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

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

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

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

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

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

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

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

2.4.1 Fehlerraten

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

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

Dabei sind

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

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

Idealerweise: Minimiere

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

Klassische Testtheorie ($m = 1$):

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

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

Fehlerraten 1. Art (type I error rates)

- PCER (per-comparison error rate):

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

  Das ist die relative Anzahl erwarteter Fehler 1. Art.

- PFER (per-family error rate):

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

  Das ist die absolute Anzahl erwarteter Fehler 1. Art.

- FWER (family-wise error rate):

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

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

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

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

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

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

Starke und schwache Kontrolle

Typischerweise gilt: Für eine *unbekannte* Teilmenge

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

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

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

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

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

2.4.2 Multiple Testprozeduren

Bonferroni-Prozedur

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

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

d.h.

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

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

Holm-Prozedur

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

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

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

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

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

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

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

Schritt 3. usw.

Es gilt:

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

Beweis:

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

Eine falsche Ablehnung liegt vor, wenn

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

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

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

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

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

□

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

Step-Down-Prozeduren

Allgemeine Struktur: Sei

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

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

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

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

Eine Alternative sind:

Step-Up-Prozeduren

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

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

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

Bemerkung.

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