\sigma^2}u} (1 - \frac{nu^2}{v})^{\frac{n-3}{2}} I_{\{v>nu^2\}}. \end{aligned}$$

Q^v_ϑ hat daher zum Lebesguemaß auf $\mathbf{R}$ die Dichte

$$q^v_\vartheta(u) = C_v(n, \mu, \sigma^2) e^{\frac{n\mu}{\sigma^2}u} (1 - \frac{nu^2}{v})^{\frac{n-3}{2}} I_{\{v>nu^2\}}. \tag{6.8}$$

Insbesondere läßt sich für festes v die Menge $\{Q^v_\vartheta : \vartheta \in \Theta\}$ als einparametrige Exponentialfamilie mit natürlichem Parameter $\frac{n\mu}{\sigma^2}$ auffassen und damit als Familie mit

isotonem Dichtequotienten. Ein schärfster Test von $\{\frac{n\mu}{\sigma^2} \leq 0\}$ gegen $\{\frac{n\mu}{\sigma^2} > 0\}$ und damit von $\{\mu \leq 0\}$ gegen $\{\mu > 0\}$ hat daher die Form (siehe Theorem 6.3.4)

$$\phi_v(x_1, \ldots, x_n) := g_v(\bar{x}) =: \begin{cases} 1, \text{ falls } \bar{x} > \tau(v)\sqrt{\frac{v}{n}} \\ 0, \text{ falls } \bar{x} < \tau(v)\sqrt{\frac{v}{n}} \end{cases} .$$

Dabei wird $\tau(v)$ durch

$$\begin{aligned} \alpha &= \int g_v(u)\, dQ^v_{0,\sigma^2}(u) = \int g_v(u)\, q^v_{0,\sigma^2}(u) du \\ &= \int_{\max(-\tau(v),\tau(v)\sqrt{\frac{v}{n}})}^{\sqrt{\frac{v}{n}}} \left(1 - \left(\frac{u}{\sqrt{v/n}}\right)^2\right)^{\frac{n-3}{2}} du \Big/ \int_{-\tau(v)}^{\sqrt{\frac{v}{n}}} \left(1 - \left(\frac{u}{\sqrt{v/n}}\right)^2\right)^{\frac{n-3}{2}} du \\ &= \int_{\max(-1,\tau(v))}^{1} (1-y^2)^{\frac{n-3}{2}} dy \Big/ \int_{-1}^{1} (1-y^2)^{\frac{n-3}{2}} dy \end{aligned}$$

bestimmt. $\tau(v) = \tau$ ist also unabhängig von v.

Wir zeigen nun, daß der durch

$$\phi(X_1, \ldots, X_n) := g(\bar{X}, V) := g_V(\bar{X})$$

bestimmte Test mit dem t-Test übereinstimmt: Für $z > 0$ sei $r(z) = \frac{z}{\sqrt{n-1+z^2}}$. $r(z)$ ist streng isoton in z, und es gilt

$$\begin{aligned} & \phi(X_1, \ldots, X_n) = 1 \\ \Leftrightarrow\ & \bar{X} > \tau\sqrt{\frac{V}{n}} \\ \Leftrightarrow\ & \sqrt{\frac{n}{S^2}}\bar{X} > \tau\sqrt{\frac{(n-1)S^2 + n\bar{X}^2}{S^2}} = \tau\sqrt{(n-1) + \left(\sqrt{\frac{n}{S^2}}\bar{X}\right)^2} \\ \Leftrightarrow\ & r\left(\sqrt{\frac{n}{S^2}}\bar{X}\right) > \tau \\ \Leftrightarrow\ & \sqrt{\frac{n}{S^2}}\bar{X} > \tau' := r^{-1}(\tau). \end{aligned}$$

Also ist ϕ ein t-Test.

Um (6.7) zu beweisen, bleibt nur noch das folgende Lemma zu zeigen:

Lemma 6.5.4 *$\phi(X_1, \ldots, X_n)$ ist schärfster Test von Γ gegen Θ_1 in $\hat{D}_\alpha$.*

Zum Beweis benötigen wir ein weiteres Lemma:

Lemma 6.5.5 *Sei $\phi' = g'(\bar{X}, V)$ ein Test, $m_{\phi'}(\vartheta) = \alpha$ für alle $\vartheta \in \Gamma$. Für $v \in \mathbf{R}$ sei $g'_v(u) := g'(u, v)$. Dann gilt für alle $\vartheta \in \Gamma$:*

$$\int g'_v(u)\, dQ^v_\vartheta(u) = \alpha \quad \text{für Lebesgue-f.a. } v > 0.$$

Beweis: Wegen (6.7) ist $\{\mathcal{N}(\mu, \sigma^2) : (\mu, \sigma^2) \in \Gamma\}$ eine einparametrige Exponentialfamilie mit natürlichem Parameter $\frac{-1}{2\sigma^2}$, für die V eine vollständige Statistik ist. Für $\vartheta \in \Gamma$ (d.h. $\mu = 0$) hängt q_ϑ^v nicht von σ^2 und damit nicht von ϑ ab, siehe (6.8). Also folgt die Behauptung wegen der Vollständigkeit von V bzgl. $\{P_\vartheta : \vartheta \in \Gamma\}$ aus

$$E_\vartheta \left[\int g'_V(u)\, dQ_\vartheta^V(u)) \right] = E_\vartheta[E_\vartheta[g'(\bar{X}, V)|V]] = E_\vartheta[g'(\bar{X}, V)] = m_{\phi'}(\vartheta) = \alpha. \qquad (6.9)$$

□

Beweis von Lemma 6.5.4: $\phi \in \hat{D}(\alpha)$, da $\phi(X_1, \ldots, X_n) = g(\bar{X}, V)$ und da für $\vartheta \in \Gamma$

$$E_\vartheta[\phi] = E_\vartheta[E_\vartheta[g(\bar{X}, V)|V]] = E_\vartheta \left[\int g(u, V)\, dQ_\vartheta^V(u) \right] = E_\vartheta[\alpha] = \alpha. \qquad (6.10)$$

Ist $\phi'(X_1, \ldots, X_n) = g'(\bar{X}, V)$ ein weiterer Test von Γ gegen H_1 mit $E_\vartheta[\phi'] = \alpha\, \forall \alpha \in \Gamma$, dann ist wegen Lemma 6.5.5 $\bar{X} \mapsto g'_v(\bar{X})$ für Lebesgue-f.a. $v > 0$ ein auf Γ α-ähnlicher Test von Γ gegen Θ_1 für die Menge $\{Q_\vartheta^v : \vartheta \in \Theta\}$ von Verteilungen auf $\mathbf{R}$. Da ϕ nach Konstruktion schärfster Test dieser Art ist, folgt

$$\int g(u, v)\, dQ_\vartheta^v(u) \geq \int g'(u, v)\, dQ_\vartheta^v(u)$$

für alle $\vartheta \in \Theta_1$, und durch Erwartungswertbildung auf beiden Seiten erhält man (vergl. (6.9) und (6.10)) $E_\vartheta[\phi] \geq E_\vartheta[\phi']$ für alle $\vartheta \in \Theta_1$. □

Literaturverzeichnis

[1] J.-R.Barra, *Mathematical Basis of Statistics*, Academic Press, 1981

[2] P.Billingsley, *Probability and Measure*, Second Edition, Wiley, 1985

[3] L.Breiman, Statistics: With a View towards Applications, Houghton Mifflin Company, Boston, 1973

[4] B.Efron, *Controversies in the foundations of statistics*, Amer.Math.Monthly **85**, pp.231-246, 1978

[5] B.Efron, *Maximum likelihood and decision theory. (The 1981 Wald Memorial Lectures)*, Annals of Statistics **10**, pp.340-356, 1982

[6] U.Krengel, *Einführung in die Wahrscheinlichkeitstheorie und Statistik*, Vieweg Verlag, Reihe "Aufbaustudium Mathematik", 1988

[7] U.Krengel, *Mathematische Statistik*, Vorlesungsskript, Göttingen, 1973/74

[8] A.Rényi, *Probability Theory*, North-Holland Series in Applied Mathematics and Mechanics, 1970

[9] S.Siegel, *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill, 1956

[10] A.N.Širjaev, *Wahrscheinlichkeit*, VEB Deutscher Verlag der Wissenschaften, 1987

[11] W.Winkler *Vorlesungen zur Mathematischen Statistik*, Teubner Studienbücher Mathematik, 1983

[12] H.Witting, *Mathematische Statistik*, x-te Auflage, Teubner Studienbücher Mathematik, 1974 (für x=2)