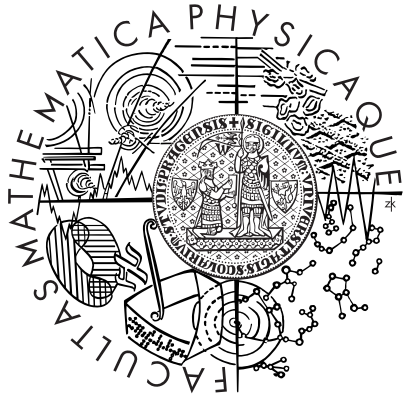
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

# BAKALÁŘSKÁ PRÁCE




Matěj Kadavý

## Mnohonásobné testování

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: Mgr. Zdeněk Hlávka, Ph.D.

Studijní program: Matematika
Studijní obor: Obecná matematika

2007

Rád bych na tomto místě poděkoval svému vedoucímu panu Mgr. Zdeňku Hlávkovi, Ph.D. za odborné vedení, ochotu a vstřícnost a za to, že mi po celý čas psaní této práce poskytoval cenné rady a připomínky.

Prohlašuji, že jsem svou bakalářskou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

V Praze dne Matěj Kadavý

# Obsah

Název práce: Mnohonásobné testování
Autor: Matěj Kadavý
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedoucí bakalářské práce: Mgr. Zdeněk Hlávka, Ph.D.
e-mail vedoucího: hlavka@karlin.cuni.cz

Abstrakt: V této práci se zaměříme na dva přístupy, jak měřit chybu, ke které dochází při chybném zamítnutí jednotlivých hypotéz v mnohonásobném testování: false discovery rate ($FDR$) a familywise error rate ($FWE$). Popíšeme procedury kontrolující $FWE$: Hochbergovu, Holmovu a Hommelovu proceduru. Ukážeme, že procedury kontrolující $FDR$ pro nezávislé testové statistiky, jako je BH procedura, jsou mnohem silnější než procedury kontrolující $FWE$. Zjistíme, že v případě většího počtu neplatných hypotéz je BH procedura příliš konzervativní. Tento problém částečně vyřeší adaptivní BH procedura. Dokážeme, že BH procedura kontroluje $FDR$ pro závislé testové statistiky mající t-rozdělení. V simulacích aplikovaných na lineární model se zaměříme na rozdíly mezi procedurami kontrolující $FDR$ a $FWE$ jak pro nezávislé, tak i pro závislé testové statistiky.

Klíčová slova: $p$-hodnota, Bonferroniho nerovnost, $FDR$, $FWE$, Hochbergova procedura, Holmova procedura, Hommelova procedura, BH procedura

Title: Multiple-hypothesis testing
Author: Matěj Kadavý
Department: Department of Probability and Mathematical Statistics
Supervisor: Mgr. Zdeněk Hlávka, Ph.D
Supervisor's e-mail address: hlavka@karlin.cuni.cz

Abstract: In this paper we introduce two multiple-hypothesis testing error measures: the false discovery rate ($FDR$) and the familywise error rate ($FWE$). The procedures controlling the $FWE$ presented here are: Hochberg's procedure, Holm's procedure and Hommel's procedure. A simple procedure called BH procedure is given here as an $FDR$ controlling procedure for independent test statistics and is shown to be much more powerful than procedures controlling the $FWE$. When some of the hypotheses are in fact false, the BH procedure is too conservative. We present an adaptive BH procedure, where the number of true null hypotheses is estimated by the given $p$-values. We prove that BH procedure controls the $FDR$ for dependent test statistics of t-distribution. In simulation part we compare the procedures controlling the $FDR$ and the $FWE$ on linear model for both: independent and dependent test statistics.

Key words: $p$-value, Bonferroni inequality, $FDR$, $FWE$, Hommel's procedure, Hochberg's procedure, Holm's procedure, BH procedure

# Kapitola 1

# Úvod

Uvažujme, že chceme testovat jednoduchou nulovou hypotézu $H_0$ proti alternativě $H_A$. Předpokládejme, že test je založený na testové statistice $T$. Pro daný kritický obor $\Gamma$ zamítneme $H_0$, když $T \in \Gamma$, a nezamítneme $H_0$, když $T \notin \Gamma$. Chyba prvního druhu nastane, když $T \in \Gamma$, ale $H_0$ platí. Chyba druhého druhu nastane, když $T \notin \Gamma$, ale $H_1$ platí. Zpravidla vyžadujeme, aby pravděpodobnost chyby prvního druhu byla rovna předem zvolenému číslu $\alpha$. Toto číslo pak nazýváme hladina testu. Když určujeme $\Gamma$, vybíráme z kritických oborů, které mají pravděpodobnost chyby prvního druhu menší nebo rovnou $\alpha$. Vybereme ten, který má nejmenší pravděpodobnost chyby druhého druhu. Tedy kritický obor je vybírán tak, aby kontroloval chybu prvního druhu.

V mnohonásobném testování je situace obtížnější. Nyní každý test má svojí vlastní chybu prvního druhu a není zcela zřejmé, jak měřit chybu, ke které dojde při chybném zamítnutí jednotlivých hypotéz. První navrhovanou veličinou, která měla řešit tento problém, je *FWE-Familywise error rate*. Veličina *FWE* je definována jako pravděpodobnost, že alespoň jedna hypotéza ze všech testovaných hypotéz je chybně zamítnuta. Tedy místo kontroly pravděpodobnosti chyby prvního druhu pro každý test na hladině $\alpha$ je kontrolováno *FWE* na hladině $\alpha$.

Další veličinou, která se zabývá kontrolou chyby v mnohonásobném testování, je *FDR-False discovery rate*. Tato veličina je definována jako střední hodnota poměru chybně zamítnutých nulových hypotéz a celkového počtu zamítnutých hypotéz. Tedy stejně jako pro *FWE* bude platit, že místo kontroly pravděpodobnosti chyby prvního druhu pro každý test na hladině $\alpha$ je kontrolováno *FDR* na hladině $q$, kde $q \leq \alpha$.

Cílem této práce bude popsat jednotlivé procedury, které kontrolují buď *FWE* nebo *FDR* na daných hladinách. Ukážeme, že procedury kontrolující *FWE* kontrolují i *FDR*. Dále ukážeme, že při větším počtu neplatných nulových hypotéz není kontrola *FWE* nezbytně nutná, a proto raději dáváme přednost procedurám kontrolující *FDR*. V případě, kdy očekáváme malý počet zamítaných hypotéz, je vhodnější dávat zase přednost procedurám kontrolujících *FWE*.

Teoretická část této práce je rozdělena do dvou kapitol. V první kapitole se zabýváme mnohonásobným testováním pro nezávislé testové statistiky a v druhé kapitole pro závislé statistiky. V odstavci 2.1 si ukážeme, jak získat $p$-hodnoty pomocí těchto statistik a v od-

stavci 2.2 použijeme Bonferroniho nerovnost ke stanovení horní hranice pro kontrolu *FWE*. V odstavci 2.3 zadefinujeme a popíšeme základní vlastosti *FDR*. V odstavcích 2.4 a 2.5 navrhneme procedury kontrolující *FWE* a *FDR* na daných hladinách. Tyto procedury jsou založené na vzestupně srovnaných $p$-hodnotách, podle kterých se pak odhaduje počet zamítaných hypotéz. Zaměříme se zejména na Benjamini-Hochberg proceduru-BH proceduru. Dokážeme, že tato procedura kontroluje *FDR* na hladině $q$. Z důkazu bude zřejmé, že pro *FDR* platí nerovnost

$$FDR \leq \frac{m_0}{m} q \leq q,$$

kde $m_0$ je počet platných nulových hypotéz. Tato nerovnost nám dá prostor pro zlepšení BH procedury. V odstavcích 2.7 a 2.8 se budeme zabývat odhadem počtu $m_0$. Na tomto odhadu pak bude založena nová adaptivní procedura, kterou si popíšeme v odstavci 2.9. V odstavcích 2.11, 2.12 a 2.13 se budeme zabývat základními vlastnostmi procedur kontrolujících *FWE*, tedy Holmovou, Hochbergovou a Hommelovou procedurou.

V další kapitole se budeme zabývat tím, jaké vlastnosti musejí splňovat závislé testové statistiky, aby BH procedura kontrolovala *FDR*. V odstavci 3.1 ukážeme, že nutnou podmínkou je, že mají vlastnost *PRDS-Positive regression dependency*. V odstavci 3.2 dokážeme, jaká rozdělení splňují vlastnost *PRDS* a nakonec se nám podaří ukázat, že testové statistiky se Studentovým rozdělením mají tuto vlastnost. V posledním odstavci této kapitoly dokážeme, že pokud sdružené rozdělení testových statistik má vlastnost *PRDS* na množině testových statistik odpovídajících nulové hypotéze, pak BH procedura kontroluje *FDR* na hladině menší nebo rovné $\frac{m_0}{m} q$.

V poslední kapitole budeme aplikovat výsledky získané z teoretické části na lineární model. Ukážeme, že procedury kontrolující *FDR* zamítají více hypotéz než procedury kontrolující *FWE*, pokud hladina *FWE* a *FDR* je stejná. Dále se budeme zabývat vlivem závislosti testových statistik na tyto procedury.

# Kapitola 2

# Nezávislé testové statistiky

## 2.1 P-hodnoty

Nechť $\mathbf{X} = (X_1, \ldots, X_m)$ označuje náhodný vektor se sdruženou distribuční funkcí $F_{\mathbf{X}}(\mathbf{x})$. Nulovou hypotézou budeme rozumět nějaké tvrzení o rozdělení určeném touto distribuční funkcí. Nulové hypotézy $H_0^1, \ldots, H_0^m$ testujeme pomocí testových statistik $T_1, \ldots, T_m$ náhodného vektoru $\mathbf{X}$.

Metody mnohonásobného testování, což je testování $m$ libovolných nulových hypotéz současně, jsou založené na vzestupně srovnaných $p$-hodnotách, které získáme z testových statistik. Proto nyní uvedu definici $p$-hodnoty pro oboustranný test s krátkým ilustrativním příkladem: $p$-hodnota je nejmenší hladina testu, při které bychom zamítli nulovou hypotézu. Ekvivalentně $p$-hodnota vyjadřuje pravděpodobnost spočítanou za platnosti nulové hypotézy, že dostaneme právě naši testovou statistiku nebo statistiku více odporující naší hypotéze. Nechť $H_0$ označuje nulovou hypotézu, pak

$$p\text{-hodnota} = \mathrm{P}(|T| \geq T_0 | H_0), \tag{2.1}$$

kde $T_0$ je spočtená hodnota testové statistiky.

*Příklad.* Uvažujme, že chceme testovat jednoduchou nulovou hypotézu $H_0$ proti oboustranné alternativě $H_A$. Test je založený na testové statistice $T$, která má Studentovo rozdělení s $n-k$ stupni volnosti. Spočtenou hodnotu testové statistiky si označme $T_0$. Dále označme $f_{n-k}$ (resp. $F_{n-k}$) hustotu (resp. distribuční funkci) Studentova rozdělení s $n-k$ stupni volnosti. Pro $T_0 \leq 0$ platí, že $p = 1$. Předpokládejme, že $T_0 > 0$, pak

$$p = \mathrm{P}(|T| \geq T_0) = \int\limits_{|T| \geq T_0} f_{n-k}(x)\,\mathrm{d}x = \int_{-\infty}^{-T_0} f_{n-k}(x)\,\mathrm{d}x + \int_{T_0}^{\infty} f_{n-k}(x)\,\mathrm{d}x \tag{2.2}$$

$$= 2\int_{T_0}^{\infty} f_{n-k}(x)\,\mathrm{d}x = 2[1 - F_{n-k}(T_0)] = 2F_{n-k}(-T_0), \tag{2.3}$$

tedy $T_0 = t_{n-k}(p)$, kde $t_{n-k}(p)$ je kvantil Studentova rozdělení s $n-k$ stupni volnosti v bodě $p$. Je-li $\alpha$ hladina testu, pak pro $T_0 > 0$ platí, že

$$T_0 = t_{n-k}(p) > t_{n-k}(\alpha/2) \tag{2.4}$$
$$F_{n-k}^{-1}(1-p/2) > F_{n-k}^{-1}(1-\alpha/2) \tag{2.5}$$
$$-p > -\alpha \tag{2.6}$$
$$p < \alpha. \tag{2.7}$$

Je tedy zřejmé, že máme-li testové statistiky $T_1, \ldots, T_m$, pro které platí

$$T_1 \geq T_2 \geq \cdots \geq T_m, \tag{2.8}$$

pak

$$p_1 \leq p_2 \leq \cdots \leq p_m. \tag{2.9}$$

## 2.2 Bonferroniho nerovnost

Předpokládejme, že máme $H_0^1, \ldots, H_0^m$ nulových hypotéz. Nechť $A_j$, $j = 1, \ldots, m$, označuje jev, že nezamítáme $j$-tou platnou nulovou hypotézu $H_0^j$ a nechť $\mathrm{P}(A_j) = 1 - \alpha_j$. Tedy pravděpodobnost chybného zamítnutí $j$-té platné nulové hypotézy, nebo-li chyba prvního druhu, je $\alpha_j$. Označme $A_j^C$ doplňkový jev k $A_j$, pak

$$1 - \delta = \mathrm{P}\left(\bigcap_{j=1}^{m} A_j\right) = 1 - \mathrm{P}\left[\left(\bigcap_{j=1}^{m} A_j\right)^C\right] = 1 - \mathrm{P}\left(\bigcup_{j=1}^{m} A_j^C\right)$$
$$\geq 1 - \sum_{j=1}^{m} \mathrm{P}\left(A_j^C\right) = 1 - \sum_{j=1}^{m} \alpha_j. \tag{2.10}$$

Nerovnost (2.10) se nazývá *Bonferroniho nerovnost*. V případě $\alpha_j = \alpha$, $j = 1, \ldots, m$, platí, že $\mathrm{P}\left(\bigcap_{j=1}^{m} A_j\right) \geq 1 - m\alpha$, tedy pravděpodobnost, že nezamítneme platné hypotézy je číslo větší než $1 - m\alpha$. Číslo $\delta$ v Bonferroniho nerovnosti je pravděpodobnost, že zamítneme alespoň jednu platnou nulovou hypotézu. Číslo $\delta$ budeme dále označovat výrazem *FWE*, což je zkratka z anglického výrazu *Familywise error rate*.

Je dobré zmínit, že velikost *FWE* závisí také na míře závislosti mezi jevy $A_j$, $j = 1 \ldots m$. Tedy pokud závislost mezi jevy $A_j$ je dostatečně malá, pak z definice podmíněné pravděpodobnosti dostáváme

$$\mathrm{P}\left(\bigcap_{j=1}^{m} A_j\right) = \mathrm{P}(A_1)\mathrm{P}(A_2|A_1) + \cdots + P(A_m|A_1, A_2, \ldots, A_{m-1}) \tag{2.11}$$
$$\cong \prod_{j=1}^{m} \mathrm{P}(A_j) = \prod_{j=1}^{m} (1 - \alpha_j) \geq 1 - \sum_{j=1}^{m} \alpha_j, \tag{2.12}$$

kde poslední nerovnost plyne z *Bernoulliho nerovnosti* pro $\alpha_j \in (0,1)$, $j = 1, \dots m$. Symbolem $\cong$ zde i dále budeme myslet *je přibližně rovno.*

Z uvedené Bonferroniho nerovnosti snadno vyplývá, že pokud chceme kontrolovat $FWE$ na hladině nejvýše $\alpha$, pak stačí každou nulovou hypotézu $H_0^j,\ j = 1, \dots, m$, testovat na hladině $\frac{\alpha}{m}$. Tato úvaha vede na základní Bonferroniho proceduru pro mnohonásobné testování.

**Bonferroniho procedura** :

Uvažujme $H_0^1, \dots, H_0^m$ nulových hypotéz s testovými statistikami $T_1, \dots, T_m$.
Nechť $p_1, \dots, p_m$ jsou $p$-hodnoty odpovídající testovým statistikám.
Pak zamítneme $H_j$, pokud $p_j \leq \frac{\alpha}{m},\ j = 1, \dots, m$.

## 2.3 False discovery rate *FDR*

Uvažujme testování $m$ nulových hypotéz. Předpokládejme, že $m_0$ hypotéz platí. Zavedeme náhodné veličiny $R, T, U, V, S$. Náhodná veličina $R$ označuje celkový počet zamítnutých hypotéz. Náhodná veličina $U$ označuje počet správně nezamítnutých nulových hypotéz. Náhodná veličina $V$ označuje počet chybně zamítnutých nulových hypotéz. Náhodná veličina $T$ označuje počet chybně nezamítnutých nulových hypotéz. Náhodná veličina $S$ označuje počet správně zamítnutých nulových hypotéz. Realizace těchto náhodných veličin budeme značit malými písmeny, tedy $r, t, u, v, s$.

Tabulka 2.1: Počet chyb při testování $m$ nulových hypotéz

| | Počet nezamítnutých | Počet zamítnutých | Celkový počet |
|---|---|---|---|
| Nulová hypotéza platí | U | V | $m_0$ |
| Alternativní hypotéza platí | T | S | $m - m_0$ |
| | $m - R$ | R | $m$ |

Náhodná veličina $R$ je pozorovatelná, protože známe celkový počet všech zamítnutých hypotéz. Náhodné veličiny $U, V, T, S$ nejsou pozorovatelné, ale platí pro ně, že $R = V + S$ a $m - R = U + T$. Testujeme-li hypotézy jednotlivě na hladině $\alpha$, pak $R(\alpha)$ je rostoucí funkcí proměnné $\alpha$. Při tomto označení platí, že $FWE$ je $\mathrm{P}(V \geq 1)$ a Bonferonniho procedura zaručí, že $\mathrm{P}(V \geq 1) \leq \alpha$. Často se nám v praxi stává, že celkový závěr z jednotlivých hypotéz není nezbytně chybný, jakmile alespoň jedna z nich chybná je. Ilustrujme si tuto myšlenku na krátkém příkladě.

Máme novou léčebnou proceduru na hepatitidu typu B a chceme ji porovnat se stávající léčebnou procedurou. Za celkovou nulovou hypotézu si vezmeme: nová procedura není lepší než stávající. Tyto dvě procedury budeme srovnávat v $m$ různých kritériích jako je výše jaterních enzymů $ALT, AST, GMT$ a výše antigenů viru $HB_SA_g, HB_eA_g$ v krvi atd. Tedy celková hypotéza se nám rozdělí na $m$ dílčích hypotéz, které si označíme $H_0^i$, $i = 1, \dots, m$.

Pokud bude nová procedura lepší než stávající, pak nastane více zamítnutí jednotlivých hypotéz než přijmutí. V tomto případě nám bohužel kontrola *FWE* nebude nic platná. Líbilo by se nám, kdybychom uměli kontrolovat poměr mezi počtem chybně zamítnutých nulových hypotéz a celkového počtu zamítnutých hypotéz. Tato úvaha nás vede k zavedení náhodné veličiny $Q = \frac{V}{R} = \frac{V}{V+S}$. Náhodná veličina $Q$ tedy vyjadřuje poměr mezi počtem chybně zamítnutých nulových hypotéz a všech zamítnutých hypotéz. Pak definujeme *FDR* jako střední hodnotu náhodné veličiny $Q$, tedy

$$FDR = \mathrm{E}Q = \mathrm{E}\left(\frac{V}{V+S}\right) = \mathrm{E}\left(\frac{V}{R}\right). \tag{2.13}$$

Určitě nás bude zajímat, jaký je vztah mezi *FWE* a *FDR*.

1. Pokud platí $m_0 = m$, pak $FDR = FWE$.
2. Pokud platí $m_0 < m$, pak $FDR \leq FWE$.

*Důkaz.* 1) Zřejmě platí $s = 0$ a $v = r$.
Pokud $v = 0$, pak $Q = 0$ a $\mathrm{P}(V \geq 1) = 0 = \mathrm{E}Q$.
Pokud $v > 0$, pak $Q = \frac{v}{v+s} = \frac{v}{v} = 1$ tedy $\mathrm{P}(V \geq 1) = \mathrm{E}Q = 1$.
2) Pokud $v > 0$, pak $\frac{v}{r} \leq 1 \Leftrightarrow I_{(v\geq 1)} \geq Q$ a aplikací střední hodnoty na obě strany dostaváme $\mathrm{P}(V \geq 1) = \mathrm{E}I_{(V\geq 1)} \geq \mathrm{E}Q$.

Z uvedených vlastností je zřejmé, že kontroluje-li procedura *FWE*, pak kontroluje i *FDR*.

*Poznámka* 2.3.1. Pokud budeme předpokládat platnost všech nulových hypotéz, tedy $m = m_0$, pak bude podíl $\frac{V}{R}$ roven identicky jedné. Tomuto se můžeme vyhnout tím, že v celé práci budeme předpokládat, že alespoň jedna nulová hypotéza není platná.

## 2.4 Procedury kontrolující *FWE*

Pro následující procedury předpokládejme, že testujeme $H_0^1, \ldots, H_0^m$ nulových hypotéz testovými statistikami $T_1, \ldots, T_m$, kterým odpovídají $p$-hodnoty $p_1, \ldots, p_m$. Srovnejme $p$-hodnoty dle velikosti, tedy platí $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$. Ke každé proceduře jsou uvedeny přepočítané $p$-hodnoty, u kterých horní index určuje příslušnost k dané proceduře. Tyto $p$-hodnoty slouží k tomu, že danou procedurou zamítneme $i$-tou hypotézu, právě když $i$-tá přepočítaná $p$-hodnota je menší než stanovená konvenční hladina $\alpha$. Tedy zamítáme hypotézy tak, jak jsme zvyklí v případě testování pouze jedné hypotézy.

**1. Jednokroková Bonferroniho** :
Zamítneme $H_0^i$ pokud $p_i \leq \frac{\alpha}{m}$,
$p$-hodnoty $p_i^{bonf} = \min\limits_{i=1,\ldots,m}(mp_i, 1)$.

**2. Holm step-down procedura** :
Zamítneme $H_0^{(i)}$ pro $i = 1, \ldots, k-1$, kde $k = \min\limits_{i=1,\ldots,m}\{i : p_{(i)} > \frac{\alpha}{m-i+1}\}$,
$p$-hodnoty $p_{(i)}^{holm} = \max\limits_{j=1,\ldots,i}\{\min\{(m-j+1)p_{(j)}, 1\}\}$.

**3. Hochberg step-up procedura** :

Zamítneme $H_0^{(i)}$ pro $i = 1, \ldots, k-1$, kde $k = \max_{i=1,\ldots,m} \{i : p_{(i)} \leq \frac{\alpha}{m-i+1}\}$,

$p$-hodnoty $p_{(i)}^{hoch} = \min_{j=i,\ldots,m} \{\min\{(m-j+1)p_{(j)}, 1\}\}$.

**4. Hommelova procedura** :

Položme $k = \max_{j=1,\ldots,m} \{j : p_{(m-j+l)} > \frac{l}{j}\alpha \text{ pro } l = 1, \ldots, j\}$.

Pokud takové $j$ neexistuje, zamítneme všechny hypotézy,

jinak zamítneme $H_{(i)},\ i = 1, \ldots, n$, pro $p_{(i)} \leq \frac{\alpha}{j}$.

## 2.5 Procedury kontrolující *FDR*

**1. Benjaminy Hochberg, BH procedura** :

Zamítneme $H_i$ pro $i = 1, \ldots, k$, kde $k = \max_{i=1,\ldots,m} \{i : p_{(i)} \leq \frac{i\alpha}{m}\}$,

$p$-hodnoty $p_{(i)}^{BH} = \min_{j=1,\ldots,i} \{\min\{\frac{mp_{(i)}}{j}, 1\}\}$.

**Benjaminy, Yekutieli, BY procedura** :

Zamítneme $H_{(i)}$ pro $i = 1, \ldots, k-1$, kde $k = \max_{i=1,\ldots,m} \{i : p_{(i)} \leq \frac{\alpha i}{m(\sum_{j=1}^{m} \frac{1}{j})}\}$,

$p$-hodnoty $p_{(i)}^{BY} = \min_{j=i,\ldots,m} \{\min\{\frac{mp_{(j)}}{j} \left(\sum_{l=1}^{m} \frac{1}{l}\right), 1\}\}$.

V následující částí budeme předpokládat, že každou hypotézu testujeme jednotlivě na hladině $q$, kde $q \in (0, 1)$. Dále budeme předpokládat, že testové statistiky odpovídající nulovým hypotézám $H_0^1, \ldots, H_0^m$ jsou nezávislé. Jako důsledek dostaneme, že BH procedura kontroluje *FDR* na hladině $q$.

**Věta 2.5.1.** *Pro nezávislé testové statistiky a pro libovolné počet platných nulových hypotéz, tedy pro libovolné $m_0$, kde $m_0 : 0 \leq m_0 \leq m$, kontroluje BH procedura FDR na hladině $q$.*

*Důkaz.* Plyne z následujícího lemmatu.

**Lemma 2.5.1.** *Pro libovolných $m_0$, $0 \leq m_0 \leq m$, nezávislých testových statistik odpovídajících $m_0$ platným nulovým hypotézám s příslušnými p-hodnotami $p_1, \ldots, p_{m_0}$ a pro $m_1 = m - m_0$ nezávislých testových statistik odpovídajících platným alternativním hypotézám s příslušnými p-hodnotami $p_{m_0+1}, \ldots, p_m$, BH procedura splňuje nerovnost*

$$\mathrm{E}(Q|P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}) \leq \frac{m_0 q}{m}. \tag{2.14}$$

*Důkaz.* Důkaz provedeme indukcí podle $m$. Pro $m = 1$ platí triviálně.
Nechť tvrzení platí pro každé $m' \leq m$, chceme ukázat platnost pro $m + 1$. Pokud $m_0 = 0$, tedy neexistuje žádná platná nulová hypotéza, je $Q = 0$ a $\mathrm{E}(Q|P_1 = p_1, \ldots, P_m = p_m) = 0 \leq \frac{m_0}{m+1} q$. Pokud $m_0 > 0$ označíme $P_i^{'},\ i = 1, \ldots, m_0$, $p$-hodnoty odpovídající platným nulovým hypotézám a nechť $P_{(m_0)}^{'}$ označuje největší z nich. Feller, viz. kapitola III část 3

v [6], dokázal, že za předpokladu nezávislosti testových statistik jsou uspořádané $p$-hodnoty odpovídající platným nulovým hypotézám, nezávislé a rovnoměrně rozdělené na intervalu $[0, 1]$.

Bez újmy na obecnosti můžeme předpokládat, že prvních $m_1$ $p$-hodnot, které odpovídají platným alternativním hypotézám, jsou uspořádány vzestupně, tedy $p_1 \leq p_2 \leq \cdots \leq p_{m_1}$. Polož

$$j_0 = \max_{j=1,\ldots,m_1} \left\{ j : p_j \leq \frac{m_0 + j}{m+1} q \right\} \tag{2.15}$$

a označíme $p'' = \frac{m_0+j_0}{m+1} q$. Podmíněním levé strany v (2.14) $P'_{(m_0)} = p$ dostaneme

$$\begin{aligned} &\mathrm{E}(Q | P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}) && (2.16)\\ &\quad = \int_0^{p''} E(Q | P'_{(m_0)} = p, P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}) f_{P'_{(m_0)}}(p)\,\mathrm{d}p && (2.17)\\ &\qquad + \int_{p''}^1 E(Q | P'_{(m_0)} = p, P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}) f_{P'_{(m_0)}}(p)\,\mathrm{d}p, && (2.18) \end{aligned}$$

kde $f_{P'_{(m_0)}}(p) = m_0 p^{m_0-1}$ je hustota odpovídající uspořádané $m_0$-té $p$-hodnotě. V prvním integrálu je $P'_{m_0} = p \leq p'' = \frac{m_0+j_0}{m+1} q$, tedy $m_0$ platných nulových a $j_0$ platných alternativních hypotéz je zamítnuto BH procedurou na hladině $q$. Tedy $Q = \frac{V}{R} = \frac{m_0}{m_0+j_0}$. Užitím nerovnosti (2.15) dostaneme

$$\begin{aligned} &\int_0^{p''} E(Q | P'_{(m_0)} = p, \ldots, P_m = p_{m_1}) m_0 p^{m_0-1}\,\mathrm{d}p && (2.19)\\ &\quad = \int_0^{p''} \frac{m_0}{m_0 + j_0} m_0 p^{m_0-1}\,\mathrm{d}p = \frac{m_0}{m_0 + j_0} \int_0^{p''} m_0 p^{m_0-1}\,\mathrm{d}p && (2.20)\\ &\quad = \frac{m_0}{m_0 + j_0} (p'')^{m_0} \leq \frac{m_0}{m_0 + j_0} \frac{m_0 + j_0}{m+1} (p'')^{m_0-1} q && (2.21)\\ &\quad = \frac{m_0}{m+1} (p'')^{m_0-1} q. && (2.22) \end{aligned}$$

Nyní se budeme zabývat druhým integrálem

$$\int_{p''}^1 E(Q | P'_{(m_0)} = p, P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}) f_{P'_{(m_0)}}(p)\,\mathrm{d}p.$$

Pro každé $j$ zvlášť, $j = 1, \ldots, m_1$, uvažujme

$$p_{j_0} < p_j \leq P'_{(m_0)} = p < p_{j+1} \ \wedge \ p_{j_0} \leq \frac{m_0 + j_0}{m+1} q < P'_{(m_0)} = p < p_{j_0+1}. \tag{2.23}$$

Kvůli způsobu, jak jsme definovali $j_0$ a $p'' = \frac{m_0+j_0}{m+1} q$, žádná hypotéza nemůže být zamítnuta pomocí $p, p_{j+1}, p_{j+2}, \ldots, p_{m_1}$. Zbytek hypotéz zamítnout můžeme. Proto jejich počet je $m - (m_1 - j + 1) = m - m_1 + j - 1 = m_0 + j - 1$. Uvažujme všechny nulové a alternativní

hypotézy dohromady. Nechť jejich $p$-hodnoty jsou srovnány vzestupně dle velikosti. Pak $H_{(i)}$ může být zamítnuta, právě když existuje $k : i \leq k \leq m_0 + j + 1$, pro které platí $p_{(k)} \leq \frac{k}{m+1}q$, což je ekvivalentní s

$$\frac{p_{(k)}}{p} \leq \frac{k}{m_0 + j + 1} \frac{m_0 + j - 1}{(m+1)p} q. \tag{2.24}$$

Za podmínky $P'_{(m_0)} = p$ jsou $\frac{P'_i}{p}$, $i = 1, \ldots, m_0 - 1$, nezávislé náhodné veličiny s rovnoměrným rozdělením na [0,1]. Čísla $\frac{p_i}{p}$, $i = 1, \ldots, j$, odpovídají platným alternativním hypotézám. Leží mezi 0 a 1, protože z (2.23) víme, že $p_j \leq P'_{(m_0)} = p$. Užitím (2.24) k testování $m' = m_0 + j - 1 \leq m$ je ekvivalentní s BH procedurou s konstantou $\frac{m_0+j-1}{(m+1)p}q$ namísto $q$ . Použitím indukčního předpokladu dostáváme

$$E(Q|P'_{(m_0)} = p, P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}) \leq \frac{m_0 - 1}{m_0 + j - 1} \frac{m_o + j - 1}{(m+1)p} q \tag{2.25}$$

$$= \frac{m_0 - 1}{(m+1)p} q. \tag{2.26}$$

Nyní už vidíme, že (2.25) závisí na $p$, ale ne pro $p \in (p_j, p_{j+1})$, pro které se integrál počítá. Dostáváme tedy

$$\begin{aligned}
&\int_0^{p''} E(Q|P'_{(m_0)} = p, P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}) f_{P'_{(m_0)}}(p)\,\mathrm{d}p \\
&\leq \int_{p''}^1 \frac{m_0 - 1}{(m+1)p} m_0 (p)^{m_0 - 1} q\,\mathrm{d}p \\
&= \frac{m_0}{m+1} q \int_{p''}^1 (m_0 - 1)(p)^{m_0 - 2}\,\mathrm{d}p \\
&= \frac{m_0}{m+1} q [1 - (p'')^{m_0 - 1}]
\end{aligned} \tag{2.27}$$

Nyní stačí použít součet nerovností (2.19) a (2.27) k dokončení důkazu lemmatu. □

*Dokončení důkazu věty 2.5.1:* stačí integrovat nerovnost (2.14) přes rozdělení testových statistik odpovídající alternativním hypotézám. Tedy označme $f_{P''}(\mathbf{p})$ hustotu sdruženého rozdělení $p$-hodnot odpovídajících alternativním hypotézám. Hustotu $f_{P''}(p)$ nedokážeme explicitně vyjádřit. Stačí ale použít horní odhad z lemma 2.5.1 na podmíněnou střední hodnotu a dále využít Fubiniho věty a faktu, že $\int_0^1 f_{P''}(p)\,\mathrm{d}p = 1$. Tedy dostáváme

$$\begin{aligned}
\mathrm{E}(Q) &= \int_{(0,1)\times\cdots\times(0,1)} \cdots \int E(Q|P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}) f_{P''}(\mathbf{p})\,\mathrm{d}\mathbf{p} \\
&\leq \frac{m_0}{m} q \prod_{i=1}^{m_1} \int_0^1 f_{P''}(p)\,\mathrm{d}p = \frac{m_0 q}{m}.
\end{aligned}$$

□

## 2.6 Adaptivní přístup ke kontrole *FDR*

Z nerovnosti $\mathrm{E}(Q|P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}) \leq \frac{m_0 q}{m} \leq q$ vyplývá, že když $m_0 < m$, tedy když předpokládáme platnost alespoň jedné alternativní hypotézy, je hladina daná BH procedurou menší než hladina, které chceme ve skutečnosti dosáhnout. Ke zvýšení síly testu je zapotřebí nějakým způsobem odhadnout počet nulových hypotéz. Pokud $m_0$ známe, pak $\mathrm{E}V = m_0 q$ a $r(q) = v + s$, kde $r, v, s$ jsou realizace náhodných veličin $R, V, S$. Odhad $\mathrm{E}Q$ je

$$\mathrm{E}Q = \frac{m_0 q}{r(q)} \tag{2.28}$$

Pokud $m_0$ neznáme, pak můžeme $V$ odhadnout pomocí odhadu $m_0$, protože $\widehat{v(q)} = \widehat{m_0} q$, kde $\widehat{v(q)}$ (resp. $\widehat{m_0}$) je odhad $v(q)$ (resp. $m_0$). Konečně odhad *FDR* je

$$\mathrm{E}\widehat{Q} = \frac{\widehat{v(q)}}{r(q)} = \frac{\widehat{m_0} q}{r(q)} \tag{2.29}$$

Nyní už můžeme vysvětlit, co pro nás znamená adaptivní přístup ke kontrole *FDR*. Jde o přístup, který je citlivý na počet platných nulových hypotéz. Následující tři body shrnují adaptivní přístup ke kontrole *FDR* v mnohonásobném testování na hladině $q$.

1. Určeme přípustnou hladinu *FDR* $q$ (závisí na typu experimentu).
2. Užijme výsledky experimentu k odhadu $m_0$ ($\widehat{m_0}$).
3. Zvolme $\alpha$, které maximalizuje $r(\alpha)$ za podmínky omezení

$$\mathrm{E}\widehat{Q} = \frac{\widehat{m_0}\alpha}{r(\alpha)} \leq q. \tag{2.30}$$

**Adaptivní BH procedura** :

Položme $k = \max_{i=1,\ldots,m} \{i : p_{(i)} \leq \frac{i}{\widehat{m_0}} q\}$.

Pokud takové $k$ neexistuje, nezamítejme žádnou hypotézu,
jinak zamítejme $H_{(i)},\ i = 1, \ldots, k$.

## 2.7 Odhad počtu nulových hypotéz, $m_0$

V tomto paragrafu se budeme snažit odhadnout počet nulových hypotéz, $m_0$. Náš odhad bude založen na nerovnosti (2.14), proto budeme opět předpokládat testování $m$ hypotéz se spojitými a vzájemně nezávislými testovými statistikami $T_1, \ldots, T_m$ s odpovídajícími $p$-hodnotami $P_1, \ldots, P_m$, které pro nás nyní budou náhodnými veličinami. Realizace náhodných veličin $P_1, \ldots, P_m$ označíme malými písmeny, tedy $p_1, \ldots, p_m$.

Pokud $m = m_0$, tj. předpokládáme platnost všech nulových hypotéz, pak množinu získaných $p$-hodnot můžeme považovat jako realizaci uspořádaného výběru z rovnoměrného

rozdělení na [0,1]. Pro střední hodnotu $i$-té nejmenší $p$-hodnoty platí, že $\mathrm{E}P_{(i)} = \frac{i}{m+1}$. Obrázek 2.2 zobrazuje graf bodů $[i, p_{(i)}],\ i = 1, \ldots, m$. Vidíme, že se chovají jako přímka procházející počátkem a bodem $[m+1, 1]$, tedy přímka procházející bodem $[0, 0]$ a směrnicí $S = \frac{1}{m+1}$.

Pokud platí $m_0 < m$, pak $p$-hodnoty odpovídající platným alternativním hypotézám jsou ve většině případů menší než $p$-hodnoty odpovídající platným nulovým hypotézám, tak že se koncentrují na levé straně grafu. Pravá strana zůstává přibližně lineární se směrnicí $S = \frac{1}{m_0+1}$, viz. obrázek 2.1. Na tomto místě musíme zmínit, že neplatí to, že všechny $p$-hodnoty odpovídající platným nulovým hypotézám musejí být vždy větší než $p$-hodnoty odpovídající platným alternativním hypotézám. Proto jsme výše uvedli ve většině případů. Toto pozorování nás vede k myšlence, že k odhadnutí $m_0$ nám bude stačit nějak odhadnout $S$. Když $\widehat{S}$ bude odhad $S$, pak $\widehat{m_0} + 1 = \frac{1}{\widehat{S}}$ bude odhad $m_0$. My ale chceme, aby $\widehat{m_0} \geq m_0$, proto položíme $\widehat{m_0} = 1/\widehat{S}$.

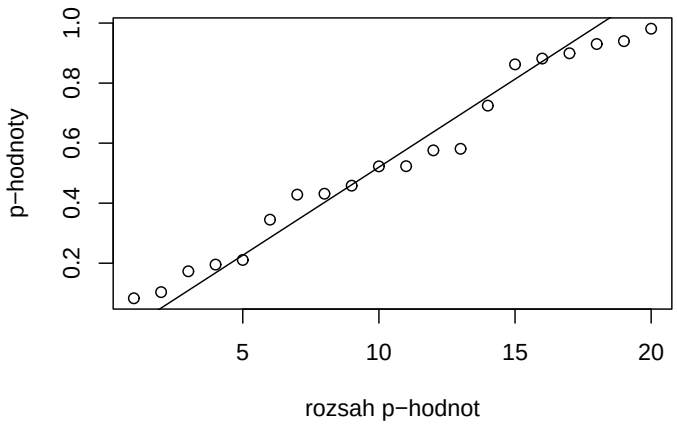
K odhadnutí $S$ použijeme následující algoritmus, který je v angličtině znám pod názvem *Lowest SLope* (*LSL*).

***LSL* algoritmus** :

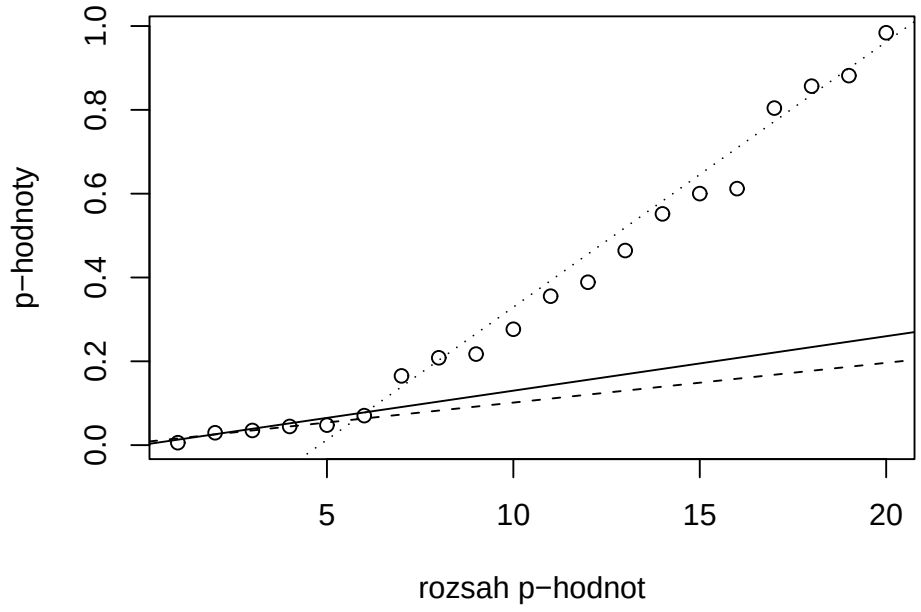
1. Počítejme $i$-tou směrnici $S_i = \frac{1-p_{(i)}}{m+1-i}$.
2. Začněme s $i = 1$ a pokračujme podle bodu 1., dokud $S_i \geq S_{i-1}$.
3. Když $S_i < S_{i-1}$ položme $\widehat{S} = S_i$ a $\widehat{m_0} = \lfloor \min(\frac{1}{S_i} + 1, m) \rfloor$,

kde symbolem $\lfloor a \rfloor$ rozumíme celou dolní část z čísla $a$.

Obrázek 2.1: *Lineární model s 20 vysvětlujícími proměnnými. Předpokládáme platnost všech nulových hypotéz, tedy případ* $m_0 = m = 20$.



Obrázek 2.2: *Lineární model s 20 vysvětlujícími proměnnými pro případ:* $m_0 = 15$ *a* $m = 20$. *Tečkovaná přímka se směrnicí* $1/(\widehat{m_0} + 1)$ *vyjadřuje lineární vztah p-hodnot odpovídajících nezamítnutým hypotézám. Čárkovaná přímka vyjadřuje lineární vztach p-hodnot odpovídajících zamítnutým hypotézám a nakonec plná přímka označuje hranici, pod kterou adaptivní BH procedura zamítá všechny hypotézy.*




## 2.8 Vlastnosti odhadu $S$ pomocí metody *LSL*

Pro tento paragraf předpokládejme to samé, co jsme předpokládali v prvním odstavci předchozího paragrafu 2.7. Nyní si ukážeme tři vlastnosti odhadu $S$ pomocí metody *LSL*:

a) Dále předpokládejme, že máme $m_0 + 1 - j$ $p$-hodnot $p_{(j)}, \ldots, p_{(m)}$ z $m_0$ $p$-hodnot odpovídajících platným nulovým hypotézám. Označme si rozdíly $g_i = p_{(i+1)} - p_{(i)}$, $i = j, \ldots, m-1$ a $g_m = 1 - p_{(m)}$. Pak $G_i$, $i = j, \ldots, m$, budou označovat náhodné veličiny, jejichž realizace jsou právě $g_i$. Pak pro $\overline{G} = \frac{1}{m+1-j} \sum_{i=j}^{m} G_i$ platí

$$\mathrm{E}\overline{G} = \frac{1}{m_0 + 1 - j} \sum_{i=j}^{m} \mathrm{E}G_i = \frac{1}{m_0 + 1 - j}[1 - \mathrm{E}P_{(j)}] \tag{2.31}$$

$$= \frac{1}{m_0 + 1 - j} - \frac{j}{(m_0 + 1)(m_0 + 1 - j)} = \frac{1}{m_0 + 1} = \mathrm{E}S_j. \tag{2.32}$$

Tedy náš odhad je nestranný.

b) Podmínka $S_i \geq S_{i-1}$ je ekvivalentní s $\frac{p_{(i)}-p_{(i-1)}}{1-p_{(i-1)}} \leq \frac{1}{m+1-(i-1)}$, neboť

$$\frac{1-p_{(i)}}{m+1-i} \geq \frac{1-p_{(i-1)}}{m+1-(i-1)} \tag{2.33}$$

$$[1-p_{(i)}][m+1-(i-1)] \geq [1-p_{(i-1)}](m+1-i) \tag{2.34}$$

$$1-p_{(i)}[m+1-(i-1)] \geq -p_{(i)}(m+1-i)+p_{(i-1)}-p_{(i)} \tag{2.35}$$

$$1-p_{(i-1)} \geq [m+1-(i-1)][p_{(i)}-p_{(i-1)}] \tag{2.36}$$

$$\frac{p_{(i)}-p_{(i-1)}}{1-p_{(i-1)}} \leq \frac{1}{m+1-(i-1)}. \tag{2.37}$$

Hodnota $\frac{1}{m+1-(i-1)}$ je střední hodnotou normovaného rozdílu splňující podmínku, že všechny $p$-hodnoty odpovídající správně nezamítnutým nulovým hypotézám jsou větší nebo rovny $p_{(j-1)}$. Tedy podmínka $S_j \leq S_{j-1}$ je ekvivalentní vypuštění malé $p$-hodnoty z odhadu $m_0$, jestliže její rozdíl od většího souseda je menší než její střední hodnota, což nás vede k podezření, že nulová hypotéza neplatí.

c) Pokud směrnice $S$ je odhadována z $p_{(i)}$ $i = j, \ldots, m$, pak pro odhad $m$ platí

$$\widehat{m_0} \geq \frac{1}{S_j} = \frac{m+1-j}{1-p_{(j)}} \geq m+1-j. \tag{2.38}$$

Tedy výsledný odhad je vždy větší než počet $p$-hodnot užitých na odhad.

## 2.9 Adaptivní BH procedura s odhadem $m_0$ pomocí směrnic

Pomocí výsledků z předchozích dvou paragrafů můžeme již zapsat adaptivní BH proceduru. Znovu předpokládáme testování $m$ hypotéz se spojitými a vzájemně nezávislými testovými statistikami $T_1, \ldots, T_m$ s odpovídajícími $p$-hodnotami $p_1, \ldots, p_m$.

**Adaptivní BH procedura** :

1. Srovnejme $p$-hodnoty podle velikosti.
2. Položme $k = \max_{i=1,\ldots,m} \{i : p_{(i)} \leq \frac{i}{m}q\}$.
   - Pokud neexistuje, nezamítejme žádnou hypotézu a skončeme.
   - Pokud existuje, počítejme směrnici $S_i = \frac{1-p_{(i)}}{m+1-i}$.
     Začněme s $i = 1$ a pokračujme, dokud $S_i \geq S_{i-1}$.

3. Pokud $S_j < S_{j-1}$, položme $\widehat{m_0} = \lfloor \min_{i=1,\dots,j} \{\frac{1}{S_j} + 1, m\} \rfloor$.
4. Položme $k = \max_{i=1,\dots,m} \{i : p_{(i)} \leq \frac{i}{\widehat{m_0}} q\}$.
5. Zamítneme $H_{(i)}$ pro $i = 1, \dots, k$.

## 2.10 Dvoufázová BH procedura

Dvoufázova BH procedura se skládá ze dvou fází, kde v první fázi provedeme základní BH proceduru s hladinou $q$ a ve druhé fázi provedeme opět BH proceduru nyní však ale již s odhadem počtu platných nulových hypotéz $m_0$. Tento počet získáme pomocí počtu zamítnutých hypotéz, které jsme získali v první fázi. Dvoufázovou proceduru můžeme motivovat následovně. Podle definice je $m_0 \leq m-(R-V)$. Pak BH procedura v prvním kroku zajistí, že $\mathrm{E}\left(\frac{V}{R}\right) \leq \frac{m_0}{m} q$. Z toho plyne, že $V$ lze shora odhadnout pomocí $q\frac{m_0}{m}R$. Tudíž

$$m_0 \leq m - \left(R - q\frac{m_0}{m}R\right), \tag{2.39}$$

což je ekvivalentní s

$$m_0 \leq \frac{m-R}{1-\frac{R}{m}q} \leq \frac{m-R}{1-q} \leq (m-R)(1+q). \tag{2.40}$$

**Dvoufázová BH procedura** :

1. Proveďme BH proceduru na hladině $q' = \frac{q}{q+1}$.
2. Nechť $r_1$ je počet zamítnutých hypotéz.
   - Pokud $r_1 = m$ zamítneme všechny hypotézy.
   - Pokud $r_1 = 0$ nezamítejme žádnou hypotézu.
3. Položme $\widehat{m_0} = (m - r_1)$.
4. Použijme BH proceduru s hladinou $q^* = \frac{m}{\widehat{m_0}} q'$.

Tato procedura kontroluje *FDR* na hladině $q$. Příslušný důkaz lze najít v [5]. Tato procedura má větší sílu, neboť počet zamítnutých hypotéz v druhém kroku nemůže být menší, než počet zamítnutých hypotéz v kroku prvním. Mohli bychom tento přístup odhadu $m_0$ dále rozšířit na $n$-fázovou proceduru, ve které by se $m_0$ počítalo pomocí $m-r_{n-1}$, dokud by se počet zamítnutých hypotéz zvyšoval.

## 2.11 Holmova a Hochbergova procedura

V této části popíšeme základní rysy Holmovy a Hochbergovy procedury zmíněné v části 2.4. Holmova procedura začíná uspořádáním $p$-hodnot, tak že $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ a

každá $p_{(i)}$ je porovnána s $\frac{\alpha}{m-i+1}$. Tedy $p_{(1)}$ je porovnána s $\frac{\alpha}{m}$, což je stejné jako u Bonferroniho procedury, ale pro $p_{(i)},\ i = 2, \ldots, m$, jsou hodnoty, se kterými porovnáváme $p_{(i)}$ větší než $\frac{\alpha}{n}$.

Holm ve své práci [7] dokázal, že jeho procedura kontroluje $FWE$ na hladině $\alpha$. Pokud testujeme hypotézy na stejné hladině, pak Holmova procedura zamítne alespoň tolik hypotéz jako Bonferroniho procedura. Tedy Holmova procedura je silnější. Pro zjednodušení položme $r_i = (m - i + 1)p_{(i)}$. Porovnávání $r_i$ a $\alpha$ je nyní ekvivalentní s porovnáváním $p_{(i)}$ a $\frac{\alpha}{m-i+1}$. Je důležité si všimnout, že $r_i$ již nemusejí být srovnány podle velikosti. Pokud $p_{(i)} = p_{(i+1)}$ pak $r_i > r_{i+1}$, ale hypotéza $H_{(i+1)}$ nemůže být zamítnuta, pokud nebude zamítnuta $H_{(i)}$. Dokonce se může stát, že $p_{(i)} < p_{(i+1)}$ ale $r_i > r_{i+1}$. Tedy test, který by zamítal hypotézu $H_{(i)}$, pokud $r_i \leq \alpha$, by byl špatný.

Holmova procedura začíná s porovnáváním od největších $p$-hodnot, dokud nenajde $i$, pro které platí, že $(m - i + 1)p_{(i)} \leq \alpha$. Pak zbývající netestované hypotézy zamítne.
Hochbergova procedura porovnává $p$-hodnoty se stejnými hodnotami jako Holmova procedura, ale začíná s porovnáváním od nejmenších $p$-hodnot. Zamítne hypotézu $H_{(i)}$, pokud $(m - j + 1)p_{(j)} \leq \alpha$ pro nějaké $j \geq i$.

Dobrým společným rysem uspořádaných $p$-hodnot založených na Holmově i Hochbergově proceduře je, že tvoří neklesající posloupnost. To znamená, že pokud hypotéza $H_j$ je více signifikantní než $H_i$, nemůže nastat, že by hypotéza $H_j$ byla méně signifikantní než hypotéza $H_i$ v $p$-hodnotách založených na Holmově nebo na Hochbergově proceduře, tj. musí platit $p^{holm}_{(j)} < p^{holm}_{(i)}$ (resp. $p^{hoch}_{(j)} < p^{hoch}_{(i)}$). Další příjemnou vlastností Hochbergových $p$-hodnot ($p^{hoch}$) je, že žádná nemůže být větší než původní $p$-hodnoty a navíc žádná nemůže být větší než jedna. To se může stát například u Bonferroniho $p$-hodnot $p^{bonf}$.

## 2.12 Úplná procedura

Hommelova procedura je založena na tzv. Úplné proceduře, proto si úplnou proceduru nyní popíšeme. Předpokládejme opět soubor nulových hypotéz $H^1_0, \ldots, H^m_0$ s nezávislými spojitými testovými statistikami $T_1, \ldots, T_m$.

Definujme nyní všechny možné kombinace těchto hypotéz takto: $H_I = \bigcap \{H^i_0 : i \in I\}$ pro všechny $I \in K$, kde $K$ je množina všech neprázdných podmnožin množiny $\{1, 2, \ldots, m\}$. Nechť pro každé $I \in K$ existuje test založený na testových statistikách $T_I$. Pro dané $\alpha$ je $H_I$ zamítnuta, pokud je na hladině $\alpha$ zamítnuta $H_I$ a každá $H_J$ pomocí testových statistik $T_J$, kde $J \in K$ a $J \supseteq I$. Pravděpodobnost, že chybně zamítneme jednu nebo více hypotéz, když testujeme všechny $H_I$, je nejvýše $\alpha$.

Tuto proceduru začneme s celkovým testem všech hypotéz $H_I = \bigcap \{H_i : i = 1, \ldots, m\}$. Pokud je test zamítnut na hladině $\alpha$, pokračujeme s testováním, stále na hladině $\alpha$, každé podmnožiny s $m - 1$ hypotéz. Pokud stále zamítáme hypotézy na hladině $\alpha$, pokračujeme s testováním, dokud nedosáhneme podmnožin o velikosti 1, tedy testu individuálních hypotéz $H_i$, $i = 1, \ldots, m$. Výše popsaná procedura se nazývá úplná procedura. V angličtině je známá pod názvem *Closed Test Procedure.*

Jak vytváříme $p$-hodnoty, podle kterých se pak rozhodujeme, zda-li hypotézu $H_I$ zamítneme,

či nikoliv, si ukážeme nyní. Ať $p_j$ jsou $p$-hodnoty odpovídající testovým statistikám $T_J$ pro všechny množiny $J \in K$, pro které platí, že $J \supseteq I$. Víme, že $H_I$ je zamítnuta, pokud $p_J \leq \alpha$ pro každé $H_J$, kde $J \supseteq I$. Proto příslušnou $p$-hodnotu pro testování $H_I$ musíme definovat jako nejmenší hodnotu z $p_J$. V nejobecnějším případě, abychom dostali příslušnou $p$-hodnotu pro jednotlivou hypotézu $H_I$, musíme získat $p$-hodnoty pro všechny možné kombinace hypotéz obsahující $H_i$. Celkový počet testů, který musíme provést je $\sum_{i=1}^{m} \binom{m}{i} = 2^m - 1$.

## 2.13 Hommelova procedura

Také v tomto paragrafu budeme předpokládat, že testujeme $m$ nulových hypotéz $H_0^1, \ldots, H_0^m$ se spojitými a vzájemně nezávislými testovými statistikami $T_1, \ldots, T_m$ s odpovídajícími $p$-hodnotami $p_1, \ldots, p_m$. Definujme celkovou hypotézu takto: $H_0 = \{H_0^1, H_0^2 \ldots, H_0^m\}$.

Hommelova procedura je založena na úplném testu, kde testujeme jednotlivé soubory hypotéz $H_I$ pomocí Simesova testu. Simes prezentoval ve své práci [10] následující proceduru. Zamítni $H_0 = \{H_0^1, H_0^2 \ldots, H_0^m\}$, pokud existuje $i$, pro které platí $p_i \leq \frac{i}{m}\alpha$. Simes dokázal, že pro nezávislé statistiky má test hladinu $\alpha$. Příslušná $p$-hodnota k $H_0$ je tedy, stejně jako v úplném testu, nejmenší hodnota z $\frac{mp_i}{i}$. Nevýhodou Simesova testu je, že neukazuje jakou konkrétní hypotézu $H_i$ zamítnout, pokud byla hypotéza $H_0$ zamítnuta. Hommelova procedura, jak již bylo zmíněno výše, používá v úplném testu Simesův test, navíc výrazně zmenšuje celkový počet testů, který se musí provést v úplném testu. Hommelovu proceduru lze zapsat následovně:

**Hommelova procedura** :

Nechť $k$ je počet hypotéz největší podmnožiny hypotéz, pro kterou Simesův test není signifikantní, tj.

$$k = \max_{j=1,\ldots,m} \left\{ j : p_{(m-j+l)} > \frac{l\alpha}{j} \text{ pro } l = 1, \ldots, j \right\} \tag{2.41}$$

Pokud takové $k$ neexistuje, pak zamítneme všechny hypotézy.
Jinak zamítneme $H_0^i$, pokud $p_i \leq \frac{\alpha}{k}$.

Počítání příslušných Hommelových $p$-hodnot si ukážeme na úplném testu. Úplný test vyžaduje, aby se pro každou hypotézu $H_i,\ i = 1, \ldots, m$, získaly Simesovy $p$-hodnoty pro každou podmnožinu hypotéz obsahující hypotézu $H_i$. Hommelova $p$-hodnota je pak největší z těchto Simesových $p$-hodnot. Navíc není třeba testovat každou podmnožinu obsahující hypotézu $H_i$, abychom získali příslušnou Hommelovu $p$-hodnotu. Pro podmnožinu obsahující $n$ z $m$ hypotéz stačí testovat podmnožinu obsahující $m - 1$ největších $p$-hodnot kromě $p_i$. Bude-li tato množina signifikantní na hladině $\alpha$, pak musejí být signifikantní i všechny ostatní podmnožiny o velikosti $n$. Tuto množinu si nazvěme *nejméně signifikantní*. Spočteme si Simesovy $p$-hodnoty pro každou nejméně signifikantní množinu o velikosti $n$ pro $n = 1, \ldots, m$. Maximum z těchto $p$-hodnot je pak příslušná Hommelova $p$-hodnota k hypotéze $H_0^i$.

# Kapitola 3

# Kontrola *FDR* pro závislé statistiky

Doposud jsme předpokládali, že testové statistiky, jejichž $p$-hodnoty jsme používali ke kontrole *FDR*, byly nezávislé. V praxi se však se závislými testovými statistikami setkáváme častěji než s nezávislými. Cílem této části práce bude odvození procedury, která kontroluje *FDR* pro závislé testové statistiky. Ukážeme si, že BH procedura kontroluje *FDR* jak pro pozitivně korelované normálně rozdělené statistiky, tak i pro Studentizované statistiky, což nám umožní zkoumat lineární model pro závislé vysvětlující proměnné. Dále budeme předpokládat, že alespoň jedna z nulových hypotéz $H_0^1, \ldots, H_0^m$ není platná, protože jinak by *FDR* bylo identicky rovno jedné.

## 3.1 Pozitivně regresní závislost na $I_0$

***Definice* 1.** Řekneme, že množina $D$ je *rostoucí*, pokud pro každé $x \in D$ a $y \geq x$ platí, že $y \in D$. Řekneme, že množina $D$ je *klesající*, pokud $D^C$ je rostoucí.

Nechť $\mathbf{X}$ je náhodný vektor o složkách $X_1, \ldots, X_m$. Nechť množina $I_0$ označuje jeho nějakou podmnožinu složek, tedy $I_0 \subseteq \{X_1, X_2, \ldots, X_m\}$.

***Definice* 2.** Řekneme, že náhodný vektor $\mathbf{X}$ má vlastnost *PRD*, pokud pro libovolnou rostoucí množinu $D$ platí, že $\mathrm{P}(\mathbf{X} \in D | X_1 = x_1, \ldots, X_m = x_m)$ je neklesající funkcí v $(x_1, \ldots, x_m)$.

***Definice* 3.** Řekneme, že náhodný vektor $\mathbf{X}$ má vlastnost *PRDS* na $I_0$, pokud pro libovolnou rostoucí množinu $D$ a pro každou náhodnou veličinu $X_i \in I_0$ platí, že $\mathrm{P}(\mathbf{X} \in D | X_i = x_i)$ je neklesající funkcí proměnné $x_i$.

Vlastnost *PRD* se nazývá pozitivní závislost a zkratka vychází z anglického výrazu *positive regression dependency*. Je zřejmé, že vlastnost *PRDS* je slabší než vlastnost *PRD*, neboť v *PRDS* podmiňujeme pouze jednou náhodnou veličinou a navíc množina $I_0$ nemusí obsahovat všechny náhodné veličiny $X_1, \ldots, X_m$. Postačující podmínkou pro *PRD* je podmínka $MTP_2$.

***Definice* 4.** Řekneme, že náhodný vektor $\mathbf{X}$ má vlastnost $MTP_2$, pokud pro každé $\mathbf{x}$, $\mathbf{y}$ platí

$$f(\mathbf{x}) \cdot f(\mathbf{y}) \leq f[\min(\mathbf{x}, \mathbf{y})] \cdot f[\max(\mathbf{x}, \mathbf{y})], \tag{3.1}$$

kde $f$ je sdružená hustota náhodného vektoru $\mathbf{X}$.

*Poznámka* 3.1.1. Maximum a minimum je v definici 4 počítáno po složkách.

Nyní uvedeme několik nutných podmínek pro vlastnost *PRDS*. Důkazy těchto implikací lze nalézt např. v [9]. Platí, že má-li vektor $\mathbf{X}$ vlastnost $MTP_2$, pak má i vlastnost *PRD* a tedy i *PRDS* na každé podmnožině $I_0 \subseteq I = \{X_1, \ldots, X_m\}$. Poslední podmínkou implikující vlastnost *PRDS*, kterou si zde uvedeme, je *podmíněná asociovanost* vektoru $\mathbf{X}$.

***Definice* 5.** Řekneme, že rozdělení náhodného vektoru $\mathbf{X}$ je *pozitivně asociované*, pokud pro libovolné buď obě rostoucí, nebo obě klesající funkce $f$ a $g$ v každé proměnné platí, že $\mathrm{cov}(f(\mathbf{X}), g(\mathbf{X})) \geq 0$.

***Definice* 6.** Řekneme, že rozdělení vektoru $\mathbf{X}$ je *podmíněně asociované*, jesliže pro libovolné dělení $(\mathbf{X}_1, \mathbf{X}_2)$ vektoru $\mathbf{X}$ a pro libovolnou funkci $h(\mathbf{X}_1)$ platí, že $\mathbf{X}$ při daném $h(\mathbf{X_1})$ je pozitivně asociované.

Nakonec si uvedeme ekvivalentní definici *PRDS*, která je na první pohled zřejmá, ale bude se nám hodit v důkaze věty 3.1.1. Řekneme, že náhodný vektor $\mathbf{X}$ má vlastnost *PRDS* na $I_0$, pokud pro libovolnou klesající množinu $C$ a pro každou náhodnou veličinu $X_i \in I_0$ platí, že $\mathrm{P}(\mathbf{X} \in C | X_i = x_i)$ je nerostoucí funkcí proměnné $x_i$. Nyní jsme schopni vyslovit hlavní větu této části.

**Věta 3.1.1.** *Pokud sdružené rozdělení testových statistik má vlastnost PRDS v podmnožině testových statistik odpovídajících platné nulové hypotéze, pak BH procedura kontroluje FDR na hladině menší nebo rovné* $\frac{m_0}{m}q$.

*Důkaz.* Důkaz bude uveden v paragrafu 3.3.

Naším dalším cílem bude najít statistiky, které mají vlastnost *PRDS*. Zejména by se nám hodilo, kdyby statistika se Studentovým rozdělením měla vlastnost *PRDS*.

## 3.2 Rozdělení mající vlastnost *PRDS*

a) **Mnohorozměrné normální statistiky**

Uvažujme náhodný vektor $\mathbf{X} = (X_1, \ldots, X_m)$ testových statistik testující nulovou hypotézu $\mu_i = 0$ proti jednostranné alternativě $\mu_i > 0$ pro $i = 1, \ldots, m$.

**Lemma 3.2.1.** *Předpokládejme, že* $\mathbf{X} \sim \mathbf{N}(\boldsymbol{\mu}, \mathbf{V})$. *Nechť* $I_0$ *je množina náhodných veličin* $X_i$, *pro které platí nulová hypotéza. Dále předpokládejme, že pro* $X_i \in I_0$, *a pro každé* $j \neq i$ *platí, že* $\mathbf{V}_{i,j} = \mathrm{cov}(X_i, X_j) \geq 0$, *pak rozdělení náhodného vektoru* $\mathbf{X}$ *má vlastnost PRDS na* $I_0$.

*Důkaz.* Vezmeme $X_i \in I_0$ a zavedeme následující označení: $\mathbf{X}_{(i)}$ je podvektor vektoru $\mathbf{X}$ vzniklý vyškrtnutím $i$-té složky a $\boldsymbol{\mu}_{(i)}$ jeho vektor středních hodnot, $\mathbf{V}_{(i,i)}$ je podmatice $\mathbf{V}$ bez $i$-tého sloupce i řádku a $\mathbf{V}_{(i),i} = \mathrm{cov}(X_i, \mathbf{X}_{(i)})$. Z věty 4.12 v [1] plyne, že $\mathbf{X}_{(i)}$ při daném $X_i = x_i$ má $\mathbf{N}(\boldsymbol{\mu}^{(i)}, \mathbf{V}^{(i)})$, kde

$$\boldsymbol{\mu}^{(i)} = \boldsymbol{\mu}_{(i)} + \mathbf{V}_{(i),i}\mathbf{V}_{i,i}{}^{-1}(x_i - \mu_i) \quad a \quad \mathbf{V}^{(i)} = \mathbf{V}_{(i,i)} - \mathbf{V}_{(i),i}\mathbf{V}_{i,i}{}^{-1}\mathbf{V}'_{(i),i}$$

Protože jde o funkci proměnné $x_i$, pak $\mathrm{E}(\mathbf{X}_{(i)}|X_i = x_i)$ je rostoucí. Tedy pokud $x_i \leq x'_i$, pak

$$\mathrm{E}[f(\mathbf{X}_{(i)})|X_i = x_i] \leq \mathrm{E}[f(\mathbf{X}_{(i)})|X_i = x'_i]$$

Tudíž vektor $\mathbf{X}$ má vlastnost *PRDS* na $I_0$.

*Poznámka* 3.2.1. Protože vybíráme $X_i$ pouze z množiny $I_0$, tedy $X_i$ odpovídající nulovým hypotézám, podmínka $\mathbf{V}_{(i),i} > 0$ nezávisí na kovariancích mezi $X_i$ odpovídající alternativním hypotézám.

b) **Absolutní hodnota z mnohorozměrného studentizovaného-rozdělení**

K odvození dalšího rozdělení splňující vlastnost *PRDS* potřebujeme následující lemma.

**Lemma 3.2.2.** *Pokud*

*a)* $\mathbf{Y}$ *je náhodný vektor se spojitým rozdělením mající vlastnost PRDS v* $I_0$*,*

*b)* $U$ *je nezávislá náhodná veličina se spojitým rozdělením,*

*c) pro složky vektoru* $\mathbf{X}$ *platí, že* $X_j = g_j(Y_j, U)$*,* $j = 1, \ldots, m$*, je rostoucí spojitá funkce proměnných* $Y_j$ *a* $U$*,*

*d) pro* $X_i \in I_0$ *mají* $U$ *a* $Y_i$ *vlastnost PRDS v proměnné* $X_i$*,*

*pak* $\mathbf{X}$ *má vlastnost PRDS na* $I_0$*.*

*Poznámka* 3.2.2. Je důležité si všimnout, že podmínka (d) závisí jak na transformaci $g_i$, tak i na rozdělení $Y_i$ a $U$.

*Příklad.* Nechť $U_0$ a $U_1$ jsou nezávislé náhodné veličiny s rozdělením $\chi^2$ a $W = U_0 \cdot U_1$. Ukážeme, že $U_1$ má vlastnost *PRDS* na $W$ pomocí podmínky $MTP_2$. Nechť $f_{U_0}$ (resp. $f_{U_1}$) označuje hustotu náhodné veličiny $U_0$ (resp. $U_1$). Pak z nezávislosti plyne, že sdružená hustota $f_{U_0,U_1}$ je $f_{U_0,U_1} = f_{U_0} f_{U_1}$. Zavedeme transformaci $g(U_0, U_1) \longmapsto (Y, W)$, kde $Y = U_0$. Jakobián transformace je $J_g = \begin{vmatrix} 1 & 0 \\ U_1 & U_0 \end{vmatrix} = U_0$, tedy $J_{g^{-1}} = \frac{1}{U_0}$. Podle věty o transformaci náhodného vektoru je hustota náhodného vektoru $(U_0, W)$

$$f_{U_0,W}(x_1, x_2) = \frac{1}{x_1} f_{U_0}(x_1) f_{U_1}(\frac{x_2}{x_1}) = \frac{C {x_2}^{(n/2-1)}}{x_1} \exp\left\{-\frac{x_2 + {x_1}^2}{2x_1}\right\}, \tag{3.2}$$

kde $C = \frac{1}{2^n \Gamma^2(n/2)}$ a $x_1, x_2 > 0$. Nyní je už snadné ověřit, že pro $f_{U_0,W}$ platí podmínka $MTP_2$, tedy $(U_0, W)$ má vlastnost *PRDS*.
Analogicky pro $U_1$. Navíc lze dokázat, že i $U_i^{-1},\ i = 1, 2$, mají vlastnost *PRDS* na $W$.

**Lemma 3.2.3.** *Nechť náhodný vektor* $\mathbf{Y}$ *má mnohorozměrné normální rozdělení a* $|\mathbf{Y}|$ *má vlastnost PRDS na* $I_0$*, kde* $I_0$ *je množina testových statistik testující* $\mu_i = 0$*. Pokud náhodná veličina* $S^2$ *má rozdělení* $\chi^2_\nu$ *a je nezávislá s* $\mathbf{Y}$*, pak* $|\mathbf{X}| = \frac{|\mathbf{Y}|}{S}$ *má vlastnost PRDS na* $I_0$*.*

*Důkaz.* Využijeme předcházejícího příkladu 3.2, položíme $U_0 = |Y_i|^2$ a $U_1 = \frac{1}{S^2}$. K dokončení důkazu stačí použít lemma 3.2.2.

Dokončení odvození vlastnosti *PRDS* pro absolutní hodnotu z mnohorozměrného studentizovaného rozdělení je už jen otázkou aplikace předchozího lemma 3.2.3. Uvažujme náhodný vektor $\mathbf{Y} \sim N(\boldsymbol{\mu}, \sigma^2\mathbf{I})$, kde $\sigma^2 > 0$ je neznámý parametr. Mějme oboustranný test: $\mu_i = 0$ proti alternativě $\mu_i \neq 0$. Testové statistiky získáme dělením $|\mathbf{Y}|$ s nezávislým odhadem $S$ parametru $\sigma^2$ s $\nu$ stupni volnosti, to znamená $\frac{\nu S^2}{\sigma^2} \sim \chi^2_\nu$. Podle lemma 3.2.3 platí, že pokud $|\mathbf{Y}|$ má vlastnost *PRDS* na množině testových statistik testující nulovou hypotézu, pak $\frac{|\mathbf{Y}|}{S}$ má také vlastnost *PRDS* na stejné množině.

c) **Studentizované mnohorozměrné normální rozdělení**

Stejně jako v případě (b) uvažujme náhodný vektor $\mathbf{Y} \sim N(\boldsymbol{\mu}, \sigma^2\mathbf{I})$, kde $\sigma^2$ je neznámý parametr a $S$ je nezávislý odhad parametru $\sigma^2$ s $\nu$ stupni volnosti. Položme $\mathbf{T} = \frac{\mathbf{Y}}{S}$. Chápeme-li $T_i$ jako funkci proměnné $S$, pak její monotónnost se mění se změnou znaménka $\operatorname{sgn} Y_i$. Není tedy splněna podmínka (c) v lemma 3.2.2. V článku [4] je dokázáno, že pokud má náhodný vektor $\mathbf{Y}$ vlastnost *PRDS* a hladina testu je $q \leq 1/2$, pak BH procedura použitá na $\frac{\mathbf{Y}}{S}$ kontroluje *FDR*.

## 3.3 Důkaz věty 3.1.1

Tato část se bude skládat pouze z důkazu věty 3.1.1.
*Důkaz.* Nejprve si označme konstanty vyskytující se v BH proceduře $q_i = \frac{i}{m}q,\ i = 1, 2 \ldots, m$. Nechť $A_{v,s}$ je jev, že BH procedura zamítne právě $v$ platných nulových a $s$ platných alternativních hypotéz. Pak *FDR* lze psát

$$\mathrm{E}Q = \sum_{s=0}^{m_1} \sum_{v=1}^{m_0} \frac{v}{v+s} \mathrm{P}(A_{v,s}), \tag{3.3}$$

kde $m_0$ je počet platných nulových hypotéz a $m_1 = m - m_0$ je počet platných alternativních hypotéz.

**Lemma 3.3.1.** *Pro libovolné $v, s \in \{1, \ldots, m\}$ platí,*

$$\mathrm{P}(A_{v,s}) = \frac{1}{v} \sum_{i=1}^{m_0} \mathrm{P}(\{P_i \leq q_{v+s}\} \cap A_{v,s}) \tag{3.4}$$

*Důkaz Lemma.* Nechť $v$ a $s$ jsou pevné a nechť $\omega \subseteq \{1, \ldots, m\}$ a $|\omega| = v$. Nechť $[A^{\omega}_{v,s}] \subseteq [A_{v,s}]$ označuje jev, že všech $v$ zamítnutých platných nulových hypotéz je v $\omega$. Platí, že $\mathrm{P}(\{P_i \leq q_{v+s}\} \cap A^{\omega}_{v,s}) = \mathrm{P}(A^{\omega}_{v,s})$ pro $i \in \omega$, jinak je nula. Pak

$$\sum_{i=1}^{m_0} \mathrm{P}(\{P_i \leq q_{v+s}\} \cap A_{v,s}) = \sum_{i=1}^{m_0} \sum_{\omega} \mathrm{P}(\{P_i \leq q_{v+s}\} \cap A^{\omega}_{v,s}) \tag{3.5}$$

$$= \sum_{\omega} \sum_{i=1}^{m_0} \mathrm{P}(\{P_i \leq q_{v+s}\} \cap A^{\omega}_{v,s}) \tag{3.6}$$

$$= \sum_{\omega} \sum_{i=1}^{m_0} \mathrm{I}(i \in \omega) \mathrm{P}(A^{\omega}_{v,s}) \tag{3.7}$$

$$= \sum_{\omega} v \mathrm{P}(A^{\omega}_{v,s}) = v \mathrm{P}(A_{v,s}), \tag{3.8}$$

Pomocí (3.3) a lemma 3.3.1 lze *FDR* psát jako

$$\mathrm{E}Q = \sum_{s=0}^{m_1} \sum_{v=1}^{m_0} \frac{v}{v+s} \left\{ \sum_{i=0}^{m_0} \frac{1}{v} \mathrm{P}(\{P_i \leq q_{v+s}\} \cap A_{v,s}) \right\} = \tag{3.9}$$

$$= \sum_{i=0}^{m_0} \left\{ \sum_{s=0}^{m_1} \sum_{v=1}^{m_0} \frac{1}{v+s} \frac{1}{v+s} \mathrm{P}(\{P_i \leq q_{v+s}\} \cap A_{v,s}) \right\}. \tag{3.10}$$

Nyní závislost $\mathrm{E}Q$ na $v$ je jen prostřednictvím $A_{v,s}$.

Nechť $\mathbf{P}^{(i)}$, $i = 1, \ldots, m$, je náhodná veličina označující zbývajících $m-1$ $p$-hodnot bez $P_i$, tedy $\mathbf{P}^{(i)} = (P_1, \ldots, P_{i-1}, P_{i+1}, \ldots, m)$. Dále ať $C^{(i)}_{v,s}$ označuje jev, kde pokud $P_i$ je zamítnuto, pak jen v případě, že je zamítnuto právě $v-1$ platných nulových hypotéz a $s$ platných alternativních hypotéz. Pak $C^{(i)}_{v,s}$ je projekce $\{P_i \leq q_{v+s}\} \cap A_{v,s}$ na podprostor generovaný $\mathbf{P}^{(i)}$. Tedy máme

$$\{P_i \leq q_{v+s}\} \cap A_{v,s} = \{P_i \leq q_{v+s}\} \cap C^{(i)}_{v,s}. \tag{3.11}$$

Označme $C^{(i)}_k = \bigcup \{C^{(i)}_{v,s} : v + s = k\}$. Navíc pro každé $i \neq j$ je $C^{(i)}_k \cap C^{(j)}_k = \emptyset$, tak že *FDR* lze psát jako

$$\mathrm{E}Q = \sum_{i=1}^{m_0} \sum_{k=1}^{m} \frac{1}{k} \mathrm{P}(P_i \leq q_k \cap C^{(i)}_k), \tag{3.12}$$

kde $\mathrm{E}Q$ už nezávisí na $v$ ani na $s$. V poslední části důkazu zkonstruujeme systém rostoucích množin, ve kterých využijeme vlastnost *PRDS*, abychom mohli omezit vnitřní součet v

(3.12) pomocí $\frac{q}{m}$. Pro tento účel definujme $D_k^{(i)} = \bigcup\{C_j^{(i)} : j \leq k\}$ pro $k = 1, \ldots, m$. Množiny $D_k^{(i)}$ lze také popsat pomocí uspořádaných $p$-hodnot náhodné veličiny $\mathbf{p^{(i)}}$, tj. $\{p_{(1)}^{(i)} \leq \cdots \leq p_{(m)}^{(i)}\}$, následujícím způsobem:

$$D_k^{(i)} = \left\{\mathbf{p} : q_{k+1} < p_{(k)}^{(i)}, q_{k+2} < p_{(k+1)}^{(i)} \ldots, q_m < p_{m-1}^{(i)}\right\} \tag{3.13}$$

pro $k = 1, \ldots, m-1$. Důvod, proč jsme si vyjadřovali $D_k^{(i)}$ v (3.13) je takový, že $D_k^{(i)}$ jsou pro každé $k$ neklesající množiny. Nyní využijeme vlastnosti *PRDS*, která tvrdí, že pro $p \leq p'$ platí:

$$\mathrm{P}(D|P_i = p) \leq \mathrm{P}(D|P_i = p'). \tag{3.14}$$

Dále je snadné dokázat, že pro $j \leq l$ a tedy $q_j \leq q_l$ platí:

$$\mathrm{P}(D|P_i = q_j) \leq \mathrm{P}(D|P_i = q_l) \tag{3.15}$$

pro libovolnou neklesající množinu $D$, nebo ekvivalentně

$$\frac{\mathrm{P}(\{P_i \leq q_k\} \cap D_k^{(i)})}{\mathrm{P}(P_i \leq q_k)} \leq \frac{\mathrm{P}(\{P_i \leq q_{k+1}\} \cap D_k^{(i)})}{\mathrm{P}(P_i \leq q_{k+1})}. \tag{3.16}$$

Využitím (3.16) a faktu, že $D_{j+1}^{(i)} = D_j^{(i)} \cup C_{j+1}^{(i)}$ platí pro každé $k \leq m-1$

$$\frac{\mathrm{P}(\{P_i \leq q_k\} \cap D_k^{(i)})}{\mathrm{P}(P_i \leq q_k)} + \frac{\mathrm{P}(\{P_i \leq q_k\} \cap C_{k+1}^{(i)})}{\mathrm{P}(P_i \leq q_{k+1})} \tag{3.17}$$

$$\leq \frac{\mathrm{P}(\{P_i \leq q_{k+1}\} \cap D_k^{(i)})}{\mathrm{P}(P_i \leq q_{k+1})} + \frac{\mathrm{P}(\{P_i \leq q_k\} \cap C_{k+1}^{(i)})}{\mathrm{P}(P_i \leq q_{k+1})} \tag{3.18}$$

$$= \frac{\mathrm{P}(\{P_i \leq q_k\} \cap D_{k+1}^{(i)})}{\mathrm{P}(P_i \leq q_{k+1})}. \tag{3.19}$$

Nyní položme $C_1 = D_1$. Opakujme hořejší nerovnost pro $i = 1, \ldots, m-1$, dokud součet na levé straně nepřejde v jeden člen, tj. dokud

$$\sum_{k=1}^{m} \frac{\mathrm{P}(\{P_i \leq q_k\} \cap C_k^{(i)})}{\mathrm{P}(P_i \leq q_k)} \leq \frac{\mathrm{P}(\{P_i \leq q_m\} \cap D_k^{(i)})}{\mathrm{P}(P_i \leq q_m)} = 1, \tag{3.20}$$

kde poslední rovnost plyne z faktu, že $D_m^{(i)}$ je celý prostor. Vratíme-li se zpět k výrazu (3.12), pak pro *FDR* platí

$$\mathrm{E}Q = \sum_{i=1}^{m_0} \sum_{k=1}^{m} \frac{1}{k} \mathrm{P}(P_i \leq q_k \cap C_k^{(i)}) \leq \sum_{i-1}^{m_0} \sum_{k=1}^{m} \frac{\mathrm{P}(\{P_i \leq q_k\} \cap C_k^{(i)})}{\mathrm{P}(P_i \leq q_k)}. \tag{3.21}$$

Víme, že za platnosti nulové hypotézy platí, že $\mathrm{P}(P_i \leq q_k) \leq q_k = \frac{k}{m}q$. Pro spojité testové statistiky dostáváme dokonce rovnost, protože pak $P_i$ jsou nezávislé, rovnoměrně rozdělené. Konečně pomocí (3.20) dostáváme, že

$$\frac{q}{m}\sum_{i=1}^{m_0}\sum_{k=1}^{m}\frac{\mathrm{P}(\{P_i \leq q_k\} \cap C_k^{(i)})}{\mathrm{P}(P_i \leq q_k)} \leq \frac{m_0}{m}q. \tag{3.22}$$

□

*Poznámka* 3.3.1. Z důkazu je zřejmé, že pokud testové statistiky jsou nezávislé, pak *FDR* lze vyjádřit:

$$\mathrm{E}Q = \sum_{i=1}^{m_0}\sum_{k=1}^{m}\frac{1}{k}\mathrm{P}\left(\left\{P_i \leq \frac{k}{m}q\right\} \cap C_k^{(i)}\right) \tag{3.23}$$

$$= \sum_{i=1}^{m_0}\sum_{k=1}^{m}\frac{1}{k}\mathrm{P}\left(P_i \leq \frac{k}{m}q\right)\mathrm{P}(C_k^{(i)}) \tag{3.24}$$

$$= \sum_{i=1}^{m_0}\frac{\alpha}{m}\sum_{k=1}^{m}\mathrm{P}(C_k^{(i)}) = \frac{m_0}{m}q, \tag{3.25}$$

což je alternativní důkaz věty 2.5.1.

# Kapitola 4

# Simulace

V této části budeme porovnávat procedury mnohonásobného testování. Zaměříme se zejména na rozdíly mezi procedurami kontrolující *FDR* a *FWE* na stejné hladině, tedy položíme $q = \alpha$. Pro všechny simulace bude platit, že hladina testu je $\alpha = 0.05$. K porovnání procedur použijeme lineární model. Budeme testovat $H_1^0, \ldots, H_m^0$ nulových hypotéz testujících nulovost regresních koeficientů jak pro nezávislé testové statistiky, tak pro závislé statistiky. Pro nezávislé statistiky jde vlastně o aplikaci věty 5.8 str. 83 v [1]. Pro závislé statistiky použijeme výsledky z části 3.1. Lineární model budeme počítat pomocí funkce $lm(y \sim x_1 + \cdots + x_m)$ ve výpočetním prostředí **R**, kde $y$ je závislá proměnná a $x_1, \ldots, x_m$ jsou vysvětlující proměnné. Ve všech zde uvedených simulacích platí, že závislá proměnná $y$ nezávisí na vysvětlujících proměnných $x_1, \ldots, x_m$ pro všechny testované hypotézy. To znamená, že $y$ není funkcí $x_1, \ldots, x_m$. Název *závislá* proměnná pochází z toho, že se ji snažíme co nejlépe aproximovat pomocí $x_1, \ldots, x_m$.

## 4.1 Porovnání procedur kontrolující *FDR* a *FWE* pro nezávislé statistiky

V této části vyžadujeme, aby testové statistiky byly vzájemně nezávislé. Nezávislost testových statistik bude splněna, pokud $X_i,\ i = 1, \ldots, m$, budou nezávislé náhodné výběry. K tomuto použijeme funkce *rnorm*, která generuje náhodný výběr z normálního rozdělení. Testované hypotézy budou tvaru:

$$\text{nulové hypotézy } H_0^i : \beta_i = 0 \text{ proti oboustranným alternativám } H_1^i : \beta_i \neq 0,$$

kde $\beta_i$ je $i$-tý regresní koeficient pro $i = 1, \ldots, m$. Víme, že platí všechny nulové hypotézy, tedy $m_0 = m$. První simulace provedeme pro pevný počet opakování, který bude roven 100 000 pro 10 a 6 vysvětlujících proměnných. Zaměříme se na dvě otázky.

1. V kolika opakováních bude existovat alespoň jedna $p$-hodnota menší než konvenční hladina $\alpha$.

2. V kolika opakováních nezamítneme ani jednu hypotézu pomocí procedur mnohonásobného testování, existuje-li alespoň jedna $p$-hodnota menší než $\alpha$.

Z teoretické části víme, že chápeme-li $p$-hodnoty jako náhodné veličiny, pak za platnosti nulové hypotézy jsou nezávislé a mají rovnoměrné rozdělení na intervalu $[0, 1]$. Nechť náhodná veličina $P_i$ označuje $i$-tou $p$-hodnotu. Pak platí $\mathrm{P}(P_i \leq \alpha) = \alpha$. Proto

$$\mathrm{P}(\nexists\, i \in \{1, \ldots, m\} :\ P_i \leq \alpha) = \mathrm{P}\left(\bigcap_{i=1}^{m} P_i > \alpha\right) = \prod_{i=1}^{m} \mathrm{P}(P_i > \alpha) = (1-\alpha)^m \tag{4.1}$$

Odpověď na naši první otázku snadno plyne z (4.1), tedy

$$\mathrm{P}(\exists\, i \in \{1, \ldots, m\}) = 1 - (1-\alpha)^m. \tag{4.2}$$

Dosadíme-li do (4.2) za $\alpha = 0.05$ a položíme-li $m = 10$ (resp. $m = 6$) dostaneme 0.4 (resp. 0.265). Teoretickou úvahu potvrzují výsledky simulací uvedených v tabulce 4.1 (resp. 4.2).

Druhou otázku zodpovíme pomocí výsledků simulací uvedených v tabulkách 4.1 a 4.2. Existuje-li alespoň jedna $p$-hodnota menší než konvenční hladina $\alpha$, pak k zamítnutí alespoň jedné hypotézy procedurami mnohonásobného testování dojde přibližně ve 19.5% pro 6 vysvětlujících proměnných. Pro 10 vysvětlujících proměnných dostáváme jen 12%. Je tedy zřejmé, že s rostoucím počtem proměnných toto číslo klesá.

V následujících simulacích se budeme zabývat rozdílem mezi procedurami kontrolujícími *FDR* a procedurami kontrolujícími *FWE*. Tento rozdíl bude nejvíce patrný, když budeme sledovat, kolik jednotlivé procedury mnohonásobného testování zamítnou hypotéz, když počet zamítnutých hypotéz na konvenční hladině $\alpha$ bude pevně daný a stejný pro všechny simulace. Tento počet bude roven 100. Při tomto počtu již všechny simulace vycházejí stejně a navíc lze poměrně snadno dosáhnout zamítnutí právě poloviny platných nulových hypotéz na konvenční hladině $\alpha$ pro $6, 8$ a 10 vysvětlujících proměnných. Padesáti procenty všech zamítnutých hypotéz budeme dále myslet to, že právě ve 100 případech dostaneme právě 5 $p$-hodnot menších než $\alpha$. Z tabulky 4.2 vidíme, že pro 50% všech zamítnutých nulových hypotéz, tedy pro 5 zamítnutých hypotéz z 10 nulových hypotéz pro 10 vysvětlujících proměnných, je v simulaci nutné provést více než 200 000 opakování. Pokud bychom chtěli dosáhnout hodnoty 60% všech zamítnutých nulových hypotéz, počet opakování by se pohyboval okolo 3 000 000.

Pro porovnání jednotlivých procedur zavedeme veličinu počítající poměr mezi počtem procedurami nezamítnutých hypotéz za podmínky, že počet zamítnutých hypotéz na hladině $\alpha$ je pevné číslo a počtem všech hypotéz. Označme si tuto veličinu jako $r_i$, kde index $i$ vyjadřuje pevný počet zamítnutých hypotéz. Tedy

$$r_i = 1 - \frac{\text{počet procedurami nezamít. h.}}{\text{počet všech h.}} \tag{4.3}$$

$$= 1 - \frac{\text{počet všech zamít. h.} - \text{počet procedurami zamít.h.}}{\text{počet všech h.}}, \tag{4.4}$$

kde $i = 1, \ldots, m$.

Jak se počítají hodnoty veličiny $r_i$ si ukážeme na následujícím příkladě.

*Příklad.* Spočteme hodnotu $r_5$ pro BH proceduru pro 10 vysvětlujících proměnných. Proto si ji označme jako $r_5^{BH}$. Nechť máme v každém ze 100 případů právě 5 z 10 $p$-hodnot menších než konvenční hladina $\alpha$. Na konvenční hladině $\alpha$ zamítneme právě 5 z 10 nulových hypotéz. Počet všech zamítnutých nulových hypotéz na konvenční hladině je $5 \cdot 100$. Počet všech nulových hypotéz je ale $10 \cdot 100$. Pak

$$r_5^{BH} = 1 - \frac{\sum_{i=1}^{10}(5-i) \cdot k_i^{BH} \cdot \operatorname{sgn}\{5-i\}}{10 \cdot 100}, \tag{4.5}$$

kde $k_i^{BH}$ je počet, kolikrát ze 100 případů BH procedura zamítla právě $i$ hypotéz.

Následující tabulky 4.3, 4.4, 4.5 zobrazují výsledky simulací jednotlivých procedur pro $6, 8$ a 10 vysvětlujících proměnných. Z uvedené teorie vyplývá, že s rostoucím počtem zamítnutých nulových hypotéz by měl růst i rozdíl mezi procedurami kontrolující *FDR* a *FWE*. Tento fakt je patrný zejména z tabulky 4.5 a grafu 4.1 pro 5 a 6 zamítnutých hypotéz při počtu 10 vysvětlujících proměnných. Dále je z tabulek zřejmé, že nejvíce hypotéz je zamítnuto adaptivní BH procedurou. Jedinou procedurou, která kontroluje *FDR* a nezamítá více hypotéz než procedury kontrolující *FWE* je BY procedura. Této proceduře jsme se ale nevěnovali ani v teoretické části, proto se s ní nebudeme zabývat ani teď. Rozdíly, které vznikly mezi BH procedurou a Hochbergovou procedurou (popř. Holmovou procedurou), vyplývají z toho, že posloupnost lineárně klesajících konstant u BH procedury je vždy větší než hyperbolicky klesající konstanty u Hochbergovy procedury, neboť platí

$$\frac{i}{m} - \frac{1}{m+1-i} = \frac{i-1}{m+1-i} - \frac{i(i-1)}{m(m+1-i)} \tag{4.6}$$

$$\geq \frac{i-1}{m+1-i} - \frac{i-1}{m+1-i} = 0 \tag{4.7}$$

Navíc u adaptivní BH pocedury může dojít k jevu, že zamítne i hypotézy, jejichž $p$-hodnoty jsou větší než hladina testu. Je to dáno tím, že máme-li větší počet hypotéz s malými $p$-hodnotami, což nás vede na podezření, že více nulových hypotéz neplatí, je číslo $\frac{i}{\widehat{m_0}}$ větší než 1 pro nějaké $i$, a tedy příslušná $p$-hodnota je porovnávána s číslem větším než je stanovená hladina testu. Tuto skutečnost demonstruje tabulka 4.6. Dále následují tabulky 4.3, 4.4, 4.5. Pro každou z nich platí, že v první řádku je počet hypotéz zamítnutých na konvenční hladině $\alpha$. V druhém řádku je počet zamítnutých hypotéz pomocí jednotlivých procedur mnohonásobného testování pro daný počet zamítnutých hypotéz na hladině $\alpha$. Na konci tohoto paragrafu jsou uvedeny grafy zobrazující počet zamítnutých hypotéz v závislosti na velikosti veličiny $r_i$.

| **Počet zamít.** | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| **Na hladině** $\alpha$ | 74175 | 22000 | 3450 | 356 | 18 | 1 | 0 |
| **BH** | 94959 | 4593 | 411 | 34 | 2 | 1 | 0 |
| **ABH** | 94959 | 4492 | 473 | 61 | 8 | 6 | 1 |
| **Holm** | 95109 | 4670 | 214 | 6 | 1 | 0 | 0 |
| **Hochberg** | 95104 | 4669 | 219 | 7 | 1 | 0 | 0 |
| **Hommel** | 95056 | 4712 | 224 | 7 | 1 | 0 | 0 |
| **BY** | 97925 | 1942 | 130 | 2 | 1 | 0 | 0 |

Tabulka 4.1: Tabulka udává počty zamítnutých hypotéz procedurami mnohonásobného testování pro 6 vysvětlujících proměnných pro pevný počet opakování, který je roven 100 000. V řádku Na hladině $\alpha$ se nacházejí počty zamítnutých hypotéz na konvenční hladině $\alpha$.

| **Počet zamít.** | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| **Na hladině** $\alpha$ | 61624 | 28910 | 7668 | 1522 | 233 | 40 | 3 | 0 |
| **BH** | 95066 | 4407 | 440 | 71 | 14 | 2 | 0 | 0 |
| **ABH** | 95066 | 4279 | 503 | 110 | 30 | 11 | 0 | 1 |
| **Holm** | 95212 | 4542 | 232 | 10 | 3 | 1 | 0 | 0 |
| **Hochberg** | 95211 | 4542 | 233 | 10 | 3 | 1 | 0 | 0 |
| **Hommel** | 95174 | 4574 | 236 | 12 | 3 | 1 | 0 | 0 |
| **BY** | 98296 | 1586 | 109 | 6 | 2 | 1 | 0 | 0 |

Tabulka 4.2: Tabulka udává počty zamítnutých hypotéz procedurami mnohonásobného testování pro 10 vysvětlujících proměnných pro pevný počet opakování, který je roven 100 000. V řádku Na hladině $\alpha$ se nacházejí počty zamítnutých hypotéz na konvenční hladině $\alpha$.

| **Počet zamít.** | 1 | | 2 | | | 3 | | | | 4 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Procedurami** | 0 | 1 | 0 | 1 | 2 | 0 | 1 | 2 | 3 | 0 | 1 | 2 | 3 | 4 |
| **BH** | 89 | 11 | 68 | 21 | 11 | 48 | 30 | 11 | 11 | 31 | 19 | 9 | 23 | 18 |
| **ABH** | 89 | 11 | 68 | 21 | 11 | 48 | 19 | 11 | 15 | 31 | 13 | 9 | 9 | 6 |
| **Holm** | 89 | 11 | 74 | 17 | 9 | 52 | 38 | 9 | 1 | 46 | 24 | 18 | 8 | 4 |
| **Hochberg** | 89 | 11 | 73 | 17 | 10 | 52 | 38 | 9 | 1 | 46 | 24 | 16 | 8 | 6 |
| **Hommel** | 89 | 11 | 71 | 19 | 10 | 51 | 37 | 11 | 1 | 40 | 28 | 18 | 8 | 6 |
| **BY** | 95 | 5 | 86 | 11 | 3 | 77 | 16 | 6 | 1 | 64 | 14 | 14 | 6 | 2 |

Tabulka 4.3: Tabulka udává počet zamítnutých hypotéz procedurami mnohonásobného testování pro 6 vysvětlujících proměnných.

| **Počet zamít.** | 1 | | 2 | | | 3 | | | | 4 | | | | | 5 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Procedurami** | 0 | 1 | 0 | 1 | 2 | 0 | 1 | 2 | 3 | 0 | 1 | 2 | 3 | 4 | 0 | 1 | 2 | 3 | 4 | 5 |
| **BH** | 88 | 12 | 73 | 24 | 3 | 63 | 23 | 11 | 1 | 38 | 22 | 21 | 14 | 5 | 24 | 10 | 18 | 6 | 34 | 8 |
| **ABH** | 88 | 12 | 73 | 24 | 3 | 63 | 25 | 11 | 3 | 38 | 14 | 13 | 14 | 19 | 24 | 8 | 12 | 6 | 24 | 24 |
| **Holm** | 88 | 12 | 74 | 25 | 1 | 68 | 26 | 6 | 0 | 50 | 28 | 19 | 2 | 1 | 52 | 42 | 12 | 4 | 0 | 0 |
| **Hochberg** | 88 | 12 | 74 | 25 | 1 | 68 | 26 | 6 | 0 | 50 | 28 | 19 | 2 | 1 | 40 | 42 | 14 | 4 | 0 | 0 |
| **Hommel** | 88 | 12 | 74 | 25 | 1 | 66 | 27 | 7 | 0 | 48 | 30 | 19 | 2 | 1 | 38 | 40 | 12 | 6 | 0 | 0 |
| **BY** | 96 | 4 | 86 | 13 | 1 | 88 | 12 | 0 | 0 | 76 | 14 | 3 | 0 | 1 | 64 | 22 | 6 | 4 | 0 | 0 |

Tabulka 4.4: Tabulka udává počet zamítnutých hypotéz procedurami mnohonásobného testování pro 8 vysvětlujících proměnných.

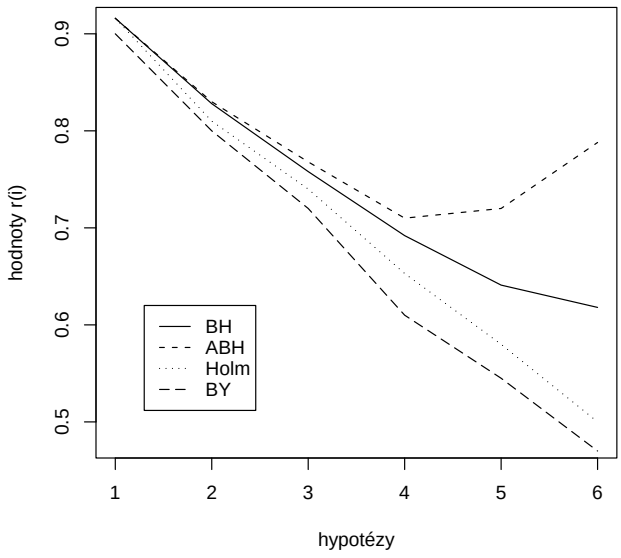
| **Počet zamít.** | 1 | | 2 | | | 3 | | | | 4 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Procedurami** | 0 | 1 | 0 | 1 | 2 | 0 | 1 | 2 | 3 | 0 | 1 | 2 | 3 | 4 |
| **HB** | 84 | 16 | 78 | 16 | 6 | 58 | 27 | 14 | 1 | 50 | 23 | 16 | 7 | 4 |
| **ABH** | 84 | 16 | 78 | 15 | 7 | 58 | 23 | 15 | 3 | 50 | 19 | 14 | 8 | 9 |
| **Holm** | 84 | 16 | 79 | 20 | 1 | 61 | 30 | 9 | 0 | 56 | 35 | 9 | 0 | 0 |
| **Hochberg** | 84 | 16 | 79 | 20 | 1 | 61 | 30 | 9 | 0 | 55 | 35 | 10 | 0 | 0 |
| **Hommel** | 84 | 16 | 79 | 20 | 1 | 61 | 30 | 9 | 0 | 54 | 36 | 10 | 0 | 0 |
| **BY** | 97 | 12 | 94 | 5 | 1 | 82 | 14 | 4 | 0 | 88 | 8 | 4 | 0 | 0 |

| **Počet zamit.** | 5 | | | | | | 6 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Procedurami** | 0 | 1 | 2 | 3 | 9 | 5 | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| **HB** | 34 | 24 | 25 | 5 | 13 | 3 | 21 | 23 | 17 | 13 | 14 | 4 | 8 |
| **ABH** | 34 | 12 | 10 | 15 | 11 | 10 | 21 | 5 | 7 | 7 | 6 | 10 | 17 |
| **Holm** | 43 | 40 | 13 | 3 | 2 | 0 | 37 | 37 | 17 | 3 | 2 | 4 | 0 |
| **Hochberg** | 43 | 39 | 13 | 3 | 2 | 0 | 37 | 37 | 17 | 3 | 2 | 4 | 0 |
| **Hommel** | 41 | 39 | 15 | 3 | 2 | 0 | 35 | 35 | 19 | 5 | 2 | 4 | 0 |
| **BY** | 72 | 17 | 7 | 2 | 2 | 0 | 63 | 23 | 5 | 3 | 2 | 4 | 0 |

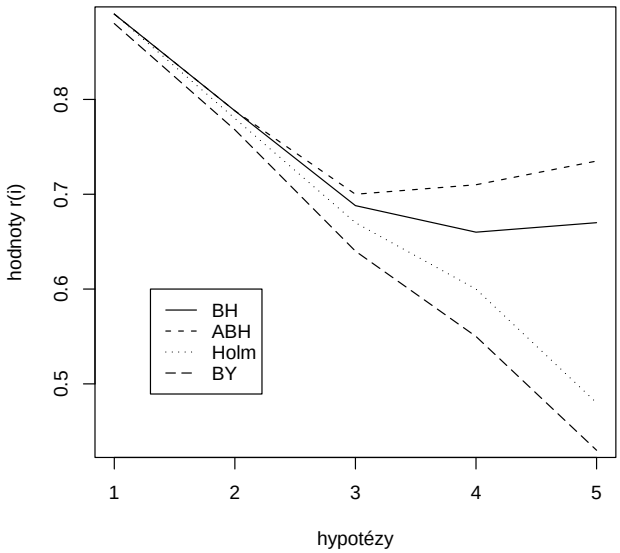
Tabulka 4.5: Tabulka udává počet zamítnutých hypotéz procedurami mnohonásobného testování pro 10 vysvětlujících proměnných. U adaptivní BH procedury součet zamítnutí nečiní 100 pro 5 a 6 zamítnutí. Je to tím, že došlo k většímu počtu zamítnutí než byl počet zamítnutí na konvenční hladině.

Tabulka 4.6: Přehled $p$-hodnot příslušných k jednotlivým procedurám mnohonásobného testování.

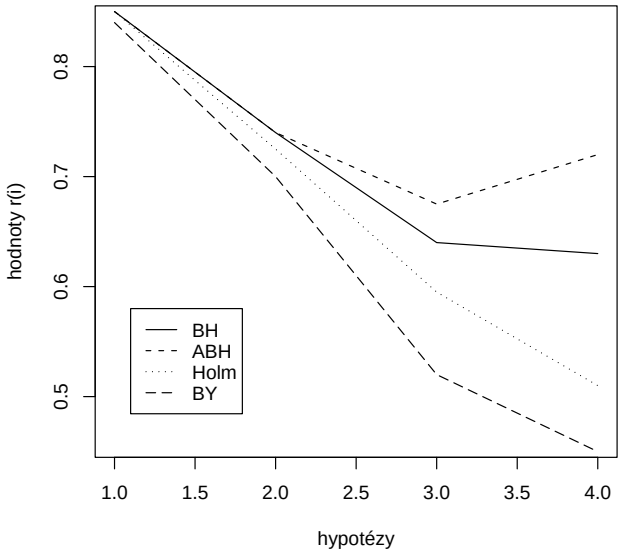
| $i$ | $p_i$ | $p_i^{bonf}$ | $p_i^{BH}$ | $p_i^{ABH}$ | $p_i^{Holm}$ | $p_i^{Hoch}$ | $p_i^{Hom}$ | $p_i^{BY}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 0,00012 | 0,00144 | 0,00144 | 0,00069 | 0,00144 | 0,00144 | 0,00144 | 0,0044 |
| 2 | 0.0027 | 0.0324 | 0.0162 | 0.0077 | 0.0297 | 0.0927 | 0.0250 | 0.0484 |
| 3 | 0.0039 | 0.0468 | 0.0156 | 0.0074 | 0.0390 | 0.0390 | 0.0312 | 0.0484 |
| 4 | 0.0085 | 0.1020 | 0.0255 | 0.0112 | 0.0765 | 0.0700 | 0.0590 | 0.0620 |
| 5 | 0.0091 | 0.1092 | 0.0218 | 0.0104 | 0.0765 | 0.0700 | 0.0637 | 0.0620 |
| 6 | 0.0100 | 0.1200 | 0.0200 | 0.0095 | 0.0765 | 0.0700 | 0.0700 | 0.0627 |
| 7 | 0.0600 | 0.7200 | 0.1030 | 0.0496 | 0.3620 | 0.3620 | 0.3620 | 0.3210 |
| 8 | 0.1480 | 1.0000 | 0.2220 | 0.1060 | 0.7410 | 0.7410 | 0.5920 | 0.6890 |
| 9 | 0.2368 | 1.0000 | 0.3150 | 0.1510 | 0.9470 | 0.7580 | 0.7580 | 0.9800 |
| 10 | 0.3847 | 1.0000 | 0.4610 | 0.2210 | 1.0000 | 0.7580 | 0.7580 | 1.0000 |
| 11 | 0.5789 | 1.0000 | 0.6310 | 0.3020 | 1.0000 | 0.7580 | 0.7580 | 1.0000 |
| 12 | 0.7584 | 1.0000 | 0.7580 | 0.3630 | 1.0000 | 0.7580 | 0.7580 | 1.0000 |




Obrázek 4.1: Vyjadřuje závislost $r_i$ na počtu zamítnutých hypotéz na konvenční hladině $\alpha$ pro 10 vysvětlujících proměnných.




Obrázek 4.2: Vyjadřuje závislost $r_i$ na počtu zamítnutých hypotéz na konvenční hladině $\alpha$ pro 8 vysvětlujících proměnných.




Obrázek 4.3: Vyjadřuje závislost $r_i$ na počtu zamítnutých hypotéz na konvenční hladině $\alpha$ pro 6 vysvětlujících proměnných.

## 4.2 Porovnání procedur kontrolující *FDR* a *FWE* pro závislé statistiky

Simulace pro závislé statistiky bude obdobná simulaci pro nezávislé statistiky, proto k ní ani neuvádím zdrojový kód. Rozdíl bude v tom, že pokud chceme, aby BH procedura kontrolovala *FDR*, musejí t-statistiky mít vlastnost *PRDS* a hladina testu musí být menší než 1/2. Druhá podmínka pro nás není nikterak omezující, proto jí zde ani nebudeme věnovat pozornost. Nutnou podmínkou pro splnění vlastnosti *PRDS* je předpoklad pozitivní kovariance mezi všemi vysvětlujícími proměnnými. My se zde omezíme na speciální případ, kdy všechny kovariance mezi náhodnými veličinami $X_i,\ i = 1, \ldots, m$, budou stejné a jednotlivé rozptyly budou rovny jedné. Varianční matice má tedy tvar

$$\begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \ldots & 1 \end{pmatrix} \tag{4.8}$$

Budeme se snažit vytvořit náhodný výběr z normálního rozdělení s varianční maticí $\mathbf{V}$. Nejdříve si vytvoříme náhodné výběry $Z_i,\ i = 1, \ldots, m$, z normálního rozdělení se střední hodnotou $\mu$ a jednotkovým rozptylem. K tomu opět použijeme funkci *rnorm* s parametry $\mu$ a 1. Pak stačí už jen položit $\mathbf{X} = \mathbf{V}^{\frac{1}{2}}\mathbf{Z}$, kde $\mathbf{Z} = (Z_1, \ldots, Z_m)$. To plyne z faktu, že $\operatorname{var}\mathbf{Z} = \mathbf{I}$, matice $\mathbf{V}^{\frac{1}{2}}$ je symetrická, tak že platí

$$\operatorname{var}\mathbf{X} = \operatorname{var}(\mathbf{V}^{\frac{1}{2}}\mathbf{Z}) = \mathbf{V}^{\frac{1}{2}}\operatorname{var}(\mathbf{Z})(\mathbf{V}^{\frac{1}{2}})' = \mathbf{V}^{\frac{1}{2}}(\mathbf{V}^{\frac{1}{2}})' = \mathbf{V}.$$

Zbývá jen spočítat odmocninovou matici $\mathbf{V}^{\frac{1}{2}}$. Nechť $\lambda_1, \ldots, \lambda_m$ jsou vlastní čísla matice $\mathbf{V}$ a $u_1, \ldots, u_m$ jsou vlastní vektory k těmto vlastním číslům. Z věty A.6 v [1] plyne, že

$$\mathbf{V}^{\frac{1}{2}} = \mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{U}',$$

kde $\mathbf{\Lambda}^{\frac{1}{2}} = \operatorname{diag}\{\lambda_1, \ldots, \lambda_m\}$ a $\mathbf{U}' = (u_1, \ldots, u_m)$. Funkci počítající odmocninovou matici lze nalézt v dodatcích.

Porovnáním tabulek 4.2, 4.3, 4.4 s tabulkami 4.5, 4.6, 4.7, 4.8, 4.9, 4.10 docházíme k závěru, že simulace pro závislé statistiky vycházejí velmi podobně jako pro nezávislé statistiky. Je zde opět patrný rozdíl mezi procedurami kontrolující *FDR* a procedurami kontrolující *FWE*, zejména při zvyšujícím se počtu zamítaných hypotéz. Nejvíce hypotéz je zamítnuto pomocí adaptivní BH procedury. Z tabulky 4.10 vyplývá, že s rostoucí korelací se zvyšuje rozdíl mezi Hochbergovou (popř. Holmovou) a Hommelovou procedurou. Hommelova procedura zamítá nejvíce hypotéz z procedur kontrolující *FWE*. Dále z grafu 4.2 je patrné, že v BH i ABH proceduře dochází k poměrně významnému nárůstu hodnot veličiny $r_i$ pro 5 a 6 zamítaných hypotéz, což nás vede k domněnce, že kontrola *FDR* při větším počtu hypotéz je vhodnější nástrojem pro kontrolu celkové chyby v mnohonásobném testování.

| **Počet zamít.** | 1 | | 2 | | | 3 | | | | 4 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Procedurami** | 0 | 1 | 0 | 1 | 2 | 0 | 1 | 2 | 3 | 0 | 1 | 2 | 3 | 4 |
| **BH** | 86 | 14 | 59 | 29 | 12 | 41 | 18 | 23 | 18 | 22 | 18 | 28 | 22 | 10 |
| **ABH** | 86 | 14 | 59 | 28 | 13 | 41 | 13 | 15 | 31 | 22 | 7 | 13 | 13 | 38 |
| **Holm** | 86 | 14 | 62 | 35 | 3 | 50 | 30 | 15 | 5 | 38 | 39 | 16 | 6 | 1 |
| **Hochberg** | 86 | 14 | 62 | 35 | 3 | 49 | 29 | 16 | 6 | 38 | 39 | 16 | 6 | 1 |
| **Hommel** | 86 | 14 | 62 | 35 | 3 | 45 | 32 | 17 | 6 | 27 | 44 | 21 | 7 | 1 |
| **BY** | 91 | 9 | 62 | 18 | 0 | 73 | 16 | 9 | 2 | 68 | 20 | 8 | 4 | 0 |

Tabulka 4.7: Tabulka udává počet zamítnutých hypotéz procedurami mnohonásobného testování pro závislé statistiky s varianční maticí (4.8) s $\rho = 0.5$ a 6 vysvětlujících proměnných.

| **Počet zamít.** | 1 | | 2 | | | 3 | | | | 4 | | | | | 5 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Procedurami** | 0 | 1 | 0 | 1 | 2 | 0 | 1 | 2 | 3 | 0 | 1 | 2 | 3 | 4 | 0 | 1 | 2 | 3 | 4 | 5 |
| **BH** | 86 | 14 | 73 | 19 | 8 | 61 | 29 | 9 | 1 | 43 | 23 | 20 | 10 | 4 | 25 | 17 | 17 | 20 | 14 | 7 |
| **ABH** | 86 | 14 | 73 | 19 | 8 | 61 | 20 | 10 | 6 | 43 | 14 | 10 | 11 | 19 | 25 | 7 | 7 | 4 | 5 | 41 |
| **Holm** | 86 | 14 | 74 | 22 | 4 | 65 | 32 | 3 | 0 | 50 | 38 | 9 | 2 | 1 | 35 | 31 | 22 | 8 | 3 | 1 |
| **Hochberg** | 86 | 14 | 74 | 22 | 4 | 65 | 32 | 3 | 0 | 50 | 38 | 9 | 2 | 1 | 35 | 31 | 22 | 8 | 3 | 1 |
| **Hommel** | 86 | 14 | 73 | 23 | 4 | 65 | 32 | 3 | 0 | 47 | 40 | 10 | 2 | 1 | 35 | 32 | 22 | 8 | 4 | 1 |
| **BY** | 96 | 4 | 92 | 7 | 1 | 88 | 9 | 0 | 0 | 76 | 18 | 3 | 0 | 1 | 60 | 18 | 13 | 5 | 3 | 1 |

Tabulka 4.8: Tabulka udává počet zamítnutých hypotéz procedurami mnohonásobného testování pro závislé statistiky s varianční maticí (4.8) s $\rho = 0.5$ a 8 vysvětlujících proměnných.

| **Počet zamít.** | 1 | | 2 | | | 3 | | | | 4 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Procedurami** | 0 | 1 | 0 | 1 | 2 | 0 | 1 | 2 | 3 | 0 | 1 | 2 | 3 | 4 |
| **BH** | 87 | 13 | 72 | 23 | 5 | 61 | 27 | 7 | 5 | 52 | 28 | 12 | 5 | 3 |
| **AHB** | 87 | 13 | 72 | 23 | 5 | 61 | 25 | 8 | 6 | 52 | 20 | 11 | 9 | 6 |
| **Holm** | 87 | 13 | 74 | 24 | 2 | 65 | 29 | 5 | 1 | 57 | 34 | 8 | 1 | 0 |
| **Hochberg** | 87 | 13 | 74 | 24 | 2 | 65 | 29 | 5 | 1 | 57 | 34 | 8 | 1 | 0 |
| **Hommel** | 87 | 13 | 74 | 24 | 2 | 63 | 31 | 5 | 1 | 56 | 35 | 8 | 1 | 0 |
| **BY** | 94 | 6 | 90 | 9 | 1 | 85 | 10 | 4 | 1 | 70 | 24 | 6 | 0 | 0 |

| **Počet zamit.** | 5 | | | | | | 6 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Procedurami** | 0 | 1 | 2 | 3 | 4 | 5 | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| **BH** | 43 | 17 | 15 | 11 | 10 | 4 | 41 | 13 | 7 | 19 | 10 | 8 | 2 |
| **AHB** | 43 | 11 | 4 | 11 | 12 | 16 | 41 | 3 | 0 | 9 | 7 | 6 | 26 |
| **Holm** | 53 | 29 | 12 | 5 | 1 | 0 | 50 | 24 | 17 | 7 | 1 | 1 | 0 |
| **Hochberg** | 53 | 29 | 12 | 5 | 1 | 0 | 50 | 24 | 17 | 7 | 1 | 1 | 0 |
| **Hommel** | 47 | 34 | 12 | 6 | 1 | 0 | 40 | 25 | 21 | 9 | 3 | 2 | 0 |
| **BY** | 70 | 18 | 7 | 4 | 1 | 0 | 70 | 13 | 11 | 4 | 1 | 1 | 0 |

Tabulka 4.9: Tabulka udává počet zamítnutých hypotéz procedurami mnohonásobného testovaní pro závislé statistiky s varianční maticí (4.8) s $\rho = 0.5$ a 10 vysvětlujících promměnných.

| **Počet zamít.** | 1 | | 2 | | | 3 | | | | 4 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Procedurami** | 0 | 1 | 0 | 1 | 2 | 0 | 1 | 2 | 3 | 0 | 1 | 2 | 3 | 4 |
| **BH** | 74 | 26 | 67 | 23 | 10 | 47 | 24 | 16 | 13 | 24 | 25 | 19 | 17 | 15 |
| **ABH** | 74 | 26 | 67 | 23 | 10 | 47 | 17 | 14 | 22 | 24 | 15 | 7 | 8 | 36 |
| **Holm** | 74 | 26 | 72 | 24 | 4 | 53 | 32 | 11 | 4 | 41 | 35 | 14 | 7 | 3 |
| **Hochberg** | 74 | 26 | 72 | 24 | 4 | 53 | 32 | 11 | 4 | 41 | 34 | 14 | 8 | 3 |
| **Hommel** | 74 | 26 | 70 | 26 | 4 | 52 | 32 | 12 | 4 | 36 | 32 | 20 | 9 | 3 |
| **BY** | 89 | 11 | 88 | 10 | 2 | 79 | 14 | 4 | 3 | 67 | 17 | 9 | 6 | 3 |

Tabulka 4.10: Tabulka udává počet zamítnutých hypotéz procedurami mnohonásobného testování pro závislé statistiky s varianční maticí (4.8) s $\rho = 0,99$ a 6 vysvětlujících proměnných.

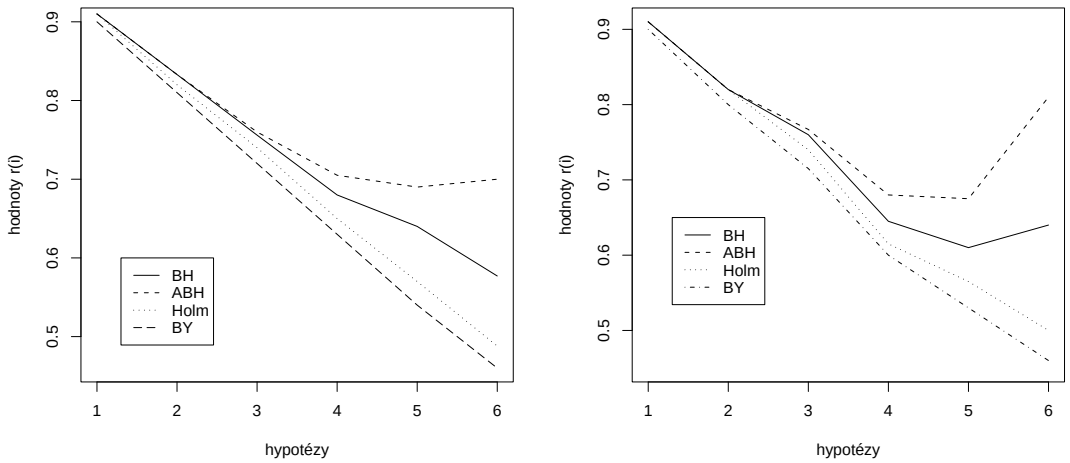
| **Počet zamít.** | 1 | | 2 | | | 3 | | | | 4 | | | | | 5 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Procedurami** | 0 | 1 | 0 | 1 | 2 | 0 | 1 | 2 | 3 | 0 | 1 | 2 | 3 | 4 | 0 | 1 | 2 | 3 | 4 | 5 |
| **BH** | 93 | 7 | 80 | 18 | 2 | 57 | 24 | 15 | 4 | 39 | 20 | 19 | 13 | 9 | 24 | 28 | 14 | 14 | 15 | 5 |
| **ABH** | 93 | 7 | 80 | 18 | 2 | 57 | 19 | 18 | 6 | 39 | 10 | 13 | 13 | 19 | 24 | 13 | 3 | 9 | 2 | 35 |
| **Holm** | 93 | 7 | 80 | 20 | 0 | 60 | 28 | 12 | 0 | 46 | 30 | 15 | 8 | 1 | 40 | 39 | 12 | 7 | 1 | 1 |
| **Hochberg** | 93 | 7 | 80 | 20 | 0 | 60 | 28 | 12 | 0 | 45 | 30 | 15 | 8 | 2 | 40 | 39 | 12 | 7 | 1 | 1 |
| **Hommel** | 93 | 7 | 80 | 20 | 0 | 60 | 28 | 12 | 0 | 43 | 29 | 18 | 8 | 2 | 36 | 37 | 15 | 10 | 1 | 1 |
| **BY** | 97 | 3 | 94 | 6 | 0 | 80 | 14 | 6 | 0 | 66 | 17 | 10 | 6 | 1 | 68 | 17 | 7 | 6 | 1 | 1 |

Tabulka 4.11: Tabulka udává počet zamítnutých hypotéz procedurami mnohonásobného testování pro závislé statistiky s varianční maticí (4.8) s $\rho = 0.99$ a pro 8 vysvětlujících proměnných.

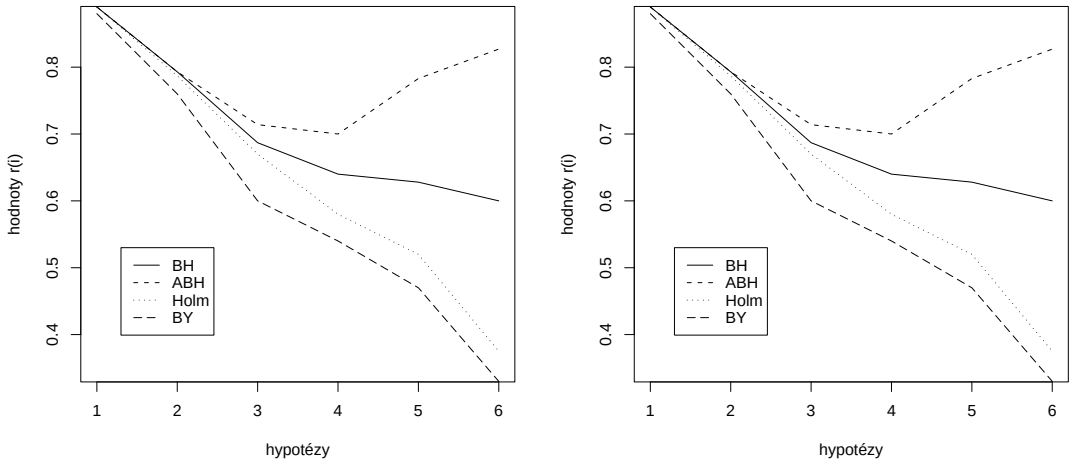
| **Počet zamít.** | 1 | | 2 | | | 3 | | | | 4 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Procedurami** | 0 | 1 | 0 | 1 | 2 | 0 | 1 | 2 | 3 | 0 | 1 | 2 | 3 | 4 |
| **BH** | 86 | 14 | 81 | 17 | 2 | 64 | 17 | 12 | 7 | 59 | 24 | 9 | 5 | 3 |
| **AHB** | 86 | 14 | 81 | 17 | 2 | 64 | 16 | 10 | 9 | 59 | 15 | 16 | 4 | 6 |
| **Holm** | 86 | 14 | 81 | 17 | 2 | 67 | 25 | 6 | 2 | 63 | 30 | 6 | 1 | 0 |
| **Hochberg** | 86 | 14 | 81 | 17 | 2 | 67 | 25 | 6 | 2 | 63 | 30 | 6 | 1 | 0 |
| **Hommel** | 86 | 14 | 81 | 17 | 2 | 66 | 25 | 7 | 2 | 63 | 30 | 6 | 1 | 0 |
| **BY** | 96 | 4 | 86 | 12 | 2 | 86 | 12 | 2 | 0 | 86 | 11 | 2 | 1 | 0 |

| **Počet zamit.** | 5 | | | | | | 6 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Procedurami** | 0 | 1 | 2 | 3 | 4 | 5 | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| **BH** | 45 | 28 | 13 | 5 | 5 | 4 | 22 | 20 | 10 | 14 | 18 | 8 | 8 |
| **AHB** | 45 | 12 | 8 | 10 | 10 | 13 | 22 | 8 | 6 | 4 | 2 | 2 | 42 |
| **Holm** | 51 | 39 | 6 | 2 | 2 | 0 | 34 | 44 | 10 | 10 | 4 | 0 | 0 |
| **Hochberg** | 51 | 39 | 6 | 2 | 2 | 0 | 34 | 44 | 10 | 10 | 2 | 0 | 0 |
| **Hommel** | 50 | 37 | 9 | 2 | 2 | 0 | 34 | 26 | 26 | 12 | 2 | 0 | 0 |
| **BY** | 78 | 18 | 1 | 2 | 1 | 0 | 64 | 20 | 8 | 6 | 2 | 0 | 0 |

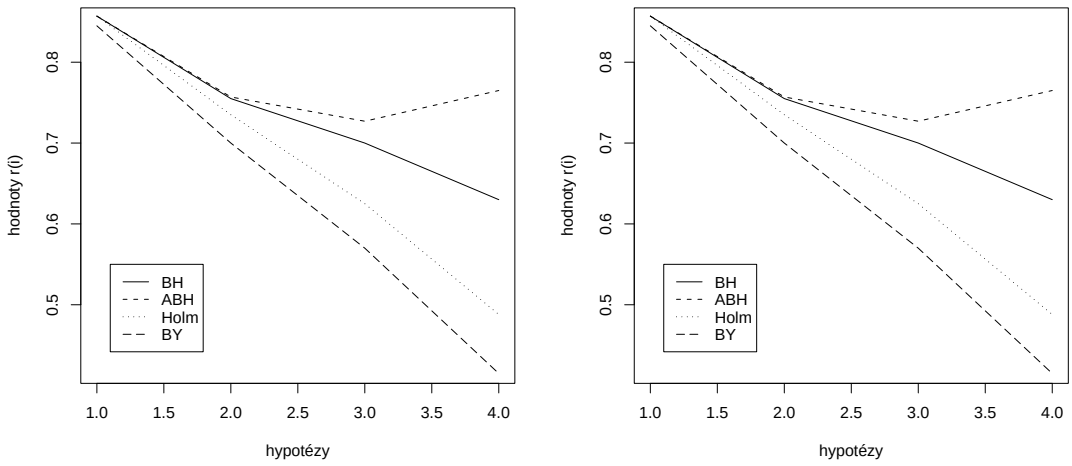
Tabulka 4.12: Tabulka udává počet zamítnutých hypotéz procedurami mnohonásobného testování pro závislé statistiky s varianční maticí (4.8) s $\rho = 0,99$ a pro 10 vysvětujících proměnných.




Obrázek 4.4: Vyjadřuje závislost $r_i$ na počtu zamítnutých hypotéz na konvenční hladině $\alpha$ pro 10 vysvětlujících proměnných pro $\rho = 0.5$ (vlevo) a $\rho = 0.99$ (vpravo).




Obrázek 4.5: Vyjadřuje závislost $r_i$ na počtu zamítnutých hypotéz na konvenční hladině $\alpha$ pro 8 vysvětlujících proměnných pro $\rho = 0.5$ (vlevo) a $\rho = 0.99$ (vpravo).




Obrázek 4.6: Vyjadřuje závislost $r_i$ na počtu zamítnutých hypotéz na konvenční hladině $\alpha$ pro 6 vysvětlujících proměnných pro $\rho = 0.5$ (vlevo) a $\rho = 0.99$ (vpravo).

# Kapitola 5

# Závěr

V této práci jsme se zabývali metodami mnohonásobného testování, které byly založeny na vzestupně srovnaných $p$-hodnotách. Vymezili jsme dva odlišné přístupy ke kontrole chyby, která vznikne chybným zamítnutím jednotlivých hypotéz. V teoretické části jsme popsali jednotlivé procedury kontrolující *FWE* a procedury kontrolující *FDR*. Dokázali jsme, že kontroluje-li procedura *FWE*, pak kontroluje i *FDR*. Dále jsme dokázali, že BH procedura kontroluje *FDR* i pro závislé testové statistiky se Studentovým rozdělením. Proto jsme mohli provést simulace i pro závislé statistiky.

Všechny simulace byly provedeny pro případ $m = m_0$. Tedy předpokládali jsme pouze platnost nulových hypotéz, proto jsme nemohli měřit sílu testu jednotlivých procedur. Místo síly testu jsme ale zavedli veličinu $r_i$, která počítala poměr mezi počtem procedurami nezamítnutých hypotéz za podmínky, že počet zamítnutých hypotéz na konvenční hladině je pevně dané číslo a počtem všech hypotéz. Tato veličina pak jednoznačně potvrdila, že procedury kontrolující *FDR* zamítají více hypotéz než procedury kontrolující *FWE*, když hladina testu pro *FWE* je stejná jako pro *FDR*. Ukázalo se, že míra závislosti mezi jednotlivými testovými statistikami má vliv pouze na procedury kontrolující *FDR*. Rozdíly mezi těmito procedurami jsou patrné pro 50% a více zamítaných hypotéz na konvenční hladině a s rostoucí korelací se ještě zvyšují.

# Dodatek A

# Dodatky

Zdrojový kód spustitelný v programu **R** ke všem procedurám.

```
PocetZamitBY<-function(pvals,hladina){
 q<- hladina pocet<-0
 p<-sort(pvals)
 delka<- length(p)
 pBY<- rep(1,delka)
 l<-1:delka
 o<-rep(1,delka)
 konst<-sum(o/l)
 for (i in (1:delka)){
      pom<- rep(1,i)
      for (j in (i:delka)) {
           pom[j]<- min((delka*p[j]*konst)/j,1)}
           pBY[i]<- min(pom)}
 for (i in (1:delka)) {
      if (pBY[i]<=q) {pocet<- pocet+1}}
PocetZamitBY<-pocet}

PocetZamitHoch<- function(pvals,hladina){
 q<- hladina pocet<-0
 p<-sort(pvals)
 delka<- length(p)
 phoch<- rep(0,delka)
 for (i in (1:delka)) {
      pom<- rep(1,i)
      for (j in (i:delka)) {
           pom[j]<-min((delka-j+1)*p[j],1)}
           phoch[i]<- min(pom)}
      for (i in (1:delka)) {
```

```
           if (phoch[i]<=q) {pocet<- pocet+1}}
PocetZamitHoch<-pocet}

PocetZamitHolm<-function(pvals,hladina){
 q<- hladina
 pocet<-0
 p<-sort(pvals)
 delka<- length(p)
 pholm<- rep(0,delka)
 for (i in (1:delka)) {
      pom<- rep(0,i+1)
      for (j in 1:i) {
           pom[j]<-min((delka-j+1)*p[j],1)}
           pholm[i]<- max(pom)}
      for (i in (1:delka)) {
           if (pholm[i]<=q) {pocet<- pocet+1}}
PocetZamitHolm<-pocet}

PocetZamitHommel<- function(pvals,hladina){
 q<-hladina
 pocet<-0
 p<- sort(pvals)
 a<-p delka<-length(p)
 k<-delka:2
 c<-rep(2,delka)
 cmin<-1
 for (m in k) {
      c<-rep(2,delka)
      for (i in(delka-m+1):delka) {
           c[i]<- (m*p[i])/(m+i-delka)}
      cmin<- min(c)
      for (i in (delka-m+1):delka) {
           if (a[i]<cmin) a[i]<-cmin}
      for (i in 1:(delka-m+1)) {
           c[i]<-min(cmin,m*p[i])
           if (a[i]<c[i]) a[i]<-c[i]}}
 phommel<-a
 for (i in (1:delka)) {
      if (phommel[i]<=q) pocet<- pocet+1}
 PocetZamitHommel<-pocet}

PocetZamitBH<- function(pvals,hladina){
 q<- hladina
```

```
 pocet<-0
 p<-sort(pvals)
 delka<- length(p)
 pBH<- rep(1,delka)
 for (i in (1:delka)){
      pom<- rep(1,i)
      pmin<-min(pBH)
      for (j in (1:i)) {
           pom[j]<- min((delka*p[i])/j,1)}
           pBH[i]<- min(pom)}
      for (i in (1:delka)) {
           if (pBH[i]<=q) {pocet<- pocet+1}}
 PocetZamitBH<-pocet}

PocetZamitABH<- function(pvals,hladina){
 q<- hladina
 pocet<-0
 p<-sort(pvals)
 delka<- length(p)
 if (p[1]>(q)/delka) {PocetZam<-0} else
    {S1<-(1-p[1])/delka
     i<-1
     repeat{
        S2<-(1-p[i])/(delka+1-i)
        i<- i+1
        if (S2>=S1) S1<-S2 else {break}}
 m_0<-min(1/S1+1,delka)}
 pABH<- rep(1,delka)
 for (i in (1:delka)){
      pom<- rep(1,i)
      for (j in (1:i)) {
           pom[j]<- min((m_0*p[i])/j,1)}
      pABH[i]<- min(pom)}
 for (i in (1:delka)) {
     if (pABH[i]<=q) pocet<- pocet+1}
 PocetZamitABH<-pocet}
```

Funkce *PocetZamitBY, PocetZamitHoch, PocetZamitHolm, PocetZamitHom, PocetZamitBH a PocetZamitABH* počítají, kolik hypotéz zamítáme pro vstupní $p$-hodnoty na dané hladině testu *hladina* pro jednotlivé procedury v tomto pořadí: BY procedura, Hochbergova procedura, Holmova procedura, Hommelova procedura, BH procedura a adaptivní BH procedura.

```
function Simulace{pozor,rozp,hladina,pocZamit)
 q<-hladina
 cBH<-rep(0,11)
 cABH<-rep(0,11)
 cHoch<-rep(0,11)
 cHolm<-rep(0,11)
 cBY<-rep(0,11)
 cHom<-rep(0,11)
 cpocet<-0
 k<-pocZamit
repeat{y<-rnorm(pozor,1.55,rozp)
       x1<-rnorm(pozor,1.5,rozp)
       x2<-rnorm(pozor,0.5,rozp)
       x3<-rnorm(pozor,1.15,rozp)
       x4<-rnorm(pozor,-0.3,rozp)
       x5<-rnorm(pozor,-0.4,rozp)
       x6<-rnorm(pozor,1.2,rozp)
       x7<-rnorm(pozor,-0.2,rozp)
       x8<-rnorm(pozor,0,rozp)
       x9<-rnorm(pozor,0,rozp)
       x10<-rnorm(pozor,-0.5,rozp)
       a<-lm(y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10)
       b<-summary(a) poc<-0
       for (i in 1:10){
            pval[i]<-b$coef[i+1,4]
            if(pval[i]<q) poc<-poc+1}
       if (poc==k) {BH<-pocetZamitBH(pval,q)
       ABH<-PocetZamitABH(pval,q)
       BY<-PocetZamitBY(pval,q)
       Holm<-PocetZamitHolm(pval,q)
       Hom<-PocetZamitHommel(pval,q)
       BH<-PocetZamitBH(pval,q)
       Hoch<-PocetZamitHoch(pval,q)
       cBH[BH+1]<-cBH[BH+1]+1
       cABH[ABH+1]<-cABH[ABH+1]+1
       cHoch[Hoch+1]<-cHoch[Hoch+1]+1
       cHolm[Holm+1]<-cHolm[Holm+1]+1
       cHom[Hom+1]<-cHom[Hom+1]+1
       cBY[BY+1]<-cBY[BY+1]+1
       cpocet<-cpocet+1}
       if (cpocet>=100) break
Simulace}
```

Pomocí funkce *Simulace* získáme počty zamítnutých hypotéz procedurami mnohonásobného testování. Funkce v každém opakování vytvoří lineární model s 10 vysvětlujícími proměnnými $x_1, \ldots, x_{10}$ a pokračuje, dokud počet zamítnutých hypotéz na dané hladině testu není roven 100.

```
V12<-function(kovariance,dimenze){
 ro<- kovariance
 d<- dimenze
 L<-rep(ro,d*d)
 dim(L)<-c(d,d)
 V<-L-diag(rep(ro-1,d))
 t<- eigen(V)
V12<-t$vectors%*%diag(sqrt(abs(t$values)))%*%t(t$vectors)}
```

Funkce *V12* vytváří odmocninovou matici typu $(n, n)$ a tvaru (4.8), kde $n$ je rovno parametru *dimenze*. Jednotlivé závislé vektory dostaneme násobením vektorů $x_1, \ldots, x_{10}$ maticí $V12$ zprava.

# Literatura


[1] Anděl J.: *Základy matematické statistiky.* Matfyzpress (2005).

[2] Benjamini Y. Hochberg Y.: *Controling the false discovery rate: a practical and powerful approach to multiple testing.*. J. J. Roy. Statist. Soc. Ser B. **25** (2000) 289–300.

[3] Benjamini Y. Hochberg Y.: *The adaptive control of the false discovery rate in multiple hypotheses testing.* J. Roy. Statist. Soc. Ser. B. **57** (1995) 289–300.

[4] Benjamini Y. Hochberg Y.: *The control of the false discovery rate in multiple testing under dependency.* Annals of Statist. **29** (2001) 1165–1188.

[5] Benjamini Y. Yekutiely D.: *Adaptive linear step-up procedures that controls the false discovery rate.*

[6] Feller W.: *An introduction to probability theory and its applications.* John Wiley and Sons. (1971).

[7] Holm S.: *A simple sequentially rejective multiple test procedure.* Skand. J. Statist. **6** (1979) 65–70.

[8] Seber G.A.F. Lee A.J.: *Linear Regression Analysis.* John Wiley and Sons. (2003).

[9] Sanat B. Sarkar K.: *Some probability inequalities for ordered $MTP_2$ random variables: a proof of the Simes conjecture.* Annals of Statist. **26** (1998) 494–504.

[10] Simes R.J.: *An imporved Bonferroni procedure for multiple tests of significance.* Biometrika **73** (1986) 751–754.

[11] Zvára K.: *Regresní analýza.* Academia Praha (1989).