2.2 Klassische Testtheorie

Ziel: Finde Test zum Niveau α mit optimaler Güte (Power) für $\theta \in \Theta_1$. Dabei ist n finit.

2.2.1 Problemstellung

- Sei Θ der Parameterraum; die Hypothesen seien

$$H_0 : \ \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta_1,$$

  mit $\Theta_0 \cap \Theta_1 = \emptyset$, d.h. Θ_0 und Θ_1 sind disjunkt. Möglicherweise, jedoch nicht notwendigerweise, gilt $\Theta_0 \cup \Theta_1 = \Theta$.

- Eine Nullhypothese heißt *einfach*, wenn sie aus einem einzelnen Element aus Θ besteht, d.h. $\Theta_0 = \{\theta_0\}$. Ansonsten spricht man von *zusammengesetzten* Hypothesen. Dabei ist Folgendes zu beachten: Etliche Nullhypothesen sind scheinbar einfach, aber tatsächlich zusammengesetzt. Dies ist häufig dann der Fall, wenn Nuisanceparameter auftauchen.

  Beispiel: Seien $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ mit μ und σ^2 unbekannt. Die Nullhypothese $H_0 : \mu = 0$ ist eine zusammengesetzte Hypothese, da

$$\Theta = \{(\mu, \sigma^2) : -\infty < \mu \leq \infty, 0 < \sigma^2 < \infty\}$$

  und

$$\Theta_0 = \{(\mu, \sigma^2) : \mu = 0, 0 < \sigma^2 < \infty\}.$$

- Ergebnisse/Aktionen:

$$\begin{aligned} A_0 : &\quad H_0 \text{ wird nicht abgelehnt} \\ A_1 : &\quad H_0 \text{ wird abgelehnt} \end{aligned}$$

- Test zum Niveau α:

$$\mathbb{P}_\theta(A_1) \leq \alpha, \quad \text{für alle } \theta \in \Theta_0$$

- **Testfunktionen** (vgl. Abschnitt 1.2.1): Tests werden oft folgendermaßen formuliert: Wähle eine Teststatistik $T(X)$, eine Stichprobe X und einen kritischen Bereich C_α. Dann lautet der Test

$$\phi(x) = \begin{cases} 1 & \text{, falls } T(x) \in C_\alpha \qquad (H_0 \text{ ablehnen}), \\ 0 & \text{, falls } T(x) \notin C_\alpha \qquad (H_0 \text{ nicht ablehnen}). \end{cases}$$

- Für die Testtheorie dieses Abschnitts werden solche Testfunktionen $\phi(x) \in \{0, 1\}$ erweitert zu *randomisierten Testfunktionen* $\phi(x) \in [0, 1]$:

  1. Für gegebene Daten $X = x$ ist $\phi(x) \in [0, 1]$.
  2. Ziehe eine (davon unabhängige) Bernoullivariable $W \sim \text{Bin}(1, \phi(x))$.
  3. Lehne H_0 genau dann ab, wenn $W = 1$.

Interpretation: $\phi(x)$ ist die Wahrscheinlichkeit für die Ablehnung von H_0 gegeben die Beobachtung $X = x$. Im Spezialfall $\phi(x) \in \{0, 1\}$ reduziert sich ein randomisierter Test auf einen üblichen, nicht randomisierten Test. Randomisierte Tests sind (für die Theorie) vor allem bei diskreten Teststatistiken relevant.

Beispiel 2.21 (Randomisierter Binomialtest)**.** *Sei* $X \sim Bin(10, \pi)$ *und*

$$H_0 : \pi \leq \frac{1}{2}, \quad H_1 : \pi > \frac{1}{2}.$$

Test: H_0 *ablehnen* $\Leftrightarrow$ $X \geq k_\alpha$*, wobei* k_α *so, dass*

$$\mathbb{P}_\pi(X \geq k_\alpha) \leq \alpha \quad \textit{für } \pi = \frac{1}{2}.$$

Es ist

$$\mathbb{P}_{0.5}(X \geq k) = \begin{cases} 0.00098 & , k = 10 \\ 0.01074 & , k = 9 \\ 0.05469 & , k = 8 \\ \dots \end{cases}$$

Für $\alpha = 0.05$ *würde die Wahl* $k_\alpha = 8$ *wegen* $0.054 > 0.05$ *nicht möglich sein. Wählt man aber* $k_\alpha = 9$*, so schöpft man* $\alpha = 0.05$ *bei weitem nicht aus, d.h. der Test ist sehr konservativ. Die Lösung ist ein randomisierter Test*

$$\phi(x) = \begin{cases} 1 & , x \in \{9, 10\} \\ 67/75 & , x = 8 \\ 0 & , x \leq 7, \end{cases}$$

d.h. ziehe bei $x = 8$ *eine bernoulliverteilte Zufallsvariable mit Wahrscheinlichkeit* $67/75$*. Wird* 1 *realisiert, so wird* H_0 *abgelehnt.*

Die Randomisierung ist ein künstlicher Vorgang, um das Signifikanzniveau α auszuschöpfen, d.h.

$$\mathbb{P}_\theta(A_1) = \alpha$$

für dasjenige θ auf dem Rand zwischen Θ_0 und Θ_1 zu erreichen. Ein randomisierter Test besitzt in der Regel folgende Struktur:

$$\phi(x) = \begin{cases} 1 & , x \in B_1 \\ \gamma(x) & , x \in B_{10} \\ 0 & , x \in B_0. \end{cases}$$

Der Stichprobenraum wird also in **drei** Teile zerlegt:

B_1 strikter *Ablehnungsbereich* von H_0, d.h. $x \in B_1 \Rightarrow$ Aktion A_1.

B_0 strikter *Annahmebereich*, d.h. $x \in B_0 \Rightarrow$ Aktion A_0.

B_{10} *Randomisierungsbereich*, d.h. $x \in B_{10}$ führt mit Wahrscheinlichkeit $\gamma(x)$ zur Ablehnung und mit Wahrscheinlichkeit $1 - \gamma(x)$ zur Annahme von H_0. B_{10} kann als *Indifferenzbereich* interpretiert werden.

In der Regel wird ein Test mit einer Teststatistik $T = T(X)$ formuliert. Dann haben randomisierte Tests oft die Form:

$$\phi(x) = \begin{cases} 1, \ T(x) > c \\ \gamma, \ T(x) = c \\ 0, \ T(x) < c \, . \end{cases}$$

Falls $T(X)$ eine stetige Zufallsvariable ist, gilt $\mathbb{P}(T(X) = c) = 0$, d.h. für stetige T reduziert sich $\phi(x)$ zu

$$\phi(x) = \begin{cases} 1, \ T(x) \geq c \\ 0, \ T(x) < c \, . \end{cases}$$

Bei diskreten Teststatistiken T wie beim exakten Binomialtest ist gewöhnlich $\gamma > 0$, da $\mathbb{P}(T(X) = c) > 0$. Der Wert c ist an der „Entscheidungsgrenze" zwischen A_1 und A_0. Dass man die Entscheidung durch eine zufällige Prozedur herbeiführt, stößt in der Praxis auf Bedenken.

Die (frequentistische) Theorie zeigt, dass die Priori-Wahrscheinlichkeit

$$\mathbb{P}_\theta(A_1) = \int_{\mathcal{X}} \underbrace{\mathbb{P}(A_1|x)}_{\phi(x)} \ \underbrace{f(x|\theta)dx}_{d\mathbb{P}_\theta} = \mathbb{E}_\theta[\phi(X)], \quad \theta \in \Theta_1$$

bei Randomisierung maximiert werden kann ($\phi(x)$ ist die bedingte Wahrscheinlichkeit, a posteriori, d.h. bei gegebener Stichprobe, für A_1 zu entscheiden). „Maximal" bezieht sich auf „durchschnittliche" Optimalität des Tests bei wiederholter Durchführung.

Subjektive Sichtweise: Man wird bei $T(x) = c$ bzw. $x \in B_{10}$ eher noch keine Entscheidung treffen („Indifferenzbereich").

Für $n \to \infty$ geht (in der Regel) $\mathbb{P}(T(X) = c)$ gegen 0, d.h. für großes n wird der Randomisierungsbereich B_{10} immer kleiner. Idee: Bei $T(x) = c$ zusätzliche Daten erheben.

Güte, Gütefunktion (power, power function)

Bei einer Testentscheidung gibt es folgende Möglichkeiten:

	A_0: H_0 beibehalten	A_1: H_1 ist signifikant
H_0 trifft zu	richtige Aussage	Fehler 1. Art
H_1 trifft zu	Fehler 2. Art	richtige Aussage

Es ist $\phi(x) = \mathbb{P}(A_1|x)$ die bedingte Wahrscheinlichkeit für A_1 gegeben die Stichprobe x. Ist $\mathbb{P}_\theta(A_1)$ die unbedingte Wahrscheinlichkeit / Priori-Wahrscheinlichkeit, dann gilt (wie oben)

$$\mathbb{P}_\theta(A_1) = \int_{\mathcal{X}} \mathbb{P}(A_1|x) f(x|\theta) \, dx = \int \phi(x) f(x|\theta) \, dx = \mathbb{E}_\theta[\phi(X)]$$

und somit auch $\mathbb{P}_\theta(A_0) = \mathbb{E}_\theta(1 - \phi(X))$ für $\theta \in \Theta$.

Definition 2.31 (Gütefunktion eines Tests ϕ)**.**

1. *Die Abbildung* $g_\phi(\theta) = \mathbb{E}_\theta[\phi(X)] = \mathbb{P}_\theta(A_1)$, $\theta \in \Theta$, *heißt* Gütefunktion *des Tests* ϕ.

$$\begin{aligned} g_\phi(\theta) &= \mathbb{P}_\theta(A_1) && \textit{Wahrscheinlichkeit für Fehler 1. Art, } \theta \in \Theta_0 \\ 1 - g_\phi(\theta) &= \mathbb{P}_\theta(A_0) && \textit{Wahrscheinlichkeit für Fehler 2. Art, } \theta \in \Theta_1 \end{aligned}$$

   Außerdem:

$$g_\phi(\theta) = \mathbb{P}_\theta(A_1) \qquad \textit{Macht (power) des Tests, } \theta \in \Theta_1$$

2. *Die Größe*

$$\alpha(\phi) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(A_1) = \sup_{\theta \in \Theta_0} g_\phi(\theta)$$

   heißt (tatsächliches) Niveau (level, size) *von* ϕ *und ist die* supremale Wahrscheinlichkeit für den Fehler 1. Art.

$$\beta(\phi) = \sup_{\theta \in \Theta_1} \mathbb{P}_\theta(A_0) = 1 - \inf_{\theta \in \Theta_1} g_\phi(\theta)$$

   ist die supremale Wahrscheinlichkeit für den Fehler 2. Art.

- Bei den „üblichen" Tests (zum Beispiel beim einseitigen Gauß-Test) gilt wegen der Monotonie und Stetigkeit von $g_\phi(\theta)$

$$\alpha(\phi) + \beta(\phi) = 1,$$

  d.h. $\alpha(\phi)$ kann nur auf Kosten von $\beta(\phi)$ klein gehalten werden (und umgekehrt).
  Allgemein gilt dagegen nur

$$\alpha(\phi) + \beta(\phi) \leq 1$$

  (bei unverfälschten Tests, siehe Definition 2.37).

- *Programm der klassischen Testtheorie*: Maximiere unter Beschränkung

$$g_\phi(\theta) \leq \alpha \text{ für alle } \theta \in \Theta_0$$

  bei fest vorgegebenem $\alpha > 0$ die Güte für $\theta \in \Theta_1$, d.h.

$$g_\phi(\theta) \geq \max_{\widetilde{\phi}} g_{\widetilde{\phi}}(\theta) \quad \text{für } \theta \in \Theta_1$$

  bei „konkurrierenden" Tests $\widetilde{\phi}$. H_0 und H_1 werden also unsymmetrisch betrachtet.

- Wegen der Beziehung $\alpha(\phi) + \beta(\phi) = 1$ muss dabei das vorgegebene Signifikanzniveau α ausgeschöpft werden, d.h.

$$\alpha(\phi) = \alpha$$

  gelten. Bei $\alpha(\phi) < \alpha$ wird automatisch

$$\beta(\phi) = 1 - \inf_{\theta \in \Theta_1} g_\theta(\phi)$$

  für $\theta \in \Theta_1$ größer als notwendig, d.h. die Güte des Tests schlechter.

- Folgende Problemstellungen werden nach diesem Konzept betrachtet:

  1. *Einfaches H_0 vs. einfaches H_1*: Neyman-Pearson-Theorem zeigt, wie bester Test zu konstruieren ist.
  2. *Einfaches H_0 vs. zusammengesetztes H_1*: Basierend auf dem Neyman-Pearson-Theorem kann für bestimmte Fälle ein „gleichmäßig bester Test" (UMP, uniformly most powerful test) konstruiert werden. In anderen Fällen existiert — zumindest ohne weitere Restriktionen — kein UMP-Test.
  3. *Zusammengesetztes H_0 vs. zusammengesetztes H_1*: Suche nach einem UMP-Test ist noch schwieriger.

2.2.2 Satz von Neyman-Pearson

Problemstellung: Einfache Nullhypothese vs. einfache Alternativhypothese, also

$$H_0 : \theta = \theta_0, \quad \text{vs.} \quad H_1 : \theta = \theta_1$$

mit $\theta_0 \neq \theta_1$. Sei $f_0(x) = f(x|\theta_0), f_1(x) = f(x|\theta_1)$. Dann heißt

$$\Lambda(x) = \frac{f_1(x)}{f_0(x)}$$

Likelihood-Quotient. Ein (bester) Test hat nach Neyman-Pearson die Form:

$$H_0 \text{ ablehnen} \Leftrightarrow \Lambda(x) > k_\alpha$$

mit k_α so gewählt, dass der Test das Niveau α einhält. Aber: Falls $\Lambda(x)$ diskret ist, gibt es ein theoretisches Problem. Dies führt zu

Definition 2.32 (Randomisierter LQ-Test)**.** *Ein Test $\phi^*(x)$ heißt* randomisierter Likelihood-Quotienten-Test*, kurz* LQ-Test *(likelihood ratio test, LRT)* $\overset{def}{\Leftrightarrow}$ *$\phi^*(x)$ hat die Struktur*

$$\phi^*(x) = \begin{cases} 1 & , f_1(x) > k f_0(x) \ \Leftrightarrow \ \Lambda(x) > k \\ \gamma(x) & , f_1(x) = k f_0(x) \ \Leftrightarrow \ \Lambda(x) = k \\ 0 & , f_1(x) < k f_0(x) \ \Leftrightarrow \ \Lambda(x) < k \end{cases}$$

mit Konstante $k > 0$ und $0 < \gamma(x) < 1$. Falls $\Lambda(X)$ stetig ist, gilt $\mathbb{P}_\theta(\Lambda(X) = k) = 0$. Dann reicht ein nicht-randomisierter Test

$$\phi^*(x) = \begin{cases} 1, \ f_1(x) > k f_0(x) \ \Leftrightarrow \ \Lambda(x) > k \\ 0, \ \textit{sonst.} \end{cases}$$

Satz 2.33 (Neyman-Pearson, Fundamentallemma)**.**

1. *Optimalität: Für jedes k und $\gamma(x)$ hat der Test ϕ^* maximale Macht unter allen Tests, deren Niveau höchstens gleich dem Niveau von ϕ^* ist.*
2. *Existenz: Zu vorgegebenem $\alpha \in (0, 1)$ existieren Konstanten k^* und γ^*, so dass der LQ-Test ϕ^* mit diesem k^* und $\gamma(x) = \gamma^*$ für alle x exakt das Niveau α besitzt.*

3. *Eindeutigkeit: Falls ein Test ϕ mit Niveau α maximale Macht (= kleinsten Fehler 2. Art) unter allen anderen Tests mit Niveau α besitzt, dann ist ϕ ein LQ-Test (eventuell mit Ausnahme einer Nullmenge $\mathcal{X}_0 \subset \mathcal{X}$ von Stichproben x, d.h. $\mathbb{P}_{\theta_0}(\mathcal{X}_0) = \mathbb{P}_{\theta_1}(\mathcal{X}_0) = 0$).*

Beweis.

1. Sei ϕ ein Test mit
$$\mathbb{E}_{\theta_0}[\phi(X)] \leq \mathbb{E}_{\theta_0}[\phi^*(X)] \tag{2.2}$$
und
$$U(x) = (\phi^*(x) - \phi(x))(f_1(x) - kf_0(x)).$$
   - Für $f_1(x) - kf_0(x) > 0$ ist $\phi^*(x) = 1$, also $U(x) \geq 0$.
   - Für $f_1(x) - kf_0(x) < 0$ ist $\phi^*(x) = 0$, also $U(x) \geq 0$.
   - Für $f_1(x) - kf_0(x) = 0$ ist $U(x) = 0$.

   Also: $U(x) \geq 0$ für alle x. Somit:
$$\begin{aligned} 0 &\leq \int U(x)dx \\ &= \int (\phi^*(x) - \phi(x))(f_1(x) - kf_0(x))\, dx \\ &= \int \phi^*(x)f_1(x)\, dx - \int \phi(x)f_1(x)\, dx + k\left(\int \phi(x)f_0(x)\, dx - \int \phi^*(x)f_0(x)\, dx\right) \\ &= \mathbb{E}_{\theta_1}[\phi^*(X)] - \mathbb{E}_{\theta_1}[\phi(X)] + \underbrace{k(\mathbb{E}_{\theta_0}[\phi(X)] - \mathbb{E}_{\theta_0}[\phi^*(X)])}_{\leq 0 \text{ wegen } (1.2)} \end{aligned}$$
   $\Rightarrow \mathbb{E}_{\theta_1}[\phi^*(X)] \geq \mathbb{E}_{\theta_1}[\phi(X)]$, d.h. die Macht von ϕ^* ist größer als die Macht von ϕ.

2. Die Verteilungsfunktion $G(k) = \mathbb{P}_{\theta_0}(\Lambda(x) \leq k)$ ist monoton steigend in k. Sie ist ferner rechtsstetig, d.h.
$$G(k) = \lim_{y \downarrow k} G(y) \quad \text{für alle } k.$$
   Betrachtet man die Gleichung
$$G(k^*) = 1 - \alpha$$
   und versucht diese bezüglich k^* zu lösen, so gibt es zwei Möglichkeiten:

   (i) Entweder ein solches k^* existiert,

   (ii) oder die Gleichung kann nicht exakt gelöst werden, aber es existiert ein k^*, so dass
$$G_-(k^*) = \mathbb{P}_{\theta_0}(\Lambda(X) < k^*) \leq 1 - \alpha < G(k^*)$$
   (das entspricht der „Niveaubedingung").

   Im ersten Fall setzt man $\gamma^* = 0$, im zweiten
$$\gamma^* = \frac{G(k^*) - (1 - \alpha)}{G(k^*) - G_-(k^*)}.$$

In diesem Fall hat der Test genau das Niveau α, wie behauptet, denn:

$$\begin{aligned}\mathbb{E}_{\theta_0}[\phi(X)] &= \mathbb{P}_{\theta_0}\left(\frac{f_1(X)}{f_0(X)} > k^*\right) + \frac{G(k^*) - 1 + \alpha}{G(k^*) - G_-(k^*)}\mathbb{P}_{\theta_0}\left(\frac{f_1(X)}{f_0(X)} = k^*\right) \\ &= (1 - G(k^*)) + \frac{G(k^*) - 1 + \alpha}{G(k^*) - G_-(k^*)}(G(k^*) - G_-(k^*)) \\ &= \alpha.\end{aligned}$$

3. Zu gegebenem α sei ϕ^* der nach 2. existierende LQ-Test definiert durch eine Konstante k und eine Funktion $\gamma(x)$. Man nehme an, ϕ ist ein anderer Test mit gleichem Niveau α und der gleichen (nach 1. maximalen) Macht wie ϕ^*. Definiert man $U(x)$ wie in 1., dann ist $U(x) \geq 0$ für alle x und $\int U(x)\, dx = 0$, da $\mathbb{E}_{\theta_1}[\phi^*(X)] - \mathbb{E}_{\theta_1}[\phi(X)] = 0$ und $\mathbb{E}_{\theta_0}[\phi^*(X)] - \mathbb{E}_{\theta_0}[\phi(X)] = 0$ nach Annahme. Daraus, dass U nicht-negativ mit Integral 0 ist, folgt, dass $U(x) = 0$ für fast alle x. Dies wiederum bedeutet, dass $\phi(x) = \phi^*(x)$ oder $f_1(x) = kf_0(x)$, d.h. $\phi(x)$ ist ein LQ-Test (für fast alle x). □

Bemerkung. *Für einfache Hypothesen H_0 und H_1 sind klassische Testtheorie und Likelihood-Quotienten-Test noch identisch. Für zusammengesetzte Hypothesen (der Praxisfall) trennen sich die Konzepte:*

- *Klassische Testtheorie sucht weiter nach optimalen Tests (für finite Stichproben).*
- *Likelihoodbasierte Tests verallgemeinern $\Lambda(x)$ bzw. sind quadratische Approximationen von $\Lambda(x)$, deren Verteilungsfunktion (unter H_0) nur asymptotisch ($n \to \infty$) gilt.*

Beispiel 2.22 (Binomialtest)**.** *Betrachte*

$$H_0 : \pi = \pi_0 \quad vs. \quad H_1 : \pi = \pi_1$$

mit $0 < \pi_0 < \pi_1 < 1$. Die Dichte (Wahrscheinlichkeitsfunktion) der i.i.d. Stichprobe $X = (X_1, \ldots, X_n)^\top$ lautet

$$f(x|\pi) = \pi^z(1 - \pi)^{n-z} \quad mit \quad z = \sum_{i=1}^{n} x_i,$$

der Likelihood-Quotient

$$\Lambda(x) = \frac{\pi_1^z(1 - \pi_1)^{n-z}}{\pi_0^z(1 - \pi_0)^{n-z}} = \left(\frac{1 - \pi_1}{1 - \pi_0}\right)^n \cdot \left(\frac{\pi_1(1 - \pi_0)}{\pi_0(1 - \pi_1)}\right)^z := \Lambda(z).$$

Da $\Lambda(x) = \Lambda(z)$ streng monoton in z ist, lässt sich $\Lambda(z) > k$ äquivalent umformen in $z > \Lambda^{-1}(k) =: c$. Der Likelihood-Quotienten-Test ϕ^* mit kritischer Zahl k und (konstanter) Randomisierung γ^* hat dann die Form

$$\phi^*(x) = \begin{cases} 1 & , Z = Z(x) > c \\ \gamma^* & , Z = Z(x) = c \\ 0 & , Z = Z(x) < c \end{cases}$$

mit der „Teststatistik" Z. Dabei können wir uns (wegen des Wertebereichs von Z) auf $c \in \{0, 1, \ldots, n\}$ beschränken. γ^* ist aus der Niveaubedingung

$$\mathbb{P}_{\pi_0}(Z > c) + \gamma^* \mathbb{P}_{\pi_0}(Z = c) \stackrel{!}{=} \alpha$$

zu bestimmen. Der Test ϕ^* hängt von π_0 ab, jedoch nicht von π_1!

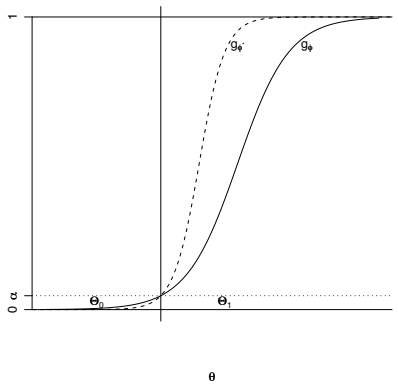
Bemerkung. *Falls H_1 wahr ist, dann bestimmt π_1 die Wahrscheinlichkeit für den „realisierten" Fehler 2. Art $\mathbb{P}_{\pi_1}(A_0)$. Je weiter π_1 von π_0 entfernt ist, umso kleiner ist die Wahrscheinlichkeit für den Fehler 2. Art und umso größer ist die Power an der Stelle $\pi = \pi_1$.*

2.2.3 Gleichmäßig beste Tests

Definition 2.34 (Gleichmäßig bester (UMP, uniformly most powerful) Test)**.** *Ein Niveau-α-Test ϕ^* heißt* gleichmäßig bester *oder* UMP Test *zum Niveau α* $\overset{def}{\Leftrightarrow}$

1. $\mathbb{E}_\theta[\phi^*(X)] \leq \alpha$ *für alle* $\theta \in \Theta_0$.
2. *Für jeden anderen Niveau-α-Test ϕ mit $\mathbb{E}_\theta[\phi(X)] \leq \alpha$ für alle $\theta \in \Theta_0$ gilt:*

$$\mathbb{E}_\theta[\phi^*(X)] \geq \mathbb{E}_\theta[\phi(X)] \text{ für alle } \theta \in \Theta_1.$$



Bemerkung. *Der Begriff „gleichmäßig" in obiger Definition bezieht sich auf die Gleichmäßigkeit der Eigenschaft $g_{\phi^*} \geq g_\phi$ auf Θ_1 für jeden anderen Test ϕ.*

Beste einseitige Tests bei skalarem θ

In Beispiel 1.22 (Binomialtest für einfache Hypothesen) hing der Test ϕ^* nicht vom speziellen $\pi_1(\equiv H_1) > \pi_0(\equiv H_0)$ ab. Daraus folgt, dass ϕ^* für alle $\pi_1 > \pi_0$ besser ist als ein anderer Test ϕ. Entscheidend dafür ist, dass der Dichte- bzw. Likelihood-Quotient monoton in z ist. Dies gilt allgemeiner und führt zu folgender Definition.

Definition 2.35 (Verteilungen mit monotonem Dichtequotienten)**.** *Die Verteilungsfamilie $\{f(x|\theta),\ \theta \in \Theta \subseteq \mathbb{R}\}$ mit skalarem Parameter θ besitzt* monotonen Dichte- *bzw.* Likelihood-Quotienten *(kurz: MLQ)* $\overset{def}{\Leftrightarrow}$ *es existiert eine Statistik T, so dass*

$$\Lambda(x) = \frac{f(x|\theta_1)}{f(x|\theta_0)}$$

monoton wachsend in $T(x)$ für je zwei $\theta_0, \theta_1 \in \Theta$ mit $\theta_0 \leq \theta_1$ ist.

Bemerkung.

1. *Monoton* wachsend *ist keine echte Einschränkung; ist $\Lambda(x)$ monoton fallend in $\widetilde{T}(x)$, so definiert man $T(x) = -\widetilde{T}(x)$.*

2. *Jede* einparametrische *Exponentialfamilie in $T(x)$ und $\gamma(\theta)$ besitzt monotonen Dichtequotienten, wenn $\gamma(\theta)$ monoton in θ ist. Letzteres gilt insbesondere für die* natürliche *Parametrisierung $\gamma(\theta) = \theta$.*

Satz 2.36 (UMP-Test bei MLQ)**.** *Gegeben sei $\mathcal{P}_\theta = \{f(x|\theta) : \theta \in \Theta \subseteq \mathbb{R}\}$ mit MLQ in $T(x)$ und die Hypothesen*

$$H_0 : \theta \leq \theta_0 \quad vs. \quad H_1 : \theta > \theta_0.$$

1. *Existenz: Es gibt einen UMP-Test ϕ^* zum Niveau α, nämlich*

$$\phi^*(x) = \begin{cases} 1, & T(x) > c \\ \gamma, & T(x) = c \\ 0, & T(x) < c. \end{cases}$$

   Dabei sind c und γ eindeutig bestimmt durch die Niveaubedingung

$$\mathbb{E}_{\theta_0}[\phi^*(X)] = \mathbb{P}_{\theta_0}(T(X) > c) + \gamma \mathbb{P}_{\theta_0}(T(X) = c) = \alpha.$$

2. *Die Gütefunktion $g_{\phi^*}(\theta)$ ist monoton wachsend in θ und sogar streng monoton wachsend für alle θ mit $0 < g_{\phi^*}(\theta) < 1$. Die maximale Wahrscheinlichkeit für den Fehler 1. Art ist $g_{\phi^*}(\theta_0) = \alpha$.*

3. *ϕ^* besitzt auch gleichmäßig minimale Wahrscheinlichkeiten für den Fehler 1. Art unter allen Tests ϕ für H_0 vs. H_1 mit $g_\phi(\theta_0) = \alpha$.*

4. *ϕ^* ist (mit Wahrscheinlichkeit 1) eindeutig bestimmt.*

Bemerkung. *Es gilt weiterhin: Ist ϕ^* der beste Test für das einfache Alternativproblem*

$$H_0 : \theta = \theta_0 \quad vs. \quad H_1 : \theta = \theta_1,$$

so ist ϕ^* auch der UMP-Test zum Niveau α für zusammengesetzte Hypothesen

$$H_0 : \theta \in \Theta_0 \quad vs. \quad H_1 : \theta \in \Theta_1,$$

wenn ϕ^* nicht von dem speziellen Wert $\theta_1 \in H_1$ abhängt und für alle $\theta \in H_0$ das Niveau α einhält.

Beispiel 2.23.

1. *Binomialtest mit $H_0 : \pi \leq \pi_0$ gegen $H_1 : \pi > \pi_0$ hat MLQ in $Z(x) =$ "Anzahl der Erfolge" (vgl. obiges Beispiel und Bemerkung). Der Binomialtest ist also UMP-Test.*

2. *Gleichverteilung*

3. *Gauß-Test*

4. *Exponentialverteilung*

5. *Poissonverteilung*

Bemerkung. *Oft existiert zwar kein UMP-Test, jedoch ein lokal bester (einseitiger) Test:* ϕ_{lok} *heißt* lokal bester Niveau α-Test $\overset{def}{\Leftrightarrow}$

$$g'_{\phi_{lok}}(\theta_0) = \frac{d}{d\theta} g_{\phi_{lok}}(\theta_0) \geq \frac{d}{d\theta} g_{\phi}(\theta_0),$$

wobei $g_{\phi_{lok}}(\theta_0) = g_{\phi}(\theta_0) = \alpha$ *gilt.*

Beste unverfälschte zweiseitige Tests bei skalarem θ

Für zweiseitige Testprobleme der Form

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq \theta_0$$

gibt es in der Regel keinen UMP-Test, insbesondere auch dann nicht, wenn MLQ vorliegt. Deshalb wird eine Restriktion auf eine kleinere Klasse von konkurrierenden Tests notwendig.

Definition 2.37 (Unverfälschter Niveau-α-Test)**.** *Ein Test* ϕ *für* H_0 *vs.* H_1 *heißt* unverfälschter (unbiased) Niveau-α-Test $\overset{def}{\Leftrightarrow}$

$$g_\phi(\theta) \leq \alpha \text{ für alle } \theta \in \Theta_0, \quad g_\phi(\theta) \geq \alpha \text{ für alle } \theta \in \Theta_1.$$

Satz 2.38 (Zweiseitige UMPU (uniformly most powerful unbiased) Tests)**.** *Sei*

$$f(x|\theta) = c(\theta) \exp(\theta T(x)) h(x)$$

eine einparametrische Exponentialfamilie mit natürlichem Parameter $\theta \in \Theta$ *(*Θ *sei ein offenes Intervall) und Statistik* $T(x)$*. Dann ist*

$$\phi^*(x) = \begin{cases} 1 & , \ T(x) < c_1 \\ \gamma_1 & , \ T(x) = c_1 \\ 0 & , \ c_1 < T(x) < c_2 \\ \gamma_2 & , \ T(x) = c_2 \\ 1 & , \ T(x) > c_2 \end{cases}$$

ein UMPU-Test zum Niveau α *unter allen unverfälschten Tests* ϕ *zum Niveau* α *für das Testproblem* $H_0 : \theta = \theta_0 \quad vs. \quad H_1 : \theta \neq \theta_0$*. Dabei werden* $c_1, c_2, \gamma_1, \gamma_2$ *aus*

$$\mathbb{E}_{\theta_0}[\phi^*(X)] = \alpha, \quad \mathbb{E}_{\theta_0}[\phi^*(X)T(X)] = \alpha \mathbb{E}_{\theta_0}[T(X)]$$

bestimmt.

Beispiel 2.24.

1. *Zweiseitiger Binomial-Test*

$$H_0 : \ \pi = \pi_0 \quad vs. \quad H_1 : \ \pi \neq \pi_0$$

ist UMPU-Test.

2. *Zweiseitiger Gauß-Test mit* $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, σ^2 *bekannt, ist für*

$$H_0 : \mu = \mu_0 \quad vs. \quad H_1 : \mu \neq \mu_0$$

UMPU-Test.

3. *Zweiseitiger Poisson-Test: Bei* $X_1, \ldots, X_n \overset{i.i.d.}{\sim} Po(\lambda)$

$$H_0 : \lambda = \lambda_0 \quad vs. \quad H_1 : \lambda \neq \lambda_0$$

liegt eine einparametrische Exponentialfamilie mit natürlichem Parameter $\theta = \log \lambda$ *vor. Äquivalente Hypothesen in* θ *sind*

$$H_0 : \theta = \theta_0 \quad vs. \quad H_1 : \theta \neq \theta_0.$$

Bestimmung der Prüfgröße:

$$\begin{aligned} f(x_i|\theta) &= h(x_i)c(\theta)\exp(\theta x_i) \\ f(x|\theta) &= f(x_1|\theta)\cdot \ldots \cdot f(x_n|\theta) \propto \exp\Big(\theta \underbrace{\sum_{i=1}^{n} x_i}_{T(x)}\Big) \end{aligned}$$

und somit

$$\phi^*(x) = \begin{cases} 1 & , \sum_{i=1}^n x_i < c_1 \\ \gamma_1 & , \sum_{i=1}^n x_i = c_1 \\ 0 & , c_1 < \sum_{i=1}^n x_i < c_2 \\ \gamma_2 & , \sum_{i=1}^n x_i = c_2 \\ 1 & , \sum_{i=1}^n x_i > c_2 \ . \end{cases}$$

4. *Zweiseitiger* χ^2*-Test auf die Varianz: Seien* $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2), \mu$ *bekannt. Getestet wird*

$$H_0 : \sigma^2 = \sigma_0^2 \quad vs. \quad H_1 : \sigma^2 \neq \sigma_0^2.$$

Mehrparametrische Verteilungsannahme

- Bislang: θ skalar.

  $\Rightarrow$ $\theta = (\mu, \sigma^2)$ ist bei $N(\mu, \sigma^2)$ Verteilung nicht in der Theorie optimaler Tests enthalten.

  $\Rightarrow$ t-Test auf μ (bei unbekanntem σ^2) und andere sind nicht erfasst.

- Idee: „Optimale" Tests lassen sich (noch) für eine skalare Komponente η von $\theta = (\eta, \xi)$, wobei ξ mehrdimensional sein darf, konstruieren. ξ ist als Stör-/Nuisanceparameter zu betrachten.

- Voraussetzung an Verteilungsfamilie: $\{f(x|\theta), \ \theta \in \Theta \subseteq \mathbb{R}^k\}$ ist eine (strikt) k-parameterische Exponentialfamilie mit natürlichem Parameter $\theta = (\eta, \xi)$ und $T = (U, V)$, U skalar. Dies führt auf die Theorie bedingter Tests.

- Passend zum Beispiel für
  - t-Test: Vergleich von μ_1, μ_2 bei unabhängigen Stichproben nur, falls $\sigma_1^2 = \sigma_2^2 = \sigma^2$ ist.
  - Test auf Signifikanz von β_1 in linearer Einfachregression.
- Bereits nicht mehr anwendbar für
  - Vergleich von μ_1, μ_2 bei $\sigma_1^2 \neq \sigma_2^2$ (Behrens-Fisher-Problem).
  - Test auf Signifikanz von β_1 im Logit- oder Poisson-Regressionsmodell.

$\Rightarrow$ (asymptotische) Likelihood-Theorie, Bayes-Inferenz.

2.3 Bereichsschätzungen und Konfidenzintervalle

2.3.1 Definition und Beurteilung der Güte

Definition 2.39 (Bereichsschätzung)**.** *Eine* Bereichsschätzung *(ein* Konfidenzbereich*)* C *für* $\tau(\theta)$, $\tau : \Theta \to \Sigma$, *zum (vorgegebenen)* Vertrauensgrad *(*Konfidenzniveau*)* $1 - \alpha$ *ist eine Abbildung des Stichprobenraums* $\mathcal{X}$ *in die Potenzmenge* $\mathcal{P}(\Sigma)$, *also* $x \to C(x) \in \mathcal{P}(\Sigma)$, *mit* $\{\tau(\theta) \in C(X)\}$ *messbar und*

$$\mathbb{P}_\theta(\tau(\theta) \in C(X)) \geq 1 - \alpha \quad \textit{für alle } \theta \in \Theta.$$

$C(X)$ ist ein zufälliger Bereich in $\mathcal{P}(\Sigma)$. Nach Beobachtung der Stichprobe $X = x$ ist $C(x)$ gegeben. Der Aussage

$$\tau(\theta) \in C(x) \quad (\text{richtig} \quad \overset{!}{\text{oder}} \quad \text{falsch})$$

wird der Vertrauensgrad $1 - \alpha$ zugeordnet. Dabei gilt die bekannte Häufigkeitsinterpretation. Ist $C(x)$ für jedes x ein Intervall, so heißt $C(x)$ *Konfidenzintervall* und C eine *Intervallschätzung*.

Eine Wahrscheinlichkeitsaussage zu

$$\tau(\theta) \in C(x)$$

bei gegebenem x ist im Rahmen der Bayes-Inferenz (ohne logische Probleme) möglich.

Die „Präzision" von $C(X)$ wird gemessen durch die erwartete Größe des Bereichs bzw. durch die Länge des Konfidenzintervalls.

Beispiel 2.25. *Seien* $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ *und*

$$C(X) = \left[\bar{X} - t_{n-1}\Big(\frac{\alpha}{2}\Big)\frac{S}{\sqrt{n}}, \bar{X} + t_{n-1}\Big(\frac{\alpha}{2}\Big)\frac{S}{\sqrt{n}}\right]$$

ein Konfidenzintervall für μ. *Die Länge*

$$L = 2\, t_{n-1}\Big(\frac{\alpha}{2}\Big)\frac{S}{\sqrt{n}}$$

von $C(X)$ ist zufällig mit Erwartungswert

$$\mathbb{E}(L) = 2\, t_{n-1}\Big(\frac{\alpha}{2}\Big)\frac{1}{\sqrt{n}}\ \mathbb{E}(S) = 2\, t_{n-1}\Big(\frac{\alpha}{2}\Big)\frac{\sigma}{\sqrt{n}}\sqrt{\frac{2}{n-1}}\frac{\Gamma(n/2)}{\Gamma((n-1)/2)}.$$

Es gilt:

$$\begin{aligned} 1-\alpha \ \textit{größer} \quad &\rightarrow \quad \mathbb{E}(L) \ \textit{größer,} \\ n \ \textit{größer} \quad &\rightarrow \quad \mathbb{E}(L) \ \textit{kleiner.} \end{aligned}$$

Bei der Beurteilung der Präzision eines Konfidenzintervalls durch die Länge ist ein Konfidenzintervall umso besser, je kürzer seine erwartete Länge ist. Allgemein wird ein Konfidenz*bereich* C durch die mittlere „Größe" beurteilt. Dazu sei π eine Verteilung (oder ein Maß) auf Θ. Dann ist

$$\pi(C(x))$$

die Größe von $C(x)$. Bei Konfidenzintervallen ergibt sich die Länge, wenn π das Lebesgue-Maß ist. Dann ist

$$\mathbb{E}_\theta(\pi(C(X)))$$

die zu erwartende Größe. Zur Beurteilung der Güte reicht die erwartete Länge bzw. Größe allein nicht aus.

Definition 2.40 (Kennfunktion eines Konfidenzbereichs)**.** *Eine* Kennfunktion *ist definiert als eine Funktion*

$$k_C(\theta, \theta') := \mathbb{P}_\theta(C(X) \ni \tau(\theta')).$$

Dabei ist θ der „wahre" Wert und θ' irgendein Wert in Θ.

Für $\theta = \theta'$ ist „$C(X) \ni \tau(\theta')$" eine Aussage, deren Wahrscheinlichkeit möglichst groß sein soll. Für $\theta \neq \theta'$ mit $\tau(\theta') \neq \tau(\theta)$ ist „$C(X) \ni \tau(\theta')$" eine Aussage, deren Wahrscheinlichkeit möglichst klein gehalten werden soll.

Im Weiteren betrachten wir den Spezialfall $\tau(\theta) = \theta$ mit skalarem θ. Dann ist

$$k_C(\theta, \theta') = \mathbb{P}_\theta(C(X) \ni \theta').$$

Definition 2.41.

1. *Ein Konfidenzbereich besitzt den* Vertrauensgrad $1-\alpha :\overset{def}{\Leftrightarrow}$

$$k_C(\theta, \theta') \geq 1-\alpha \quad \textit{für alle } \theta' = \theta.$$

2. *Ein Konfidenzbereich zum Vertrauensgrad $1-\alpha$ heißt* unverfälscht $:\overset{def}{\Leftrightarrow}$

$$k_C(\theta, \theta') \leq 1-\alpha \quad \textit{für } \theta' \neq \theta.$$

3. *Ein [unverfälschter] Konfidenzbereich C_0 zum Vertrauensgrad $1-\alpha$ heißt* gleichmäßig bester (trennscharfer) *[bzw. gleichmäßig bester unverfälschter]* Konfidenzbereich $:\overset{def}{\Leftrightarrow}$ *für alle $\theta' \neq \theta$ und alle [unverfälschten] Konfidenzbereiche C zum Vertrauensgrad $1-\alpha$ gilt*

$$k_{C_0}(\theta, \theta') \leq k_C(\theta, \theta').$$

Lemma 2.42. *Jeder gleichmäßig beste Konfidenzbereich besitzt auch die kleinste zu erwartende Größe (aber nicht umgekehrt).*

Beweis.

$$\begin{aligned}
\int_{\mathcal{X}} \pi(C(x)) d\mathbb{P}_\theta(x) &= \int_{\mathcal{X}} \int_{\Theta} I_{C(x)}(\theta') d\pi(\theta') d\mathbb{P}_\theta(x) \\
&= \int_{\Theta} \int_{\mathcal{X}} I_{C(x)}(\theta') d\mathbb{P}_\theta(x) d\pi(\theta') \qquad \text{(Fubini)} \\
&= \int_{\Theta} \underbrace{\mathbb{P}_\theta(\{x : C(x) \ni \theta'\})}_{k_C(\theta,\theta')} d\pi(\theta').
\end{aligned}$$

Für jedes „wahre" θ gilt also

$$\underbrace{\int_{\mathcal{X}} \pi(C(x)) d\mathbb{P}_\theta(x)}_{\text{erwartete Größe}} = \underbrace{\int_{\Theta} k_C(\theta, \theta') d\pi(\theta')}_{\substack{\text{erwarteter Wert der Kenn-} \\ \text{funktion des Konfidenzbereichs}}} .$$

□

2.3.2 Dualität zwischen Konfidenzbereichen und Tests

Wir legen den oben beschriebenen Spezialfall $\tau(\theta) = \theta$ mit skalarem θ zugrunde.

Zu jedem festen θ betrachten wir einen Niveau-α-Test $\phi_\theta(x)$ für die Nullhypothese $H_0 = \{\theta\}$ gegen die Alternative $H_1 = \Theta \backslash H_0$. Die Tests sollen nicht randomisiert sein, so dass sie durch die Festlegung einer Prüfgröße $T_\theta = T_\theta(x)$ und eines kritischen Bereichs (Ablehnbereichs) K_θ bestimmt werden:

$$\phi_\theta(x) = \begin{cases} 1 & \text{für } T_\theta(x) \in K_\theta, \\ 0 & \text{sonst.} \end{cases}$$

Die Nullhypothese „Der unbekannte Parameter hat den Wert θ" wird nach Beobachtung von $X = x$ genau dann nicht abgelehnt — durch die Beobachtung „bestätigt" — wenn

$$T_\theta\ (x)\ \in\ \bar{K}_\theta\ =\ \text{Annahmebereich des Tests } \phi_\theta$$

gilt. Daher ist es naheliegend, als einen Konfidenzbereich nach der Beobachtung $X = x$ den Bereich

$$C(x) := \{\theta \in \Theta : T_\theta(x) \in \bar{K}_\theta\}$$

zu definieren; dem entspricht vor der Beobachtung der zufällige Bereich

$$C(X) = \{\theta \in \Theta : T_\theta(X) \in \bar{K}_\theta\}$$

bzw.

$$C(X) = \{\theta \in \Theta : \phi_\theta(X) = 0\}$$

Eine Bestätigung dieser Vorgangsweise ist der folgende Satz.

Satz 2.43 (Korrespondenzsatz)**.**

1. *Ist* $\{\phi_\theta\}$ *eine Menge von Tests* ϕ_θ *für* $H_0 = \{\theta\}$ *gegen* $H_1 = \Theta\backslash\{\theta\}$ *zum Niveau* α*, so ist* $C(X) := \{\theta \in \Theta : \phi_\theta(X) = 0\}$ *ein Konfidenzbereich zum Vertrauensgrad* $\gamma = 1 - \alpha$*.*

2. *Ist* $\{\phi_\theta\}$ *eine Menge gleichmäßig bester [unverfälschter] Tests, so ist auch* $C(X)$ *ein gleichmäßig bester [unverfälschter] Konfidenzbereich.*

Beweis. Der Beweis zu 1. ergibt sich aus

$$\mathbb{P}_\theta(C(X) \ni \theta) = \mathbb{P}_\theta(\phi_\theta(X) = 0) = 1 - \alpha \quad \text{für alle } \theta \in \Theta,$$

derjenige für 2. aus der Beziehung

$$\begin{aligned} k_C(\theta, \theta') &= \mathbb{P}_\theta(C(X) \ni \theta') = \mathbb{P}_\theta(\phi_{\theta'}(X) = 0) \\ &= 1 - \mathbb{P}_\theta(\phi_{\theta'}(X) = 1) = 1 - g_{\phi_{\theta'}}(\theta) \end{aligned}$$

für alle $\theta, \theta' \in \Theta$. Dabei bezeichnet $g_{\phi_{\theta'}}$ die Gütefunktion des Tests $\phi_{\theta'}$. □

Der Korrespondenzsatz lässt sich verallgemeinern auf die Situation, in der man gegenüber bestimmten Fehlschätzungen besonders empfindlich ist; man hat dazu eine Testfamilie solcher Tests zugrunde zu legen, die die entsprechenden Hypothesen testen, also nicht mehr Tests mit zweiseitiger Fragestellung. Darüber hinaus gilt der im Korrespondenzsatz enthaltene Zusammenhang zwischen Tests und einem Konfidenzbereich auch dann, wenn randomisierte Tests zugelassen werden, so dass man auf diese Weise zu einem randomisierten Konfidenzbereich kommt: $C(x)$ ist die Menge aller θ, die bei der Beobachtung x von dem Test ϕ_θ (auch nach Randomisierung) nicht abgelehnt werden.

Auf diese Weise lässt sich die Theorie der Bereichsschätzungen auf die Testtheorie zurückführen bis auf das folgende Problem: Damit ein „vernünftiger" Konfidenzbereich (vernünftig im topologischen Sinn, also zum Beispiel ein Konfidenzintervall) aus der Testfamilie konstruierbar ist, muss die Testfunktion $\phi_\theta(x)$, besser noch die Prüfgröße $T_\theta(x)$ *als Funktion in* θ (für jedes feste θ) „gutartig" sein (im Idealfall monoton in θ); außerdem darf die Verteilung von $T_\theta(X)$ nicht von θ abhängen, zusammen bedeutet dies: $T_\theta(X)$ muss eine *Pivotgröße* sein, die auf „einfache" (zum Beispiel monotone) Weise von θ abhängt: Gesucht sind einfach strukturierte Pivotgrößen.

2.4 Multiples Testen

Literatur:

- Lehmann & Romano, Kapitel 9
- Dudoit, Shaffer & Boldrick (2003): *Multiple Hypothesis Testing in Microarray Experiments*, Statistical Science (**18**), Seiten 71-103

Problem: Eine endliche Menge von (Null-) Hypothesen $H_1, ..., H_m$ soll mit Hilfe *eines* Datensatzes simultan getestet werden.

Beispiele:

- *Varianzanalyse:* Vergleich mehrerer Behandlungsarten mit Kontrolle (zum Beispiel Placebo oder „übliche" Therapie). Ein simultaner Test der Form

$$H_0 : \theta_1 = \ldots = \theta_m = 0 \quad \text{vs.} \quad H_{\text{alter}} : \quad \text{wenigstens ein } \theta_j \neq 0$$

ist oft nicht ausreichend: Wenn H_0 abgelehnt wird, möchte man wissen, welche θ_j's signifikant von 0 verschieden sind. Hierzu können (simultan) die einzelnen Hypothesen

$$H_j := H_{0j} : \theta_j = 0$$

für $j = 1, \ldots, m$ getestet werden. In der Regel ist m vergleichsweise klein; es können „klassische" multiple Testverfahren verwendet werden.

- *Microarray-Experimente:* Seien $X_1, \ldots, X_m$ (normalisierte log-) Expressionen von Genen $1, \ldots, m$ auf Microarrays, $X_j \overset{\text{a}}{\sim} N(\mu_j, \sigma_j)$ für $j = 1, \ldots, m$ und m von der Größenordnung 1000 bis 10000. Es soll untersucht werden, welche Gene signifikanten Einfluss auf einen Phänotyp, zum Beispiel eine bestimmte Krankheit, haben. In einem naiven Ansatz könnte dies wie oben durch simultane Tests untersucht werden. Wenn m und die Anzahl m_0 richtiger Hypothesen jedoch groß ist, werden mit hoher Wahrscheinlichkeit eine oder mehr Hypothesen fälschlicherweise abgelehnt. Für unabhängige Teststatistiken $T_1, \ldots, T_m$ gilt zum Beispiel folgende Tabelle.

m	1	2	5	10	50
$\mathbb{P}$(mindestens eine falsche Ablehnung)	0.05	0.10	0.23	0.40	0.92

Es werden „neue" multiple Testverfahren gesucht, um Fehlerraten zu kontrollieren.

2.4.1 Fehlerraten

Die Situation bei m vorgegebenen Hypothesen kann wie folgt beschrieben werden:

	Anzahl nicht abgelehnter Nullhypothesen	Anzahl abgelehnter Nullhypothesen	
Anzahl richtiger Nullhypothesen	U	V	m_0
Anzahl falscher Nullhypothesen	T	S	m_1
	$m - R$	R	

Dabei sind

- m_0 die (unbekannte) Anzahl richtiger Nullhypothesen,
- $m_1 = m - m_0$ die (unbekannte) Anzahl falscher Nullhypothesen,
- R eine beobachtbare Zufallsvariable,
- S, T, U, V unbeobachtbare Zufallsvariablen.

In der Microarray-Analyse bedeutet das Ablehnen von H_j, dass das Gen j „differentiell exprimiert" ist.

Idealerweise: Minimiere

- Anzahl V von Fehlern 1. Art (falsch positiv),
- Anzahl T von Fehlern 2. Art (falsch negativ).

Klassische Testtheorie ($m = 1$):

$$\begin{aligned} \mathbb{P}(\text{Fehler 1. Art}) &\leq \alpha \\ \mathbb{P}(\text{Fehler 2. Art}) &\to \min \end{aligned}$$

Verschiedene Verallgemeinerungen zur Kontrolle der Fehlerraten sind bei multiplem Testen möglich.

Fehlerraten 1. Art (type I error rates)

- PCER (per-comparison error rate):

  $$\text{PCER} = \frac{\mathbb{E}(V)}{m}$$

  Das ist die relative Anzahl erwarteter Fehler 1. Art.

- PFER (per-family error rate):

  $$\text{PFER} = \mathbb{E}(V)$$

  Das ist die absolute Anzahl erwarteter Fehler 1. Art.

- FWER (family-wise error rate):

  $$\text{FWER} = \mathbb{P}(V \geq 1)$$

  Das ist die Wahrscheinlichkeit für mindestens einen Fehler 1. Art.

- FDR (false discovery rate; Benjamini & Hochberg, 1995):

  $$\text{FDR} = \mathbb{E}(Q) \qquad \text{mit} \qquad Q = \begin{cases} \frac{V}{R} & \text{für } R > 0, \\ 0 & \text{für } R = 0. \end{cases}$$

  Das ist die erwartete relative Häufigkeit von Fehlern 1. Art unter den R abgelehnten Hypothesen.

Es gilt PCER $\leq$ FDR $\leq$ FWER $\leq$ PFER (FDR = FWER bei $m = m_0$).

Starke und schwache Kontrolle

Typischerweise gilt: Für eine *unbekannte* Teilmenge

$$\Lambda_0 \subseteq \{1, \ldots, m\}$$

sind die Hypothesen $H_j, j \in \Lambda_0$, richtig, für den Rest falsch. *Starke* Kontrolle liegt vor, wenn eine Fehlerrate für *jede* Teilmenge Λ_0 nach oben durch α beschränkt wird, zum Beispiel

$$\text{FWER} \leq \alpha$$

gilt. *Schwache* Kontrolle liegt vor, wenn die Fehlerrate kontrolliert wird, falls *alle* Nullhypothesen richtig sind.

Klassische Ansätze (zum Beispiel Bonferroni- und Holm-Prozedur, siehe folgender Abschnitt) kontrollieren *stark*. Der FDR-Ansatz von Benjamini und Hochberg kontrolliert die FDR *schwach* und ist (deshalb) weniger konservativ.

2.4.2 Multiple Testprozeduren

Bonferroni-Prozedur

Lehne für $j = 1, \ldots, m$ die Hypothesen H_j ab, falls für den p-Wert gilt: $p_j \leq \frac{\alpha}{m}$. Es gilt:

$$\text{FWER} \leq \alpha \quad \textit{stark},$$

d.h.

$$\mathbb{P}\left(V \geq 1 \,\Big|\, \bigcap_{j \in \Lambda_0} H_j \right) \leq \alpha.$$

Nachteil: Das Niveau α/m der individuellen Tests wird bei großem m und üblichem α extrem klein. Bei Microarrays bleiben relevante Gene deshalb mit hoher Wahrscheinlichkeit unentdeckt.

Holm-Prozedur

Ordne die p-Werte $p_j, j = 1, \ldots, m$, der individuellen Tests $H_1, \ldots, H_m$ der Größe nach an. Dann ist

$$p_{(1)} \leq \ldots \leq p_{(m)}$$

mit den entsprechend sortierten Hypothesen $H_{(1)}, \ldots, H_{(m)}$. Als nächstes erfolgt *schrittweise* folgende Prozedur:

Schritt 1. Falls $p_{(1)} \geq \frac{\alpha}{m}$, akzeptiere $H_1, \ldots, H_m$.

Falls $p_{(1)} < \frac{\alpha}{m}$, lehne $H_{(1)}$ ab und teste die verbleibenden $m-1$ Hypothesen zum Niveau $\frac{\alpha}{m-1}$.

Schritt 2. Falls $p_{(1)} < \frac{\alpha}{m}$, aber $p_{(2)} \geq \frac{\alpha}{m-1}$, akzeptiere $H_{(2)}, \ldots, H_{(m)}$ und stoppe.

Falls $p_{(1)} < \frac{\alpha}{m}$ und $p_{(2)} < \frac{\alpha}{m-1}$, lehne nach $H_{(1)}$ auch $H_{(2)}$ ab und teste die verbleibenden $m-2$ Hypothesen zum Niveau $\frac{\alpha}{m-2}$.

Schritt 3. usw.

Es gilt:

$$\text{FWER} \leq \alpha \quad \textit{stark}.$$

Beweis:

Sei j^* der kleinste (zufällige) Index mit $p_{(j^*)} = \min_{j \in \Lambda_0} p_j$.

Eine falsche Ablehnung liegt vor, wenn

$$p_{(1)} \leq \alpha/m, p_{(2)} \leq \alpha/(m-1), \ldots, p_{(j^*)} \leq \alpha/(m-j^*+1).$$

Da $j^* \leq m - m_0 + 1$ gelten muss, folgt daraus

$$\min_{j \in \Lambda_0} p_j = p_{(j^*)} \leq \alpha/(m-j^*+1) \leq \alpha/m_0.$$

Damit ist die Wahrscheinlichkeit für eine falsche Ablehnung ($V \geq 1$) nach oben beschränkt durch

$$FWER \leq \mathbb{P}(\min_{j \in \Lambda_0} p_j \leq \alpha/m_0) \leq \sum_{j \in \Lambda_0} \mathbb{P}(p_j \leq \alpha/m_0) \leq \alpha.$$

□

Die Holm-Prozedur ist eine spezielle Form folgender Step-Down-Prozeduren:

Step-Down-Prozeduren

Allgemeine Struktur: Sei

$$\alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_m.$$

Falls $p_{(1)} \geq \alpha_1$, akzeptiere alle Hypothesen. Sonst lehne für $r = 1, \ldots, m$ die Hypothesen $H_{(1)}, \ldots, H_{(r)}$ ab, falls

$$p_{(1)} < \alpha_1, \ldots, p_{(r)} < \alpha_r.$$

Die Holm-Prozedur benutzt $\alpha_j = \alpha/(m - j + 1)$.

Eine Alternative sind:

Step-Up-Prozeduren

Falls $p_{(m)} < \alpha_m$, verwerfe alle Hypothesen. Sonst lehne für $r = 1, \ldots, m$ die Hypothesen $H_{(1)}, \ldots, H_{(r)}$ ab, falls

$$p_{(m)} \geq \alpha_m, \ldots, p_{(r+1)} \geq \alpha_{r+1},$$

aber $p_{(r)} < \alpha_r$.

Bemerkung.

- *Aussagen über* starke *Kontrolle finden sich zum Beispiel in Lehmann & Romano, Kapitel 9.*
- *Für* $m \sim 100, 1000$ *und größer: Immer noch geringe Power, deutlich weniger als für die Einzeltests. Benjamini & Hochberg (1995) raten, die false discovery rate FDR zu kontrollieren. Die Eigenschaften von Multiplen Testprozeduren sind weiterhin Gegenstand aktueller Forschung.*
- *Für* $m \sim 100, 1000$ *und größer: Immer noch geringe Power, deutlich weniger als für die Einzeltests. Benjamini & Hochberg (1995) raten, die false discovery rate FDR zu kontrollieren. Die Eigenschaften von Multiplen Testprozeduren sind weiterhin Gegenstand aktueller Forschung.*
- *Die diversen Prozeduren lassen sich teils günstig mit Hilfe von* adjustierten *p-Werten* $\widetilde{p}_j$ *formulieren, siehe Dudoit, Shaffer & Boldrick (2003).*
- *Resampling Methoden (Bootstrap, Permutationen, ...) sind notwendig, um (adjustierte) p-Werte zu berechnen.*
- *Software:* `www.bioconductor.org`.