Multiples Testen

Vorlesungsskript

Thorsten Dickhaus
Humboldt-Universität zu Berlin
Sommersemester 2010
Version: 24. September 2010

Vorbemerkungen

Die Kapitel 1, 3 und 4 dieses Skripts sind im Wesentlichen aus den Vorlesungsskripten über Multiples Testen von Helmut Finner und Iris Pigeot übernommen. Beiden gilt mein herzlicher Dank. Sollten sich in diesen Kapiteln Fehler finden, so bin dafür natürlich ich verantwortlich. Lob und positive Kritik gebührt indes den Original-AutorInnen.

Für die Manuskripterstellung danke ich Mareile Große Ruse und Jens Stange.

Übungsaufgaben und R-Programme zu diesem Kurs stelle ich auf Anfrage gerne zur Verfügung. Einige Referenzen dazu finden sich im Text an den zugehörigen Stellen.

Verzeichnis der Abkürzungen und Symbole

AORC	Asymptotically Optimal Rejection Curve
$B(p,q)$	Betafunktion, $B(p,q) = \Gamma(p)\Gamma(q)/\Gamma(p+q)$
$\lceil x \rceil$	Kleinste ganze Zahl größer oder gleich x
χ^2_ν	Chi-Quadrat Verteilung mit ν Freiheitsgraden
$\complement M$	Komplement der Menge M
cdf.	Cumulative distribution function
δ_a	Dirac-Maß im Punkte a
ecdf.	Empirical cumulative distribution function
$=^d$	Gleichheit in Verteilung
F_X	Verteilungsfunktion einer reellwertigen Zufallsvariable X
FDR	False Discovery Rate
FWER	Family Wise Error Rate
$\lfloor x \rfloor$	Größte ganze Zahl kleiner oder gleich x
$\Gamma(\cdot)$	Gammafunktion, $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt,\ x > 0$
im(X)	Bildbereich einer Zufallsgröße X
iid.	independent and identically distributed
$\mathbf{1}_M$	Indikatorfunktion einer Menge M
$\mathcal{L}(X)$	Verteilungsgesetz einer Zufallsvariable X

LFC	Least Favorable Configuration
MTP_2	Multivariate total positivity of order 2
$\mathcal{N}(\mu, \sigma^2)$	Normalverteilung mit Parametern μ und σ^2
Φ	Verteilungsfunktion der $\mathcal{N}(0,1)$-Verteilung
$\varphi(\cdot)$	Verteilungsdichte der $\mathcal{N}(0,1)$-Verteilung
PRDS	Positive regression dependency on subsets
pdf.	Probability density function
SD	Step-down
SU	Step-up
SUD	Step-up-down
UNI$[a,b]$	Gleichverteilung auf dem Intervall $[a,b]$

Inhaltsverzeichnis

Kapitel 1

Einführung und Beispiele

1.1 Grundlagen aus der Statistik

Bezeichne X eine Zufallsgröße, die den möglichen Ausgang eines Experimentes beschreibt.[1]
Sei Ω der zu X gehörige Stichprobenraum, d. h., die Menge aller möglichen Realisierungen von X und $\mathcal{A} \subseteq 2^\Omega$ eine σ-Algebra über Ω. Die Elemente von $\mathcal{A}$ heißen messbare Teilmengen von Ω oder Ereignisse.
Bezeichne $\mathbb{P}^X$ die Verteilung von X. Es gelte $\mathbb{P}^X \in \mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$.

Definition 1.1 (Statistisches Experiment / Modell)
Ein Tripel $(\Omega, \mathcal{A}, \mathcal{P})$ mit $\Omega \neq \emptyset$ eine nichtleere Menge, $\mathcal{A} \subseteq 2^\Omega$ eine σ-Algebra über Ω und $\mathcal{P} = \{\mathbb{P}_\vartheta : \vartheta \in \Theta\}$ eine Familie von Wahrscheinlichkeitsmaßen auf $\mathcal{A}$ heißt statistisches Experiment bzw. statistisches Modell.
Falls $\Theta \subseteq \mathbb{R}^k, k \in \mathbb{N}$, so heißt $(\Omega, \mathcal{A}, \mathcal{P})$ parametrisches statistisches Modell, $\vartheta \in \Theta$ Parameter und Θ Parameterraum.

Statistische Inferenz beschäftigt sich damit, Aussagen über die wahre Verteilung $\mathbb{P}^X$ bzw. den wahren Parameter ϑ zu gewinnen. Speziell: Entscheidungsprobleme, insbesondere Testprobleme.
Testprobleme: Gegeben zwei disjunkte Teilmengen $\mathcal{P}_0, \mathcal{P}_1$ von $\mathcal{P}$ mit $\mathcal{P}_0 \cup \mathcal{P}_1 = \mathcal{P}$ ist eine Entscheidung darüber gesucht, ob $\mathbb{P}^X$ zu $\mathcal{P}_0$ oder $\mathcal{P}_1$ gehört. Falls $\mathcal{P}$ durch ϑ eineindeutig identifiziert ist, kann die Entscheidungsfindung auch vermittels ϑ und Teilmengen Θ_0 und Θ_1 von Θ mit $\Theta_0 \cap \Theta_1 = \emptyset$ und $\Theta_0 \cup \Theta_1 = \Theta$ formalisiert werden.
Formale Beschreibung des Testproblems:

$$H_0 : \vartheta \in \Theta_0 \quad \text{versus} \quad H_1 : \vartheta \in \Theta_1 \quad \text{oder}$$

$$H_0 : \mathbb{P}^X \in \mathcal{P}_0 \quad \text{versus} \quad H_1 : \mathbb{P}^X \in \mathcal{P}_1.$$

[1] Witting (1985): „Wir denken uns das gesamte Datenmaterial zu einer „Beobachtung“ x zusammengefasst.“

Die $H_i, i = 1, 2$ nennt man Hypothesen. H_0 heißt Nullhypothese, H_1 Alternativhypothese / Alternative. Oft interpretiert man H_0 und H_1 auch direkt selbst als Teilmengen des Parameterraums, d. h., $H_0 \cup H_1 = \Theta$ und $H_0 \cap H_1 = \emptyset$. Zwischen H_0 und H_1 ist nun aufgrund von $x \in \Omega$ eine Entscheidung zu treffen. Dazu benötigt man eine Entscheidungsregel. Diese liefert ein statistischer Test.

Definition 1.2 (Statistischer Test)
Ein (nicht-randomisierter) statistischer Test ist eine messbare Abbildung

$$\varphi : (\Omega, \mathcal{A}) \to (\{0, 1\}, 2^{\{0,1\}}).$$

Konvention:

$$\varphi(x) = 1 \iff \textit{Nullhypothese wird verworfen, Entscheidung für } H_1,$$

$$\varphi(x) = 0 \iff \textit{Nullhypothese wird nicht verworfen.}$$

$\{x \in \Omega : \varphi(x) = 1\}$ *heißt Ablehnbereich (oder auch kritischer Bereich) von* φ*, kurz:* $\{\varphi = 1\}$.
$\{x \in \Omega : \varphi(x) = 0\}$ *heißt Annahmebereich von* φ*, kurz:* $\{\varphi = 0\} = \complement\{\varphi = 1\}$.

Problem: Testen beinhaltet mögliche Fehlentscheidungen.
Fehler 1. Art (α-Fehler, type I error): Entscheidung für H_1, obwohl H_0 wahr ist.
Fehler 2. Art (β-Fehler, type II error): Nicht-Verwerfung von H_0, obwohl H_1 wahr ist.
In der Regel ist es nicht möglich, die Wahrscheinlichkeiten für die Fehler 1. und 2. Art gleichzeitig zu minimieren. Daher: Asymmetrische Betrachtungsweise von Testproblemen.

(i) Begrenzung der Fehlerwahrscheinlichkeit 1. Art durch eine vorgegebene obere Schranke α (Signifikanzniveau, englisch: level),

(ii) Unter der Maßgabe (i) Minimierung der Wahrscheinlichkeit für Fehler 2. Art $\Rightarrow$ „optimaler“ Test.

Eine (zum Niveau α) statistisch abgesicherte Entscheidung kann also immer nur zu Gunsten von H_1 getroffen werden $\Rightarrow$ Merkregel: „Was nachzuweisen ist stets als Alternative H_1 formulieren!“.

Bezeichnungen 1.3

(i) $\beta_\varphi(\vartheta) = \mathbb{E}_\vartheta[\varphi] = \mathbb{P}_\vartheta(\varphi(X) = 1) = \int_\Omega \varphi d\mathbb{P}_\vartheta$ *bezeichnet die Ablehnwahrscheinlichkeit eines vorgegebenen Tests* φ *in Abhängigkeit von* $\vartheta \in \Theta$*. Für* $\vartheta \in \Theta_1$ *heißt* $\beta_\varphi(\vartheta)$ *Gütefunktion von* φ *an der Stelle* ϑ*. Für* $\vartheta \in \Theta_0$ *ergibt* $\beta_\varphi(\vartheta)$ *die Typ I-Fehlerwahrscheinlichkeit von* φ *unter* $\vartheta \in \Theta_0$.

Für $\alpha \in (0, 1)$ *vorgegeben heißt*

(ii) ein Test φ mit $\beta_\varphi(\vartheta) \leq \alpha$ für alle $\vartheta \in H_0$ Test zum Niveau α,

(iii) ein Test φ zum Niveau α unverfälscht, falls $\beta_\varphi(\vartheta) \geq \alpha$ für alle $\vartheta \in H_1$.

(iv) ein Test φ_1 zum Niveau α besser als ein zweiter Niveau-α Test φ_2, falls $\beta_{\varphi_1}(\vartheta) \geq \beta_{\varphi_2}(\vartheta)$ für alle $\vartheta \in H_1$ und $\exists \vartheta^* \in H_1$ mit $\beta_{\varphi_1}(\vartheta^*) > \beta_{\varphi_2}(\vartheta^*)$.

1.2 Motivation und Beispiele

Bislang: Klärung *einer* Fragestellung (formuliert als statistisches Hypothesenpaar) anhand der Beobachtung $x \in \Omega$.
Im Folgenden: Klärung *mehrerer* Fragen gleichzeitig anhand von $x \in \Omega$.
$\rightarrow$ simultane statistische Inferenz, *statistische Mehrentscheidungsprobleme.*

Statistische Mehrentscheidungsverfahren:

(i) Multiple Tests (Englisch oft: multiple comparisons)

(ii) Simultane Konfidenzbereiche

(iii) Selektionsverfahren

(iv) Partitionsverfahren

(v) Rankingverfahren

Beispiel 1.4 (Mehrgruppenvergleiche im balancierten ANOVA-Design)
Wir betrachten Beobachtungen der Form $X = (X_{ij})_{i=1,\ldots,k,\ j=1,\ldots,n}$, wobei $X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$ stochastisch unabhängige Zufallsvariablen auf $\mathbb{R}$, $\mu_i \in \mathbb{R}\ \forall 1 \leq i \leq k, \sigma^2 > 0$ (unbekannte) Varianz, $k \geq 3, n \geq 2, \nu := k(n-1)$ (Freiheitsgrade)[2]. Dieses Modell wird in der Praxis häufig benutzt und Gegenstand vieler weiterer Untersuchungen sein. Zum Beispiel könnten die μ_i die mittleren Erträge von k unterschiedlichen Getreidesorten oder die mittlere Wirksamkeit von k unterschiedlichen Medikamenten beschreiben. Formal erhalten wir als statistisches Modell $\Omega = \mathbb{R}^{k \cdot n}$, $\mathcal{A} = \mathbb{B}^{k \cdot n}$, $\vartheta = (\mu_i, \ldots, \mu_k, \sigma^2) \in \mathbb{R}^k \times [0, \infty) = \Theta$.

***Problem (i):** Die μ_i sollen paarweise auf Unterschiede getestet werden.*

$$\textit{Hypothesen: } H_{ij} : \{\mu_i = \mu_j\} \ \textit{ vs. } \ K_{ij} : \{\mu_i \neq \mu_j\}, 1 \leq i < j \leq k.$$

[2]Wir bezeichnen durchgängig Dimensionalitäten von Parametern mit k, Anzahlen unabhängiger Wiederholungen (Stichprobenumfänge) mit n und Anzahlen simultan zu prüfender Hypothesen mit m.

Es sind also $m = \binom{k}{2} = k(k-1)/2$ *Hypothesen zu testen (am gleichen Datenmaterial!). Hinweis: Die klassische Varianzanalyse testet nur die Globalhypothese* $H_0 = \bigcap_{1 \le i < j \le k} H_{ij}$. *Falls diese abgelehnt wird, lässt sich nicht lokalisieren, wo Unterschiede liegen!*
Würde man jede Hypothese H_{ij} *mit einem* t*-Test zum Niveau* α *prüfen, so wäre die Wahrscheinlichkeit für irgendeinen Fehler 1. Art im Allgemeinen wesentlich größer als* α*, falls mehrere der* H_{ij} *wahr sind. Dies impliziert die Notwendigkeit von simultanen Typ I-Fehlermaßen. Zum Beispiel ist in Keuls (1952)* $k = 13$ *und damit* $m = 78$.

Problem (ii): *Es sollen Konfidenzintervalle* $C_{ij}(x)$ *für alle paarweisen Differenzen* $\theta_{ij} = \mu_i - \mu_j$, $1 \le i < j \le k$, *angegeben werden. Auch hier steht man vor dem Problem, dass sich die Wahrscheinlichkeiten für Nichtüberdeckungen aufsummieren können. Ein Ausweg ist hier die Einführung eines simultanen Konfidenzniveaus, also die Forderung*

$$\forall \mu \in \mathbb{R}^k : \forall \sigma^2 > 0 : \mathbb{P}_{\mu,\sigma^2}(C_{ij}(X) \ni \theta_{ij} \; \forall 1 \le i < j \le k) \ge 1 - \alpha. \tag{1.1}$$

Problem (iii): *Es sollen die besten (oder die beste) Behandlung(en) bzw. Sorte(n) (*μ_i*:* i*-ter Behandlungs-/Sortenmittelwert) gefunden (selektiert) werden.* $\rightarrow$ *Selektions-, Auswahlverfahren.*
Hierzu gibt es eine Vielzahl von Ansätzen. Häufig wird die Menge der guten Behandlungen charakterisiert durch

$$G(\vartheta) = \{i : \max_{1 \le j \le k} \mu_j - \mu_i \le \varepsilon\sigma\} \text{ für ein } \varepsilon \ge 0.$$

Dann gilt also. Behandlung i *ist „gut“ genau dann, wenn* $i \in G(\vartheta)$. *Dies läuft auf das Rechnen mit Orderstatistiken hinaus. Die Theorie reicht zurück zu Bechhofer (1954) und Gupta (1956). Ein mögliches Zielkriterium ist, eine Selektionsregel (-menge)* $S = S(X)$ *minimaler Mächtigkeit zu finden, so dass*

$$\mathbb{P}_\vartheta(S(X) \cap G(\vartheta) \ne \emptyset) \ge P^* \; \forall \vartheta \in \Theta^*$$

für ein vorgegebenes PCS (Probability of a Correct Selection)-Niveau P^* *und eine „geeignete“ Teilparametermenge* Θ^*.

Problem (iv): *Die Menge aller Behandlungen soll in Teilmengen mit vordefinierten Eigenschaften zerlegt (partitioniert) werden* $\rightarrow$ *Partitionsverfahren.*
Ist zum Beispiel $i = k$ *eine Kontroll- oder Standardbehandlung bzw. -sorte, so könnte ein Ziel sein, die anderen Behandlungen in gute, schlechte und äquivalente (im Vergleich mit dem Standard) einzuteilen:*

$$\begin{aligned} G(\vartheta) &= \{i : \mu_i - \mu_k > \delta_2\sigma\} \quad \text{„gute“}, \\ B(\vartheta) &= \{i : \mu_k - \mu_i > \delta_2\sigma\} \quad \text{„schlechte“}, \\ E(\vartheta) &= \{i : |\mu_i - \mu_k| \le \delta_1\sigma\} \; \text{„äquivalente“}, 0 < \delta_1 < \delta_2. \end{aligned}$$

Ein mögliches Kriterium hier lautet: Finde eine Partitionsregel $S = S(X) = (S_1, S_2, S_3)$ mit

$$\forall \vartheta \in \Theta : \mathbb{P}_\vartheta(G(\vartheta) \subseteq S_1(X), B(\vartheta) \subseteq S_2(X), E(\vartheta) \subseteq S_3(X)) \geq P^*$$

unter der Nebenbedingung $S_1 \cup S_2 \cup S_3 = \{1, \ldots, k-1\}$.

***Problem (v):** Den Behandlungen soll entsprechend ihrer Qualität ein Rang zugeordnet werden, d. h., die Menge $\{1, \ldots, k\}$ soll in beste, zweitbeste, . . ., schlechteste angeordnet werden*
$\rightarrow$ Rankingverfahren.
Ein mögliches Zielkriterium hier:

$$\forall \vartheta \in \Theta^* : \mathbb{P}_\vartheta(\text{„korrektes Ranking“}) \geq P^*,\ \Theta^* \subset \Theta \text{ „geeignet“}.$$

In dieser Vorlesung: Beschränkung auf Probleme der Form (i) und (ii)!

Beispiel 1.5 (Multiple Endpunkte, Pocock et al., 1987)
Der Prüfplan klinischer Studien enthält in aller Regel mehrere zu untersuchende Zielkriterien. Damit kann folgenderweise umgegangen werden:

- *ein primäres Zielkriterium auswählen und mit einem statistischen Test überprüfen, den Rest nur explorativ untersuchen (oft nicht möglich, da Zielkriterien gleichwertig sind)*
- *alle Zielkriterien mit statistischen Tests absichern $\rightarrow$ multiples Tesproblem*

Konkret bei Pocock et al.: chronische Atemwegserkrankungen

- *randomisierte Doppelblindstudie, Cross-Over Design*
- 17 *Patienten mit Asthma oder chronischer obstruktiver Atemwegserkrankung*
- *Behandlung mit a) Inhalationsmittel, b) Placebo an jedem Patienten in zufälliger Reihenfolge, jeweils über vier Wochen*

Zu messende Standardatemwegsparameter:

(i) peak expiratory flow rate (PEFR)

(ii) forciertes Ausatmungsvolumen

(iii) forcierte Vitalkapazität

Frage: Existieren Unterschiede zwischen Placebo und Medikament hinsichtlich dieser Parameter?
Dazu: Multiplen statistischen Test verwenden, Signifikanzniveau pro Test adjustieren.

Ohne Kenntnis der Abhängigkeitsstruktur zwischen den Messwerten zu den Parametern (i)-(iii): Signifikanzniveau α dritteln $\rightarrow$ kann konservativ sein (später mehr!)

Das Verfahren der Autoren arbeitet Vorkenntnisse über die Abhängigkeitsstruktur in die Testprozedur ein.

Fazit: Die Auswertung eines einzelnen Datensatzes anhand mehrerer statistischer Tests ist eine nichttriviale Erweiterung, denn

1. Prüfgrößen der Einzeltests sind im Allgemeinen nicht stochastisch unabhängig.
2. Ihre gemeinsame Verteilung ist schwer bzw. gar nicht bestimmbar.
3. Wird jeder Einzeltest zum Niveau α durchgeführt, kann die Irrtumswahrscheinlichkeit für die Gesamtaussage unüberschaubar werden (wenngleich genau diese letztendlich interessiert).

Diese Gesamtaussage, d. h., die Verbindung der einzelnen Testentscheidungen, ist nur dann statistisch valide, wenn sie ebenfalls durch ein vorgegebenes Kriterium für die Wahrscheinlichkeit möglicher Fehlentscheidungen abgesichert ist. D. h., ein multipler Test sollte die Wahrscheinlichkeit kontrollieren, dass bei der Gesamtheit aller Einzeltests eine oder mehrere Nullhypothesen fälschlicherweise verworfen werden und dennoch die vorhandenen Abweichungen von den Nullhypothesen mit möglichst hoher Güte aufdecken können.
Außerdem sollte er insgesamt zu einer „vernünftigen“ Testentscheidung führen.

$\Longrightarrow$ **Eigene Theorie multipler Tests notwendig!**

Bemerkung 1.6
Die Theorie multipler Tests hat noch viele weitere nützliche Anwendungen. besonders beliebt sind zum Beispiel multiple Tests als Modellselektionsverfahren, d. h., zur Festlegung der Anzahl und Auswahl der Prädiktoren / Variablen ohne Techniken wie Kreuzvalidierung. Vgl. Bauer et al. (1988).

1.3 Begriffe und Notation, multiples Niveau

Definition 1.7 (Multiples Testproblem)
Seien $(\Omega, \mathcal{A}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und $I \neq \emptyset$ eine beliebige Indexmenge mit $|I| \geq 2$. Seien $\emptyset \neq H_i \subset \Theta$ paarweise verschieden für $i \in I$ und $K_i := \Theta \setminus H_i$. Dann heißt

a) Die Menge $\mathcal{H} := \{H_i, i \in I\}$ ein Hypothesensystem. Ist I endlich, so heißt $\mathcal{H}$ ein endliches Hypothesensystem.

b) H_i wahr $\Longleftrightarrow \vartheta \in H_i$,
H_i falsch $\Longleftrightarrow \vartheta \in K_i$, falls $\vartheta \in \Theta$ der wahre Parameter ist.

c) $I_0 \equiv I_0(\vartheta) := \{i \in I : \vartheta \in H_i\}$ *Indexmenge der wahren (Null-)Hypothesen und* $I_1 \equiv I_1(\vartheta) := I \setminus I_0 = \{i \in I : \vartheta \in K_i\}$ *Indexmenge der falschen (Null-) Hypothesen.*

d) *das Tupel* $(\Omega, \mathcal{A}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta}, \mathcal{H})$ *ein multiples Testproblem. Für* $|I| < \infty$ *heißt* $(\Omega, \mathcal{A}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta}, \mathcal{H})$ *ein endliches (finites) multiples Testproblem.*

Anmerkung: Im Folgenden werden fast ausschließlich endliche multiple Testprobleme betrachtet. Dies vermeidet Messbarkeitsprobleme.

Definition 1.8 (Multipler Test)
Gegeben sei ein multiples Testproblem $(\Omega, \mathcal{A}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta}, \mathcal{H})$. $\varphi = (\varphi_i : i \in I)$ *heißt (nicht-randomisierter) multiper Test (für* $\mathcal{H}$*), falls*

$$\forall i \in I : \varphi_i : (\Omega, \mathcal{A}) \to (\{0,1\}, 2^{\{0,1\}}) \text{ messbare Abbildung.}$$

Es gilt die Konvention

$$\varphi_i(x) = 1 \iff H_i \text{ wird verworfen, Entscheidung für } K_i,$$

$$\varphi_i(x) = 0 \iff H_i \text{ wird nicht verworfen.}$$

Für $|I| = m \in \mathbb{N}$ *ordnet* φ *also jeder Beobachtung* $x \in \Omega$ *einen* m*-dimensionalen Vektor von Nullen und Einsen zu.*

Bevor wünschenswerte Eigenschaften multipler Tests formuliert werden können, sind strukturierte Hypothesensysteme zu betrachten.

Definition 1.9 (strukturierte Hypothesensysteme)
Sei $\mathcal{H} := \{H_i, i \in I = \{1, \dots, m\}\}$ *ein endliches Hypothesensystem.*

a) *Eine Hypothese* $H_i \in \mathcal{H}$ *heißt Obermenge (Implikation) von* $H_j \in \mathcal{H}$ *(in Zeichen:* $H_i \supseteq H_j$*,* $H_i, H_j \in \mathcal{H}$*), falls aus der Richtigkeit von* H_j *die Richtigkeit von* H_i *folgt.* H_i *heißt echte Obermenge von* H_j *(*$H_i \supset H_j$*), falls* $H_i \supseteq H_j$ *und* $H_i \neq H_j$*.* H_i *heißt direkte Obermenge von* H_j*, falls* $H_i \supset H_j$ *und* $\nexists H_k \in \mathcal{H}, k \neq i, k \neq j$ *mit* $H_i \supset H_k \supset H_j$*.*

b) $H_i \in \mathcal{H}$ *heißt Elementarhypothese, falls sie nicht als Durchschnitt ihrer echten Obermengen darstellbar ist.*

c) $H_i \in \mathcal{H}$ *heißt Schnitthypothese, falls sie Durchschnitt ihrer echten Obermengen ist, d. h., falls* $H_i = \bigcap_{j\in I: H_j \supset H_i} H_j$*.*

d) $H_i \in \mathcal{H}$ *heißt Globalhypothese, falls sie der nicht-leere Durchschnitt aller Elementarhypothesen aus* $\mathcal{H}$ *ist.*

e) $H_i \in \mathcal{H}$ *heißt Minimalhypothese, falls sie in* $\mathcal{H}$ *keine echte Obermenge besitzt.*

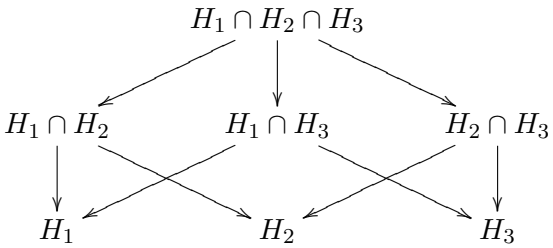
f) $H_i \in \mathcal{H}$ *heißt Maximalhypothese, falls sie keine echte Obermenge irgendeiner Hypothese in* $\mathcal{H}$ *ist.*

g) $\mathcal{H}$ *heißt durchschnittsabgeschlossen, falls* $\forall \emptyset \neq J \subseteq I : H_J := \cap_{j \in J} H_j = \emptyset$ *oder* $H_J \in \mathcal{H}$.

h) $\mathcal{H}$ *heißt hierarchisch, falls mindestens ein* $H_i \in \mathcal{H}$ *eine echte Obermenge in* $\mathcal{H}$ *besitzt.*
$\mathcal{H}$ *heißt streng hierarchisch, falls* $\mathcal{H}$ *nur genau eine Minimalhypothese enthält und falls jede nicht-Minimalhypothese in* $\mathcal{H}$ *genau eine direkte Obermege in* $\mathcal{H}$ *besitzt.*

Schema 1.10
Seien $m = 3$ *und* $\mathcal{H} = \{H_1, H_2, H_3\}$.

a)



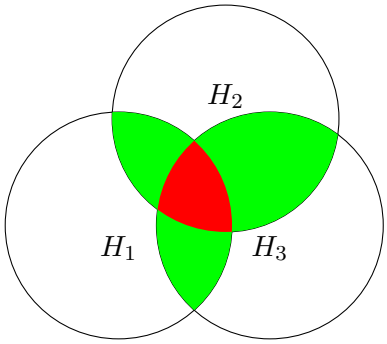
Global- und Maximalhypothese

Schnitthypothesen

Elementar- und Minimalhypothesen

Die Pfeilspitzen zeigen auf die Hypothesen, die die jeweiligen Obermengen (Implikationen) sind.

b)



Die folgenden Definitionen und Lemmata (1.11 bis 1.18) beschreiben (wünschenswerte) Eigenschaften multipler Tests in strukturierten Hypothesensystemen. Der Rest des Abschnitts thematisiert dann mögliche Fehler multipler Tests.

Definition 1.11 (Lehmann, 1957a)
Ein multipler Test $\varphi = (\varphi_i : i \in I)$ *für das multiple Testproblem* $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$ *heißt kompatibel oder allgemein widerspruchsfrei, falls*

$$\forall x \in \Omega : \bigcap_{i \in I : \varphi_i(x) = 0} H_i \cap \bigcap_{i \in I : \varphi_i(x) = 1} K_i \neq \emptyset.$$

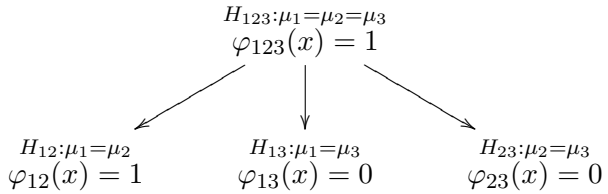
Anmerkung: Allgemeine Widerspruchsfreiheit zu fordern ist sehr restriktiv! Man kann auch „allgemein widerspruchsfreie Entscheidung“ für ein beobachtetes $x^* \in \Omega$ definieren, wobei die Bedingung in Definition 1.11 für x^* erfüllt sein muss. Viele bekannte multiple Tests sind nicht allgemein widerspruchsfrei. Da eine echte Entscheidung nur im Falle $\varphi_i(x) = 1$ getroffen wird, existiert die folgende Abschwächung.

Definition 1.12 (Lehmann, 1957b)
Ein multipler Test $\varphi = (\varphi_i : i \in I)$ für $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$ heißt kompatibel 1. Art oder widerspruchsfrei 1. Art, falls

$$\forall x \in \Omega : \bigcap_{i \in I : \varphi_i(x) = 1} K_i \neq \emptyset.$$

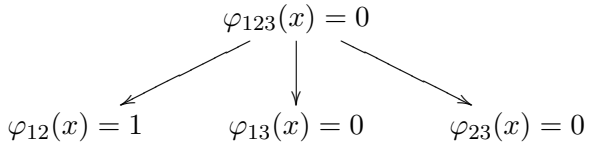
Beispiel 1.13
Betrachte das ANOVA-Modell aus Beispiel 1.4 mit $k = 3$ Gruppen und $\mathcal{H} = \{H_{ij} : \mu_i = \mu_j, 1 \leq i < j \leq 3\} \cup H_{123} : \mu_1 = \mu_2 = \mu_3 \Rightarrow m = 4$. Folgende Situationen sind denkbar:

(a)



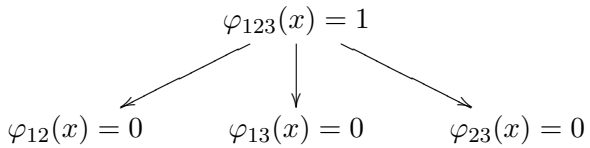
$\Rightarrow \varphi$ ist nicht allgemein widerspruchsfrei, liefert aber eine widerspruchsfreie Entscheidung 1. Art für das beobachtete x.

(b)



Testentscheidung zwar widerspruchsfrei 1. Art, aber inkohärent.

(c)



Testentscheidung zwar widerspruchsfrei 1. Art (für das beobachtete x), aber dissonant.

Aus den Beispielen folgt, dass weder allgemeine Widerspruchsfreiheit (zu restriktiv) noch Widerspruchsfreiheit 1. Art (nicht restriktiv genug) überzeugende Konzepte sind. Daher nun zur formalen Definition von Kohärenz und Konsonanz.

Definition 1.14 (Gabriel, 1969)
Ein multipler Test φ für $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$ heißt kohärent, falls

$$\forall i, j \in I \text{ mit } H_i \subseteq H_j : \{\varphi_j = 1\} \Rightarrow \{\varphi_i = 1\}.$$

Mit H_j lehnt ein kohärenter multipler Test auch jede Hypothese H_i ab, von der H_j Obermenge ist. Anderenfalls heißt φ inkohärent.

Definition 1.15 (Gabriel, 1969)
Ein multipler Test $\varphi = (\varphi_i : i \in I = \{1, \ldots, m\})$ für $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$ heißt konsonant, falls

$$\forall i \in I \text{ mit } \exists j \in I : H_i \subset H_j : \{\varphi_i = 1\} \subseteq \bigcup_{j: H_j \supset H_i} \{\varphi_j = 1\}.$$

Wird H_i von einem konsonanten multiplen Test φ abgelehnt und gibt es echte Obermengen H_j von H_i in $\mathcal{H}$, so wird auch mindestens eine dieser Obermengen von φ abgelehnt. Anderenfalls heißt φ dissonant.

Bemerkung 1.16
Eine Konsonanz von φ verhindert nicht notwendigerweise einen Widerspruch allgemeiner Art; ihre Forderung kann sogar einen solchen erzwingen!

Lemma 1.17
Sei $\varphi = (\varphi_i : i \in I = \{1, \ldots, m\})$ ein allgemein widerspruchsfreier multipler Test für $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$. Dann ist φ auch kohärent.

Beweis: Zur Übung. ■

Lemma 1.18 (Sonnemann, EDV in Medizin und Biologie (1982), bzw. Sonnemann, 2008)
Sei φ ein multipler Test für $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$. Dann gilt

a) Kohärenz von φ ist äquivalent zu

(i)

$$\forall j \in I : \{\varphi_j = 1\} = \bigcap_{i: H_i \subseteq H_j} \{\varphi_i = 1\},$$

(ii)

$$\forall i \in I : \{\varphi_i = 1\} = \bigcup_{j: H_j \supseteq H_i} \{\varphi_j = 1\},$$

(iii)

$$\forall i \in I : \{\varphi_i = 0\} = \bigcap_{j: H_j \supseteq H_i} \{\varphi_j = 0\}.$$

b) φ *ist kohärent und konsonant genau dann, wenn*

$$\forall i \in I : \{\varphi_i = 1\} = \bigcup_{j:H_j \supset H_i} \{\varphi_j = 1\}$$

Beweis: Teil a) zur Übung. Für b) ist zu zeigen: Mit

$$\begin{array}{lll} [1] & \forall i,j \in I \text{ mit } H_i \subseteq H_j : \{\varphi_j = 1\} \Rightarrow \{\varphi_i = 1\} & \text{und} \\ [2] & \forall i \in I \text{ mit } \exists j \in I : H_i \subset H_j : \{\varphi_i = 1\} \subseteq \bigcup_{j:H_j \supset H_i} \{\varphi_j = 1\} & \text{gilt:} \\ [1] \text{ und } [2] & \Longleftrightarrow \forall i \in I : \{\varphi_i = 1\} = \bigcup_{j:H_j \supset H_i} \{\varphi_j = 1\}. & [3] \end{array}$$

„$\Rightarrow$“: Aus [1] folgt nach 1.18a(ii), dass $\bigcup_{j:H_j \supset H_i} \{\varphi_j = 1\} \subseteq \{\varphi_i = 1\}$. Zusammen mit [2] ergibt sich $\{\varphi_i = 1\} \subseteq \bigcup_{j:H_j \supset H_i} \{\varphi_j = 1\} \subseteq \{\varphi_i = 1\}$, also $\{\varphi_i = 1\} = \bigcup_{j:H_j \supset H_i} \{\varphi_j = 1\}$, d. h. [3].
„$\Leftarrow$“: [3] $\Rightarrow$ [2] ist trivial. Ferner folgt aus [3], dass $\bigcup_{j:H_j \supset H_i} \{\varphi_j = 1\} \subseteq \{\varphi_i = 1\}$. Daraus folgt für alle i, j mit $H_i \subset H_j : \{\varphi_j = 1\} \subseteq \{\varphi_i = 1\}$, was wiederum [1] impliziert. ■

Definition 1.19
Sei $\varphi = (\varphi_i : i \in I = \{1, \ldots, m\})$ *ein multipler Test für* $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$.

a) φ *heißt ein multipler Test zum lokalen Niveau* $\alpha \in (0,1)$ *in der Komponente* $\varphi_i, i \in I$*, falls* $\mathbb{P}_\vartheta(\{\varphi_i = 1\}) \leq \alpha$ *für alle* $\vartheta \in H_i$*.*

b) φ *heißt ein multipler Test zum (allgemeinen) lokalen Niveau* $\alpha \in (0,1)$*, falls* $\mathbb{P}_\vartheta(\{\varphi_i = 1\}) \leq \alpha$ *für alle* $\vartheta \in H_i$ *für alle* $i \in I$*.*

Bemerkung 1.20
Ist φ *ein multipler Test zum lokalen Niveau* α *und* $|I_0(\vartheta)| = m_0$*, alle* $\varphi_i(X)$ *stochastisch unabhängig mit* $\mathbb{P}_\vartheta(\{\varphi_i = 1\}) = \alpha$ *für alle* $\vartheta \in H_i$*, für alle* $i \in I_0(\vartheta)$*, so folgt*

$$\forall \vartheta \in \Theta : \mathbb{P}_\vartheta(\bigcup_{i \in I_0(\vartheta)} \{\varphi_i = 1\}) = 1 - (1-\alpha)^{m_0} \underset{(m_0 \to \infty)}{\to} 1,$$

d. h., die Wahrscheinlichkeit für irgendeinen Fehler 1. Art strebt mit wachsendem m_0 *gegen* 1*. Das kann nicht wünschenswert sein!*

Eine erste Möglichkeit zur Kopplung der Komponenten ist das Kozept des globalen Niveaus. Hierbei wird das Hypothesenpaar $H_0 = \bigcap_{i=1}^m H_i$ gegen $\complement H_0 = \bigcup_{i=1}^m K_i$ getestet.

Definition 1.21
Ein multipler Testi $\varphi = (\varphi_i : i \in I = \{1, \ldots, m\})$ *für* $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$ *heißt multipler Test zum globalen Niveau* $\alpha \in (0,1)$*, falls*

$$\mathbb{P}_\vartheta(\bigcup_{i=1}^m \{\varphi_i = 1\}) \leq \alpha \text{ für alle } \vartheta \in H_0 = \bigcap_{i=1}^m H_i.$$

Anmerkung:

(i) Die implizite Aufteilung von Θ in zwei disjunkte Teilmengen entspricht nicht der Idee multipler Tests, da diese nach $\{0,1\}^m$ abbilden (vgl. Definition 1.8). Dies impliziert eine Aufteilung von Θ in 2^m Teilmengen. Eine „richtige“ Entscheidung liefert φ also dann, wenn $\varphi(x) = (\varepsilon_1, \ldots, \varepsilon_m)$, $\varepsilon_j \in \{0,1\}$ für alle $j = 1, \ldots, m$ und $\vartheta \in \bigcap_{j:\varepsilon_j=0} H_j \cap \bigcap_{\ell:\varepsilon_\ell=1} K_\ell$. Ein multipler Fehler 1. Art ist dann gegeben, falls $\exists j \in \{1, \ldots, m\} : \vartheta \in H_j \wedge \varphi_j(x) = 1$.

(ii) Die Globalhypothese und das globale Niveau werden im Kontext des Simes-Tests (siehe Simes, 1986) und der FDR in Kapitel 5 noch einmal wichtig.

Definition 1.22
Sei φ ein multipler Test für $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$ und $\mathcal{H} = \{H_i, i \in I = \{1, \ldots, m\}\}$. Dann ergibt φ einen

a) multiplen Fehler 1. Art, falls $\exists j \in I : \vartheta \in H_j$ und $\varphi_j(x) = 1$.

b) multiplen Fehler 2. Art, falls $\exists j \in I : \vartheta \in K_j$ und $\varphi_j(x) = 0$.

Anmerkung: Bei einem multiplen Testproblem können beiderlei Fehler gleichzeitig auftreten!

Definition 1.23
Ein multipler Test $\varphi = (\varphi_i : i \in I = \{1, \ldots, m\})$ heißt multipler Test zum multiplen Niveau $\alpha \in (0,1)$, falls

$$\forall \vartheta \in \Theta : \mathbb{P}_\vartheta\Big(\bigcup_{i \in I_0(\vartheta)} \{\varphi_i = 1\}\Big) \leq \alpha, \text{ wobei } \bigcup_{i \in \emptyset} \{\varphi_i = 1\} := \emptyset.$$

Anmerkung: Ein multipler Test zum multiplen Niveau α beschränkt („kontrolliert“) die Wahrscheinlichkeit für irgendeinen multiplen Fehler 1. Art durch α, gleichgültig, wie viele und welche der H_i wahr sind.

Bezeichnungen im Englischen: (für $\mathbb{P}_\vartheta(\bigcup_{i \in I_0(\vartheta)} \{\varphi_i = 1\})$)

- (Type I) Family-Wise Error Rate (FWER)
- Experiment-Wise Error Rate

Beispiel 1.24 (Bonferroni-Test, vgl. Bonferroni, 1936)
Sei $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$ ein multiples Testproblem mit $\mathcal{H} = \{H_i, i \in I = \{1, \ldots, m\}\}$. Sei $\varphi = (\varphi_i, i \in I)$ ein multipler Test für $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$ mit der Eigenschaft

$$\mathbb{P}_\vartheta(\{\varphi_i = 1\}) \leq \alpha/m \text{ für alle } \vartheta \in H_i \text{ für alle } i \in I, \tag{1.2}$$

d. h., ein multipler Test zum allgemeinen lokalen Niveau α/m für $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$. Dann ist φ ein multipler Test zum multiplen Niveau α, denn für alle $\vartheta \in \Theta$ gilt

$$\begin{aligned} \mathit{FWER}_\vartheta(\varphi) &= \mathbb{P}_\vartheta\Big(\bigcup_{i\in I_0(\vartheta)} \{\varphi_i = 1\}\Big) \\ &\leq \sum_{i\in I_0(\vartheta)} \mathbb{P}_\vartheta(\{\varphi_i = 1\}) \\ &\underset{(1.2)}{\leq} m_0\alpha/m \leq \alpha. \end{aligned}$$

Die Ungleichung $\mathbb{P}(\bigcup_{i=1}^m A_i) \leq \sum_{i=1}^m \mathbb{P}(A_i)$ heißt auch Bonferroni-Ungleichung und ein gemäß (1.2) konstruierter multipler Test ein Bonferroni-Test.
<u>Nachteil:</u> α/m ist sehr klein für großes $m \Rightarrow$ geringe Güte von Bonferroni-Tests.

Beispiel 1.25 (Šidák-Test, vgl. Šidák, 1967)
Sei $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$ ein multiples Testproblem mit $\mathcal{H} = \{H_i, i \in I = \{1, \ldots, m\}\}$. und $\varphi = (\varphi_i, i \in I)$ ein multipler Test für $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$ mit den folgenden Eigenschaften.

(i) Die Zufallsvariablen $\varphi_i(X), i \in I$, sind stochastisch unabhängig.

(ii) Für alle $i \in I$ gilt für alle $\vartheta \in H_i$

$$\mathbb{P}_\vartheta(\{\varphi_i = 1\}) \leq 1 - (1-\alpha)^{1/m} =: \alpha_m. \tag{1.3}$$

Dann ist φ ein multipler Test zum multiplen Niveau $\alpha \in (0, 1)$, denn für alle $\vartheta \in \Theta$ gilt

$$\begin{aligned} \mathit{FWER}_\vartheta(\varphi) &= \mathbb{P}_\vartheta\Big(\bigcup_{i\in I_0(\vartheta)} \{\varphi_i = 1\}\Big) \\ &= 1 - \mathbb{P}_\vartheta\Big(\bigcap_{i\in I_0(\vartheta)} \{\varphi_i = 0\}\Big) \\ &\underset{(i)}{=} 1 - \prod_{i\in I_0(\vartheta)} \mathbb{P}_\vartheta(\{\varphi_i = 0\}) \\ &\underset{(ii)}{\leq} 1 - \prod_{i\in I_0(\vartheta)} (1-\alpha)^{1/m} \\ &= 1 - (1-\alpha)^{m_0/m} \\ &\leq 1 - (1-\alpha) = \alpha. \end{aligned}$$

<u>Anmerkung:</u>

- Für alle $m \in \mathbb{N}$ gilt $\alpha/m < 1 - (1-\alpha)^{1/m}$.
- Asymptotisch gilt: $m\alpha_m \underset{(m_0\to\infty)}{\to} -\ln(1-\alpha) \underset{\forall\alpha\in(0,1)}{>} \alpha \equiv m\alpha/m$.

- Allerdings gilt auch für die Šidák-Korrektur: $\alpha_m \underset{(m_0 \to \infty)}{\to} 0$.

Lemma 1.26

Äquivalente Bedingungen für die Kontrolle des multiplen Niveaus $\alpha \in (0,1)$ eines multiples Tests $\varphi = (\varphi_i, i \in I = \{1, \dots, m\})$ für $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$ mit $\mathcal{H} = \{H_i, i \in I\}$ sind gegeben durch

(a) $\inf\limits_{\vartheta \in \Theta} \mathbb{P}_\vartheta(\bigcap\limits_{i \in I_0(\vartheta)} \{\varphi_i = 0\}) \geq 1 - \alpha.$

(b) $\forall \emptyset \neq J \subseteq I : \forall \vartheta \in H_J = \bigcap\limits_{j \in J} H_j : \mathbb{P}_\vartheta(\bigcup\limits_{j \in J} \{\varphi_j = 1\}) \leq \alpha.$

Beweis:

zu (a):

$$\begin{aligned}
\inf_{\vartheta \in \Theta} \mathbb{P}_\vartheta(\bigcap_{i \in I_0(\vartheta)} \{\varphi_i = 0\}) &\geq 1 - \alpha \\
\iff 1 - \inf_{\vartheta \in \Theta} \mathbb{P}_\vartheta(\bigcap_{i \in I_0(\vartheta)} \{\varphi_i = 0\}) &\leq \alpha \\
\iff \sup_{\vartheta \in \Theta} \left[1 - \mathbb{P}_\vartheta(\bigcap_{i \in I_0(\vartheta)} \{\varphi_i = 0\})\right] &\leq \alpha \\
\iff \forall \vartheta \in \Theta : 1 - \mathbb{P}_\vartheta(\bigcap_{i \in I_0(\vartheta)} \{\varphi_i = 0\}) &\leq \alpha \\
\iff \forall \vartheta \in \Theta : \mathbb{P}_\vartheta\left(\complement\left(\bigcap_{i \in I_0(\vartheta)} \{\varphi_i = 0\}\right)\right) &\leq \alpha \\
\underset{\text{de Morgan}}{\iff} \mathbb{P}_\vartheta(\bigcup_{i \in I_0(\vartheta)} \{\varphi_i = 1\}) &\leq \alpha.
\end{aligned}$$

zu (b):

$$\begin{aligned}
&[1] \quad \forall \vartheta \in \Theta : \mathbb{P}_\vartheta(\textstyle\bigcup_{i \in I_0(\vartheta)} \{\varphi_i = 1\}) &&\leq \alpha \\
&[2] \quad \forall \emptyset \neq J \subseteq I : \forall \vartheta \in H_J = \textstyle\bigcap_{j \in J} H_j : \mathbb{P}_\vartheta(\bigcup_{j \in J} \{\varphi_j = 1\}) &&\leq \alpha
\end{aligned}$$

zu zeigen: $[1] \iff [2]$.

„$\Leftarrow$“: Trivial, da für alle $\vartheta \in \Theta$ $I_0(\vartheta) \subseteq I$ und $\vartheta \in H_{I_0(\vartheta)}$.

„$\Rightarrow$“: $\vartheta \in H_J \Rightarrow J \subseteq I_0(\vartheta) \Rightarrow \mathbb{P}_\vartheta(\bigcup_{j \in J} \{\varphi_j = 1\}) \leq \mathbb{P}_\vartheta(\bigcup_{i \in I_0(\vartheta)} \{\varphi_i = 1\}) \underset{[1]}{\leq} \alpha.$ ■

Bemerkung 1.27

(i) Ein multipler Test zum multiplen Niveau α ist auch ein multipler Test zum globalen Niveau α (setze under (b) $J = I = \{1, \dots, m\}$).

(ii) Unter Beachtung von Messbarkeitsbedingungen lassen sich die obigen Begriffe auf abzählbare und überabzählbare Hypothesensysteme ausdehnen.

Satz 1.28
Sei $\mathcal{H} = \{H_i, i \in I\}$ ein durchschnittsabgeschlossenes Hypothesensystem und $\varphi = (\varphi_i, i \in I)$ ein kohärenter multipler Test für $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$ zum (allgemeinen) lokalen Niveau α.
Dann ist φ ein multipler Test zum multiplen Niveau α für $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$.

Beweis: Sei $\vartheta \in \Theta$ mit $I_0(\vartheta) \neq \emptyset$. Wegen der Durchschnittsabgeschlossenheit von $\mathcal{H}$ existiert ein $i \in I$ mit $H_i = \bigcap_{j \in I_0(\vartheta)} H_j$ und offensichtlich ist $\vartheta \in H_i$. Also ist für alle $j \in I_0(\vartheta) : H_j \supseteq H_i$. Da φ kohärent ist, folgt nach Lemma 1.18a(ii), dass $\{\varphi_i = 1\} \supseteq \bigcup_{j \in I_0(\vartheta)} \{\varphi_j = 1\}$. Folglich ist

$$\text{FWER}_\vartheta(\varphi) = \mathbb{P}_\vartheta(\bigcup_{j \in I_0(\vartheta)} \{\varphi_j = 1\}) \leq \mathbb{P}_\vartheta(\{\varphi_i = 1\}) \leq \alpha,$$

da φ ein multipler Test zum allgemeinen lokalen Niveau α ist. ■

Satz 1.29 (Closure Principle, siehe Marcus, Peritz, and Gabriel (1976);
Abschlussprinzip, siehe Sonnemann, 2008)
Sei $\mathcal{H} = \{H_i, i \in I\}$ ein durchschnittsabgeschlossenes Hypothesensystem und $\varphi = (\varphi_i, i \in I)$ ein (beliebiger) multipler Test für $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$ zum (allgemeinen) lokalen Niveau α.
Definiere den zu φ gehörigen Abschlusstest (closed multiple test procedure) $\bar{\varphi} = (\bar{\varphi}_i, i \in I)$ durch

$$\forall i \in I : \bar{\varphi}_i(x) = \min_{j : H_j \subseteq H_i} \varphi_j(x).$$

Dann gilt:

(a) $\bar{\varphi}$ ist ein Test zum multiplen Niveau α.

(b) $\forall \emptyset \neq I' \subset I : \bar{\varphi}' := (\bar{\varphi}_i, i \in I')$ ist ein Test zum multiplen Niveau α für $\mathcal{H}' = \{H_i, i \in I'\}$.

(c) $\bar{\varphi}$ und $\bar{\varphi}'$ sind kohärent.

Beweis: Sind $i, j \in I$ mit $H_i \subset H_j$ und $x \in \Omega$, so ist $\bar{\varphi}_i(x) = \min_{k : H_k \subseteq H_i} \varphi_k(x) \geq \min_{k : H_k \subseteq H_j} \varphi_k(x) = \bar{\varphi}_j(x)$, also ist (c) gezeigt.
Da für alle $i \in I$ $\bar{\varphi}_i \leq \varphi_i$ gilt und φ das allgemeine lokale Niveau α kontrolliert, ist auch $\bar{\varphi}$ ein Test zum allgemeinen lokalen Niveau α. Zusammen mit (c) und Satz 1.28 folgt (a).
Nun ist (b) trivial. ■

Bemerkung 1.30

(a) Der zu einen multiplen Test φ (zum lokalen Niveau α) gehörige Abschlusstest $\bar{\varphi}$ lehnt eine Hypothese $H_i \in \mathcal{H}$ genau dann ab, wenn φ sowohl H_i als auch alle Hypothesen $H_j \in \mathcal{H}$, von denen H_i Obermenge ist, ablehnt.

(b) Ist $\mathcal{H}$ nicht durchschnittsabgeschlossen, so kann man hilfsmäßig alle fehlenden Schnitthypothesen zu $\mathcal{H}$ hinzunehmen. Sind ℓ Elementarhypothesen zu testen, so besteht das erzeugte durchschnittsabgeschlossene Hypothesensystem $\bar{\mathcal{H}}$ aus bis zu $2^\ell - 1$ Hypothesen. Wie wir in Kapitel 4 sehen werden, muss aber in aller Regel nicht für alle Hypothesen in $\bar{\mathcal{H}}$ ein Test zum lokalen Niveau α explizit durchgeführt werden.

(c) Satz 1.28 zeigt, dass unter gewissen Voraussetzungen ein multipler Test zum (allgemeinen) lokalen Niveau α auch ein multipler Test zum multiplen Niveau α ist. Die Umkehrung gilt selbstverständlich unbedingt.

(d) Falls $\mathcal{H}$ disjunkt ist, d. h., $\forall i, j \in I, i \neq j : H_i \cap H_j = \emptyset$, und φ ein multipler Test für $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$ zum lokalen Niveau α ist, so ist φ automatisch ein multipler Test zum multiplen Niveau α (da φ kohärent ist und $\mathcal{H}$ durchschnittsabgeschlossen). Es existieren oft viele Möglichkeiten, Θ in disjunkte Teilmengen zu partitionieren ($\longrightarrow$ Partitionsprinzip, Finner and Strassburger, 2002). Ist z. B. speziell $I = \Theta$ und $H_\vartheta = \{\vartheta\}$ für alle $\vartheta \in \Theta$ und $\varphi = (\varphi_\vartheta : \vartheta \in \Theta)$ ein Test zum (allgemeinen) lokalen Niveau α, so ist φ ein Test zum multiplen Niveau α.

Beispiel 1.31 (Zweigruppen t-Test)
Modell: $X = (X_{ij}), i = 1, 2, j = 1, \ldots, n_i$, *alle* $X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$ *stochastisch unabhängig,* σ^2 *unbekannt. Teste* $H_= : \{\mu_1 = \mu_2\}$. *Dazu sei*

$$T(X) = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{\bar{X}_{1.} - \bar{X}_{2.}}{S}, \text{ wobei } S^2 = \frac{1}{\nu} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})^2, \nu = n_1 + n_2 - 2.$$

Der zweiseitige t-Test für $H_=$ lautet damit

$$\varphi_=(x) = \left\{ \begin{array}{ccc} 1 & & > \\ & |t| := |T(x)| & \\ 0 & & \leq \end{array} \begin{array}{c} \\ t_{\nu;\alpha/2} \\ \\ \end{array} \right\}, \alpha \in (0, 1/2).$$

Sollte $H_=$ durch $\varphi_=$ abgelehnt werden, so ist es verlockend, sich im Falle $t < -t_{\nu;\alpha/2}$ (bzw. $t > t_{\nu;\alpha/2}$) für $\mu_1 < \mu_2$ (bzw. $\mu_1 > \mu_2$) zu entscheiden.
Frage: Ist dies zulässig? Es könnte ein sogenannter Fehler III. Art (directional error) auftreten, d. h., Entscheidung für $\mu_1 < \mu_2$ (bzw. $\mu_1 > \mu_2$), obwohl in Wahrheit $\mu_1 > \mu_2$ (bzw. $\mu_1 < \mu_2$) gilt.
Formale mathematische Lösung: Abschlussprinzip!
Wir fügen die beiden Hypothesen $H_\leq : \{\mu_1 \leq \mu_2\}$ und $H_\geq : \{\mu_1 \geq \mu_2\}$ hinzu. Damit ist $H_= = H_\leq \cap H_\geq$. Lokale Niveau α-Tests für $H_\leq$ und $H_\geq$ sind gegeben durch

$$\varphi_\leq(x) = \left\{ \begin{array}{ccc} 1 & > & \\ & t & t_{\nu;\alpha} \\ 0 & \leq & \end{array} \right\} \text{ und } \varphi_\geq(x) = \left\{ \begin{array}{ccc} 1 & < & \\ & t & -t_{\nu;\alpha} \\ 0 & \geq & \end{array} \right\}.$$

Bilde den Abschlusstest $\bar{\varphi} = (\bar{\varphi}_{\leq}, \bar{\varphi}_{=}, \bar{\varphi}_{\geq})$ *mit* $\bar{\varphi}_{=} = \varphi_{=}$, $\bar{\varphi}_{\leq} = \varphi_{=}\varphi_{\leq}$, $\bar{\varphi}_{\geq} = \varphi_{=}\varphi_{\geq}$.

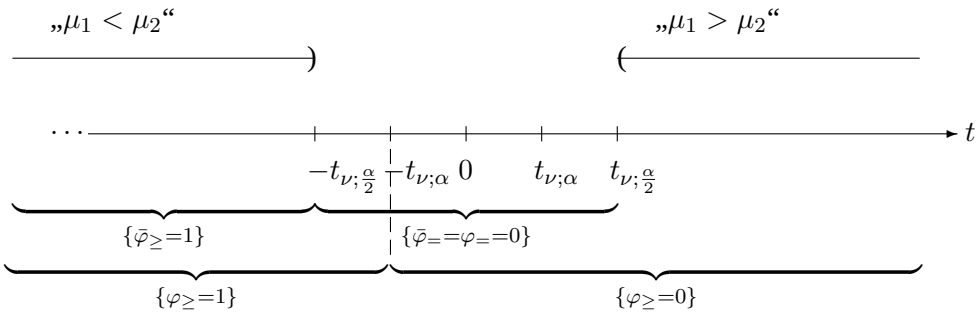


Abbildung 1.1: Abschlusstest für $\{H_{=}, H_{\leq}, H_{\geq}\}$

$\Longrightarrow$ *Typ III-Fehler automatisch mit kontrolliert!*
Man kann sogar noch mehr aus der Realisierung t *inferieren (siehe Übung).*

1.4 Weitere Typ I-Fehlerkonzepte, multiple Gütemaße

Um weitere Typ I-Fehlerkonzepte und Gütemaße kompakt darstellen zu können, führen wir zunächst die folgenden summarischen Zufallsgrößen ein.

Bezeichnungen 1.32
Seien das multiple Testproblem $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$ *und der multiple Test* $\varphi = (\varphi_i, i \in I = \{1, \ldots, m\})$ *für* $\mathcal{H} = (H_i, i \in I)$ *fest vorgegeben. Dann bezeichnen*

a) m *die Anzahl aller zu prüfender Hypothesen,*
$m_0 \equiv m_0(\vartheta)$ *die Anzahl wahrer Nullhypothesen in* $\mathcal{H}$*,*
$m_1 \equiv m_1(\vartheta) = m - m_0(\vartheta)$ *die Anzahl falscher Nullhypothesen in* $\mathcal{H}$*.*

b) $R(\vartheta) = \sum_{i=1}^{m} \varphi_i$ *die (zufällige) Anzahl verworfener Hypothesen.*

c) $V(\vartheta) = \sum_{i \in I_0(\vartheta)} \varphi_i$ *die (zufällige) Anzahl fälschlicherweise verworfener Hypothesen,*
$S(\vartheta) = \sum_{i \in I_1(\vartheta)} \varphi_i$ *die (zufällige) Anzahl korrekterweise verworfener Hypothesen,*
also $V(\vartheta) + S(\vartheta) = R(\vartheta)$*.*

Zusammenfassend:

	Testentscheidung		
Hypothesen	0	1	
wahr	$m_0 - V(\vartheta)$	$V(\vartheta)$	$m_0(\vartheta)$
falsch	$m_1 - S(\vartheta)$	$S(\vartheta)$	$m_1(\vartheta)$
	$m - R(\vartheta)$	$R(\vartheta)$	m

Tabelle 1.1: Summarische Größen einer multiplen Testprozedur

Anmerkung:

(i) Von den Größen in Tableau 1.1 sind in der Praxis nur m und $R(\vartheta)$ beobachtbar.

(ii) FWER$(\varphi) = \sup_{\vartheta \in \Theta} \mathbb{P}_\vartheta(V(\vartheta) > 0)$.

Definition 1.33 (Hochberg and Tamhane, 1987)
Seien das multiple Testproblem $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$ und der multiple Test $\varphi = (\varphi_i, i \in I)$ für $\mathcal{H} = (H_i, i \in I)$ fest vorgegeben.

a) Die Per Family Error Rate (PFER) von φ für gegebenes $\vartheta \in \Theta$ ist definiert als PFER$_\vartheta(\varphi) = \mathbb{E}_\vartheta[V(\vartheta)]$.

b) Die Per Comparison Error Rate (PCER) von φ für gegebenes $\vartheta \in \Theta$ ist definiert als PCER$_\vartheta(\varphi) = \mathbb{E}_\vartheta[V(\vartheta)]/m$.

Anmerkung:

(i) Die Begriffe in Definition 1.33 sind nicht allgemeine Konvention. Insbesondere die Bezeichung „PFER“ ist umstritten, da $\mathbb{E}_\vartheta[V(\vartheta)]$ nicht notwendigerweise in $[0, 1]$ liegt. Alternativ werden „Expected Number of False Rejections (ENFR)“ für die PFER und „Expected Error Rate (EER)“ für die PCER verwendet.

(ii) Mehr zu den Größen in Definition 1.33: vgl. Aufgabe 3 von Übungsblatt 1.

Speziell für extrem hochdimensionale Probleme wie in der Genetik, bei Proteomanalysen oder in der Kosmologie wird wie Kontrolle des multiplen Niveaus α häufig als zu restriktiv empfunden, speziell dann, wenn es sich um explorative Analysen handelt. Daher nun einige „aufgeweichte“ Typ I-Fehlerkriterien.

Definition 1.34 (vgl. Hommel and Hoffmann, 1988)
Voraussetzungen wie unter Definition 1.33.
Die k-FWER von φ ist definiert als die Wahrscheinlichkeit, mehr als $0 \leq k < m$ wahre Nullhypothesen fälschlicherweise zu verwerfen, also

$$k\text{-FWER}_\vartheta(\varphi) = \mathbb{P}_\vartheta(V(\vartheta) > k).$$

Definition 1.35 (vgl. Benjamini and Hochberg (1995), Storey, 2002a)
Voraussetzungen wie unter Definition 1.33.

a) Die Zufallsvariable

$$FDP_\vartheta(\varphi) = \frac{V(\vartheta)}{R(\vartheta) \vee 1}$$

heißt die False Discovery Proportion (FDP) von φ.

b) Die False Discovery Rate (FDR) von φ ist definiert als der erwartete Anteil von Typ I-Fehlern unter allen Verwerfungen von φ, also

$$FDR_\vartheta(\varphi) = \mathbb{E}_\vartheta[FDP_\vartheta(\varphi)].$$

c) Die positive False Discovery Rate (pFDR) von φ ist definiert durch

$$pFDR_\vartheta(\varphi) = \mathbb{E}_\vartheta\left[\frac{V(\vartheta)}{R(\vartheta)}\middle| R(\vartheta) > 0\right].$$

Anmerkung:

(i) Die pFDR ist nur im Bayesianischen Kontext sinnvoll interpretierbar.

(ii) FDP, FDR, pFDR und verwandte Größen spielen in Kapitel 5 die Hauptrollen.

Zur vollständigen Bewertung und zum Vergleich multipler Tests brauchen wir noch geeignete Typ II-Fehlerkonzepte bzw. Gütemaße. Analog zu Definition 1.19 starten wir komponentenweise.

Definition 1.36
Sei $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$ ein endliches multiples Testproblem und $\Phi = \{\varphi : \Omega \to \{0,1\}^m \text{ messbar}\}$ die Menge aller zugehörigen multiplen Tests. Seien $\varphi^{(1)} = (\varphi_i^{(1)}, i \in I)$ und $\varphi^{(2)} = (\varphi_i^{(2)}, i \in I)$ zwei multiple Tests aus Φ zum allgemeinen lokalen Niveau $\alpha \in (0,1)$. Dann heißt

(a) $\varphi^{(1)}$ in der j-ten Komponente nicht schlechter als $\varphi^{(2)}$, falls

$$\forall \vartheta \in K_j : \mathbb{P}_\vartheta(\varphi_j^{(1)} = 1) \geq \mathbb{P}_\vartheta(\varphi_j^{(2)} = 1). \tag{1.4}$$

(b) $\varphi^{(1)}$ komponentenweise nicht schlechter als $\varphi^{(2)}$, falls (1.4) für alle $j \in I$ gilt.

(c) $\varphi^{(1)}$ in der j-ten Komponente besser als $\varphi^{(2)}$, falls $\varphi^{(1)}$ in der j-ten Komponente nicht schlechter als $\varphi^{(2)}$ ist und $\exists \vartheta^* \in K_j : \mathbb{P}_{\vartheta^*}(\varphi_j^{(1)} = 1) > \mathbb{P}_{\vartheta^*}(\varphi_j^{(2)} = 1)$.

(d) $\varphi^{(1)}$ komponentenweise besser als $\varphi^{(2)}$, falls $\varphi^{(1)}$ in jeder Komponente $j \in I$ besser als $\varphi^{(2)}$ ist.

Definition 1.37 (multiple Gütemaße, vgl. Maurer and Mellein, 1988)
Sei $\varphi = (\varphi_i, i \in I$ ein multipler Test für $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$ und $\vartheta \in \Theta$ derart, dass $I_1(\vartheta) \neq \emptyset$ gilt. Dann bezeichnet

(a)

$$\begin{aligned} SG_\varphi : \bigcup_{i\in I} K_i &\to [0,1] \\ \vartheta &\mapsto \mathbb{P}_\vartheta\Big(\bigcap_{i\in I_1(\vartheta)} \{\varphi_i = 1\}\Big) \end{aligned}$$

die simultane Güte („total power") von φ.

(b)

$$\begin{aligned} EG_\varphi : \bigcup_{i\in I} K_i &\to \mathbb{R}_{\geq 0} \\ \vartheta &\mapsto \mathbb{E}_\vartheta[S(\vartheta)] \end{aligned}$$

die erwartete Anzahl korrekterweise von φ verworfener Hypothesen.

(c)

$$\begin{aligned} MEG_\varphi : \bigcup_{i\in I} K_i &\to [0,1] \\ \vartheta &\mapsto \mathbb{E}_\vartheta[S(\vartheta)]/m_1(\vartheta) \end{aligned}$$

den erwarteten Anteil korrekterweise von φ verworfener Hypothesen.

Anmerkung: Falls $\varphi^{(1)} \in \Phi$ komponentenweise besser ist als $\varphi^{(2)} \in \Phi$, so ist die simultane Güte von $\varphi^{(1)}$ nicht notwendigerweise größer als die von $\varphi^{(2)}$.

Kapitel 2

Das Konzept der p-Werte

Viele gängige multiple Testverfahren lassen sich kompakt mit Hilfe sogenannter „p-Werte“ p_i für die einzelnen Hypothesenpaare H_i vs. K_i, $i \in I$, darstellen. Deswegen schieben wir dieses kurze Kapitel ein, das noch einmal einen Aspekt der „gewöhnlichen“, eindimensionalen Testtheorie zum Thema hat.

Definition 2.1 (p-Wert)
Sei $(\Omega, \mathcal{A}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und sei φ ein Test für das Hypothesenpaar $\emptyset \neq H \subset \Theta$ versus $K = \Theta \setminus H$, der auf einer Prüfgröße $T : \Omega \to \mathbb{R}$ basiert. φ sei charakterisiert durch die Angabe von Ablehnbereichen $\Gamma_\alpha \subset \mathbb{R}$ für jedes Signifikanzniveau $\alpha \in (0, 1)$, so dass $\varphi(x) = 1 \iff T(x) \in \Gamma_\alpha$ für $x \in \Omega$ gilt. Dann ist der p-Wert einer Realisierung $x \in \Omega$ bezüglich φ definiert als

$$p_\varphi(x) = \inf_{\{\alpha: T(x)\in\Gamma_\alpha\}} \mathbb{P}^*(T(X) \in \Gamma_\alpha),$$

wobei das Wahrscheinlichkeitsmaß $\mathbb{P}^*$ so gewählt ist, dass

$$\mathbb{P}^*(T(X) \in \Gamma_\alpha) = \sup_{\vartheta\in H} \mathbb{P}_\vartheta(T(X) \in \Gamma_\alpha)$$

gilt, falls H eine zusammengesetzte Nullhypothese ist.

Bemerkung 2.2

(i) Falls H einelementig („einfach“) und $\mathbb{P}_H \equiv \mathbb{P}_{\vartheta_0}$ ein stetiges Wahrscheinlichkeitsmaß ist, so gilt (in aller Regel)

$$p_\varphi(x) = \inf\{\alpha : T(x) \in \Gamma_\alpha\}.$$

(ii) p-Werte werden häufig auch als „beobachtete Signifikanzniveaus“ bezeichnet.

(iii) Als Wahrscheinlichkeitsausdrücke liegen p-Werte stets in $[0, 1]$, unabhängig vom Wertebereich von T. Das erleichtert die Bearbeitung von multiplen Testproblemen mit unterschiedlichen Messskalen für die Einzeltests über p-Werte.

(iv) Sei Ω^{-1} der Urbildraum von X. Die Abbildung $p_\varphi(X) : \Omega^{-1} \to [0,1], \omega \mapsto p_\varphi(X(\omega))$, lässt sich als Zufallsvariable auffassen. Leider wird sie dennoch üblicherweise mit Kleinbuchstabe bezeichnet, um Verwechslungen mit (indizierten) Wahrscheinlichkeitsmaßen vorzubeugen. Es muss also häufig aus dem Kontext heraus interpretiert werden, ob $p_\varphi \equiv p$ einen realisierten Wert aus $[0,1]$ oder eine Zufallsvariable meint.

Definition 2.3

Unter den Voraussetzungen von Definition 2.1 sei die Teststatistik $T(X)$ derart, dass die Monotoniebedingung

$$\forall \vartheta_0 \in H : \forall \vartheta_1 \in K : \forall c \in \mathbb{R} : \mathbb{P}_{\vartheta_0}(T(X) > c) \leq \mathbb{P}_{\vartheta_1}(T(X) > c) \tag{2.1}$$

gilt. Dann heißt φ ein Test vom (verallgemeinerten) Neyman-Pearson Typ, falls für alle $\alpha \in (0,1)$ eine Konstante c_α existiert, so dass

$$\varphi(x) = \begin{cases} 1, & T(x) > c_\alpha, \\ 0, & T(x) \leq c_\alpha. \end{cases}$$

Bemerkung 2.4

(a) Die Monotoniebedingung (2.1) wird häufig so umschrieben, dass „die Teststatistik unter Alternativen zu größeren Werten neigt“.

(b) Die zu einem Test vom Neyman-Pearson (N-P) Typ gehörigen Ablehnbereiche sind gegeben als $\Gamma_\alpha = (c_\alpha, \infty)$.

(c) Die Konstanten c_α werden in der Praxis bestimmt über $c_\alpha = \inf\{c \in \mathbb{R} : \mathbb{P}^*(T(X) > c) \leq \alpha\}$ mit $\mathbb{P}^*$ wie in Definition 2.1 („am Rande der Nullhypothese“). Ist H einelementig und $\mathbb{P}_H$ stetig, so gilt $c_\alpha = F_T^{-1}(1-\alpha)$, wobei F_T die Verteilungsfunktion von $T(X)$ unter H bezeichnet.

(d) Fundamentallemma der Testtheorie von Neyman und Pearson: Unter (leicht verschärftem) (2.1) ist ein Test vom N-P Typ gleichmäßig (über alle $\vartheta_1 \in K$) bester Test für H versus K.

Lemma 2.5

Sei φ ein Test vom N-P Typ und $\mathbb{P}^*$ unabhängig von α. Dann gilt für die Berechnung des p-Wertes einer Realisierung $x \in \Omega$ bezüglich φ, dass

$$p_\varphi(x) = \mathbb{P}^*(T(X) > t^*) \;\; \textit{mit} \;\; t^* := T(x).$$

Beweis: Die Ablehnbereiche $\Gamma_\alpha = (c_\alpha, \infty)$ sind geschachtelt. Demnach wird $\inf\{\alpha : T(x) \in \Gamma_\alpha\}$ offensichtlich in (t^*, ∞) angenommen. Aufgrund der Struktur dieses Ablehnbereiches gilt ferner $\mathbb{P}^*(T(X) \in (t^*, \infty)) = \mathbb{P}^*(T(X) > t^*)$. ■

Anmerkung: Ist H einelementig, $\mathbb{P}_H$ stetig und φ vom N-P Typ, so gilt mit den Bezeichnungen aus Bemerkung 2.4 und Lemma 2.5 für alle $x \in \Omega$, dass $p_\varphi(x) = 1 - F_T(t^*)$.

Satz 2.6 (Testen mit dem p-Wert)
*Sei $\alpha \in (0,1)$ ein **fest vorgegebenes** Signifikanzniveau und $\mathbb{P}^*$ stetig.Dann gilt die Dualität*

$$\varphi(x) = 1 \Longleftrightarrow p_\varphi(x) < \alpha.$$

Nur für Tests vom N-P Typ. Da die Funktion $t \mapsto \mathbb{P}^*(T(X) > t^*)$ monoton fallend in t ist und aufgrund der Konstruktion von c_α (siehe 2.4.c) $\mathbb{P}^*(T(X) > c_\alpha) \leq \alpha$ sowie für alle $\mathbb{R} \ni c < c_\alpha$: $\mathbb{P}^*(T(X) > c) > \alpha$ gelten muss, ist $p_\varphi(x) < \alpha$ gleichbedeutend mit $t^* > c_\alpha$. Das führt bei einem Test vom N-P Typ aber gerade zur Ablehnung von H. ■

Bemerkung 2.7

(i) Der Vorteil von p-Werten für das Testen ist, dass sie unabhängig von einem a priori festgesetzten Signifikanzniveau α ausgerechnet werden können. Dies ist der Grund, warum alle gängigen Statistik-Softwaresysteme statistische Hypothesentests über die Berechnung von p-Werten implementieren. Aus puristischer Sicht birgt das jedoch Probleme, da man mit dieser Art des Testens tricksen kann. Hält man aich nämlich nicht an die gute statistische Praxis, alle Rahmenbedingungen des Experimentes (einschließlich des Signifikanzniveaus!) vor Erhebung der Daten festzulegen, so kann man der Versuchung erliegen, α erst a posteriori (nach Durchführung des Experimentes und Anschauen des resultierenden p-Wertes) zu setzen, um damit zu einer intendierten Schlussfolgerung zu kommen. Deswegen lehnen viele Statistiker die in satz 2.6 gezeigte Art des Testens strikt ab.

(ii) Die Interpretation des p-Wertes ist zu bedenken. Der p-Wert gibt eine Antwort auf die Frage: „Wie wahrscheinlich sind die gemessenen Daten, gegeben dass die Nullhypothese stimmt?“ und nicht auf die Frage „Wie wahrscheinlich ist es, dass die Nullhypothese wahr ist, gegeben die gemessenen Daten?“, obschon letztere Frage manchmal interessanter erscheinen mag und Praktiker ab und an dazu tendieren, den p-Wert dahingehend umzudeuten.

Satz 2.8
Ist unter den Voraussetzungen von Definition 2.1 H einelementig, $\mathbb{P}_H$ stetig und φ ein Test vom N-P Typ, so folgt

$$p_\varphi(X) \underset{H}{\sim} \mathit{UNI}([0,1]).$$

Beweis: Folgt unmittelbar aus dem Prinzip der Quantilstransformation (vgl. Wahrscheinlichkeits-

theorie), denn für $t \in [0,1]$ gilt

$$\begin{aligned}
\mathbb{P}_H(p_\varphi(X) \leq t) &= \mathbb{P}_H(1 - F_T(T(X)) \leq t) \\
&= \mathbb{P}_H(F_T(T(X)) \geq 1 - t) \\
&= \mathbb{P}(U \geq 1 - t) = 1 - \mathbb{P}(U \leq 1 - t) \\
&= 1 - (1 - t) = t,
\end{aligned}$$

wobei U eine UNI($[0,1]$)-verteilte Zufallsvariable bezeichnet. ■

Anmerkung:

(i) Unbedingt gilt, dass $p_\varphi(X)$ stochastisch nicht kleiner als eine auf $[0,1]$ gleichverteilte Zufallsvariable ist, also $\forall \vartheta \in H : \mathbb{P}_\vartheta(p_\varphi(X) \leq t) \leq t$, $t \in [0,1]$.

(ii) Da für viele in der statistischen Praxis angewendeten Modelle die Voraussetzungen von Satz 2.8 erfüllt sind, können im Rahmen eines mutiplen Testproblems heterogene Einzel-Testprobleme H_i versus K_i, $i \in I$, elegant mit Hilfe von p-Werten p_i, $i \in I$, einheitlich statistisch behandelt werden.

Bemerkung 2.9 (Resampling-basierte Schätzung von p-Werten)
Oftmals können selbst unter der Nullhypothese keine exakten Verteilungseigenschaften von $T(X)$ bestimmt werden. Der p-Wert $p_\varphi(x)$ wird in solchen Fällen mit Hilfe von Computersimulationen unter Verwendung sogenannter Resamplingverfahren approximiert. Für Zweistichprobenprobleme sind typischerwesie Permutationstests geeignete Werkzeuge dafür. In Einstichprobenproblemen wird gerne das sogenannte Bootstrapverfahren (B. Efron) verwendet (Münchhausen-Methode der Statistik).
Betrachten wir zur Illustration den parametrischen Einstichproben-Bootstrap für Lageprobleme. Dazu sei $(\mathbb{R}, \mathfrak{B}(\mathbb{R}), (\mathbb{P}_\vartheta)_{\vartheta \in \Theta \subseteq \mathbb{R}})$ ein parametrisches statistisches Modell und $(\mathbb{P}_\vartheta)_{\vartheta \in \Theta}$ derart, dass für je zwei Parameter $\vartheta_1 \neq \vartheta_2$ aus Θ die zugehörigen Verteilungsfunktionen die Gleichung $F_{\vartheta_1}(x - \vartheta_1) = F_{\vartheta_2}(x - \vartheta_2)$ für alle $x \in \mathbb{R}$ erfüllen. Über die genaue Gestalt von F sei nur bekannt, dass $\mathbb{P}_\vartheta$ für alle $\vartheta \in \Theta$ eine endliche Varianz ungleich Null habe. Zu testen sei das Hypothesenpaar $H : \{\vartheta = 0\}$ versus $K : \{\vartheta > 0\}$.
Ohne Beweis und Herleitung ergibt sich dann das folgende Resamplingschema zur Schätzung eines p-Wertes.

Resamplingschema:

(A) Erhebe eine Zufallsstichprobe vom Umfang n mit $X_1, \ldots, X_n$ i.i.d. $\sim X$.

(B) Bilde die Teststatistik $T_n = \bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ der Originalvariablen $X_1, \ldots, X_n$.

(C) Generiere durch Ziehen mit Zurücklegen B „bootstrap Datensätze“ $((X^*_{1,b}, \ldots, X^*_{n,b}))_{b=1,\ldots,B}$. Es wird also für alle $1 \leq i \leq n, 1 \leq b \leq B$ $X^*_{i,b}$ zufällig (gleichverteilt) aus $\{X_1, \ldots, X_n\}$ gezogen.

(D) Berechne für $b = 1, \ldots, B$ die „bootstrap Teststatistiken“ $T_{n,b} = \sqrt{\frac{n}{n-1}} \cdot (\bar{X}^*_{n,b} - \bar{X}_n)$.

(E) Ermittle den approximativen p-Wert für das Testproblem H versus K beruhend auf der Stichprobe $X_1, \ldots, X_n$ als

$$\hat{p}(x) = \frac{|\{T_{n,b} : T_{n,b} > T_n\}| + 1}{B + 1}.$$

Damit der Nenner $B + 1$ in der Approximation glatt wird und um somit eine möglichst rundungsfehlerfreie Schätzung des p-Wertes zu erhalten, sind $B = 999$, $B = 4999$ oder $B = 9999$ beliebte Wahlen für die Simulationsanzahl B.

Heuristische Herleitung des obigen Resamplingschemas:

Nach zentralem Grenzwertsatz (ZGWS) folgt (F_X hinreichend glatt), dass

$$\mathcal{L}\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}\right) \underset{n\to\infty}{\to} \mathcal{N}(0,1),$$

wobei $\mu = \mathbb{E}[X]$ und $\sigma^2 = Var(X)$. Dies gilt nach Slutsky auch noch, falls σ ersetzt wird durch $V_n^{1/2}$ mit $V_n = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$.

Wenden wir den ZGWS nun auf eine bootstrap Stichprobe $X^*_{1,b}, \ldots, X^*_{n,b}$ an, so ergibt sich

$$\mathcal{L}\left(\frac{\sqrt{n}\left(\bar{X}^*_{n,b} - \mathbb{E}[X^*_{1,b}]\right)}{\sqrt{Var(X^*_{1,b})}}\right) \underset{n\to\infty}{\to} \mathcal{N}(0,1).$$

Man zeigt leicht, dass $\mathbb{E}[X^*_{1,b}] = \bar{X}_n$ und $Var(X^*_{1,b}) = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = (n-1)V_n/n$. Also haben unter $\mu = 0$ die Statistiken

$$\frac{\sqrt{n}\bar{X}_n}{V_n^{1/2}} \quad \text{und} \quad \frac{\sqrt{n}\left(\bar{X}^*_{n,b} - \bar{X}_n\right)}{\sqrt{\frac{n-1}{n}} V_n^{1/2}}$$

die gleiche Limesverteilung. „Kürzen“ von $\sqrt{n}$ und $V_n^{1/2}$ rechtfertigt nun den Vergleich von $T_n = \bar{X}_n$ und $T_{n,b} = \sqrt{\frac{n}{n-1}} \cdot (\bar{X}^*_{n,b} - \bar{X}_n)$.

Bemerkung 2.10

Beruht ein multipler Test $\varphi = (\varphi_i, i \in I)$ auf p-Werten p_i, $i \in I$ für die einzelnen Hypothesenpaare H_i versus K_i, $i \in I$, so ist φ ein multipler Test zum allgemeinen lokalen Niveau α, falls $\varphi_i(x) = 1 \Longleftrightarrow p_i(x) < \alpha$ für alle $i \in I$ (folgt direkt aus Satz 2.6).

Kapitel 3

Simultan verwerfende multiple Testprozeduren

3.1 Allgemeine Theorie und der erweiterte Korrespondenzsatz

Simultane Testprozeduren werden in einem Schritt durchgeführt, d.h. die Überprüfung einer Hypothese hängt nicht vom Testergebnis bereits überprüfter Hypothesen ab $\Rightarrow$ „simultan“ (englisch: single-step test). Für jeden Einzeltest $\varphi_i, i \in I$, wird derselbe kritische Wert zum Vergleich mit dem realisierten Wert einer Teststatistik $T_i, i \in I$, verwendet. Leider ist dies noch keine mathematische Definition eines Simultantests.

Hochberg and Tamhane (1987):
„[Ein Simultantest ist] characterized by a collection of test statistics $Z_i, i \in I$, and a common critical constant ξ such that the procedure rejects H_i if $Z_i > \xi$, $i \in I$.“
Mit $Z_i := \varphi_i(X)$, $i \in I$ und $\xi = 0$ wäre dann aber jeder multiple Test $\varphi = (\varphi_i, i \in I)$ ein Simultantest.
Wir folgen der Theorie von Gabriel (1969). Hierbei wird $\xi = c_\alpha$ so bestimmt, dass der zum Testen der Globalhypthese $H_0 = \bigcap_{i \in I} H_i$ verwendete Test ein vorgegebenes Signifikanzniveau $\alpha \in (0, 1)$ einhält. Unter gewissen Annahmen an die Struktur von $\mathcal{H}$ und die Teststatistiken $\{T_i, i \in I\}$ liefert dies einen Test zum multiplen Niveau α.

Definition 3.1
Gegeben sei ein multiples Testproblem $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H} = \{H_i, i \in I^* := \{0, 1, \ldots, m\}\})$ und ein multipler Test $\varphi = (\varphi_i, i \in I^*)$ für $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$ basierend auf Teststatistiken $T_i, i \in I^*$, die (im Sinne der Monotoniebedingung (2.1) aus Definition 2.3) unter der Alternative K_i zu größeren Werten tendieren als unter H_i. Dann heißt

1. *$(\mathcal{H}, \mathcal{T})$ mit $\mathcal{T} = \{T_i, i \in I^*\}$ eine Testfamilie.*

2. $\varphi = (\varphi_i, i \in I^*)$ *eine simultan verwerfende Testprozedur, falls*

$$\varphi_i = \begin{cases} 1, & \textit{falls} \quad T_i > c_\alpha, \\ 0, & \textit{falls} \quad T_i \leq c_\alpha, \ \textit{wobei} \end{cases}$$

$$\forall \vartheta \in H_0 : \mathbb{P}_\vartheta \left(\{\varphi_0 = 1\} \right) = \mathbb{P}_\vartheta \left(\{T_0 > c_\alpha\} \right) \leq \alpha. \tag{3.1}$$

Anmerkung:

(i) Oft ist man gar nicht daran interessiert, die Globalhypothese H_0 zu testen. Sie wird nur hilfsmäßig zu $\mathcal{H}$ hinzugenommen, um c_α festzulegen.

(ii) Im Folgenden wird es darum gehen, Strukturannahmen herauszuarbeiten, sodass der durch (3.1) bestimmte multiple Test φ ein Test zum multiplen Niveau α ist.

Definition 3.2

(a) Eine Testfamilie $(\mathcal{H}, \mathcal{T})$ heißt monoton, falls für alle $i, j \in I^*$ mit $H_i \subseteq H_j$ und für alle $x \in \Omega$ gilt: $T_i(x) \geq T_j(x)$.

(b) $(\mathcal{H}, \mathcal{T})$ heißt streng monoton, falls für alle Nicht-Elementarhypothesen H_i und für jedes $x \in \Omega$ gilt: $T_i(x) = \max_{j: H_i \subset H_j} T_j(x)$

(c) $(\mathcal{H}, \mathcal{T})$ heißt gemeinsam, falls $\forall\, J \subseteq I \quad \forall \vartheta \in \bigcap_{j \in J} H_j$ die gemeinsame Verteilung von $\{T_j, j \in J\}$ dieselbe ist.

(d) $(\mathcal{H}, \mathcal{T})$ heißt abgeschlossen, falls $\mathcal{H}$ durchschnittsabgeschlossen ist.

Bemerkung 3.3

(i) Die Monotoniebedingung 3.2 (a) kann zu „μ-fast sicher“ abgeschwächt werden, wobei μ ein σ-finites, dominierendes Maß ist.

(ii) Obwohl Monotonie im Sinne von 3.2 (a) sehr restriktiv anmutet, ist sie zum Beispiel erfüllt, falls die $T_i, i \in I^*$, Likelihood-Ratio-Teststatistiken sind (siehe Übung). Damit ist die Theorie der Simultantests auf die große Klasse der Likelihood-Ratio-basierten Testfamilien anwendbar.

Satz 3.4

(a) Eine simultane Testprozedur basierend auf einer Testfamilie $(\mathcal{H}, \mathcal{T})$ ist genau dann kohärent, wenn $(\mathcal{H}, \mathcal{T})$ monoton ist.

(b) Eine simultane Testprozedur basierend auf einer Testfamilie $(\mathcal{H}, \mathcal{T})$ ist genau dann kohärent und konsonant, wenn $(\mathcal{H}, \mathcal{T})$ streng monoton ist.

Beweis:
zu (a): Seien $i, j \in I^*$ mit $H_i \subseteq H_j$ beliebig, aber fest gewählt. Dann gilt

$$\begin{aligned} & \forall\, x \in \Omega : T_i(x) \geq T_j(x) \\ \Leftrightarrow \forall \alpha \in (0,1) : \quad & \{x \in \Omega :\, T_j(x) > c_\alpha\} \subseteq \{x \in \Omega :\, T_i(x) > c_\alpha\} \\ \Leftrightarrow \quad & \{\varphi_j = 1\} \subseteq \{\varphi_i = 1\}. \end{aligned}$$

(Für die mittlere Äquivalenz muss streng genommen gelten, dass der Wertebereich der T_i durch $\{c_\alpha :\, \alpha \in (0,1)\}$ überdeckt wird.)

zu (b): Nach Lemma 1.18 (b) ist Kohärenz und Konsonanz (gemeinsam) äquivalent zu:
$\forall i \in I$, für die H_i echte Obermengen in $\mathcal{H}$ besitzt, gilt: $\{\varphi_i = 1\} = \bigcup_{j:H_j \supset H_i} \{\varphi_j = 1\}$.
Dies wiederum ist (bei Simultantests) äquivalent dazu, dass für alle $\alpha \in (0,1)$ gilt:

$$\begin{aligned} & \{x \in \Omega :\, T_i(x) > c_\alpha\} = \bigcup_{j:H_j \supset H_i} \{x \in \Omega :\, T_j(x) > c_\alpha\} \\ \Leftrightarrow \quad & T_i(x) = \max_{j:H_j \supset H_i} T_j(x) \qquad (*) \\ \Leftrightarrow \quad & (\mathcal{H}, \mathcal{T}) \text{ ist streng monoton.} \end{aligned}$$

(Für $(*)$ muss dieselbe Forderung an $\{c_\alpha :\, \alpha \in (0,1)\}$ gestellt werden wie unter dem Beweis zu (a)). ■

Satz 3.5
Sei $(\mathcal{H}, \mathcal{T})$ eine Testfamilie und $\varphi = (\{\varphi_i, i \in I^* = \{0, 1, \ldots, m\})$ eine simultane Testprozedur für das multiple Testproblem $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H} = \{H_i, i \in I^*\})$ basierend auf $(\mathcal{H}, \mathcal{T})$.
Es gelte

(a) $(\mathcal{H}, \mathcal{T})$ ist monoton

(b) $(\mathcal{H}, \mathcal{T})$ ist abgeschlossen oder gemeinsam.

Dann ist φ ein multipler Test zum multiplen Niveau α für $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$.

Beweis:
Wegen (a) ist φ nach Satz 3.4 kohärent. Ferner ist φ unter (a) wegen (3.1) ein Test zum allgemeinen lokalen Niveau α. Ist $(\mathcal{H}, \mathcal{T})$ abgeschlossen, so ist φ nach Satz 1.28 ein Test zum multiplen Niveau α.
Ist $(\mathcal{H}, \mathcal{T})$ nicht abgeschlossen, aber gemeinsam, so liefert die Monotonie von $(\mathcal{H}, \mathcal{T})$ zusammen mit Lemma 1.26 (b), dass φ Test zum multiplen Niveau α ist (detaillierter Beweis in Gabriel, 1969). ■

Viele Simultantests werden aus simultanen Konfidenzbereichen konstruiert. Daher vor den konkreten Beispielen noch kurz etwas zur Korrespondenz von Tests und Konfidenzbereichen.

Definition 3.6
Gegeben sei ein statistisches Modell $(\Omega, \mathcal{A}, \mathcal{P} = \{P_\vartheta : \vartheta \in \Theta\})$. *Dann heißt* $\mathcal{C} = (C(x) : x \in \Omega)$ *mit* $C(x) \subseteq \Theta \, \forall \, x \in \Omega$ *eine Familie von Konfidenzbereichen zum Konfidenzniveau* $1 - \alpha$ *für* $\vartheta \in \Theta :\Longleftrightarrow \forall \, \vartheta \in \Theta : \mathbb{P}_\vartheta \left(\{x : C(x) \ni \vartheta\} \right) \geq 1 - \alpha$.

Satz 3.7 (Korrespondenzsatz, siehe z.B. Lehmann and Romano (2005) oder Witting, 1985)

(a) Liegt für jedes $\vartheta \in \Theta$ *ein Test* φ_ϑ *zum Niveau* α *vor und wird* $\varphi = (\varphi_\vartheta, \, \vartheta \in \Theta)$ *gesetzt, so ist* $\mathcal{C}(\varphi)$, *definiert über* $C(x) = \{\vartheta \in \Theta : \varphi_\vartheta(x) = 0\}$, *eine Familie von Konfidenzbereichen zum Konfidenzniveau* $1 - \alpha$.

(b) Ist $\mathcal{C}$ *eine Familie von Konfidenzbereichen zum Konfidenzniveau* $1 - \alpha$ *und definiert man* $\varphi = (\varphi_\vartheta, \, \vartheta \in \Theta)$ *über* $\varphi_\vartheta(x) = 1 - \mathbf{1}_{C(x)}(\vartheta)$, *so ist* φ *ein Test zum allgemeinen lokalen Niveau* α *(und nach Bemerkung 1.30 (d) sogar ein Test zum multiplen Niveau* α*).*

Beweis:
Sowohl in (a) als auch in (b) erhält man $\forall \, \vartheta \in \Theta \quad \forall \, x \in \Omega : \varphi_\vartheta(x) = 0 \Longleftrightarrow \vartheta \in C(x)$. Also ist φ ein Test zum allgemeinen lokalen Niveau α genau dann, wenn

$$\begin{aligned} & \forall \, \vartheta \in \Theta : \quad \mathbb{P}_\theta \left(\{\varphi_\vartheta = 0\} \right) \geq 1 - \alpha \\ \Leftrightarrow \quad & \forall \, \vartheta \in \Theta : \quad \mathbb{P}_\vartheta \left(\{x : C(x) \ni \vartheta\} \right) \geq 1 - \alpha \\ \Leftrightarrow \quad & \mathcal{C} \text{ ist Familie von Konfidenzbereichen zum Konfidenzniveau } 1 - \alpha. \end{aligned}$$

■

Bemerkung 3.8 *(a) Die Dualität* $\varphi_\vartheta(x) = 0 \Leftrightarrow \vartheta \in C(x)$ *lässt sich schön grafisch veranschaulichen, falls* Ω *und* Θ *eindimensional sind.*

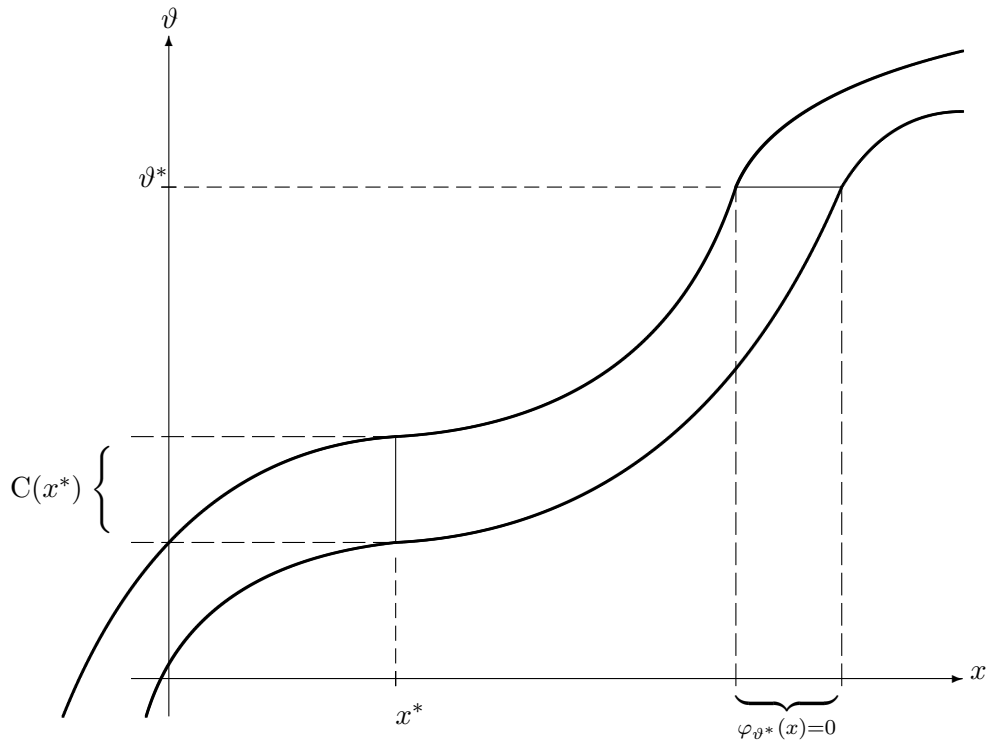


Abbildung 3.1: Dualität $\varphi_\vartheta(x) = 0 \Leftrightarrow \vartheta \in C(x)$

(b) Ein einzelner Test φ zum Niveau α für eine Hypothese H kann interpretiert werden als $(1-\alpha)$-Konfidenzbereich. Setze dazu

$$C(x) = \begin{cases} \Theta\,, & \textit{falls } \varphi(x) = 0, \\ K = \Theta \backslash H\,, & \textit{falls } \varphi(x) = 1. \end{cases}$$

Umgekehrt liefert jeder Konfidenzbereich $C(x)$ einen Test zum Niveau α für eine Hypothese $H \subset \Theta$.
Setze hierzu $\varphi(x) = \mathbf{1}_K(C(x))$, wobei

$$\mathbf{1}_B(A) := \begin{cases} 1\,, & \textit{falls } A \subseteq B, \\ 0\,, & \textit{sonst.} \end{cases}$$

für beliebige Mengen A und B.

Frage: Lässt sich ein Korrespondenzsatz für multiple Tests zum multiplen Niveau α für allgemeine Hypothesensysteme $\mathcal{H}$ beweisen?

Satz 3.9 (Erweiterter Korrespondenzsatz, Finner, 1994)
Sei $\mathcal{H} = \{H_i, i \in I\}$ eine beliebige Hypothesenfamilie und $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$ ein multiples Testproblem.

(a) Ist $\varphi = (\varphi_i, i \in I)$ *ein multipler Test zum multiplen Niveau* α *für* $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$ *und* $C(x) := \bigcap_{j:\varphi_j(x)=1} K_j \quad \forall\, x \in \Omega$, *wobei* $\bigcap_{j\in\emptyset} K_j := \Theta$, *so ist* $\mathcal{C} := \mathcal{C}(\varphi) = (C(x),\, x \in \Omega)$ *eine Familie von Konfidenzbereichen zum Konfidenzniveau* $1-\alpha$ *für* $\vartheta \in \Theta$.

(b) Liegt eine Familie $\mathcal{C} = (C(x),\, x \in \Omega)$ *von Konfidenzbereichen zum Konfidenzniveau* $1-\alpha$ *für* $\vartheta \in \Theta$ *vor und sei* $\varphi(\mathcal{C}) = (\varphi_i,\, i \in I)$ *definiert durch* $\varphi_i(x) = \mathbf{1}_{K_i}(C(x))\ \forall\, x \in \Omega$, $\forall\, i \in I$, *so ist* $\varphi(\mathcal{C})$ *ein Test zum multiplen Niveau* α *für* $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$.

(c) Ist φ *ein Test zum multiplen Niveau* α *für* $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$ *und* $\varphi_\vartheta := \max_{i\in I_0(\vartheta)} \varphi_i\ \forall\, \vartheta \in \Theta$ $\Rightarrow (\varphi_\vartheta,\, \vartheta \in \Theta)$ *ist Test zum multiplen Niveau* α *für* $(\Omega, \mathcal{A}, \mathcal{P}, \Theta)$.

(d) Falls $(\varphi_\vartheta,\, \vartheta \in \Theta)$ *ein multipler Test zum multiplen Niveau* α *für* $(\Omega, \mathcal{A}, \mathcal{P}, \Theta)$ *ist und* $\varphi = (\varphi_i,\, i \in I)$ *definiert wird über* $\varphi_i = \min_{\vartheta\in H_i} \varphi_\vartheta \quad \forall\, i \in I$, *so ist* φ *ein Test zum multiplen Niveau* α *für* $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$.

Beweis: Übung. ■

Anmerkung:
Gilt in Satz 3.9 speziell $\mathcal{H} = \Theta$, so reduziert sich der erweiterte Korrespondenzsatz wieder zum „gewöhnlichen“ Korrespondenzsatz 3.7. Unter (a) gilt dann nämlich zum Beispiel

$$C(x) = \bigcap_{j:\varphi_j(x)=1} K_j = \bigcap_{\vartheta:\varphi_\vartheta=1} \Theta\backslash\{\vartheta\} = \{\vartheta \in \Theta :\ \varphi_\vartheta(x) = 0\}.$$

3.2 Spezielle Methoden im Kontext der Varianzanalyse

Modell 3.10 (Einfache Varianzanalyse, vgl. Beispiel 1.4)
Es seien $X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2), i = 1, \ldots, k\,, j = 1, \ldots, n_i\,, \mu_i \in \mathbb{R} \forall i, \sigma^2 > 0$, X_{ij} *stochastisch unabhängig und* $n_\cdot := \sum_{i=1}^k n_i > k \geq 3$ $(X = (X_{ij})_{i=1,\ldots,k;j=1,\ldots,n_i})$.
Mathematisch formal:
$(\Omega, \mathcal{A}, \mathcal{P}) = (\mathbb{R}^{n_\cdot}, \mathcal{B}^{n_\cdot}, \mathcal{P})$, *wobei*

$$\mathcal{P} = \left\{ \mathcal{N}\left(\begin{pmatrix} \mu_1 \mathbf{1}_{n_1} \\ \vdots \\ \mu_k \mathbf{1}_{n_k} \end{pmatrix}, \sigma^2 I_{n_\cdot} \right) \right\} \tag{3.2}$$

Hierbei bezeichnet $\mathbf{1}_{n_j}, j = 1, \ldots, k$, *einen* n_j*-dimensionalen Vektor aus lauter Einsen und* $I_{n_\cdot}$ *die Einheitsmatrix im* $\mathbb{R}^{n_\cdot \times n_\cdot}$.

In dem Modell (3.2) sollen (zum multiplen Niveau α) alle paarweisen Mittelwertvergleiche durchgeführt werden. Es liegen also $\binom{k}{2} = k(k-1)/2$ Hypothesenpaare der Form

$$H_{ij} :\ \{\mu_i = \mu_j\} \quad \text{vs.} \quad K_{ij} :\ \{\mu_i \neq \mu_j\}, \quad 1 \leq i < j \leq k \tag{3.3}$$

vor. Die Globalhypothese lautet

$$H_0 \quad = \bigcap_{1\le i<j\le k} H_{ij} = \{\mu_1 = \mu_2 = \ldots = \mu_k\}$$

Geeignete Teststatistiken zum Prüfen der H_{ij} sind gegeben durch

$$T_{ij}(X) = \sqrt{\frac{n_i n_j}{n_i + n_j}}\,\frac{|\overline{X}_{i.} - \overline{X}_{j.}|}{S}, \quad 1 \le i < j \le k$$

$$\text{mit} \quad \overline{X}_{i.} = \frac{1}{n_i}\sum_{\ell=1}^{n_i} X_{i\ell} \quad \text{und} \quad S^2 = \frac{1}{n_. - k}\sum_{i=1}^{k}\sum_{\ell=1}^{n_i}(X_{i\ell} - \overline{X}_{i.})^2.$$

Beispiel 3.11 (Bonferroni t-Test)
Obwohl streng genommen nach Definition 3.1 kein Simultantest, lässt sich natürlich für das durch (3.2) *und* (3.3) *gegebene multiple Testprobelm ein Bonferroni-Test $\varphi^{Bonf.}$ durchführen. Dieser ist gegeben als*

$$\varphi_{ij}^{Bonf.}(x) = \begin{cases} 1\,, & \textit{falls} \quad t_{ij} > \quad t_{n.-k;\alpha/(k(k-1))}, \\ 0\,, & \textit{falls} \quad t_{ij} \le \quad t_{n.-k;\alpha/(k(k-1))}. \end{cases}$$

und $\varphi^{Bonf.} = (\varphi_{ij}^{Bonf.},\, 1 \le i < j \le k)$.
Gemäß Beispiel 1.24 ist $\varphi^{Bonf.}$ ein multipler Test zum multiplen Niveau α für $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H} = \{H_{ij}, 1 \le i < j \le k\})$. Wie bereits in Beispiel 1.24 erwähnt, kann $\varphi^{Bonf.}$ sehr konservativ sein und schlechte Güteeigenschaften haben.

Anmerkung:
Leider ist $(\mathcal{H}, \mathcal{T})$ weder abgeschlossen noch gemeinsam (falls $k > 3$ und nicht alle n_i identisch).

Satz 3.12 (Scheffé, 1953)
Unter Modell 3.10 bezeichne

$$\mathcal{L} = \left\{ \sum_{j=1}^{q} h_j a^{(j)} : \quad h_j \in \mathbb{R}\ \forall j = 1, \ldots, q;\ a^{(1)}, \ldots, a^{(q)} \in \mathbb{R}^k \textit{ linear unabhängige Vektoren} \right\}$$

eine Menge von Linearkombinationen ($\rightarrow$ linearer Unterraum des $\mathbb{R}^k$ der Dimension q). Evidenterweise ist $q \le k$.
Dann gilt für alle $\mu \in \mathbb{R}^k$ und $\sigma^2 > 0$:

$$\mathbb{P}_{(\mu,\sigma^2)}\left(c^T\mu \in \left[c^T\hat{\mu} - \sqrt{q\widehat{Var}(c^T\hat{\mu})F_{q,n.-k;\alpha}}\,,\ c^T\hat{\mu} + \sqrt{q\widehat{Var}(c^T\hat{\mu})F_{q,n.-k;\alpha}} \right] \quad \forall c \in \mathcal{L} \right)$$

$$= 1 - \alpha, \ \textit{wobei } \hat{\mu} = (\overline{X}_{1.}, \ldots, \overline{X}_{k.})^T \textit{ und } \widehat{Var}(c^T\hat{\mu}) = s^2\sum_{i=1}^{k}\frac{c_i^2}{n_i}. \tag{3.4}$$

Wir haben also $\forall \mu \in \mathbb{R}^k \quad \forall \sigma^2 > 0$ simultane Konfidenzbereiche ($\forall c \in \mathcal{L}$ gleichzeitig) zum Konfidenzniveau $1-\alpha$ für den linearen Kontrast $c^T\mu$.
Für unser Hypothesensystem $\mathcal{H}$ betrachten wir nun speziell den Unterraum

$$\tilde{\mathcal{L}} = \{(c_1, \ldots, c_k)^T : \sum_{j=1}^{k} c_j = 0\}$$

der Dimension $(k-1)$. Die Paarhypothesen H_{ij} lassen sich dann ausdrücken als
$H_{ij} : \{c_{(ij)}^T \mu = 0\}$ mit

$$\tilde{\mathcal{L}} \ni c_{(ij)} = (\underbrace{0}_{1}, \ldots, 0, \underbrace{+1}_{i}, 0, \ldots, 0, \underbrace{-1}_{j}, 0, \ldots, \underbrace{0}_{k})^T.$$

Dies führt auf den Scheffé-Test.

Beispiel 3.13 (Scheffé-Test)
Unter Modell 3.10 ist der Scheffé-Test für $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$ gegeben durch $\varphi^{\text{Scheffé}} = (\varphi_{ij}^{\text{Scheffé}}$, $1 \le i < j \le k)$ mit

$$\varphi_{ij}^{\text{Scheffé}}(x) = \begin{cases} 1\,, \textit{ falls} & F_{ij}(x) > F_{k-1,n.-k;\alpha} \\ 0\,, \textit{ falls} & F_{ij}(x) \le F_{k-1,n.-k;\alpha} \end{cases}$$

und

$$F_{ij}(x) = \frac{n_i n_j}{n_i + n_j} \frac{(\overline{x}_{i.} - \overline{x}_{j.})^2}{(k-1)s^2}.$$

Bemerkung 3.14

(a) Wegen (3.4) *mit $q = k-1$ und dem erweiterten Korrespondenzsatz 3.9 (b) ist der Scheffé-Test ein multipler Test zum multiplen Niveau α.*

(b) $\varphi^{\text{Scheffé}}$ ist ein Simultantest, denn für die Teststatistik F_0 des üblichen F-Tests für die Globalhypothese gilt $F_0 = \sup_{c \in \tilde{\mathcal{L}}} F_c$ sowie $F_0 \underset{H_0}{\sim} F_{k-1,n.-k}$.

(c) Mit den Setzungen $\mathcal{H}^* := \mathcal{H} \cup \{H_0\}$ sowie $\mathcal{F}^* := (F_0, F_{12}, \ldots, F_{k-1,k})$ ist $(\mathcal{H}^*, \mathcal{F}^*)$ eine monotone Testfamilie. Damit ist $\varphi^* = (\varphi_0, \varphi_{12}^{\text{Scheffé}}, \ldots, \varphi_{k-1,k}^{\text{Scheffé}})$ nach Satz 3.4 (a) kohärent.

(d) Wegen $F_0 > \max_{1 \le i < j \le k} F_{ij}$ ist $(\mathcal{H}^*, \mathcal{F}^*)$ nicht streng monoton und φ^* daher nach Satz 3.4 (b) nicht konsonant. Es kann also vorkommen, dass $F_0 > c_\alpha$, aber $F_{ij} \le c_\alpha \; \forall\, 1 \le i < j \le k$ gilt.

(e) Der Scheffé-Test ist dem Bonferroni t-Test nicht zwingend überlegen. Es gilt mit $\nu = n. - k$

$$\sqrt{(k-1)F_{k-1,\nu;\alpha}} \begin{cases} < t_{\nu, \frac{\alpha}{k(k-1)}} & \Rightarrow \quad \varphi^{\text{Scheffé}} \textit{ besser} \\ > t_{\nu, \frac{\alpha}{k(k-1)}} & \Rightarrow \quad \varphi^{\text{Bonf.}} \textit{ besser} \end{cases}$$

bezüglich aller drei Gütemaße in Definition 1.37. Da obige Bedingung unabhängig von x ist, kann die bessere Methode schon in der Experimentplanung ausgewählt werden.

(f) Umgekehrt betrachtet kann man im Scheffé-Testprinzip ein Beispiel für das allgemeine Vereinigungs-Durchschnitts-Prinzip (union-intersection method, nach Roy, 1953) zum Testen einer mehrdimensionalen Hypothese $H \subseteq \mathbb{R}^k$ sehen (beim Scheffé-Test: $H = H_0$):

(1) Zerlege H in einfachere Hypothesen $H_a, a \in A$, so dass $H = \bigcap_{a\in A} H_a$ (∩-Prinzip)

(2) Konstruiere Tests φ_a zum Niveau $\alpha_{loc.}$ für H_a

(3) Konstruiere aus (2) einen Test φ für H : $\varphi = \max_{a\in A} \varphi_a$ bzw. $\{\varphi = 1\} = \bigcup_{a\in A} \{\varphi_a = 1\}$ (∪-Prinzip)

(4) Wähle $\alpha_{loc.}$ so, dass φ Test zum Niveau α ist, d.h. $\sup_{\vartheta\in H} \mathbb{P}_\vartheta(\bigcup_{a\in A}\{\varphi_a = 1\}) \leq \alpha$.

Als letzte Beispiele für einen Simultantest für $\mathcal{H} = (H_{ij},\ 1 \leq i < j \leq k)$ aus Modell 3.10 sollen nun noch der Tukey-Test und verwandte Prozeduren betrachtet werden.
Für den original Tukey-Test (siehe Tukey, 1953) ist unter Modell 3.10 zusätzlich noch ein balanciertes Design, also

$$n_1 = n_2 = \ldots = n_k =: n \tag{3.5}$$

notwendig. Der Tukey-Test beruht auf folgender Äquivalenz:

$$[\forall\, 1 \leq i < j \leq k : T_{ij}(x) \leq c_\alpha] \Longleftrightarrow \max_{1\leq i<j\leq k} T_{ij}(x) \leq c_\alpha.$$

Gesucht: Verteilung von $\max_{1\leq i<j\leq k} T_{ij}(X)$

Definition 3.15
Seien $Y_1, \ldots, Y_k$ iid. Zufallsvariablen.

(i) $R(Y_1, \ldots, Y_k) = \max_{1\leq i\leq k} Y_i - \min_{1\leq j\leq k} Y_j$ heißt die Spannweiten- (Range-) Statistik von $Y_1, \ldots, Y_k$.

(ii) Gilt zusätzlich $Y_i \sim \mathcal{N}(0, 1)$ und ist S^2 stochastisch unabhängig von $(Y_1, \ldots, Y_k)$ mit $\nu S^2 \sim \chi^2_\nu$ ($\nu \in \mathbb{N}$), so heißt $R(Y_1/S, \ldots, Y_k/S)$ studentisierte Spannweiten- (studentized range) Statistik von $Y_1, \ldots, Y_k$. Die Verteilung von $R(Y_1/S, \ldots, Y_k/S)$ mit Parametern k und ν wird mit $q_{k,\nu}$ bezeichnet und ihre Fraktile $q_{k,\nu;\alpha}$ sind in Software-Systemen verfügbar oder auch z.B. in Pearson and Hartley (1966) oder Hochberg and Tamhane (1987) tabelliert.

Lemma 3.16
Mit $\tilde{T}_{ij}(X) := \sqrt{2}T_{ij}(X),\ 1 \leq i < j \leq k$, gilt unter Modell 3.10 mit (3.5)*, dass*

$$\max_{1\leq i<j\leq k} \tilde{T}_{ij}(X) \underset{H_0}{\sim} q_{k,k(n-1)}.$$

Beweis: zur Übung (einfach!). ■

Beispiel 3.17 (Tukey-Test, Tukey, 1953)
Gegeben sei das multiple Testproblem $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H} = (H_{ij},\, 1 \le i < j \le k))$ *aus Modell 3.10 mit* (3.5)*. Definiere* $\tilde{T}_{ij}(x) = \sqrt{n}|\overline{x}_{i.} - \overline{x}_{j.}|/s$ *für* $x \in \mathbb{R}^{k\times n}$ *und* $i \le i < j \le k$*. Dann ist der Tukey-Test* $\varphi^{Tukey} = (\varphi_{ij}^{Tukey}, 1 \le i < j \le k)$ *für* $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$ *definiert durch*

$$\varphi_{ij}^{Tukey}(x) = \begin{cases} 1\,, & \textit{falls} \quad \tilde{T}_{ij}(x) > \quad q_{k,k(n-1);\alpha}, \\ 0\,, & \textit{falls} \quad \tilde{T}_{ij}(x) \le \quad q_{k,k(n-1);\alpha}. \end{cases}$$

Bemerkung 3.18

(a) φ^{Tukey} *ist eine simultane Testprozedur, da* $c_\alpha = q_{k,k(n-1);\alpha}$ *unter* H_0 *bestimmt wird über die Verteilung von* $\max_{1\le i<j\le k} T_{ij}(X)$*.*

(b) φ^{Tukey} *ist ein Test zum multiplen Niveau* α*, denn* $\forall \emptyset \ne I_0(\vartheta) \subseteq \{(1,2), \ldots, (k-1,k)\}$ *gilt* $\forall\, \vartheta \in \bigcap_{(i,j)\in I_0(\vartheta)} H_{ij}$ *:*

$$\begin{aligned} \mathbb{P}_\vartheta\left(\forall (i,j) \in I_0(\vartheta) :\, \varphi_{ij}^{Tukey}(X) = 0\right) &= \mathbb{P}_\vartheta\left(\max_{(i,j)\in I_0(\vartheta)} \tilde{T}_{ij}(X) \le c_\alpha\right) \\ &\ge \mathbb{P}_{H_0}\left(\max_{1\le i<j\le k} \tilde{T}_{ij}(X) \le c_\alpha\right) \\ &= 1-\alpha. \qquad \text{(nach Lemma 3.16)} \end{aligned}$$

(c) Betrachtet man wieder $\mathcal{H}^* = \mathcal{H} \cup \{H_0\}$ *und*

$$\varphi_0^{Tukey} = \begin{cases} 1\,, & \textit{falls} \quad \max\limits_{1\le i<j\le k} \tilde{T}_{ij}(x) > q_{k,k(n-1);\alpha}, \\ 0\,, & \textit{falls} \quad \max\limits_{1\le i<j\le k} \tilde{T}_{ij}(x) \le q_{k,k(n-1);\alpha}, \end{cases}$$

so ist $\varphi^* = (\varphi_0^{Tukey}, \varphi_{12}^{Tukey}, \ldots, \varphi_{k-1,k}^{Tukey})$ *wegen Satz 3.4 und der strengen Monotonie von* $(\mathcal{H}^*, \mathcal{T}^* = (\tilde{T}_0, \tilde{T}_{12}, \ldots, \tilde{T}_{k-1,k}))$ *kohärent und konsonant.*

(d) Gemäß Satz 3.9 (a) sind simultane Konfidenzintervalle für alle Differenzen $\mu_i - \mu_j$*,* $1 \le i < j \le k$*, gegeben durch:* $\forall\, \mu \in \mathbb{R}^k \quad \forall \sigma^2 > 0$*:*

$$\mathbb{P}_{(\mu,\sigma^2)}\left(\mu_i - \mu_j \in (\overline{X}_{i.} - \overline{X}_{j.}) \pm \frac{1}{\sqrt{n}} S\, q_{k,\nu;\alpha} \quad \forall 1 \le i < j \le k\right) = 1-\alpha.$$

(e) φ^{Tukey} *ist besser als* $\varphi^{Scheffé}$ *und besser als* $\varphi^{Bonf.}$ *bezüglich aller drei Gütemaße in Definition 1.37.*

Beispiel 3.19 (Tukey-Kramer-Test, Tukey (1953), Kramer (1956), Kramer, 1957)
Gilt unter Modell 3.10 die Annahme balancierten Designs gemäß (3.5) *nicht, so hängt die Verteilung von* $\max_{1\le i<j\le k} T_{ij}$ *von* $(n_1,\dots,n_k)$ *ab und ist sehr aufwändig zu berechnen. Man kann jedoch auch in diesem unbalancierten Fall einen multiplen Test zum multiplen Niveau* α *konstruieren, der auf* $q_{k,\nu}$ *mit* $\nu = n. - k$ *beruht. Hayter (1984) hat nämlich eine alte Vermutung von Kramer aus den 1950er Jahren (Kramer (1956), Kramer, 1957) bewiesen. Er konnte zeigen, dass* $\forall\,(n_1,\dots,n_k)$ *gilt:*

$$\mathbb{P}_{(\mu,\sigma^2)}\left(\max_{1\le i<j\le k} T_{ij}(X) \le q_{k,\nu;\alpha}/\sqrt{2}\right) \ge 1-\alpha.$$

Der resultierende Tukey-Kramer Test ist jedoch konservativ, d. h. er schöpft im unbalancierten Fall α *in der Regel nicht aus und hat demzufolge auch schlechte Güteeigenschaften.*
Eine andere Möglichkeit zur Bestimmung eines kritischen Wertes liefert die Šidák-Ungleichung.

Lemma 3.20 (Šidák, 1967)
Sind $Z_1,\dots,Z_r$ *beliebig stochastisch abhängige,* $\mathcal{N}(0,1)$*-verteilte Zufallsvariablen und* $S \sim \sqrt{\nu^{-1}\chi^2_\nu}$ *stochastisch unabhängig von* $Z_1,\dots,Z_r$*, dann gilt für* $c_i \in \mathbb{R}_+$ *beliebig,* $i = 1,\dots,r$

$$\mathbb{P}\left(\frac{|Z_1|}{S} \le c_1,\dots,\frac{|Z_r|}{S} \le c_r\right) \ge \mathbb{P}\left(\frac{|\tilde{Z}_1|}{S} \le c_1,\dots,\frac{|\tilde{Z}_r|}{S} \le c_r\right).$$

Dabei sind $\tilde{Z}_1,\dots,\tilde{Z}_r, S$ *stochastisch unabhängig und* $\tilde{Z}_i \sim \mathcal{N}(0,1)$, $i = 1,\dots,r$.

Definition 3.21
Seien $Y_1,\dots,Y_r, S$ *stochastisch unabhängige Zufallsvariablen mit* $Y_i \sim \mathcal{N}(0,1)$, $i = 1,\dots,r$ *und* $S \sim \sqrt{\nu^{-1}\chi^2_\nu}$*. Dann heißt die Verteilung von*

$$\max_{1\le i\le r} \frac{|Y_i|}{S}$$

die „studentized maximum modulus“-Verteilung mit Parametern r *und* ν*, in Zeichen* $|m|_{r,\nu}$*.*
Auch ihre Fraktile $|m|_{r,\nu;\alpha}$ *sind vertafelt oder per Statistiksoftware verfügbar.*

Beispiel 3.22 (GT2-Methode von Hochberg, Hochberg, 1974)
Unter Modell 3.10 mit $t_{ij} := T_{ij}(x)$*,* $1 \le i < j \le k$ *für eine Realisierung* $x \in \mathbb{R}^{n\cdot}$ *ist der GT2-Test von Hochberg für* $(\Omega,\mathcal{A},\mathcal{P},\mathcal{H} = (H_{ij},\, 1\le i<j\le k))$ *definiert durch* $\varphi^{GT2} = (\varphi^{GT2}_{ij},\, 1\le i<j\le k)$ *mit*

$$\varphi^{GT2}_{ij}(x) = \begin{cases} 1\,, & \textit{falls} \quad t_{ij} > |m|_{\binom{k}{2},n.-k;\alpha}\,, \\ 0\,, & \textit{falls} \quad t_{ij} \le |m|_{\binom{k}{2},n.-k;\alpha}. \end{cases}$$

Satz 3.23

Unter Modell 3.10 ist φ^{GT2} eine simultane Testprozedur zum multiplen Niveau α für $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H} = \{H_{ij},\, 1 \le i < j \le k\})$.

Beweis:

Sei $\emptyset \neq I_0(\vartheta) \subseteq \{(1,2), \ldots, (k-1,k)\}$. Dann gilt $\forall\, \vartheta \in \bigcap_{(i,j)\in I_0(\vartheta)} H_{ij}$ und mit $\nu := n. - k$, dass

$$\begin{aligned}\mathbb{P}_\vartheta\left(\forall\,(i,j) \in I_0(\vartheta):\, \varphi_{ij}^{GT2}(X) = 0\right) &= \mathbb{P}_\vartheta\left(\forall\,(i,j) \in I_0(\vartheta):\, T_{ij}(X) \le |m|_{\binom{k}{2},\nu;\alpha}\right)\\ &\ge \mathbb{P}_{H_0}\left(\forall\, 1 \le i < j \le k:\, \sqrt{\tfrac{n_i n_j}{n_i+n_j}}\tfrac{|\overline{X}_{i.} - \overline{X}_{j.}|/\sigma}{S/\sigma} \le |m|_{\binom{k}{2},\nu;\alpha}\right) =: (*).\end{aligned}$$

Ersetzt man nun die stochastisch abhängigen $\mathcal{N}(0,1)$-verteilten Zufallsvariablen

$$\sqrt{\frac{n_i n_j}{n_i + n_j}}\frac{(\overline{X}_{i.} - \overline{X}_{j.})}{\sigma} =: Z_{ij}$$

durch stochastisch unabhängige $\tilde{Z}_{ij}$'s, $1 \le i < j \le k$, so gilt nach Šidák-Ungleichung, dass

$$\begin{aligned}(*) &\ge \mathbb{P}_{H_0}\left(\forall\, 1 \le i < j \le k:\, \frac{|\tilde{Z}_{ij}|}{S/\sigma} \le |m|_{\binom{k}{2},\nu;\alpha}\right)\\ &= \mathbb{P}_{H_0}\left(\max_{1\le i<j\le k}\frac{|\tilde{Z}_{ij}|}{S/\sigma} \le |m|_{\binom{k}{2},\nu;\alpha}\right) =: (**).\end{aligned}$$

Da die $\tilde{Z}_{ij} \sim \mathcal{N}(0,1)$ iid. und stochastisch unabhängig von S sind und $\nu S^2/\sigma^2 \sim \chi^2_\nu$ ist, ist nach Definition 3.21

$$\max_{1\le i<j\le k}\frac{|\tilde{Z}_{ij}|}{S/\sigma} \sim |m|_{\binom{k}{2},\nu}$$

und daher $(**) = 1 - \alpha$, woraus die Behauptung folgt. ■

Zum Abschluss dieses Kapitels behandeln wir nun noch den Dunnett-Test zum Vergleich mit einer Kontrolle.

Anwendungsbeispiele:

- Agrarwissenschaften: Vergleiche verschiedene neue Düngemittel mit
  a) altem, bewährten Dünger bzgl. des Ertrags.
  b) Ertrag auf ungedüngtem Boden.
  c) verschiedene chemische Dünger mit einem biologischen.
- Medizin:
  a) Vergleiche verschiedene Dosierungen eines Präparates in ihrer Wirkung mit Placebo.

b) Vergleiche neue Präparate in ihrer Wirkung mit Placebo oder einem etablierten Standardpräparat.

$\rightarrow$ „Many-to-one“ Vergleiche.

Formal lässt sich dieses Problem beschreiben, indem unter Modell 3.10 das Hypothesensystem $\mathcal{H}$ eingeschränkt wird. Sei also eine der k Gruppen in Modell 3.10 als Kontrollgruppe ausgezeichnet. O.B.d.A. sei es die für $i = 1$. Wir definieren

$$\tilde{\mathcal{H}} = (\tilde{\mathcal{H}}_i,\ 2 \leq i \leq k) \text{ mit } \tilde{H}_i : \{\mu_i = \mu_1\} \text{ versus } \tilde{K}_i : \{\mu_i \neq \mu_1\}. \tag{M-1}$$

Es sind also hier $(k-1)$ Nullhypothesen zu prüfen. Da $\tilde{\mathcal{H}} \subset \mathcal{H}$ ist, könnte man natürlich die bisher besprochenen Verfahren für Paarvergleiche auch dazu benutzen, einen multiplen Test zum multiplen Niveau α für $(\Omega, \mathcal{A}, \mathcal{P}, \tilde{\mathcal{H}})$ durchzuführen. Dieses Vorgehen wäre jedoch (insbesondere für größere Werte von k) extrem konservativ. Daher diskutieren wir ein spezielles Verfahren, das von Dunnett (Dunnett (1955) bzw. Dunnett, 1964). Wir beschränken uns auf balancierte Designs gemäß (3.5). Das Konstruktionsprinzip folgt dem des Tukey-Tests.

Bezeichnungen 3.24
Unter Modell 3.10 mit (3.5) *wird die Verteilung von*

$$\max_{2 \leq i \leq k} \sqrt{\frac{n}{2}}\, \frac{|\overline{X}_{i.} - \overline{X}_{1.}|}{S}$$

mit $|d|_{k-1,k(n-1)}$ *bezeichnet. Ihre Fraktile sind in den Arbeiten von Dunnett (Dunnett (1955) bzw. Dunnett, 1964) tabelliert.*

Anmerkung:

$$\sqrt{\frac{n}{2}}\, \frac{|\overline{X}_{i.} - \overline{X}_{1.}|}{S} \underset{H_i}{\sim} t_\nu \text{ mit } \nu = k(n-1) \text{ Freiheitsgraden }, i = 2, \ldots, k.$$

Beispiel 3.25 („Many-one t-Test“ von Dunnett)
Gegeben sei das multiple Testproblem $(\Omega, \mathcal{A}, \mathcal{P}, \tilde{\mathcal{H}})$ *mit* $(\Omega, \mathcal{A}, \mathcal{P})$ *wie in Modell 3.10 mit* (3.5) *und* $\tilde{\mathcal{H}}$ *gemäß* (M-1). *Dann heißt* $\varphi^{Dunnett} = (\varphi_2^{Dunnett}, \ldots, \varphi_k^{Dunnett})$ *mit*

$$\varphi_i^{Dunnett} = \begin{cases} 1\,, & \textit{falls} \quad \sqrt{n/2}\, \frac{|\overline{x}_{i.} - \overline{x}_{1.}|}{s} > |d|_{k-1,\nu;\alpha}, \\ 0\,, & \textit{falls} \quad \sqrt{n/2}\, \frac{|\overline{x}_{i.} - \overline{x}_{1.}|}{s} \leq |d|_{k-1,\nu;\alpha}, \end{cases}$$

$\nu := k(n-1)$, $i = 2, \ldots, k$, *Many-one t-Test.*

Bemerkung 3.26

(a) $\varphi^{Dunnett}$ *ist eine simultane Testprozedur zum multiplen Niveau* α *(Beweis analog zu 3.18).*

(b) Wie $\varphi^{Scheffé}$ und φ^{Tukey} kann auch $\varphi^{Dunnett}$ zum Testen allgemeiner linearer Kontraste benutzt werden.

(c) Mit $\varphi_0^{Dunnett} := \max_{2\leq i\leq k} \varphi_i^{Dunnett}$ ist $(\varphi_0^{Dunnett}, \ldots, \varphi_k^{Dunnett})$ kohärent und konsonant, da die zugehörige Testfamilie streng monoton ist.

Kapitel 4

Mehrschrittige multiple Testprozeduren (zum multiplen Niveau)

Im Gegensatz zu den in Kapitel 3 betrachteten Simultantests zeichnen sich mehrschrittige Verfahren (schrittweise verwerfende multiple Tests, englisch: stepwise (rejective) multiple tests) dadurch aus, dass

a) die Entscheidung über eine Hypothese H_i durch die Hintereinanderausführung von Tests gewonnen wird und somit die Überprüfung jeder einzelnen Hypothese abhängig vom Testergebnis bereits überprüfter Hypothesen ist.

b) die verwendeten Einzeltestes im Allgemeinen nicht den selben (unter der Globalhypothese ermittelten) kritischen Wert verwenden.

Schrittweise verwertete multiple Tests sind häufig Verbesserungen von Simultantests (siehe Abschnitt 4.3) und bauen auf diesen auf (vermittels des in Satz 1.29 diskutierten Abschlussprinzips). Um die kritischen Werte für die Einzeltests unabhängig von speziellen Verteilungsannahmen angeben zu können, werden schrittweise verwerfende Testprozeduren gerne mit Hilfe von p-Werten $p_i, i \in I$ für die marginalen Testprobleme H_i vs. $K_i, i \in I$, formuliert.

Erinnerung 4.1 (an Kapitel 2)
Sei $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H} = (H_i, i \in I)$ ein multiples Testproblem.

a) Ist H_i einelementig $(i \in I)$, $\mathbb{P}_{H_i}$ stetig, und liegt φ_i eine Teststatistik $T_i(x)$ zu Grunde, die unter der Alternative K_i zu größeren Werten tendiert als unter H_i, so bezeichnet $p_i(x) = 1 - F_i(t_i^*)$ mit $x \in \Omega$, wobei F_i : die zu $\mathbb{P}_{H_i}$ gehörige Verteilungsfunkion von T_i, $t_i^* = T_i(x)$, den p-Wert von x bezüglich des marginalen Testproblems H_i versus K_i.

b) $p_i(x)$ ist eine monoton fallende Funktion von t_i^*.

c) Unter den Annahmen von Teil a) ist die Zufallsvariable $p_i(X)$ unter H_i gleichverteilt (rechteckverteilt) auf dem Intervall $[0,1]$. Ist H_i zusammengesetzt oder $\mathbb{P}_{H_i}$ nicht stetig, so ist $p_i(x)$ stochastisch nicht kleiner als UNI[0,1]

d) Ein Test φ_i zum Niveau α für H_i versus K_i ist gegeben durch $\varphi_i(x) = 1 \Leftrightarrow p_i(x) < \alpha$.

4.1 Historische Beispiele

Beispiel 4.2 (LSD-Methode von Fisher (1935), Section 24)
Die Least-Significant-Difference (LSD)-Methode ist ein zweischrittiger multipler Test. Er geht auf Fisher (1935, The Design of Experiments) zurück und ist einer der ersten schrittweisen multiplen Tests. Obwohl die LSD-Methode auf jedes experimentelle Design mit (ausschließlich) festen Effekten sowie auch auf multiple lineare Regressionsmodelle angewendet werden kann, stellen wie sie hier anhand von Modell 3.10 (alle Paarvergleiche im einfaktoriellen ANOVA-Design) dar. Betrachte also unter dem statistischen Modell (3.2) *das Hypothesensystem*

$$\mathcal{H} = \{H_{ij}, 1 \leq i < j \leq k\}$$

mit

$$H_{ij} : \{\mu_i = \mu_j\}, 1 \leq i < j \leq k$$

sowie

$$\mathcal{H}^* = \mathcal{H} \cup \{H_0\} \text{ mit } H_0 = \bigcap_{1 \leq i < j \leq k} H_{ij} = \{\mu_1 = \mu_2 = \cdots = \mu_k\}.$$

Dann ist der (α-)Least-Significant-Difference-Test von Fisher φ^{LSD} für $\mathcal{H}^*$ gegeben durch

$$\varphi^{LSD} = (\varphi_0^{LSD}, \varphi_0^{LSD} \cdot \varphi_{12}^{LSD}, \ldots, \varphi_0^{LSD} \cdot \varphi_{k-1,k}^{LSD}) \text{ mit}$$

$$\varphi_0^{LSD}(x) = \left\{ \begin{array}{lll} 1 & & > \\ & \frac{n_\cdot - k}{k-1} \cdot \frac{\sum_{i=1}^{k} n_i (\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i\cdot})^2} & F_{k-1,\nu;\alpha} \\ 0 & & \leq \end{array} \right\}, \text{ wobei } \nu = n_\cdot - k$$

$$\text{und } \varphi_{ij}^{LSD}(x) = \left\{ \begin{array}{lll} 1 & & > \\ & \sqrt{\frac{n_i n_j}{n_i + n_j}} \cdot \frac{|\bar{x}_{i\cdot} - \bar{x}_{j\cdot}|}{s} & t_{\nu;\frac{\alpha}{2}} \\ 0 & & \leq \end{array} \right\}, 1 \leq i < j \leq k.$$

Bemerkung 4.3

a) Prinzipiell (für allgemeinere Kontraste) lässt sich die LSD-Vorgehensweise wie folgt charakterisieren:

1.Stufe: *Teste die Globalhypothese H_0 mit einem geeignetem F-Test zum Niveau α. Falls dieser nicht verwirft, lehne keine einzige Hypthese aus $\mathcal{H}^*$ ab und beende die Analyse. Anderenfalls gehe zur 2. Stufe über.*
2.Stufe: *Teste jede Paarhypothese $H_{ij}, 1 \leq i < j \leq k$ mit einem geeignetem t-Test zum Niveau α. H_{ij} wird also genau dann abgelehnt, wenn die zum Testen von H_0 verwendete F-Statistik F_0 und die t-Statistik t_{ij} (zum Niveau α) signifikant groß sind.*

b) φ^{LSD} hat seinem Namen daher, dass auf der zweiten Stufe der kritische Wert $t_{\nu;\frac{\alpha}{2}}$ verwendet wird. Dies ist der kleinste Wert (least value), den eine Mittelwertsdifferenz (für sich genommen) überschreiten muss, um als signifikant groß zu gelten. Im Gegensatz dazu wird die Tukey-Methode auch als „wholly significant difference test“ oder als „honestly significant difference test“ bezeichnet, da hierbei Signifikanz bezüglich der ganzen Gruppe von Mittelwertdifferenzen gefordert wird.

Satz 4.4 (Eigenschaften von φ^{LSD})
Es gilt:

a) φ^{LSD} ist ein Test zum globalen Niveau α.

b) Für $k = 3$ ist φ^{LSD} ein Test zum multiplen Niveau α.

c) Für $k > 3$ ist φ^{LSD} kein Test zum multiplen Niveau α.

d) φ^{LSD} ist kohärent.

e) φ^{LSD} ist nicht konsonant.

Beweis: a) und **b)**: Übung!

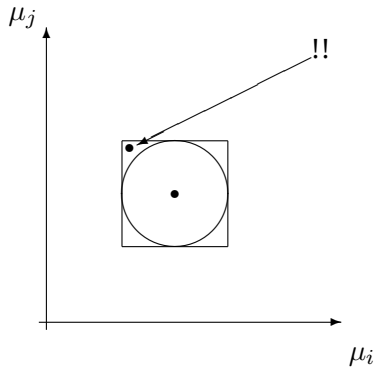
zu c): Sei O.B.d.A. $k = 4$ und bezeichne $\vartheta = (\vec{\mu}, \sigma^2)$ mit $\vec{\mu} = (\mu_1, \ldots, \mu_4)^{\mathrm{T}}$. Wegen der Stetigkeit der F-Verteilung existiert ein ϑ^* mit den folgenden Eigenschaften:
(i) $\mu_1 = \mu_2$, (ii) $\mu_3 = \mu_4$, (iii) $\mathbb{P}_{\vartheta^*}(\varphi_0^{LSD} = 0) < \alpha(1-\alpha)$ für $\alpha \in (0,1)$.

Dann gilt:

$$
\begin{aligned}
&\mathbb{P}_{\vartheta^*}(\varphi_0^{LSD} \cdot \varphi_{12}^{LSD} = 0 \ \wedge \ \varphi_0^{LSD} \cdot \varphi_{34}^{LSD} = 0) \\
&= \mathbb{P}_{\vartheta^*}(\varphi_0^{LSD} = 0 \ \vee \ (\varphi_0^{LSD} = 1 \ \wedge \ \varphi_{12}^{LSD} = 0 \ \wedge \ \varphi_{34}^{LSD} = 0)) \\
&= \mathbb{P}_{\vartheta^*}(\varphi_0^{LSD} = 0) + \mathbb{P}_{\vartheta^*}(\varphi_0^{LSD} = 1 \ \wedge \ \varphi_{12}^{LSD} = 0 \ \wedge \ \varphi_{34}^{LSD} = 0) \\
&\leq \mathbb{P}_{\vartheta^*}(\varphi_0^{LSD} = 0) + \mathbb{P}_{\vartheta^*}(\varphi_{12}^{LSD} = 0 \ \wedge \ \varphi_{34}^{LSD} = 0) \\
&= \mathbb{P}_{\vartheta^*}(\varphi_0^{LSD} = 0) + \mathbb{P}_{\vartheta^*}(\varphi_{12}^{LSD} = 0) \cdot \mathbb{P}_{\vartheta^*}(\varphi_{34}^{LSD} = 0) \\
&\underset{(iii)}{<} \alpha(1-\alpha) + (1-\alpha)^2 = 1 - \alpha.
\end{aligned}
$$

zu d): Trivial, da für alle $1 \leq i < j \leq k$ die Globalhypothese H_0 die einzige Teilmenge von H_{ij} in $\mathcal{H}^*$ ist und nach Konstruktion H_{ij} nur abgelehnt werden kann, falls $\varphi_0^{LSD} = 1$ ist.

zu e): Der Beweis kann über die Dualität von Konfidenzbereichen und statistischen Tests geführt werden. Dann ist zu zeigen, dass es Punkte außerhalb des $(1-\alpha)$-Konfidenzbereiches zu F_0, aber innerhalb der $(1-\alpha)$-Konfidenzbereiche zu den T_{ij} gibt. Dies ist wahr, wie die folgende Skizze suggeriert (Quadrat und Kreis um $(\hat{\mu}_i, \hat{\mu}_j)$):



■

Beispiel 4.5 (Muliple-Range-Test von Newman (1939) und Keuls, 1952)
Wir betrachten wieder Modell 3.10 mit balanciertem Design und dem multiplen Testproblem $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H}^*)$. *Bezeichne* $\bar{x}_{[1]} \leq \bar{x}_{[2]} \leq \cdots \leq \bar{x}_{[k]}$ *die Orderstatistiken der Stichprobenmittel und* $\mu_{[1]}, \ldots, \mu_{[k]}$ *die zugehörigen, entsprechend umsortierten Erwartungswerte. Definiere*

$$c_j := q_{j,k(n-1);\alpha} \cdot \frac{s}{\sqrt{n}} \ \textit{für} \ j = 2, \ldots, k$$

mit $q_{j,k(n-1);\alpha}$ *wie in Definition 3.15.*
Dann lässt sich der Multiple-Range-Test von Newman und Keuls (MRNK-Test) wie folgt beschreiben:

1. Schritt: *Falls* $\bar{x}_{[k]} - \bar{x}_{[1]} \leq c_k$*, so akzeptiere* $H_0 : \{\mu_1 = \mu_2 = \cdots = \mu_k\}$ *und beende die Analyse. Anderenfalls gehe zum 2.Schritt.*

2. Schritt:

(i) Falls $\bar{x}_{[k-1]} - \bar{x}_{[1]} \leq c_{k-1}$*, so schlussfolgere* $\mu_{[1]} = \cdots = \mu_{[k-1]}$ *und untersuche diese Gruppe nicht weiter. Ansonsten spalte diese Gruppe in zwei Gruppen mit jeweils* $(k-2)$ *Mittelwerten auf und gehe zum 3. Schritt.*

(ii) Analog zu (i), aber unter der Verwendung von $\bar{x}_{[k]} - \bar{x}_{[2]}$.

j-ter Schritt $(3 \leq j \leq k-1)$**:** *Vergleiche alle verbliebenen Mittelwertdifferenzen* $\bar{x}_{[j_1]} - \bar{x}_{[j_2]}$*, wobei* $j_1 - j_2 = k - j + 1$ *gilt, mit* c_{k-j+1} *und fälle Entscheidungen wie oben.*

Das Verfahren endet (spätestens) nach Abarbeiten des $(k-1)$*-ten Schrittes bzw. früher, falls in einem Schritt* j *mit* $1 \leq j < k-1$ *keine signifikant große Mittelwertdifferenz mehr aufgetreten ist.*

Bemerkung 4.6

a) Der MRNK-Test ist (hinsichtlich Güte) eine Verbesserung des Tukey-Tests. Die Differenz von zwei Mittelwerten wird nämlich als signifikant groß beurteilt, wenn die Spannweite von je ℓ *Mittelwerten , die diese beiden enthalten, (zum Niveau* α*) signifikant groß ist für jedes* $\ell = 2, \ldots, k$*. Beim Tukey -Test (vgl. Beispiel 3.17) wird nur der Paarvergleich durchgeführt, dieser allerdings zu einem (konservativen) Signifikanzniveau, das dem Vergleich von* k *Mittelwerten entspricht.*

b) Der MRNK-Test ist kohärent nach Konstruktion, aber im Allgemeinen nicht konsonant.

c) Der MRNK-Test ist ein Test zum globalen Niveau α*, aber nicht zum multiplen Niveau* α *für* $k > 3$ *(„Reparatur“ folgt in Abschnitt 4.3).*

Anwendung 4.7 (Durchführung des MRNK-Tests)
Sei $k = n = 5$ *(*$\Rightarrow k(n-1) := \nu = 20$*),* $\alpha = 0,05$*,* $s/\sqrt{n} = 1.2$ *sowie*

$$\bar{x}_{1.} = 20,7;\ \bar{x}_{2.} = 17,0;\ \bar{x}_{3.} = 16,1;\ \bar{x}_{4.} = 21,1 \text{ und } \bar{x}_{5.} = 26,5.$$

Gerundete kritische Werte:

ℓ	2	3	4	5
$q_{\ell,20;0,05}$	3,0	3,6	4,0	4,2
$q_{\ell,\nu;\alpha\cdot\frac{s}{\sqrt{n}}}$	3,6	4,3	4,8	5,1

Bei der Durchführung des MRNK-Tests ist es hilfreich, sich als Gedankenstütze die entsprechend umsortierten Erwartungswerte in einer Zeile aufzuschreiben und Gruppen mit nicht signifikant verschiedenen Spannweiten durch Striche zu kennzeichnen.

Hier:

$$\mu_3 \quad \mu_2 \quad \mu_1 \quad \mu_4 \quad \mu_5$$

1. Schritt $(\ell = k = 5)$

$$\bar{x}_{[5]} - \bar{x}_{[1]} = 26,5 - 16,1 = 10,4 > 5,1 \Rightarrow \textit{2. Schritt}$$

2. Schritt $(\ell = k - 1 = 4)$

$$\bar{x}_{[5]} - \bar{x}_{[2]} = 26,5 - 17,0 = 9,5 > 4,8 \Rightarrow \textit{3. Schritt}$$

$$\bar{x}_{[4]} - \bar{x}_{[1]} = 21,1 - 16,1 = 5,0 > 4,8 \Rightarrow \textit{3. Schritt}$$

3. Schritt $(\ell = k - 2 = 3)$

$$\bar{x}_{[5]} - \bar{x}_{[3]} = 36,5 - 20,7 = 5,8 > 4,3 \Rightarrow \textit{4. Schritt}$$

$$\bar{x}_{[4]} - \bar{x}_{[2]} = 21,1 - 17,0 = 4,1 < 4,3 \Rightarrow \quad \textit{Strich}$$

$$\bar{x}_{[3]} - \bar{x}_{[1]} = 20,7 - 16,1 = 4,6 > 4,3 \Rightarrow \textit{4. Schritt}$$

4. Schritt $(\ell = k - 3 = 2)$

$$\bar{x}_{[5]} - \bar{x}_{[4]} = 26,5 - 21,1 = 5,4 > 3,6$$

$$\bar{x}_{[2]} - \bar{x}_{[1]} = 17,0 - 16,1 = 0,9 < 3,6$$

Schlussfolgerung (Testentscheidung):

$$\mu_5 \neq \mu_j \ \ \forall j = 1, \ldots, 4$$

$$\mu_3 \neq \mu_j \ \ \textit{für} j = 1, 4, 5$$

Beispiel 4.8 (Bonferroni-Holm-Test, siehe Holm (1977) und Holm, 1979)
Sei $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H} = \{H_i, i \in I = \{1, \ldots, m\}\})$ *ein (beliebiges) endliches multiples Testproblem. O.B.d.A. sei*

$$\{i \in I : H_i \textit{ ist Elementarhypothese in } \mathcal{H}\} = \{1, \ldots, k\} \ \textit{ mit } 1 \leq k \leq m.$$

Für jedes marginale Testproblem H_i *versus* $K_i, i \in \{1, \ldots, k\}$*, sei ein* p*-Wert* p_i *verfügbar. Bezeichne* $p_{[1]} \leq p_{[2]} \leq \cdots \leq p_{[k]}$ *die Orderstatistiken dieser* k p*-Werte und* $H_{[1]}, \ldots, H_{[k]}$ *die entsprechend umsortierten Elementarhypothesen von* $\mathcal{H}$*. Setze für* $i \in \{1, \ldots, k\} =: I_k$

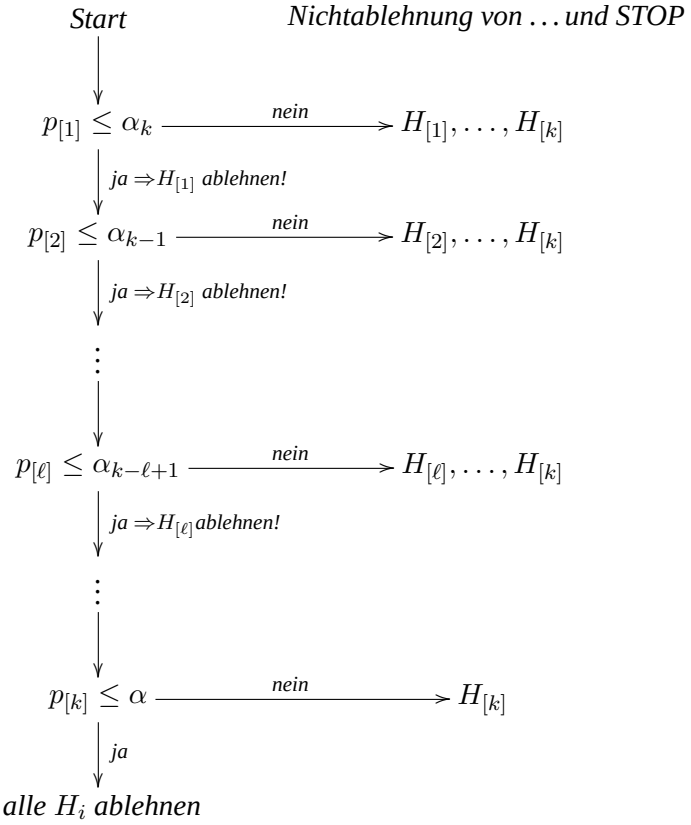
$$\alpha_i = \begin{cases} \alpha/i, & \textit{falls die } p_i(X), i \in I_k \textit{ beliebig stochastisch abhhängig sind } \textbf{\textit{(Fall I)}} \\ 1 - (1-\alpha)^{1/i}, & \textit{falls die } p_i(X), i \in I_k \textit{ stochastisch unabhängig sind } \textbf{\textit{(Fall II)}} \end{cases}$$

Dann lehnt der zugehörige Bonferroni-Holm Test φ^{BH} (genau) die Elementarhypothesen $H_{[1]}, \ldots, H_{[i^*]}$ ab, wobei

$$i^* = \max\{i \in I_k : p_{[j]} \leq \alpha_{k-j+1} \; \forall j = 1, \ldots, i\}.$$

Für die Schnitthypothesen $H_\ell, \ell \in I \setminus I_k$ gilt: H_ℓ wird genau dann abgelehnt, falls mindestens eine der zur Schnittbildung herangezogenen Elementarhypothesen verworfen wird.

Bemerkung 4.9 *a) Die Testvorschrift von φ^{BH} (für die Elementarhypothesen) lässt sich gut anhand eines Ablaufdiagramms illustrieren:*



b) φ^{BH} ist ein sogenannter step-down Test (vgl. Abschnitt 4.2), da mit der „signifikantesten“(größten) Teststatistik zu testen begonnen wird und sich φ^{BH} dann schrittweise, „nach unten“ (zu den kleineren Teststatistiken) vorarbeitet.

c) φ^{BH} ist nach Konstruktion kohärent und konsonant.

d) φ^{BH} ist (hinsichtlich Güte) eine gleichmäßige Verbesserung des entsprechenden Bonferroni- (Fall I) bzw. Šidák-Tests (Fall II).

Satz 4.10

φ^{BH} ist ein multipler Test zum multiplen Niveau α für $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H}= \{H_i, i \in I = \{1, \ldots, m\}\})$.

Beweis (per „ Nachrechnen“): Seien O.B.d.A. $H_1, \ldots, H_r$ wahr für $1 \leq r \leq k$ und $H_{r+1}, \ldots, H_k$ falsch, d.h.,

$$\vartheta \in H_1 \cap \cdots \cap H_r \cap K_{r+1} \cap \cdots \cap K_k \text{ und damit } I_0(\vartheta) \cap I_k = \{1, \ldots, r\}.$$

Eine wahre Hypothese wird von von φ^{BH} höchstens dann abgelehnt, wenn mindestens ein p_i mit $i \in \{1, \ldots, r\}$ kleiner oder gleich α_r ist.
Demnach ist

$$\mathbb{P}_\vartheta(\text{„irgendein Fehler 1. Art“}) \leq \mathbb{P}_\vartheta(\min_{1 \leq i \leq r} p_i \leq \alpha_r) = \mathbb{P}_\vartheta(\bigcup_{1 \leq i \leq r} \{p_i \leq \alpha_r\})$$

$$\begin{cases} \leq \sum_{i=1}^{r} \mathbb{P}_\vartheta(p_i \leq \alpha/r) \leq r\frac{\alpha}{r} = \alpha, & \textbf{(Fall I)} \\ = 1 - \mathbb{P}_\vartheta(\bigcap_{1 \leq i \leq r} \{p_i > \alpha_r\}) = 1 - \prod_{i=1}^{r} (1 - \mathbb{P}_\vartheta(p_i \leq \alpha_r)) \leq 1 - (1-\alpha_r)^r = \alpha. & \textbf{(Fall II)} \end{cases}$$

■

Beweis (per Abschlussprinzip): Sei $\bar{\mathcal{H}} = \{H_J : J \in \bar{I}\}$ die durch $\mathcal{H}$ erzeugte $\cap$-abgeschlossene Hypothesenfamilie mit geeignetem $\bar{I} \subseteq 2^{\{1,\ldots,k\}}$ und mit $H_J = \bigcap_{j \in J} H_j$. Damit gilt

$$H_r \supseteq H_J \Rightarrow r \in J \text{ und } \mathcal{H} \subseteq \bar{\mathcal{H}}.$$

Definiere $\varphi = (\varphi_J, J \in \bar{I})$ als

$$\varphi_J(x) = \begin{cases} 1, & \text{falls } \min_{j \in J} \leq \alpha_{|J|}, \\ 0, & \text{falls } \min_{j \in J} > \alpha_{|J|}. \end{cases}$$

Dann ist φ ein Test zum allgemeinen lokalen Niveau α für $\bar{\mathcal{H}}$. Der zu φ gehörige Abschlusstest $\bar{\varphi}$ ist definiert über

$$\bar{\varphi}_J(x) = \begin{cases} 1, & \text{falls } \forall L \supseteq J, L \in \bar{I} : \varphi_L(x) = 1, \\ 0, & \text{sonst.} \end{cases}$$

Sei nun O.B.d.A. $p_1 \leq \cdots \leq p_k$ und setze $\bar{\varphi}_i := \bar{\varphi}_{\{i\}}$. Dann gilt:

$$\bar{\varphi}_i = 1 \Leftrightarrow \forall J \in \bar{I} \text{ mit } i \in J : \min_{j \in J} p_j \leq \alpha_{|J|}$$

$$\Leftrightarrow \forall r \in \{1, \ldots, k\} : \forall J \in \bar{I} \text{ mit } i \in J \text{ und } |J| = r : \min_{j \in J} p_j \leq \alpha_r.$$

Dies folgt aber aus der Testvorschrift von φ^{BH}, denn falls $p_i \leq \alpha_{k-i+1}$ für alle $i = 1, \ldots, i^*$, so gilt offenbar:
$\forall i = 1, \ldots, i^* : \forall r = 1, \ldots, k : \forall J \in \bar{I}$ mit $i \in J$ und $|J| = r$:

$$\min_{j \in J} p_j = p_{\min\{j \in J\}} \leq \alpha_{k - \min\{j \in J\} + 1} \leq \alpha_r,$$

da $\min\{j \in J\} \leq k - r + 1$ und α_ℓ fallend in ℓ ist. ■

Anmerkung:

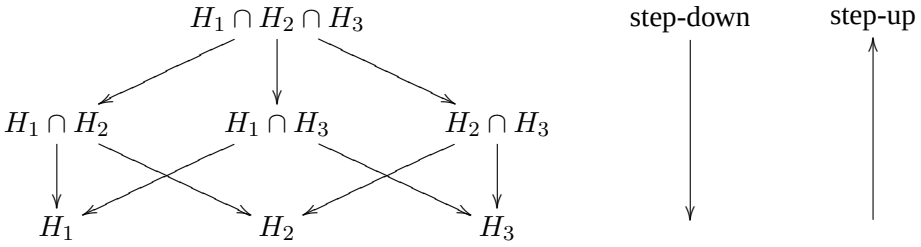
Es kann passieren, dass der Abschlusstest Hypothesen $H_i, i \in I_k$ ablehnt, die der Bonferroni-Holm-Test nicht ablehnt (siehe Übungsaufgabe). Man „erkauft“ sich also die Konsonanz von φ^{BH} mit einer gewissen Konservativität. Ist $\mathcal{H}$ „komplett“, also gilt $|\bar{\mathcal{H}}| = 2^k - 1$, so stimmen die Ergebnisse beider Testprozeduren überein.

4.2 Allgemeine Theorie von step-up und step-down Tests

Mehrschrittige Testprozeduren erfordern die Festlegung einer Reihenfolge, in der die zu testenden Hypothesen abzuarbeiten sind bzw. eine Ordnungsrelation auf $\mathcal{H}$.

- In hierarchischen Hypothesensystemen ist diese in natürlicher Weise vermittels der Obermengenrelation gegeben.
- In nicht-hierarchischen Hypothesensystemen $\mathcal{H}= (H_i, i \in I)$ kann diese Ordnung der H_i vermittels Anordnung der zugehörigen Teststatistiken bzw. p-Werte definiert werden, wie z.B. beim Bonferroni-Holm-Test (vgl Beispiel 4.8).

Zwei prinzipielle Teststrategien sind dann „ step-up“ und „ step-down“:



In größeren (mächtigeren) Hypothesensystemen besteht auch die Möglichkeit, auf einer mittleren Hierachieebene zu beginngen und dann - abhängig von getroffenen Testentscheidungen - in step-up- oder step-down-Sinne weiter zu testen (step-up-down Tests, vgl. auch Kapitel 5). In strukturierten Hypothesensystemen spricht man hierbei auch von der „focus-level“-Methode (Mansmann, Goeman). Andere Unterscheidungskriterien zwischen step-up und step-down können über Stoppregeln formuliert werden (Dunnett and Tamhane, 1992). Solche Stoppregeln liefern gleichzeitig eine Vorschrift, wie mit Hypothesen verfahren wird, die nicht explizit getestet werden.

Bemerkung 4.11

a) Step-down-Prozeduren stoppen, sobald eine Hypothese nicht verworfen wird. Alle bezüglich einer festgelegten Ordnung größeren Hypothesen dürfen ebenfalls nicht abgelehnt werden.

Speziell in hierarchischen Hypothesensystemen: Wird H_j nicht abgelehnt, so stoppt der Test und alle H_i mit $H_i \supseteq H_j$ werden ohne weitere Prüfung nicht verworfen.

b) *Step-up-Prozeduren stoppen, sobald eine Hypothese verworfen wird. Alle bezüglich einer festgelegten Ordnung kleineren Hypothesen gelten ebenfalls als verworfen. Speziell in hierarchischen Hypothesensystemen: Wird zum ersten Mal ein H_j verworfen, so stoppt der Test und führt zur Ablehnung aller H_i mit $H_i \subseteq H_j$ (ohne weitere Prüfung).*

Mathematisch formalisiert:

Definition 4.12
Seien $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H}= \{H_i, i \in I = \{1, \ldots, m\}\})$ ein endliches multiples Testproblem, $(\mathcal{H}, \preceq)$ eine Ordnungsrelation und $\varphi = (\varphi_i, i \in I)$ ein (beliebiger) multipler Test für $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$. Dann heißt

a) $\varphi^{SD}(\varphi) := (\varphi_i^{SD}, i \in I)$ *step-down Prozedur zu φ, falls $\forall x \in \Omega, \forall i \in I$:*

$$\begin{aligned} \varphi_i^{SD}(x) = \min_{j \in I: H_j \preceq H_i} \varphi_j(x) &= \prod_{j \in I: H_j \preceq H_i} \varphi_j(x) \\ &= \begin{cases} 1, & \text{falls } \varphi_j(x) = 1 \; \forall j \in I : H_j \preceq H_i, \\ 0, & \text{sonst.} \end{cases} \end{aligned} \tag{4.1}$$

b) $\varphi^{SU}(\varphi) := (\varphi_i^{SU}, i \in I)$ *step-up Prozedur zu φ, falls $\forall x \in \Omega, \forall i \in I$:*

$$\begin{aligned} \varphi_i^{SU}(x) = \max_{j \in I: H_j \succeq H_i} \varphi_j(x) &= 1 - \prod_{j \in I: H_j \succeq H_i} (1 - \varphi_j(x)) \\ &= \begin{cases} 0, & \text{falls } \varphi_j(x) = 0 \; \forall j \in I : H_j \succeq H_i, \\ 1, & \text{sonst.} \end{cases} \end{aligned} \tag{4.2}$$

Bemerkung 4.13

a) *Falls $\mathcal{H}$ in Definition 4.12 durchschnittsabgeschlossen, die Ordnungsrelation als Teilmengenbeziehung gewählt und φ ein Test zum allgemeinen lokalen Niveau α ist, so ist $\varphi^{SD}(\varphi)$ der zu φ gehörige Abschlusstest (vgl Satz 1.29). Dieser ist kohärent und ein Test zum multiplen Niveau α.*

b) *$\varphi^{SU}(\varphi)$ ist komponentenweise nicht kleiner als $\varphi^{SD}(\varphi)$.*

c) $\varphi^{SU}(\varphi)$ *und* $\varphi^{SD}(\varphi)$ *wie in Definition 4.12 sind jeweils kohärent (Beweis analog zu Satz 1.29c)).*

d) Im Allgemeinen sind weder $\varphi^{SU}(\varphi)$ *noch* $\varphi^{SD}(\varphi)$ *konsonant. Zusätzliche Bedingungen können jedoch Konsonanz erzwingen (siehe Bemerkung 4.14).*

e) Im Falle dreier Elementarhypothesen H_1, H_2, H_3 *und*

$$\mathcal{H} = \{H_1, H_2, H_3, H_1 \cap H_2, H_1 \cap H_3, H_2 \cap H_3, H_1 \cap H_2 \cap H_3\}$$

mit Teilmengenordnung (also im Beispiel aus der Einleitung) lassen sich die möglichen Entscheidungsmuster einer step-down Prozedur wie folgt illustrieren:

0	1	1	1	1	1	1	1	1
000	000	100	110	110	111	111	111	111
000	000	000	000	100	000	100	110	111
$\boxed{1}$	2	3	4	$\boxed{5}$	6	7	$\boxed{8}$	$\boxed{9}$

(O.B.d.A. sei hier $H_1 \cap H_2 \subseteq H_1 \cap H_3 \subseteq H_2 \cap H_3$ *und* $H_1 \subseteq H_2 \subseteq H_3$ *gesetzt.) Konsonanz ist hier nur für die Entscheidungsmuster 1,5,8 und 9 erfüllt.*

Bemerkung 4.14

Geht man von $\varphi^{SD}(\varphi)$ *zu* $\bar{\varphi}^{SD} = (\bar{\varphi}_i^{SD}, i \in I)$ *über mit* $\forall x \in \Omega, \forall i \in I$*:*

$$\bar{\varphi}_i^{SD}(x) = \begin{cases} \varphi_i^{SD}(x), & \text{falls } \nexists H_j \in \mathcal{H} \text{ mit } H_i \subset H_j, \\ \tilde{\varphi}_i^{SD}(x), & \text{falls } \exists H_j \in \mathcal{H} \text{ mit } H_i \subset H_j, \end{cases}$$

wobei

$$\tilde{\varphi}^{SD}(x) = \begin{cases} 1, & \text{falls } \exists H_j \supset H_i : \bar{\varphi}_j^{SD}(x) = 1, \\ 0, & \text{sonst,} \end{cases}$$

so gilt offensichtlich $\forall i \in I$*:*

$$\{\bar{\varphi}_i^{SD} = 1\} = \bigcup_{j: H_j \supset H_i} \{\bar{\varphi}_j^{SD} = 1\},$$

das heißt der „abgeschlossene step-down Test“ $\bar{\varphi}^{SD}$ *ist kohärent und konsonant. In dem Tableau unter Bemerkung 4.13 e) werden damit* $(2) \to (1)$, $(3) \to (1)$, $(4) \to (1)$, $(6) \to (1)$ *und* $(7) \to (5)$*. Man erkennt eindeutig, dass Erzwingung von Konsonanz hier zur Verschlechterung von* φ^{SD} *hinsichtlich Güte führt.*

Hinreichende Bedingungen für die Einhaltung des multiplen Niveaus α von step-up (und damit auch step-down) Tests liefert der folgende Satz.

Satz 4.15 (Alt, 1988)
Sei $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H}= \{H_i, i \in I = \{1, \ldots, m\}\})$ *ein endliches multiples Testproblem mit streng hierarchischem* $\mathcal{H}$*. Sei* $\alpha \in (0,1)$*. Sei* $\varphi = (\varphi_i, i \in I)$ *ein multipler Test für* $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$ *mit*

$$(i)\ \mathbb{P}_\vartheta(\varphi_i = 1) \leq \alpha_i \quad \forall \vartheta \in H_i\ \forall i \in I \ \text{ sowie } \ (ii)\ \alpha_i \geq 0\ \forall i \in I \text{ und } \sum_{i=1}^{m} \alpha_i \leq \alpha.$$

Dann ist der step-up Test $\varphi^{SU} = \varphi^{SU}(\varphi)$ *ein multipler Test zum multiplen Niveau* α *für* $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$.

Beweis: Sei O.B.d.A. $H_1 \subset H_2 \subset \cdots \subset H_m$. Sei $\emptyset \neq I_0(\vartheta) = \{i^*(\vartheta), \ldots, m\}$. Wegen der Kohärenz von $\varphi^{SU}(\varphi)$ gilt nach Lemma 1.18.a(ii) $\forall \vartheta \in \Theta$:

$$\mathbb{P}_\vartheta(\varphi^{SU}_{i^*(\vartheta)} = 1) = \mathbb{P}_\vartheta\Big(\bigcup_{j: H_j \supseteq H_{i^*(\vartheta)}} \{\varphi_j^{SU} = 1\}\Big) = \mathbb{P}_\vartheta\Big(\bigcup_{j \in I_0(\vartheta)} \{\varphi_j^{SU} = 1\}\Big) = \text{FWER}_\vartheta(\varphi^{SU}).$$

Ferner ist nach Konstruktion von $\varphi^{SU}(\varphi)$

$$\begin{aligned}\forall \vartheta \in \Theta : \mathbb{P}_\vartheta(\varphi^{SU}_{i^*(\vartheta)} = 1) &= \mathbb{P}_\vartheta\Big(\bigcup_{j: H_j \supseteq H_{i^*(\vartheta)}} \{\varphi_j = 1\}\Big) = \mathbb{P}_\vartheta\Big(\bigcup_{j=i^*}^{m} \{\varphi_j = 1\}\Big) \\ &\leq \sum_{j=i^*}^{m} \mathbb{P}_\vartheta(\{\varphi_j = 1\}) \underset{(i)}{\leq} \sum_{j=1}^{m} \alpha_j \underset{(ii)}{\leq} \alpha.\end{aligned}$$

■

Beispiel 4.16 (zum Abschlussprinzip)
Gegeben sei Modell 3.10 mit $k = 3$ *und dem durchschnittsabgeschlossenen Hypothesensystem* $\mathcal{H}^* = \{H_{12}, H_{13}, H_{23}, H_0\}$.

a) Testet man H_0 *mit einem* F*-Test zum Niveau* α *und im Fall der Ablehnung von* H_0 *alle* $H_{ij}, 1 \leq i < j \leq 3$ *mit* t*-Tests zum Niveau* α*, so ist dies der LSD-Test von Fisher und gemäß Abschlussprinzip ein kohärenter multipler Test zum multiplen Niveau* α*.*

b) Testet man H_0 *mit dem studentized range-Test zum Niveau* α *und verfährt weiter wie unter a), so ergibt sich der Newman-Keuls-Test. Auch hier folgt multiples Niveau* α *und Kohärenz nach Abschlussprinzip.*

c) Die beiden Verfahren unter a) und b) stehen also in Konkurrenz und es ist bislang nicht vollständig geklärt, in welchen Fällen welches Verfahren besser ist.

Für step-down Tests lassen sich allgemeine Konstruktionsprinzipien unter weniger strikten Bedingungen an das Hypothesensystem angeben.

Satz 4.17 (Abschluss in Gruppen, Shortcut, nach Sonnemann, 2008)
Sei $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H}= \{H_i, i \in I = \{1, \ldots, m\}\})$ *ein endliches multiples Testproblem und* $\mathcal{H}$ *durchschnittsabgeschlossen. O.B.d.A. seien* $H_1, \ldots, H_k$ *mit* $2 \leq k \leq m$ *die Elementarhypothesen von* $\mathcal{H}$*. Außerdem gelte* $\forall i = k+1, \ldots, m$ *(mit* $I_k = \{1, \ldots, k\}$*)*

$$H_i = \bigcap_{j \in I_k : H_j \supset H_i} H_j =: \bigcap_{j \in J(i)} H_j =: H_{J(i)}$$

sowie $m = 2^k - 1$ *(d.h.,* $\mathcal{H}$ *ist komplett). Ferner werde* $J(i) := \{i\}$ *für* $i \in I_k$ *gesetzt. Für alle* $i \in I_k$ *sei ein p-Wert* p_i *verfügbar. Seien lokale Niveau* α*-Tests für alle* $H_{J(i)}, i = 1, \ldots, m$*, gegeben als*

$$\varphi_{J(i)}(x) \quad = \begin{cases} 1, & \min_{j \in J(i)} p_j(x) \leq c_{\alpha, J(i)}, \\ 0, & \textit{sonst,} \end{cases} \tag{4.3}$$

d.h., es seien Konstanten $c_{\alpha,J(i)}, i \in I$ *gegeben mit der Eigenschaft* $\forall \vartheta \in H_{J(i)}$*:* $\mathbb{P}_\vartheta(\varphi_{J(i)} = 1) \leq \alpha \quad \forall i \in I$*. Außerdem seien die* $\varphi_{J(i)}$ *konsonant in dem Sinne*

$$\forall 1 \leq i, j \leq m : H_{J(i)} \subset H_{J(j)} \Rightarrow c_{\alpha,J(i)} \leq c_{\alpha,J(j)}. \tag{4.4}$$

Dann ist ein konsonanter und kohärenter step-down Test zum multiplen Niveau α *für* $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$ *gegeben durch* $\varphi^{SD}(\varphi) = (\varphi^{SD}_{J(i)}, i = 1, \ldots, m)$ *mit*

$$\varphi^{SD}_{J(i)}(x) = \begin{cases} 1, & \textit{falls } x \in \bigcap_{j : J(j) \supseteq J(i)} \{\varphi_{J(j)} = 1\}, \\ 0, & \textit{sonst.} \end{cases} \tag{4.5}$$

Beweis: Einhaltung des multiplen Niveaus und Kohärenz folgen aus dem Abschlussprinzip. Zum Nachweis der Konsonanz sei o.B.d.A. $p_1 \leq p_2 \leq \cdots \leq p_k$. Falls $p_1(x) > c_{\alpha,I_k} \underset{(4.3)}{\Longrightarrow} \varphi_{I_k}(x) = 0 \underset{(4.5)}{\Longrightarrow} \varphi^{SD}_{J(i)}(x) = 0 \; \forall i \in I$, da $J(i) \subseteq I_k \; \forall i \in I$. Gilt indes für alle $1 \leq i \leq j < k$, dass

$$p_i \leq c_{\alpha,\{i,\ldots,k\}} \qquad [*], \text{ sowie } p_{j+1} > c_{\alpha,\{j+1,\ldots,k\}} \qquad [**],$$

so folgt aus $[**] \quad \varphi_{\{j+1,\ldots,k\}} \underset{(4.3)}{=} 0 \underset{(4.5)}{\Longrightarrow} \; \forall \emptyset \neq J(\ell) \subseteq \{j+1, \ldots, k\} : \varphi^{SD}_{J(\ell)} = 0.$

Außerdem folgt aus $[*]$ und (4.3), dass $\varphi_{\{i,\ldots,k\}}(x) = 1$ und dies impliziert wegen (4.4) wiederum $\forall J(\ell) \subseteq \{1, \ldots, k\} : [J(\ell) \cap \{i, \ldots, j\} \neq \emptyset \Rightarrow \varphi_{J(\ell)}(x) = 1]$, denn $J(\ell) \cap \{1, \ldots, j\} \neq \emptyset \Rightarrow b := \min_{J(\ell)} \leq j$ und damit $\min_{a \in J(\ell)} p_a = p_b \leq c_{\alpha,\{b,\ldots,k\}} \leq c_{\alpha,J(\ell)}$. Zusammen mit (4.5) ergibt sich damit $\forall J(\ell) \subseteq \{1, \ldots, k\} : [J(\ell) \cap \{1, \ldots, j\} \neq \emptyset \Rightarrow \; \varphi^{SD}_{J(\ell)}(x) = 1]$, und daraus folgt die Konsonanz von φ^{SD}. ■

Beispiel 4.18

Wählt man φ wie in (4.3) mit $c_{\alpha,J(i)} = \alpha/|J(i)|$, so erfüllt φ nach Bonferroni-Ungleichung offenbar:

$$\forall i \in I : \forall \vartheta \in H_{J(i)} : \ \mathbb{P}_\vartheta(\varphi_{J(i)} = 1) \leq \alpha.$$

Ferner ist die Konsonanzbedingung (4.4) erfüllt. Für $m = 2^k - 1$ ist der durch die Gleichung (4.5) definierte step-down Test $\varphi^{SD}(\varphi)$ gerade der Bonferroni-Holm Test aus Beispiel 4.8, Fall I. Der Bonferroni-Holm Test wird deswegen manchmal auch als „abgeschlossener Bonferroni-Test" bezeichnet.

4.3 Tukey- und Scheffé-basierte step-down Tests zum multiplen Niveau

In Abschnitt 4.1 wurden mit dem Newman-Keuls Test sowie dem LSD-Test von Fisher zwei historische Beispiele von schrittweisen multiplen Testprozeduren für $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$ aus Modell 3.10 behandelt. Diese lassen sich formal als step-down Tests ansehen, auch wenn die Autoren damals noch nicht in diesen Kategorien gedacht haben. Bezüglich Güte stellen der Newman-Keuls Test eine Verbesserung des Tukey-Tests und der LSD-Test von Fisher eine Verbesserung des Scheffé-Tests dar. Beide sind jedoch nur für $k = 3$ Tests zum multiplen Niveau. Vermittels der mächtigen Konstruktionsprinzipien (Abschlussprinzip, Shortcut) aus Abschnitt 4.2 für kohärente step-down Tests zum multiplen Niveau lassen sich die Ideen von Newman und Keuls bzw. Fisher nun verwenden. um verwandte step-down Tests mit den gewünschten Eigenschaften herzuleiten. Heuristisch lässt sich anhand der Struktur von $\bar{\mathcal{H}}$ leicht erkennen, welche zusätzlichen Schritte zur „Reparatur" nötig sind. Sei dazu zunächst $k = 4$. $\bar{\mathcal{H}}$ lässt sich dann schematisch wie folgt darstellen:

$$H_{1234}$$

$$H_{123} \quad H_{124} \quad H_{134} \quad H_{234} \quad H_{12,34} \quad H_{13,24} \quad H_{14,23}$$

$$H_{12} \quad H_{13} \quad H_{14} \quad H_{23} \quad H_{24} \quad H_{34}$$

Insbesondere die mittlere Hierarchieebene (die φ^{LSD} vollkommen unberücksichtigt lässt!) ist genau zu studieren. Hier treten nämlich neben den vier „Homogenitätshypothesen" H_{123}, H_{124}, H_{134} und H_{234} noch die sogenannten „Partitionshypothesen" $H_{12,34}, H_{13,24}$ und $H_{14,23}$ auf, die der MRNK-Test wiederum zu testen „vergisst". Damit ist in beiden Fällen das Abschlussprinzip verletzt.

Anmerkung: $\bar{\mathcal{H}}$ ist nicht komplett. Die Mächtigkeit $|\bar{\mathcal{H}}_k|$ des durch $\mathcal{H}$ im Falle von k Gruppen erzeugten, durchschnittsabgeschlossenen Hypothesensystems lässt sich wie folgt bestimmen.

Bezeichne

$$s_k^{(m)} = \sum_{r=1}^{m} \frac{(-1)^{m-r}}{m!} \binom{m}{r} r^k$$

die Stirlingschen Zahlen 2.Art.

Dann gilt:

$$|\bar{H}_k| = \sum_{m=1}^{k-1} s_k^{(m)} = \sum_{m=1}^{k-1} \frac{1}{m!} \sum_{r=1}^{m} (-1)^{m-r} \binom{m}{r} r^k = \sum_{r=1}^{k-1} \frac{r^k}{r!} \sum_{m=0}^{k-1-r} \frac{(-1)^m}{m!}.$$

Es ergibt sich z.B. $|\bar{\mathcal{H}}_4| = 14$, $|\bar{\mathcal{H}}_5| = 51$, $|\bar{\mathcal{H}}_{10}| = 115.974$, $|\bar{\mathcal{H}}_{13}| = 27.644.436$, $|\bar{\mathcal{H}}_{20}| = 51.724.158.235.371$.

Dies kann für (sehr) kleines k dazu verwendet werden, zu überprüfen, dass man keine Partitionshypothesen unberücksichtigt gelassen hat. Für größere k legt die Formel nache, nach Shortcuts Ausschau zu halten.

Definition 4.19

Gegeben sei $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H} = \{H_{ij}, 1 \leq i < j \leq k\})$ aus Modell 3.10. Sei $I_k = \{1, \ldots, k\}$. Dann heißt

a) $H_J : \{\mu_i = \mu_j \; \forall i, j \in J\}$, wobei $J \subseteq I_k$ mit $|J| \geq 2$ ist, Homogenitätshypothese.

b) $H_{(J_1,\ldots,J_\ell)} = \bigcap_{i=1}^{\ell} H_{J_i}$ mit $|J_i| \geq 2$, $J_{i_1} \cap J_{i_2} = \emptyset$ für $i_1 \neq i_2$ und $\emptyset \neq \sum_{i=1}^{\ell} J_i \subseteq I_k$ Partitionshypothese.

Anmerkung: Homogenitätshypothesen lassen sich als Spezialfälle von Partitionshypothesen darstellen ($\ell = 1$).

Zum Testen von Homogenitätshypothesen können F-Tests oder studentized range-Tests benutzt werden. Zur Durchführung des Abschlusstests für $\bar{\mathcal{H}}_k$ mit $k \geq 4$ müssen demnach nur noch geeignete Tests für allgemeinere Partitionshypothesen angegeben werden.

Satz 4.20 (Finner, 1988)

Gegeben sei das multiple Testproblem $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H} = \{H_{ij}, 1 \leq i < j \leq k\})$ aus Modell 3.10 mit balanciertem Design sowie den Partitionshypothesen $H_{(J_1,\ldots,J_\ell)}$ aus Definiton 4.19. O.B.d.A. sei $\sigma^2 = 1$ und $\mu_i = 0 \; \forall i \in I_k$ unter H_0. Dann ist ein Niveau α-Test für $H_{(J_1,\ldots,J_\ell)}$ gegeben durch

$$\varphi_{(J_1,\ldots,J_\ell)}(x) = \left\{ \begin{array}{lll} 1 & & > \\ & \max\limits_{1 \leq i \leq \ell} \max\limits_{r,s \in J_i} t_{rs} & \\ 0 & & \leq \end{array} \right\} q_{(J_1,\ldots,J_\ell),\nu;\alpha}.$$

Dabei werden die kritischen Werte $q \equiv q_{(J_1,\ldots,J_\ell),\nu;\alpha}$ über die folgende Bestimmungsgleichung ermittelt. Sei dazu abkürzend $\sqrt{n}\bar{X}_{i\cdot} =: Z_i \; \forall i = 1, \ldots, k$. Damit gilt unter H_0, dass die $(Z_i)_{i \in I_k}$ i.i.d. standardnormalverteilt sind. Nun berechnet sich q über

$$1-\alpha \stackrel{!}{=} \mathbb{P}_{\vartheta_0}(\max_{1\le i\le \ell}\max_{r,t\in J_i}\frac{|Z_r-Z_t|}{S}\le q) = \int \mathbb{P}_{\vartheta_0}(\max_{1\le i\le \ell}\max_{r,t\in J_i}|Z_r-Z_t|\le qs)\,d\mathbb{P}^S(s)$$

$$= \int \prod_{i=1}^{\ell}\mathbb{P}_{\vartheta_0}(\max_{r,t\in J_i}|Z_r-Z_t|\le qs)\,d\mathbb{P}^S(s) = \int\prod_{i=1}^{\ell}G_{|J_i|}(qs)\,d\mathbb{P}^S(s),$$

wobei $G_p(\cdot)$ die Verteilungsfunktion der (nicht-studentisierten) Spannweitenstatistik von p standardnormalverteilten Zufallsvariablen bezeichnet. Diese kann durch die Verteilungsfunktion von $\mathcal{N}(0,1)$ dargestellt werden.

Korollar 4.21

Ein Abschlusstest $\bar\varphi = (\bar\varphi_{ij}, 1\le i<j\le k)$ für $(\Omega,\mathcal{A},\mathcal{P},\mathcal{H}=\{H_{ij},1\le i<j\le k\})$ aus Modell 3.10 mit balanciertem Design ist gegeben durch

$$\bar\varphi_{ij}=1 \;\Leftrightarrow\; \varphi_{(J_1,\dots,J_\ell)}=1 \;\forall H_{(J_1,\dots,J_\ell)}\subseteq H_{ij},\;\; \textit{mit}\;\; \varphi_{(J_1,\dots,J_\ell)}\;\; \textit{wie in Satz 4.20.}$$

Bemerkung 4.22

a) *Die kritischen Werte $q_{(J_1,\dots,J_\ell),\nu;\alpha}$ hängen nur von den $|J_i|, i=1\dots,\ell$ ab und micht von den in den J_i enthaltenen Indizes.*

b) *Die Anzahl der benötigten kritischen Werte entspricht der Anzahl aller Zerlegungen von k in natürliche Zahlen ohne Berücksichtigung der Reihenfolge und ohne die Zerleung $(1,\dots,1)$. Formelmäßig lässt sie sich berechnen als $g(k)=h(k)-1,\; k\ge 3$ mit*

$$h(0)=0 \;\textit{und}\; h(k)=\sum_{1\le\frac{3\ell^2+\ell}{2}\le k}(-1)^\ell h(k-\frac{3\ell^2\pm\ell}{2}), k\ge 1.$$

Beispielsweise gilt $g(10)=41$, $g(13)=100$, $g(20)=626$, $g(50)=204225$.

c) *Es gilt $\forall 1\le i<j\le k$:*

$$\varphi_{ij}^{Tukey}\le\bar\varphi_{ij}\le\varphi_{ij}^{MRNK}.$$

Da $\varphi^{Tukey}(x)$ und $\varphi^{MRNK}(x)$ einfach (über Standard-Statistiksoftware) zu bestimmen sind, ist nur noch zu prüfen, ob $\bar\varphi_{ij}(x)=1$ oder $\bar\varphi_{ij}(x)=0$ ist, falls $\varphi_{ij}^{Tukey}(x)=0$ und $\varphi_{ij}^{MRNK}(x)=1$ gilt. Die entsprechenden Hypothesen H_{ij} werden als „strittig“ bezeichnet. Dies reduziert den Rechenaufwand weiter. (Selbst von den strittigen Hypothesen können wegen Ordnungseigenschaften der $q_{|J_i|,\nu;\alpha}$ einige unberücksichtigt gelassen werden.)

Beispiel 4.23 (REGWQ-Test, vgl. Tukey (1953), Ryan (1960), Einot and Gabriel (1975), Welsch, 1972)

Eine einfachere Variante einer Verbesserung des Tukey-Tests geht auf Ryan, Einot, Gabriel und Welsch zurück. Sei dazu

$$\alpha_p = \begin{cases} \alpha, & p = k \text{ oder } p = k-1, \\ 1-(1-\alpha)^{p/k}, & p = 2, \ldots, k-2 \end{cases}$$

und $c_p = q_{p,\nu;\alpha_p}$*. Falls die* c_p *nicht steigend in* p *sind, ersetze* c_p *durch* $c'_p = \max_{2 \le r \le p} c_r$*,* $p = 2, \ldots, k$*. Dann ist der REGWQ Test* $\varphi^{REGWQ} = (\varphi_{ij}^{REGWQ}, 1 \le i < j \le k)$ *definiert durch*

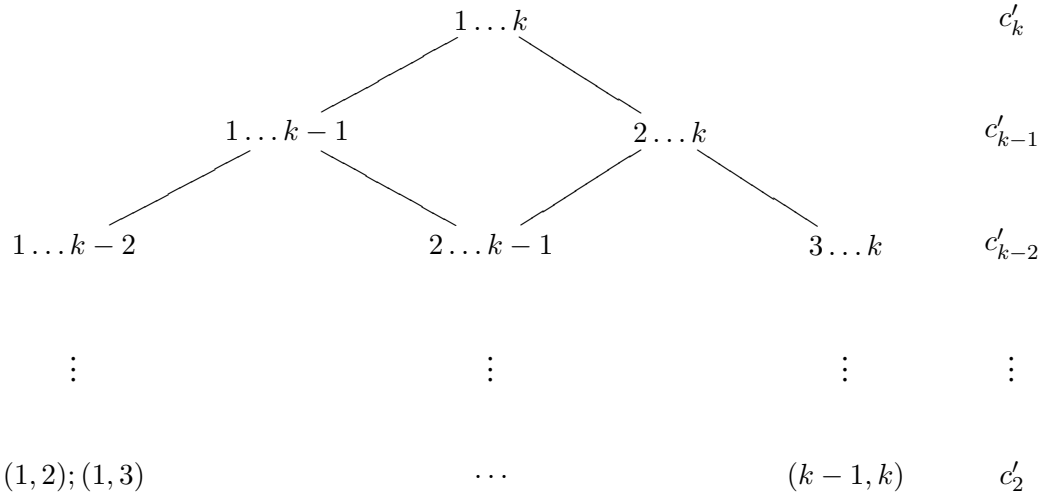
$$\varphi_{ij}^{REGWQ}(x) = 1 \quad \Leftrightarrow \quad \forall J \supseteq \{i,j\} : \max_{r,s \in J} |t_{rs}| > c'_{|J|} \forall x \in \Omega,\ \forall 1 \le i < j \le k, \text{ bzw.}$$

$$\varphi_J^{REGWQ}(x) = 1 \quad \Leftrightarrow \quad \forall L \supseteq J : \max_{i,j \in L} |t_{ij}| > c'_{|L|}.$$

Bemerkung 4.24

a) φ^{REGWQ} *ist ein multipler Test zum multiplen Niveau* α *für* $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$ *aus Modell 3.10 mit balanciertem Design.*

b) Das Testschema von φ^{REGWQ} *lässt sich wie folgt illustrieren:*



Dies ähnelt dem Newman-Keuls-Verfahren; allerdings werden bei φ^{REGWQ} *die Partitionshypothesen implizit durch Verwendung „adjustierter Signifikanzniveaus“* $\alpha_p, p = 2, \ldots, k$ *mit berücksichtigt.*

c) Für jede Hierarchieebene von $\bar{\mathcal{H}}$ *wird genau ein kritischer Wert für die Durchführung von* φ^{REGWQ} *benötigt; es handelt sich bei* φ^{REGWQ} *also um einen Shortcut.*

d) Es ist auch möglich, das Abschlussverfahren bzw. das REGW-Verfahren unter Verwendung von F*-Statistiken, also in Anlehnung an Scheffé, durchzuführen. Für* $H_{(J_1,\ldots,J_\ell)}$ *ist die Zählerstatistik des zugehörigen* F*-Tests dann gegeben als* $S_{(J_1,\ldots,J_\ell)} = \sum_{r=1}^{\ell} S_{J_r}$*, wobei* S_{J_r} *die*

Zählerstatistik des üblichen F-Tests für $H_{J_r}, 1 \leq r \leq \ell$, bezeichnet. Die Freiheitsgrade zu $S_{(J_1,\ldots,J_\ell)}$ berechnen sich als $\sum_{r=1}^{\ell} |J_r| - \ell$. SAS hatte eine Zeit lang eine REGF-Prozedur, die allerdings fehlerhaft programmiert war und nach Aufdeckung des Fehlers aus dem SAS-System ersatzlos gestrichen wurde. Die Prozedur REGWQ existiert jedoch (hoffentlich richtig programmiert) bis heute.

4.4 Step-up Tests zum multiplen Niveau unter Unabhängigkeit

Wie bereits in Bemerkung 4.13 b) angesprochen sind step-up Tests im Vergleich zu step-down Test anti-konservativ, lehnen also mehr (bzw. nicht weniger) Hypothesen ab, falls sie mit dem gleichen Satz an kritischen Werten durchgeführt werden. Bei einem step-up Test versucht man, im ersten Schritt bereits alle m Hypothesen abzulehnen. Gelingt dies nicht, wird im zweiten Schritt versucht $m-1$ Hypothesen abzulehnen, etc. Will man die FWER dennoch mit φ^{SU} kontrollieren, so macht dieses „unvorsichtige“ Vorgehen die Verwendung kleinerer kritischer Werte (für p-Werte) im Vergleich zu φ^{SD} und /oder restriktivere Modellannahmen nötig.Wir thematisieren step-up Tests hier nur für stochastisch unabhängige p-Werte. Ausgangspunkt dieser Ansätze ist der Globaltest von Simes, siehe Simes (1986).

Lemma 4.25 (Simes, 1986)
Seien $U_1, \ldots, U_m \sim$ UNI[0,1] stochastisch unabhängig, $U_{1:m} \leq \cdots \leq U_{m:m}$ die zugehörigen Orderstatistiken und $\alpha_{i:m} = i\alpha/m, i = 1, \ldots, m$ für $\alpha \in [0,1]$. Dann gilt

$$\mathbb{P}(U_{1:m} > \alpha_{1:m}, \ldots, U_{m:m} > \alpha_{m:m}) = 1 - \alpha. \tag{4.6}$$

Beweis: Per Induktion über m. ■

Korollar 4.26 (Simes-Test)
Gegeben sei ein multiples Testproblem $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H} = \{H_i, i \in I = \{1, \ldots, m\}\})$ und marginale p-Werte p_i, $i \in I$, die unter $H_0 = \bigcap_{i=1}^m H_i$ i.i.d. sind. Dann ist ein Niveau α-Test für H_0, der sogenannte Simes-Test φ^{Simes}, gegeben durch

$$\varphi^{Simes}(x) = 1 \quad \Leftrightarrow \quad \exists i \in I : p_{i:m} \leq \frac{i}{m}\alpha.$$

Anmerkung: p-Werte sind unter Nullhypothesen stets stochastisch nicht kleiner als UNI[0,1] und „=“ in dem Ausdruck (4.6) wird zu „$\geq$“, falls die Verteilung der U_i's stochastisch größer als UNI[0,1] wird.

Bemerkung 4.27
Unter den Voraussetzungen von Lemma 4.25 kann der Simes-Test natürlich auch als Grundlage

für einen Abschlusstest $\bar{\varphi} = (\bar{\varphi}_i, i \in I)$ für $\mathcal{H}$ verwendet werden. Seien dazu für Durchschnittshypothesen $H_J = \bigcap_{j \in J} H_j$ $p_{1:J} \leq \cdots \leq p_{|J|:J}$ die zu p_j, $j \in J$ gehörenden, geordneten p-Werte. Formal erhält man dann

$$\bar{\varphi}_i(x) = 1 \quad \Leftrightarrow \quad \forall J \ni i : \exists j \in J : p_{j:J} \leq \alpha_{j:|J|}.$$

Beispiel 4.28 (Hommel, 1988)
Unter den Voraussetzungen von zuvor gelte o.B.d.A. $p_1 \leq \cdots \leq p_m$. Sei

$$J^* = \{i \in I : p_{m-i+k} > \frac{k\alpha}{i} \ \forall k = 1, \ldots, i\}.$$

Setze

$$j^* = \begin{cases} \max\{i : i \in J^*\}, & \textit{falls } J^* \neq \emptyset, \\ 1, & \textit{falls } J^* = \emptyset. \end{cases}$$

Dann ist der Hommel-Test $\varphi^{Hommel} = (\varphi_i^{Hommel}, i \in I)$ gegeben durch

$$\varphi_i^{Hommel}(x) = \begin{cases} 1, & i \leq m^* := \max\{j : p_j \leq \alpha / j^*\}, \\ 0, & \textit{sonst} \end{cases} \qquad \forall x \in \Omega \, \forall i \in I.$$

Anmerkung:

(i) Es ist bemerkenswert, dass $\varphi_i^{Hommel} = \bar{\varphi}_i \ \forall i \in I$ ist.

(ii) Das Verhalten von Hommel reduziert den Rechenaufwand gegenüber dem „Durchtesten“ aller Schritthypothesen gemäß Bemerkung 4.27 (ggfs. drastisch).

(iii) φ^{Hommel} ist ein step-up Test (von der Entscheidungstruktur her).

Beispiel 4.29 (Hochberg, 1988)
Unter den Voraussetzungen von zuvor sei

$$\tilde{m} = \max\{i \in I : p_{i:m} \leq \frac{\alpha}{m - i + 1}\}.$$

Dann ist der Hochberg step-up Test $\varphi^{Hochberg} = (\varphi_i^{Hochberg}, i \in I)$ gegeben durch

$$\varphi_i^{Hochberg}(x) = 1 \ \Leftrightarrow \ p_i \leq p_{\tilde{m}:m}, \ \ i \in I, x \in \Omega.$$

Bemerkung 4.30

a) $\varphi^{Hochberg}$ ist komponentenweise nicht größer als φ^{Hommel}. Insbesondere ist $\varphi^{Hochberg}$ damit (unter Unabhängigkeit) ein Test zum multiplen Niveau α.

b) $\varphi^{Hochberg}$ *verwendet die selben kritischen Werte wie der Bonferroni-Test im Falle beliebiger Abhängigkeiten zwischen den* p*-Werten („Fall I"), braucht aber als step-up Test für FWER-Kontrolle schärfere Modellannahmen (nämlich Unabhängigkeit).*

Zum Abschluss leiten wir noch einen „echten" step-up Test mit exakter FWER-Kontrolle für unabhängige p-Werte her. Dazu zunächst ein wohlbekanntes Resultat zur rekursiven Berechnung der gemeinsamen Verteilungsfunktion von Orderstatistiken.

Lemma 4.31
Seien $X_1, \ldots, X_m$ *stochastisch unabhängig, identisch verteilte Zufallsvariablen mit Verteilungsfunktion* $F(x) = \mathbb{P}(X_1 \leq x),\ x \in \mathbb{R}$*. Seien* $X_{1:m}, \ldots, X_{m:m}$ *die zugehörigen Orderstatistiken. Weiter seien* $c_1 \leq \cdots \leq c_m$ *reelle Konstanten und* $\alpha_j := 1 - F(c_j),\ j = 1, \ldots, m$*. Sei* $F_j(c_1, \ldots, c_j) = \mathbb{P}(X_{1:j} \leq c_1, \ldots, X_{j:j} \leq c_j)$*,* $j = 1, \ldots, m$*, sowie* $F_0 \equiv 1$*. Dann gilt*

$$F_m(c_1, \ldots, c_m) = 1 - \sum_{j=0}^{m-1} \binom{m}{j} F_j(c_1, \ldots, c_j) \alpha_{j+1}^{m-j}.$$

Beweis:

$$\begin{aligned} F_m(c_1, \ldots, c_m) &= \mathbb{P}(X_{1:m} \leq c_1, \ldots, X_{m:m} \leq c_m) = 1 - \mathbb{P}(\exists j \in \{1, \ldots, m\} : X_{j:m} > c_j) \\ &= 1 - \sum_{j=0}^{m-1} \mathbb{P}(X_{1:j} \leq c_1, \ldots, X_{j:j} \leq c_j, X_{j+1:m} > c_{j+1}) \\ &= 1 - \sum_{j=0}^{m-1} F_j(c_1, \ldots, c_j) \alpha_{j+1}^{m-j} \binom{m}{j}. \end{aligned}$$

■

Dieses Resultat kann nun benutzt werden, um (unter Unabhängigkeit) exakte kritische Werte für geordnete p-Werte $p_i, i \in I$, für einen step-up Test rekursiv auszurechnen. Die Existenz einer Lösung garantiert der folgende Satz.

Satz 4.32 (Dalal and Mallows, 1992)
Sei F *eine stetige Verteilungsfunktion auf* $\mathbb{R}$ *und* $(X_m)_{m \in \mathbb{N}}$ *eine i.i.d. Folge von Zufallsvariablen mit* $X_1 \sim F$*.*
Dann gilt:
$\forall \alpha \in (0,1) : \exists (c_m)_{m \in \mathbb{N}}$ *mit* $c_i < c_{i+1}$ *für alle* $i \in \mathbb{N}$ *mit der Eigenschaft* $F_m(c_1, \ldots, c_m) = 1 - \alpha$*, wobei* F_m *wie in Lemma 4.31 die gemeinsame Verteilungsfunktion der Orderstatistiken* $X_{1:m}, \ldots, X_{m:m}$ *bezeichne.*

Nun zur Berechnung der kritischen Werte $\alpha_j, j \in I$, für die unabhängigen p-Werte.

Trivialerweise ist $\alpha_1 = \alpha$. Aus $F_m(c_1, \ldots, c_m) \stackrel{!}{=} 1-\alpha$ und $F_j(c_1, \ldots, c_j) \stackrel{!}{=} 1-\alpha \; \forall j = 1, \ldots, m-1 \; (*)$ folgt mit der Rekursionsformel aus Lemma 4.31, dass

$$\begin{aligned}
1-\alpha &= 1 - \sum_{j=0}^{m-1} \binom{m}{j} F_j(c_1, \ldots, c_j) \alpha_{j+1}^{m-j}, \; m \geq 2 \\
\iff 1-\alpha &= 1 - \alpha^m - \sum_{j=1}^{m-1} \binom{m}{j} F_j(c_1, \ldots, c_j) \alpha_{j+1}^{m-j} \\
\iff \alpha - \alpha^m &= \sum_{j=1}^{m-1} \binom{m}{j} F_j(c_1, \ldots, c_j) \alpha_{j+1}^{m-j} \\
\underset{(*)}{\iff} \frac{\alpha - \alpha^m}{1-\alpha} &= \sum_{j=1}^{m-1} \binom{m}{j} \alpha_{j+1}^{m-j} \\
\iff \sum_{j=1}^{m-1} \alpha^j &= \sum_{j=1}^{m-2} \binom{m}{j} \alpha_{j+1}^{m-j} + m\alpha_m \\
\iff \alpha_m &= \frac{1}{m} \left[\sum_{j=1}^{m-1} \alpha^j - \sum_{j=1}^{m-2} \binom{m}{j} \alpha_{j+1}^{m-j} \right].
\end{aligned}$$

Es gilt zum Beispiel $\alpha_2 = \alpha/2$, $\alpha_3 = \alpha/3 + \alpha^2/12$, $\alpha_4 = \alpha/4 + \alpha^2/12 + \alpha^3/24 - \alpha^4/96$. Diese (für einen step-up Test unter Unabhängigkeit exakten) kritischen Werte gehen auf Rom (1990) zurück und heißen daher „Rom-Werte“.

Beispiel 4.33 (Exakter step-up Test unter Unabhängigkeit)
Sei $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H} = \{H_i, i \in I = \{1, \ldots, m\}\})$ ein endliches multiples Testproblem, $p_1, \ldots, p_m$ stochastisch unabhängige p-Werte für die marginalen Testprobleme H_i versus $K_i, i \in I$, $p_{1:m}, \ldots, p_{m:m}$ die geordneten p-Werte und $H_{1:m}, \ldots, H_{m:m}$ die entsprechend umsortierten Hypothesen. Seien $\alpha = \alpha_1 \geq \cdots \geq \alpha_m$ die Rom-Werte. Dann wird durch folgende Testvorschrift ein Test zum multiplen Niveau α für $\mathcal{H}$ definiert: Lehne $H_{1:m}, \ldots, H_{m^*:m}$ ab, wobei

$$m^* = \max\{i \in I : p_{i:m} \leq \alpha_{m-i+1}\}.$$

Anmerkung:
Es gilt $\alpha/i \leq \alpha_i \leq 1 - (1-\alpha)^{1/i} \;\; \forall i \in I$. Man hat also einen Widerstreit zwischen Konservativität der Teststruktur, Größe der kritischen Werte und Strukturannahmen über die gemeinsame Verteilung der p-Werte.

Kapitel 5

False Discovery Rate (FDR)

Erinnerung 5.1 (an Definition 1.35)
Seien $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H} = \{H_i, i \in I\})$ *sowie* $\varphi = (\varphi_i, i \in I)$ *für* $\mathcal{H}$ *fest vorgegeben bzw. ausgewählt.*

(a) Die Zufallsvariable

$$FDP_\vartheta(\varphi) = \frac{V(\vartheta)}{R(\vartheta) \vee 1}$$

heißt False Discovery Proportion (FDP) von φ*.*

(b) $FDR_\vartheta(\varphi) = \mathbb{E}_\vartheta[FDP_\vartheta(\varphi)]$ *heißt False Discovery Rate (FDR) von* φ*.*

(c) Der multiple Test φ *heißt FDR-kontrollierend zum Niveau* $\alpha \in (0, 1)$*, falls*

$$FDR(\varphi) := \sup_{\vartheta \in \Theta} FDR_\vartheta(\varphi) \leq \alpha.$$

(d) $pFDR_\vartheta(\varphi) = \mathbb{E}_\vartheta[\frac{V(\vartheta)}{R(\vartheta)} | R(\vartheta) > 0]$ *heißt positive False Discovery Rate (pFDR) von* φ*.*

5.1 Allgemeine Theorie und der lineare step-up Test

Lemma 5.2 (Verhältnis von FWER, FDR und pFDR)

(a) $FDR_\vartheta(\varphi) = pFDR_\vartheta(\varphi) \cdot \mathbb{P}_\vartheta(R(\vartheta) > 0)$

(b) Ist $I = \{1, \ldots, m\}$ *endlich und* $\vartheta \in \Theta$ *derart, dass* $m_0(\vartheta) = m$ *gilt, so ist* $FDR_\vartheta(\varphi) = FWER_\vartheta(\varphi)$*. Unbedingt (für alle* $\vartheta \in \Theta$*) gilt* $FDR_\vartheta(\varphi) \leq FWER_\vartheta(\varphi)$*.*

Beweis:

zu (a):

$$\begin{aligned}\text{FDR}_\vartheta(\varphi) &= \mathbb{E}_\vartheta\left[\frac{V(\vartheta)}{R(\vartheta)\vee 1}\right] \\ &= \mathbb{E}_\vartheta\left[\frac{V(\vartheta)}{R(\vartheta)\vee 1}|R(\vartheta)>0\right]\cdot\mathbb{P}_\vartheta\left(R(\vartheta)>0\right) \\ &\quad + \mathbb{E}_\vartheta\left[\frac{V(\vartheta)}{R(\vartheta)\vee 1}|R(\vartheta)=0\right]\cdot\mathbb{P}_\vartheta\left(R(\vartheta)=0\right) \\ &= \text{pFDR}_\vartheta(\varphi)\cdot\mathbb{P}_\vartheta\left(R(\vartheta)>0\right)+0.\end{aligned}$$

zu (b):

Für $m_0 = m$ ist $V(\vartheta) = R(\vartheta)$. Damit gilt $\text{pFDR}_\vartheta(\varphi) \equiv 1$ und $\text{FDR}_\vartheta(\varphi) = \mathbb{P}_\vartheta\left(R(\vartheta)>0\right) = \mathbb{P}_\vartheta\left(V(\vartheta)>0\right) = \text{FWER}_\vartheta(\varphi)$. Unbedingt gilt $\text{FDP}_\vartheta(\varphi) \leq \mathbf{1}_{\{V(\vartheta)>0\}}$ und damit $\mathbb{E}_\vartheta\left[\text{FDP}_\vartheta(\varphi)\right] \leq \mathbb{E}_\vartheta\left[\mathbf{1}_{\{V(\vartheta)>0\}}\right] \Longleftrightarrow \text{FDR}_\vartheta(\varphi) \leq \text{FWER}_\vartheta(\varphi)$. ■

Bemerkung 5.3

(i) Aus dem Beweis von Lemma 5.2 (b) sieht man leicht, dass die pFDR zum Niveau $\alpha \in (0,1)$ (im frequentistischen Sinne) nicht zu kontrollieren ist. Sie hat Bayesianische Interpretationen (später mehr).

(ii) Lemma 5.2 (b) zeigt zwei Dinge.
Erstens kann durch Verwendung des FDR-Kriteriums nur dann eine Verbesserung (hinsichtlich Güte) erreicht werden (ein Test φ mit mehr erwarteten Ablehnungen ausgewählt werden), falls die Existenz falscher Nullhypothesen vorausgesetzt wird.
Zweitens (allgemeiner gedacht) ist der zu erwartende Gütezugewinn (wenn man Typ-I-Fehler mit der FDR statt der FWER misst) umso größer, je mehr falsche Nullhypothesen es in $\mathcal{H}$ gibt. Gilt im Extremfall zum Beispiel $m_0(\vartheta) < m\alpha$, so können durch φ alle m Nullhypothesen abgelehnt werden ($R(\vartheta) = m$), ohne dass die FDP von φ den Wert α überschreitet. Insbesondere gilt für einen schrittweise verwerfenden, FDR-kontrollierenden multiplen Test stets $\text{FDR}_\vartheta(\varphi) \leq \min(\alpha, m_0(\vartheta)/m)$ und $\text{FDR}_\vartheta(\varphi)$ ist (im Gegensatz zu $\text{FWER}_\vartheta(\varphi)$) nicht notwendigerweise eine wachsende Funktion von $\mathbb{E}_\vartheta[R(\vartheta)]$. Dadurch tragen FDR-basierte statistische Analysen eher Screeningcharakter, während FWER-basierte Analysen typischerweise als konfirmatorisch angesehen werden.

Algorithmus 5.4 (Benjamini and Hochberg, 1995)
Sei $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H} = \{H_i, i \in I = \{1,\ldots,m\}\})$ ein endliches multiples Testproblem. Dann ist der lineare step-up Test $\varphi^{LSU} = (\varphi_i^{LSU}, i \in I)$ von Benjamini und Hochberg wie folgt definiert.
Unter der Voraussetzung, dass für die marginalen Testprobleme H_i versus K_i, $i \in I$, p-Werte p_i

verfügbar sind, bezeichne $p_{1:m} \leq p_{2:m} \leq \ldots \leq p_{m:m}$ die Orderstatistiken von $(p_i, i \in I)$ und $H_{1:m}, \ldots, H_{m:m}$ die entsprechend unsortierten Hypothesen in $\mathcal{H}$. Sei $k := \max\{i \in I : p_{i:m} \leq i\alpha/m\}$. Dann lehnt φ^{LSU} genau die Hypothesen $H_{1:m}, \ldots, H_{k:m}$ ab. Falls das Maximum in I nicht exisitiert, so wird keine einzige Hypothese in $\mathcal{H}$ abgelehnt.

Bemerkung 5.5

(i) φ^{LSU} ist ein step-up Test mit Simes' kritischen Werten (vgl. Simes-Test aus Lemma 4.25 und Korollar 4.26).

(ii) Die Entscheidungsregel von φ^{LSU} lässt sich wie folgt grafisch veranschaulichen:

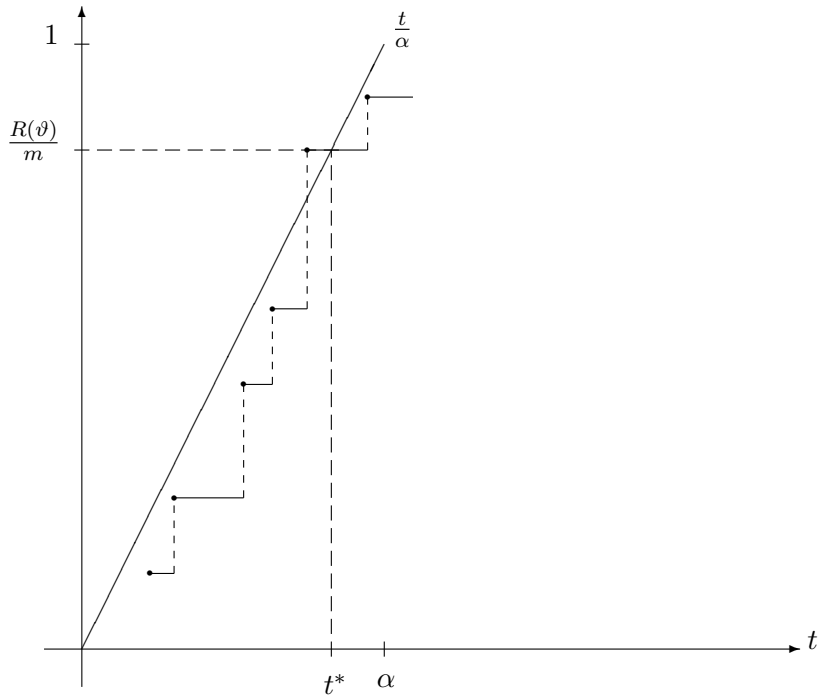


Abbildung 5.1: Grafische Veranschaulichung von φ^{LSU}

Dabei heißt $t \mapsto t/\alpha$ Simes-Gerade und $t^* = \sup\{t \in [0, \alpha] : \hat{F}_m(t) \geq t/\alpha\}$, wobei $\hat{F}_m$ die empirische Verteilungsfunktion von $(p_i, i \in I)$ bezeichnet.

(iii) Unter gewissen Voraussetzungen (siehe Definition 5.6) kontrolliert φ^{LSU} die FDR zum Niveau α.

(iv) Wie das Abschlussprinzip und daraus resultierende step-down Prozeduren die Basis für FWER-Kontrolle sind, so ist φ^{LSU} die Basis für viele FDR-kontrollierende step-up Tests.

Man findet daher häufig die Assoziationen

$$\textit{step-down} \quad \leftrightarrow \quad \textit{FWER-Kontrolle},$$

$$\textit{step-up} \quad \leftrightarrow \quad \textit{FDR-Kontrolle.}$$

Definition 5.6 (generelle Voraussetzungen)
Zum Beweis der FDR-Kontrolle eines multiplen Tests $\varphi = (\varphi_i,\, i \in I = \{1, \ldots, m\})$ für ein endliches multiples Testproblem $(\Omega, \mathcal{A}, \mathcal{P}, \mathcal{H})$ mit $\mathcal{H} = \{H_i,\, i \in I\}$ nehmen wir an, es seien marginale p-Werte $(p_i,\, i \in I)$ verfügbar und die Testentscheidung von φ komme durch Vergleich der $p_{i:m}$ mit kritischen Werten $\alpha_{1:m} \leq \alpha_{2:m} \leq \ldots \leq \alpha_{m:m}$ zustande. Dann definieren wir die folgenden Eigenschaften.

(STEP) φ *ist eine schrittweise verwerfende Testprozedur (step-up, step-down oder step-up-down).*

(D1) $\forall\, \vartheta \in \Theta : \forall j \in I : \forall i \in I_0(\vartheta) : \mathbb{P}_\vartheta(R(\vartheta) \geq j | p_i \leq t)$ *ist nicht-wachsend in* $t \in (0, \alpha_{j:m}]$.

(D2) $\forall\, \vartheta \in \Theta :\quad \forall i \in I_0(\vartheta) :\, p_i \sim$*UNI*$([0,1])$.

(I1) $\forall\, \vartheta \in \Theta :$ *Die* $p_i(X), i \in I_0(\vartheta)$*, sind iid.*

(I2) $\forall\, \vartheta \in \Theta$*:* $(p_i(X) : i \in I_0(\vartheta))$ *und* $(p_i(X) : i \in I_1(\vartheta))$ *sind stochastisch unabhängige Zufallsvektoren.*

Anmerkung:

(i) (D1) und (D2) sind Verteilungs- (distributional) Eigenschaften. (D1) ist eine positive Abhängigkeitseigenschaft, die unter anderem multivariate Verteilungen erfüllen, die multivariat total positiv der Ordnung 2 (MTP_2) oder positiv regressionsabhängig auf der Teilmenge (PRDS) $I_0(\vartheta)$ sind.

(ii) Über die Abhängigkeiten zwischen den $p_i(X), i \in I_1(\vartheta)$, werden keinerlei Annahmen gemacht.

Theorem 5.7 (Benjamini and Yekutieli (2001), Finner and Roters (2001), Sarkar, 2002)
Unter (D1) gilt

$$\forall\, \vartheta \in \Theta :\quad \textit{FDR}_\vartheta(\varphi^{LSU}) \leq \frac{m_0(\vartheta)}{m}\alpha.$$

Unter (D2)-(I2) gilt

$$\forall\, \vartheta \in \Theta :\quad \textit{FDR}_\vartheta(\varphi^{LSU}) = \frac{m_0(\vartheta)}{m}\alpha.$$

Anmerkung:

(i) Insbesondere ist (D1) eine hinreichende Bedingung dafür, dass φ^{LSU} die FDR zum Niveau α kontrolliert.

(ii) Unter (D2)-(I2) hängt $\text{FDR}_\vartheta(\varphi^{LSU})$ nicht von der „Größe“ von ϑ selbst, sondern lediglich von $m_0(\vartheta)$ ab.

(iii) Die Voraussetzungen aus Definition 5.6 können zum Nachweis der FDR-Kontrolle allgemeiner step-up-down Tests mit geeigneten kritischen Werten in Anlehnung an den nachstehenden Beweis für φ^{LSU} verwendet werden.

Beweis: (Finner, Dickhaus, and Roters, 2009). Sei $\vartheta \in \Theta$ beliebig, aber fest vorgegeben. Dann berechnet sich $\text{FDR}_\vartheta(\varphi^{LSU})$ als

$$\begin{aligned}
\text{FDR}_\vartheta(\varphi^{LSU}) &= \sum_{i\in I_0(\vartheta)} \sum_{j=1}^{m} \frac{1}{j} \mathbb{P}\left(R(\vartheta)=j, \varphi_i=1\right) \\
&\underset{\text{(STEP)}}{=} \sum_{i\in I_0(\vartheta)} \sum_{j=1}^{m} \frac{1}{j} \mathbb{P}_\vartheta\left(p_i \leq \alpha_{j:m}\right) \mathbb{P}_\vartheta\left(R(\vartheta)=j | p_i \leq \alpha_{j:m}\right) \\
&\overset{(*)}{\leq} \sum_{i\in I_0(\vartheta)} \sum_{j=1}^{m} \frac{\alpha_{j_m}}{j} \mathbb{P}_\vartheta\left(R(\vartheta)=j | p_i \leq \alpha_{j:m}\right) \\
&\overset{(**)}{\leq} \sum_{i\in I_0(\vartheta)} \{\alpha_{1:m} \mathbb{P}_\vartheta\left(R(\vartheta)\geq 1 | p_i \leq \alpha_{1:m}\right) \\
&\qquad + \sum_{j=2}^{m} \left(\frac{\alpha_{j:m}}{j} - \frac{\alpha_{j-1:m}}{j-1}\right) \mathbb{P}_\vartheta\left(R(\vartheta)\geq j | p_i \leq \alpha_{j:m}\right) \} \\
&= \frac{m_0(\vartheta)}{m}\alpha.
\end{aligned}$$

In $(*)$ gilt „=“ unter (D2) und in $(**)$ gilt „=“ unter (D2)-(I2). ■

Hinweise:

- Für schrittweise verwerfende Testprozeduren gilt:
  $\forall\, j \in I : R(\vartheta)=j \Longrightarrow [\forall\, i \in I : \varphi_i = 1 \Leftrightarrow p_i \leq \alpha_{j:m}]$.
- $\mathbb{P}_\vartheta(R(\vartheta)=j | p_i \leq \alpha_{j:m})$ wird in der vierten Beweiszeile „teleskopiert“ zu $\mathbb{P}_\vartheta(R(\vartheta)\geq j | p_i \leq \alpha_{j:m}) - \mathbb{P}_\vartheta(R(\vartheta)\geq j+1 | p_i \leq \alpha_{j:m})$ und für $j=2,\ldots,m$ wird (nach (D1) „erlaubt“!) $\mathbb{P}_\vartheta(R(\vartheta)\geq j | p_i \leq \alpha_{j-1:m})$ abgeschätzt durch $\mathbb{P}_\vartheta(R(\vartheta)\geq j | p_i \leq \alpha_{j:m})$. Für step-up Tests gilt sogar „konstant“ anstelle von „nicht-wachsend“ in (D1).

Bemerkung 5.8

Kann positive Abhängigkeit zwischen den marginalen p-Werten im Sinne von (D1) nicht angenommen werden, so ist φ^{LSU} im Allgemeinen keine FDR-kontrollierende Testprozedur. Dies kann „repariert" werden, indem die kritischen Werte von $\alpha_{i:m} = i\alpha/m$ zu $\alpha_{i:m} = \frac{i\alpha}{m\sum_{k=1}^{m} k^{-1}}, i \in I$, geändert werden, also α durch $\frac{\alpha}{\sum_{k=1}^{m} k^{-1}}$ ersetzt wird. Die resultierende Benjamini-Yekutieli Prozedur (Benjamini and Yekutieli, 2001) ist jedoch im Allgemeinen sehr konservativ. Bis heute ist das Problem „FDR-Kontrolle unter negativer bzw. nicht spezifizierter Abhängigkeitsstruktur" nicht befriedigend im Rahmen der Theorie schrittweiser multipler Tests gelöst worden.

5.2 Explizite Adaptionstechniken und die Storey-Prozedur

Auch 15 Jahre nach seiner Propagation durch Benjamini und Hochberg ist der lineare step-up Test die am weitesten verbreitete Prozedur zur FDR-basierten statistischen Datenanalyse. [Das Überprüfen von (D1) ist dabei ein Problem, das häufig nicht systematisch als Bestandteil des Analysevorgangs berücksichtigt wird.] Selbst unter (D1)-(I2) hat φ^{LSU} jedoch den Nachteil, das FDR-Niveau α im Falle von $m_0 < m$ nicht auszuschöpfen. Wäre m_0 bekannt, so könnte man φ^{LSU} zum adjustierten FDR-Niveau $\frac{m\alpha}{m_0}$ durchführen und erhielte eine (unter (D2)-(I2)) exakt die FDR kontrollierende Testprozedur (eine sogenannte Orakel-Prozedur). Da $m_0 = m_0(\vartheta)$ jedoch eine Funktion des unbekannten Parameters ϑ ist, ist ein solches Vorgehen in der Praxis nicht durchführbar. Explizite (daten-) adaptive Teststrategien schätzen daher m_0 in einem ersten Schritt vor und verwenden die Schätzung $\hat{m}_0$ dann für die Konstruktion von kritischen Werten bzw. Ablehnkurven. Für die FDR-Kontrolle ist dabei wichtig, dass $\hat{m}_0$ den unbekannten Wert m_0 „konservativ" schätzt, also (im statistischen Mittel / fast sicher) überschätzt. Es existieren mehrere konkurrierende Schätzprinzipien. So kann zum Beispiel eine unadjustierte multiple Testprozedur in einem ersten Schritt durchgeführt und ihre beobachtete Anzahl an Ablehnungen zur Konstruktion von $\hat{m}_0$ benutzt werden (so in Benjamini, Krieger, and Yekutieli (2006) oder in Blanchard and Roquain, 2008). Eine andere Schätzmethode geht bereits auf Schweder and Spjøtvoll (1982) zurück und benutzt die empirische Verteilung der (beobachteten) marginalen p-Werte.

Definition 5.9 (Schweder-Spjøtvoll-Schätzer)

Zur Schätzung von m_0 schlagen Schweder und Spjøtvoll den folgenden Schätzer vor.

$$\hat{m}_0^{Schweder} \equiv \hat{m}_0^{Schweder}(\lambda) = m\frac{1 - \hat{F}_m(\lambda)}{1 - \lambda},$$

wobei $\lambda \in (0, 1)$ ein Tuningparameter ist und $\hat{F}_m$ wie in Bemerkung 5.5 die empirische Verteilungsfunktion der marginalen p-Werte $(p_i,\ i \in I = \{1, \ldots, m\})$ bezeichnet.

Anmerkung:

Es gibt zwei grafische Veranschaulichungen des Schweder-Spjøtfall-Schätzers. Die ursprünglich

von den Autoren gewählte macht von $\hat{F}_m$ Gebrauch, eine alternative benutzt ein Histogramm der beobachteten marginalen p-Werte.

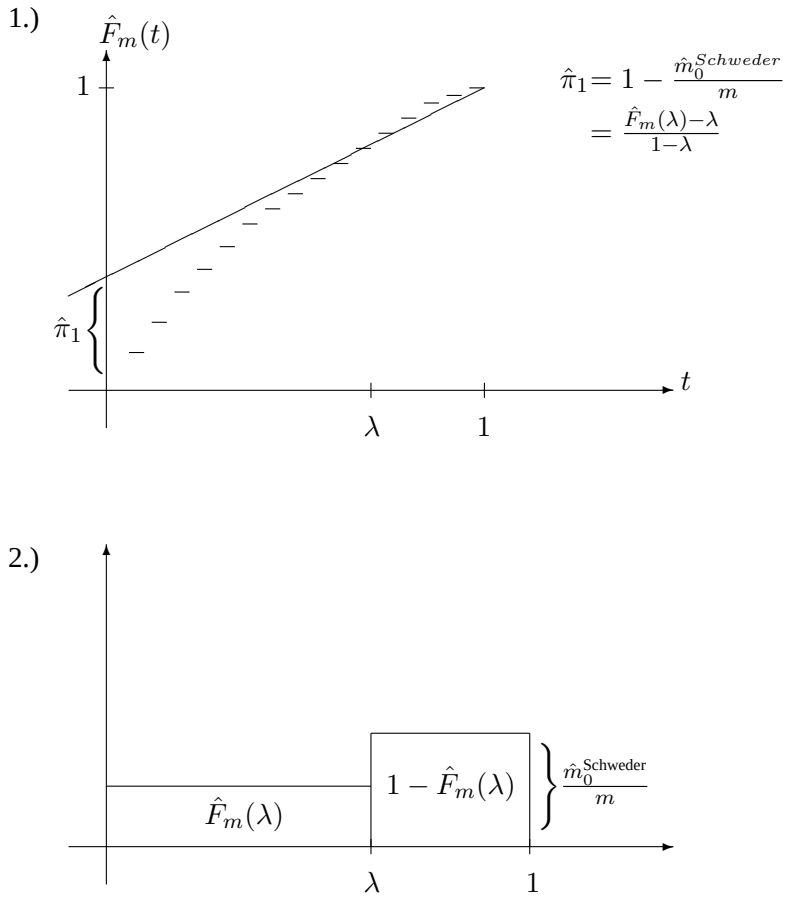


Abbildung 5.2: Grafische Veranschaulichungen des Schweder-Spjøtvoll-Schätzers

Definition 5.10 (Storey-Schätzer)
Storey (2002a), Storey (2002b) und Storey et al. (2004) schlagen den folgenden, leicht modifizierten Schätzer für $\pi_0 := m_0/m$ vor:

$$\hat{\pi}_0^{Storey} \equiv \hat{\pi}_0^{Storey}(\lambda) = \frac{1 - \hat{F}_m(\lambda) + 1/m}{1-\lambda}.$$

Satz 5.11 (Storey, Taylor, and Siegmund, 2004)
Unter (I1) und (I2) aus Definition 5.6 kontrolliert ein modifizierter step-up Test, bei dem gegenüber φ^{LSU} aus Algorithmus 5.4 α durch $\alpha/\hat{\pi}_0^{Storey}(\lambda)$ ersetzt wird, die FDR zum Niveau α.

Beweis: Mit empirischer Prozess-, Martingaltheorie und optimalem Stoppen (Stoppsatz); sehr lehrreich! ■

Abbildung 5.12

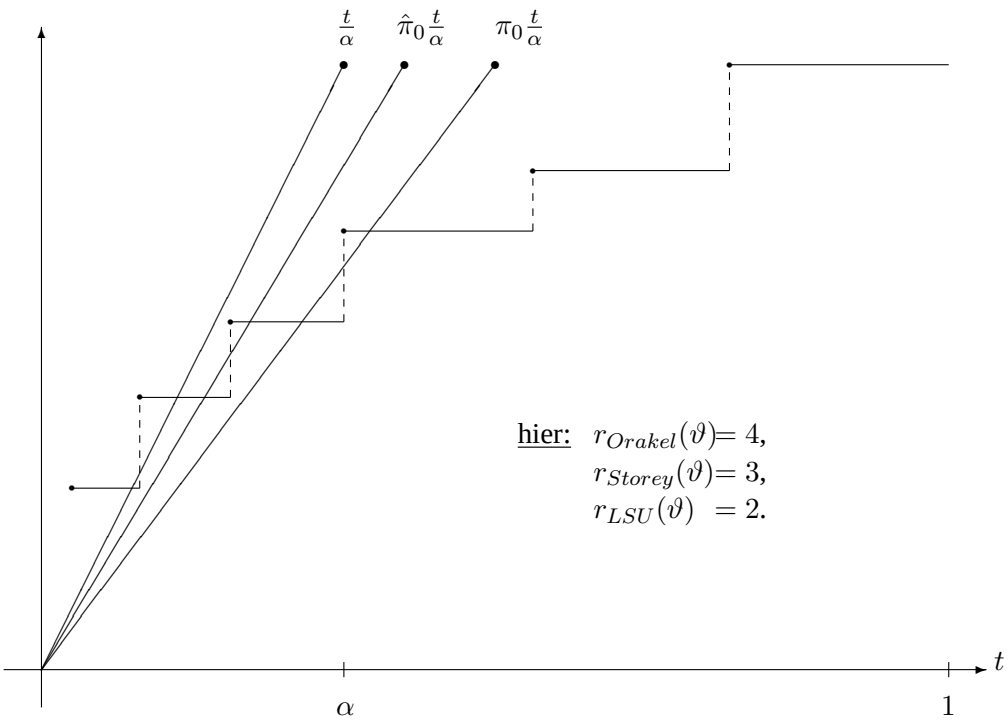


Abbildung 5.3: Nach Storey et al. adjustierte Ablehngerade

Anmerkung:
Die Annahmen (I1) und (I2) aus Satz 5.11 können für sehr großes m (asymptotische Betrachtungen) zum Konzept der „weak dependence“ abgeschwächt werden. Dabei wird nur noch gefordert, dass die empirischen Verteilungsfunktionen der p-Werte $(p_i,\, i \in I_0(\vartheta))$ und $(p_i,\, i \in I_1(\vartheta))$ jeweils gegen eine Grenzverteilungsfunktion (im Glivenko-Cantelli Sinne) konvergieren (für $m \to \infty$). Dann kann asymptotisch auch der original Schweder-Spjøtvoll-Schätzer $\hat{\pi}_0^{\text{Schweder}}$ statt $\hat{\pi}_0^{\text{Storey}}$ verwendet werden.

Bemerkung 5.13 (implizite Adaption)
Die moderne Methode der „impliziten Adaption“ vermeidet eine Vorschätzung von π_0 bzw. m_0 und versucht stattdessen, durch geschickte Wahl von kritischen Werten bzw. Ablehnkurven (zum Vergleich mit $\hat{F}_m$) flexibel auf mögliche Werte von π_0 reagieren zu können. Das Konstruktionsverfahren besteht aus zwei Schritten:

1) *Gegeben ein statistisches Modell, eine Klasse von möglichen Testverfahren sowie π_0, ermittle die ungünstigste Parameter-Konfiguration (least favorable configuration, LFC) für $\vartheta \in \Theta$ unter Alternativen, d.h., finde $\vartheta^* \in \Theta$, sodass die FDR bei gegebenem π_0 maximal wird.*

2) *Bestimme kritische Werte oder eine Ablehnkurve so, dass die FDR unter ϑ^* genau gleich $\min(\alpha, \pi_0)$ beträgt.*

Die Berechnungen unter Schritt 2) können dabei sehr kompliziert werden und manchmal nur asymptotisch (für $m \to \infty$ und unter der Annahme $m_0/m \to \pi_0 \in (0,1]$) durchgeführt werden. Unter (I1) und (I2) liefert der folgende Satz ein hilfreiches Resultat zur Bestimmung von LFCs für eine Klasse von step-up Tests.

Theorem 5.14 (Benjamini and Yekutieli, 2001)
Unter (I1) und (I2) sei φ^{SU} ein step-up Test mit kritischen Werten $\alpha_{1:m} \leq \alpha_{2:m} \leq \ldots \leq \alpha_{m:m}$ mit der Eigenschaft

$$\frac{\alpha_{j:m}}{j} \quad \textit{ist wachsend in } j. \tag{5.1}$$

Dann wächst die FDR von φ^{SU}, wenn die Verteilung von $(p_i,\ i \in I_1(\vartheta))$ stochastisch kleiner wird.

Demnach ist der LFC ϑ^* hier also dadurch gegeben, dass $\forall\, i \in I_1(\vartheta^*)$ gilt: $p_i(X) \sim \delta_0$ (Dirac-Verteilung mit Punktmasse 1 im Punkt 0). Unter (D2)-(I2) wird eine solche Konfiguration ϑ^* „Dirac-uniform Konfiguration“ $DU_{m_0,m}$ genannt.

Lemma 5.15
Unter $DU_{m_0,m}$ und angenommen, $m_0/m \to \pi_0 \in (0,1]$, gilt

$$\hat{F}_m(t) \underset{(m\to\infty)}{\longrightarrow} \pi_1 + \pi_0 \cdot t \ \textit{ gleichmäßig in } t \in [0,1]\ \mathbb{P}_{m_0,m}\textit{ - f.s.}\,,$$

wobei $\pi_1 := 1 - \pi_0$ und $\mathbb{P}_{m_0,m}$ das zu $DU_{m_0,m}$ gehörige Wahrscheinlichkeitsmaß bezeichnet.

Beweis: Satz von Glivenko-Cantelli. ■

Heuristik 5.16 (Asymptotisch optimale Ablehnkurve, Finner, Dickhaus, and Roters, 2009)
Sei $\varphi = (\varphi_i,\ i \in I = \{1,\ldots,m\})$ ein Einschritttest, d.h. $\forall\, i \in I : \varphi_i(x) = 1 \iff p_i(x) \leq t$ für ein fest vorgegebenes $t \in (0,1)$. Dann ist unter $DU_{m_0,m}$ für $m \to \infty$ die FDR von φ gegeben durch

$$\lim_{m\to\infty} \textit{FDR}_{m_0,m}(\varphi) = \frac{\pi_0 t}{\pi_1 + \pi_0 t}.$$

Ein asymptotisch optimaler Schwellenwert $t^*(\pi_0)$ erfüllt $\forall\, \pi_0 \in (\alpha, 1)$ die Beziehung

$$\begin{aligned} \lim_{m\to\infty} \textit{FDR}_{m_0,m}(\varphi) = \alpha \quad &\Leftrightarrow \quad \frac{\pi_0 t^*(\pi_0)}{\pi_1 + \pi_0 t^*(\pi_0)} = \alpha \\ &\Leftrightarrow \quad t^*(\pi_0) = \frac{\alpha \pi_1}{\pi_0(1-\alpha)}. \end{aligned}$$

Eine asymptotisch optimale Ablehnkurve f_α ist demnach gegeben durch

$$f_\alpha(t^*(\pi_0)) = \pi_1 + \pi_0 \cdot t^*(\pi_0) \iff f_\alpha(u) = \frac{u}{\alpha + (1-\alpha)u}\,, \quad u \in [0,1].$$

In Finner, Dickhaus, and Roters (2009) konnte bewiesen werden, dass für auf f_α basierende step-up-down Tests φ^{SUD} unter $DU_{m_0,m}$ mit $m_0/m \to \pi_0 \in [\alpha, 1)$ tatsächlich $\lim_{m\to\infty} FDR_{m_0,m}(\varphi^{SUD}) = \alpha$ gilt.

Abbildung 5.17 (φ^{SUD} basierend auf f_α)

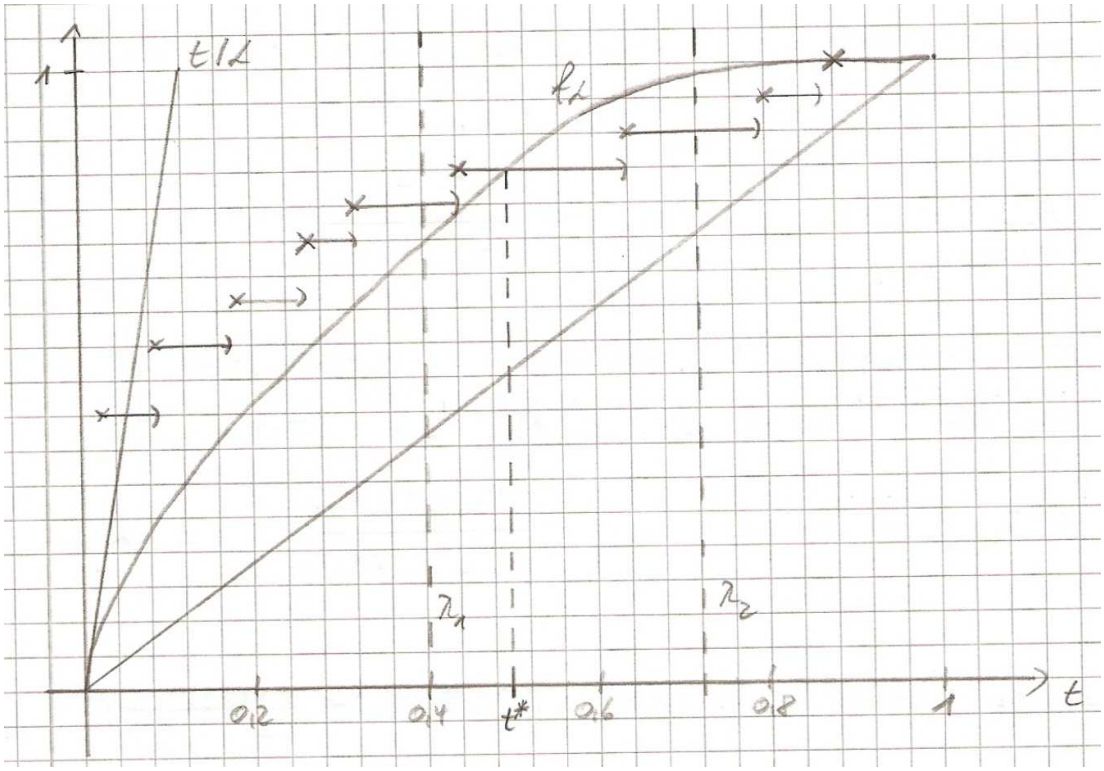


Abbildung 5.4: φ^{SUD} basierend auf der asymptotisch optimalen Ablehnkurve

Wegen $f_\alpha(1) = \hat{F}_m(1) = 1$ ist die Wahl $\lambda = 1$ unzulässig, da dann immer alle Hypothesen abgelehnt würden. Jede Wahl des Tuningparameters $\lambda \in [0, 1)$ ist jedoch asymptotisch valide (Finner, Dickhaus, and Roters, 2009).

5.3 Bayesianische Interpretationen, pFDR

Den Überlegungen in diesem Abschnitt liegt die generelle Annahme zu Grunde, dass sehr viele analoge Vergleiche simultan durchzuführen seien. Die Modellbildung hat ihren Ursprung in genetischen Microarray-Analysen, in denen geprüft werden soll, ob Gen j bzw. SNP j verantwortlich für ein erhöhtes Krankheitsrisiko ist für $j = 1, \ldots, m$ und m „sehr groß“ ($m \sim 10.000$ Gene, $m \sim 500.000$ SNPs). Dabei wird eine Gleichartigkeit der Einwirkung jedes einzelnen Gens / jedes einzelnen SNPs auf das Erkrankungsrisiko modelliert.

Modell 5.18 (Zweiklassen-Mischmodell)
Es sei $(T_1, H_1), \ldots, (T_m, H_m)$ eine iid. Folge von Teststatistiken und Indikatorvariablen. Dabei gelte die Konvention

$$H_i = 0: \quad \Leftrightarrow \textit{ i-te Nullhypothese ist wahr},$$

$$H_i = 1: \quad \Leftrightarrow \textit{ i-te Alternativhypothese ist wahr}, i = 1, \ldots, m.$$

Es sei $T_j|H_j \sim (1 - H_j)F_0 + H_j F_1$, wobei F_0 und F_1 stetige Verteilungsfunktionen mit zugehörigen Dichtefunktionen f_0 und f_1 bezeichnen. Es bezeichne $\pi_0 = \mathbb{P}(H_0) := \mathbb{P}(H_i = 0)$ die a priori-Wahrscheinlichkeit für die Richtigkeit der i-ten Nullhypothese und $\pi_1 := 1 - \pi_0$.

Ist unter Modell 5.18 ein multipler Test $\varphi = (\varphi_i,\, i \in I = \{1, \ldots, m\})$ charakterisiert über einen Ablehnbereich $\Gamma \equiv \Gamma_\alpha$ für alle $T_i, i \in I$, so gilt (nach Theorem 1 in Storey, 2003), dass

$$\text{pFDR}(\Gamma) = \mathbb{E}\left[\frac{V(\Gamma)}{R(\Gamma)}|R(\Gamma) > 0\right] = \mathbb{P}\left(H_0|T \in \Gamma\right)$$

sich als a posteriori-Wahrscheinlichkeit für die Richtigkeit der Nullhypothese, gegeben die zugehörige Teststatistik fällt in den Ablehnbereich, interpretieren lässt. Gilt speziell $\Gamma = (-\infty, t]$, so gilt

$$\text{Fdr}(t) := \text{pFDR}(\Gamma) = \pi_0 \frac{F_0(t)}{F(t)} = \mathbb{P}\left(H_0|T \leq t\right) \text{ mit } F(t) = \pi_0 F_0(t) + \pi_1 F_1(t).$$

Die Notation $\text{Fdr}(t)$ geht auf Bradley Efron zurück. Unter Verwendung der Bayes-Formel für Dichten lässt sich analog

$$\text{fdr}(t) = \pi_0 \frac{f_0(t)}{f(t)} = \mathbb{P}(H_0|T = t),$$

die sogenannte „local fdr“, bilden.
Die beiden Größen stehen in der Beziehung

$$\text{Fdr}(t) = \mathbb{E}_f\left[\text{fdr}(T)|T \leq t\right].$$

Wie immer in Bayesianischen Betrachtungsweisen hat die obige Interpretationswelt den Charme, ohne Signifikanzniveau auszukommen und gut interpretierbare a posteriori-Ausdrücke aus den erhobenen Daten zu produzieren. Statistische Inferenz kann in diesem Rahmen über „empirische Bayes“-Methoden betrieben werden:

a) Spezifiziere π_0 und f_0.

b) Schätze f durch $\hat{f}$ aus den erhobenen Daten.

c) Bilde $\widehat{\text{fdr}}(t_i) = \pi_0 f_0(t_i)/\hat{f}(t_i)$ und ziehe Schlüsse aus diesem Wert.

Alternativ ist es auch möglich, $\text{Fdr}(t)$ aus $\hat{F}$ zu schätzen und sogenannte q-Werte zu berechnen:

$$q\text{-value}(t_i) = \inf_{\alpha: t_i \in \Gamma_\alpha} \text{pFDR}(\Gamma).$$

Diese lassen sich als „a posteriori p-Werte“ auffassen.

Tabellenverzeichnis

Abbildungsverzeichnis

Literaturverzeichnis

Alt, R. (1988). Hierarchical test problems and the closure principle. In *P. Bauer, G. Hommel, E. Sonnemann (eds.): Multiple Hypothesenprüfung - Multiple Hypotheses Testing. Symposium Gerolstein 1987.*, pp. 162–176. Berlin: Springer. Medizinische Informatik und Statistik 70.

Bauer, P., B. M. Pötscher, and P. Hackl (1988). Model selection by multiple test procedures. *Statistics 19*(1), 39–44.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol. 57*(1), 289–300.

Benjamini, Y., A. M. Krieger, and D. Yekutieli (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika 93*(3), 491–507.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat. 29*(4), 1165–1188.

Blanchard, G. and E. Roquain (2008). Two simple sufficient conditions for FDR control. *Electron. J. Statist. 2*, 963–992.

Bonferroni, C. E. (1936). *Teoria statistica delle classi e calcolo delle probabilita. Pubbl. d. R. Ist. Super. di Sci. Econom. e Commerciali di Firenze 8*. Firenze: Libr. Internaz. Seeber.

Dalal, S. and C. Mallows (1992). Buying with exact confidence. *Ann. Appl. Probab. 2*(3), 752–765.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *J. Am. Stat. Assoc. 50*, 1096–1121.

Dunnett, C. W. (1964). New tables for multiple comparisons with a control. *Biometrics 20*, 482–491.

Dunnett, C. W. and A. C. Tamhane (1992). A step-up multiple test procedure. *J. Am. Stat. Assoc. 87*(417), 162–170.

Einot, I. and K. Gabriel (1975). A study of the powers of several methods of multiple comparisons. *J. Am. Stat. Assoc. 70*(351), 574–583.

Finner, H. (1988). Closed multiple range tests. In *P. Bauer, G. Hommel, E. Sonnemann (eds.): Multiple Hypothesenprüfung - Multiple Hypotheses Testing. Symposium Gerolstein 1987.*, pp. 10–32. Berlin: Springer. Medizinische Informatik und Statistik 70.

Finner, H. (1994). *Testing Multiple Hypotheses: General Theory, Specific Problems, and Relationships to Other Multiple Decision Procedures*. Habilitationsschrift. Fachbereich IV, Universität Trier.

Finner, H., T. Dickhaus, and M. Roters (2009). On the false discovery rate and an asymptotically optimal rejection curve. *Ann. Stat. 37*(2), 596–618.

Finner, H. and M. Roters (2001). On the false discovery rate and expected type I errors. *Biom. J. 43*(8), 985–1005.

Finner, H. and K. Strassburger (2002). The partitioning principle: a powerful tool in multiple decision theory. *Ann. Stat. 30*(4), 1194–1213.

Fisher, R. A. (1935). *The Design of Experiments*. Oliver & Boyd, Edinburgh and London.

Gabriel, K. R. (1969). Simultaneous test procedures - some theory of multiple comparisons. *Ann. Math. Stat. 40*, 224–250.

Hayter, A. J. (1984). A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative. *Ann. Stat. 12*, 61–75.

Hochberg, Y. (1974). Some generalizations of the T-method in simultaneous inference. *J. multivariate Analysis 4*, 224–234.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika 75*(4), 800–802.

Hochberg, Y. and A. C. Tamhane (1987). *Multiple comparison procedures.* Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics. New York etc.: John Wiley & Sons, Inc.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat., Theory Appl. 6*, 65–70.

Holm, S. A. (1977). *Sequentially rejective multiple test procedures. Statistical Research Report No. 1977-1*. Institute of Mathematics and Statistics, University of Umåa.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika 75*(2), 383–386.

Hommel, G. and T. Hoffmann (1988). Controlled uncertainty. In *P. Bauer, G. Hommel, E. Sonnemann (eds.): Multiple Hypothesenprüfung - Multiple Hypotheses Testing. Symposium Gerolstein 1987.*, pp. 154–161. Berlin: Springer. Medizinische Informatik und Statistik 70.

Keuls, M. (1952). The use of the „ Studentized Range “ in connection with an analysis of variance. *Euphytica 1*, 112–122.

Kramer, C. (1956). Extensions of multiple range tests to group means with unequal numbers of replications. *Biometrics 12*, 307–310.

Kramer, C. (1957). Extensions of multiple range tests to group correlated adjusted means. *Biometrics 13*, 13–18.

Lehmann, E. (1957a). A theory of some multiple decision problems. I. *Ann. Math. Stat. 28*, 1–25.

Lehmann, E. (1957b). A theory of some multiple decision problems. II. *Ann. Math. Stat. 28*, 547–572.

Lehmann, E. L. and J. P. Romano (2005). *Testing statistical hypotheses. 3rd ed.* Springer Texts in Statistics. New York, NY: Springer.

Marcus, R., E. Peritz, and K. R. Gabriel (1976). On closed test procedures with special reference to ordered analysis of variance. *Biometrika 63*(3), 655–660.

Maurer, W. and B. Mellein (1988). On new multiple tests based on independent p-values and the assessment of their power. In *P. Bauer, G. Hommel, E. Sonnemann (eds.): Multiple Hypothesenprüfung - Multiple Hypotheses Testing. Symposium Gerolstein 1987.*, pp. 48–66. Berlin: Springer. Medizinische Informatik und Statistik 70.

Newman, D. (1939). The distribution of range in samples from a normal population, expressed in terms of an independent estimate of standard deviation. *Biometrika 31*, 20–30.

Pearson, E. S. and H. O. Hartley (1966). *Biometrika Tables for Statisticians, Vol. I.* Cambridge University Press, England.

Pocock, S. J., N. L. Geller, and A. A. Tsiatis (1987). The analysis of multiple endpoints in clinical trials. *Biometrics 43*, 487–498.

Rom, D. M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika 77*, 663–665.

Roy, S. (1953). On a heuristic method of test construction and its use in multivariate analysis. *Ann. Math. Stat. 24*, 220–238.

Ryan, T. A. (1960). Significance tests for multiple comparison of proportions, variances, and other statistics. *Psychol. Bull. 57*, 318–328.

Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Ann. Stat. 30*(1), 239–257.

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika 40*, 87–110.

Schweder, T. and E. Spjøtvoll (1982). Plots of P-values to evaluate many tests simultaneously. *Biometrika 69*, 493–502.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *J. Am. Stat. Assoc. 62*, 626–633.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika 73*, 751–754.

Sonnemann, E. (2008). General solutions to multiple testing problems. Translation of „Sonnemann, E. (1982). Allgemeine Lösungen multipler Testprobleme. EDV in Medizin und Biologie 13(4), 120-128“. *Biom. J. 50*, 641–656.

Storey, J. D. (2002a). A direct approach to false discovery rates. *J. R. Stat. Soc., Ser. B, Stat. Methodol. 64*(3), 479–498.

Storey, J. D. (2002b). *False Discovery Rates. Theory and Applications to DNA microarrays. Ph. D. dissertation.* Stanford University, Department of Statistics.

Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann. Stat. 31*(6), 2013–2035.

Storey, J. D., J. E. Taylor, and D. Siegmund (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Stat. Soc., Ser. B, Stat. Methodol. 66*(1), 187–205.

Tukey, J. W. (1953). *The problem of multiple comparisons*. Mimeographed monograph.

Welsch, R. (1972). A modification of the Newman-Keuls procedure for of multiple comparisons. *Working Paper, Sloan School of Management, M.I.T., Boston, MA. 612-72*.

Witting, H. (1985). *Mathematische Statistik I: Parametrische Verfahren bei festem Stichprobenumfang.* Stuttgart: B. G. Teubner.