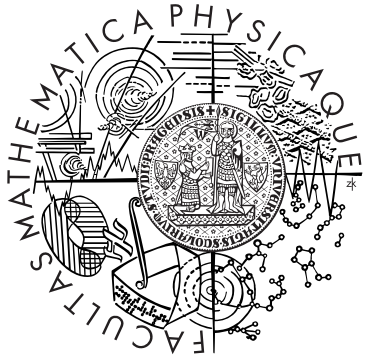
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Gabriel Lendel

Statistické testy pro víc jak dva náhodné výběry

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: Mgr. Miriam Marušiaková
Studijní program: Matematika, obecná matematika

2009

Rád by som vyjadril svoje poďakovanie vedúcej mojej bakalárskej práce Mgr. Miriam Marušiakovej za návrh témy, podnetné konzultácie, motiváciu, a tiež za hodnotné rady pri spisovaní mojej práce. Najmä, chcem však poďakovať svojim rodičom, ktorí mi umožnili štúdium na vysokej škole.

Prehlasujem, že som svoju bakalársku prácu napísal samostatne a výhradne s použitím citovaných prameňov. Súhlasím so zapožičiavaním a jej zverejňovaním.

V Prahe dňa 19. VII. 2009 Gabriel Lendel

Obsah

Názov práce: Statistické testy pro víc jak dva náhodné výběry
Autor: Gabriel Lendel
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedúci bakalárskej práce: Mgr. Miriam Marušiaková
e-mail vedúceho: maruskay@gmail.com

Abstrakt: Daných je $m > 2$ náhodných výberov, postupne z rozdelení $N[\mu_1, \sigma^2]$, $N[\mu_2, \sigma^2], \ldots, N[\mu_m, \sigma^2]$, kde σ^2 je neznámy parameter. Úlohou je testovať hypotézu o rovnosti všetkých m stredných hodnôt. V práci je popísaná metóda jednoduchej analýzy rozptylu (ANOVA), ktorá daný problém vhodne rieši. Ak hypotézu zamietneme, je potrebné rozhodnúť, ktoré súbory sa od seba signifikantne líšia. To sa vykoná pomocou metód mnohonásobného porovnávania. V praci študujeme a porovnávame LSD, Bonferroniho, Tukeyho, Scheffého, Student-Newman-Keulsov a Duncanov test. V závere práce aplikujeme jednoduchú ANOVA a metódy mnohonásobného porovnávania na reálne dáta, ktoré vyhodnotíme pomocou štatistického programu R. Na rýchle prevedenia testov v programe R je možné použiť funkcie, ktoré sa nachádzajú na priloženom CD.

Kľúčové slová: jednoduchá analýza rozptylu, metódy mnohonásobného porovnávania, vyhodnocovanie dát pomocou programu R

Title: Statistical tests for more than two random samples

Author: Gabriel Lendel
Department: Department of Probability and Mathematical Statistics
Supervisor: Mgr. Miriam Marušiaková
Supervisor's e-mail address: marusky@gmail.com

Abstract: We have $m > 2$ random samples from distributions $N[\mu_1, \sigma^2]$, $N[\mu_2, \sigma^2], \ldots, N[\mu_m, \sigma^2]$,where σ^2 is an unknown parameter. Our task is to test the hypothesis about equality of all m means. In the presented work we study one-way analysis of variance (ANOVA) which was demonstrated to be the right tool that solves the problem. If the hypothesis is to be rejected, it is necessary to decide which populations are significantly different. This is done by multiple comparison methods. In the work we study and compare LSD, Bonferroni, Tukey, Scheffé, Student-Newman-Keuls and Duncan tests. Finally, we apply one-way ANOVA and multiple comparison methods to real data. Program R is used to evaluate statistical tests.

Keywords: one-way analysis of variance, multiple comparison methods, testing in R

Úvod

Predpokladajme, že máme dva navzájom nezávislé náhodné výbery. Prvý z rozdelenia $N[\mu_1, \sigma^2]$ a druhý z $N[\mu_2, \sigma^2]$, pričom je parameter σ^2 neznámy. Testujeme hypotézu $\mu_1 = \mu_2$. Táto úloha sa štandardne rieši pomocou dvojvýberového t-testu, pričom pravdepodobnosť že neoprávnene zamietneme hypotézu sa nastaví na α. Túto úlohu je však potrebné rozšíriť, pretože sa často stáva, že je daných viac ako 2 základých súborov, pre ktoré potrebujeme testovať hypotézu o rovnosti všetých stredných hodnôt. Majme teda $m > 2$ nezávislých výberov postupne z rozdelení $N[\mu_1, \sigma^2],\ N[\mu_2, \sigma^2], \ldots, N[\mu_m, \sigma^2]$, pričom σ^2 je neznámy parameter a chceme testovať hypotézu $H_0 : \mu_1 = \cdots = \mu_m$. Myšlienka postupne aplikovať na každú dvojicu výberov dvojvýberový t-test sa ukázala byť nevhodná z hľadiska kontroly chyby 1.druhu. Ako uvidíme v kapitole 2.1, ak každý dvojvýberový t-test vykonáme na hladine α bude celková chyba 1. druhu oveľa väčšia. Z tohto dôvodu sa preto na test hypotézy H_0 používa iná metóda, ktorá udrží hladinu α. O tejto metóde založenej na testovaní submodelov pojednáva kapitola 1. Poznatky na napísanie tejto kapitoly pochádzajú prevažne z knihy [1]. Ak je v prvom kroku hypotéza H_0 zamietnutá, zaujíma nás ešte, medzi ktorými základnými súbormi sa vyskytujú rozdiely. Na riešenie tohto problému sa používajú metódy mnohonásobného porovnávania, ktoré sú popísané v 2.kapitole. V práci sa venujeme LSD, Bonferroni-LSD, Tukeyho, Scheffého, Student-Newman-Keulsovmu a Duncanovmu testu. Poznatky na napísanie tejto sekcie pochádzajú hlavne z kníh [3] a [4]. Na záver tejto kapitoly sú stručne popísané rozdiely medzi zmienenými metódami. V 3. kapitole pristúpime k aplikácii testov na reálne dáta. Na vyhodnotenie dát použijeme program R [5]. Praktické je využiť funkcie LSDCI, BonferroniCI, TukeyCI, ScheffeCI. Ide o funkcie získané z [6], ktoré boli jemne upravené, aby lepšie vyhovovali našim účelom. Je možné ich nájsť v prílohe k práci na CD.

Kapitola 1

Analýza rozptylu

Náš problém vieme interpretovať nasledovne. Na skupine jedincov pozorujeme určitý charakteristický znak, ktorý môže nadobúdať $m \geq 2$ hodnôt. Tento znak potom slúži k rozdeleniu jedincov do m porovnávacích skupín. Úlohou je overiť, či na hodnotu náhodnej veličiny pre určitého jedinca má významný vplyv hodnota znaku. Metóda, ktorá rieši daný problém, sa nazýva jednoduchá alebo jednočiniteľová analýza rozptylu (ANalysis Of VAriance, ďalej len ANOVA). Vo všeobecnosti ANOVA pripúšťa viacero znakov(často 2 alebo 3).

Uvažujeme teda $m \geq 2$ základných súborov. Pre i-ty súbor máme náhodný výber $Y_{i,j}$ $(j = 1, \ldots, n_i)$, kde n_i je rozsah výberu. Ak je rozsah všetkých súborov rovnaký, hovoríme o balancovanom modeli. Pre rozdelenie veličín predpokladáme: $Y_{i,j} \sim N[\mu_i, \sigma^2]$, pričom všetky parametre sú neznáme. Predpokladom je teda rovnosť rozptylov vo všetkych súboroch. Testujeme hypotézu

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_m,$$

teda inými slovami: Existujú v súboroch významné rozdiely stredných hodnôt, ktoré súvisia s efektom znaku? Alternatívou k hypotéze H_0 bude tvrdenie, že minimálme jedna dvojica stredných hodnôt sa vzájomne líši.

Pre všeobecný prípad ANOVA sa skonštruujú dvojice modelov. Zložitejší model predpokladá, že významný vplyv má viacero znakov (rôzna hodnota znaku rôzne ovplyvňuje hodnotu náhodnej veličiny). Jednoduchší model predpokladá vplyv žiadneho alebo menšieho počtu znakov. Následne sa vytvorí špeciálna varianta F-testu, ktorá modely vhodne porovná. Celá metóda je založená na testovaní submodelov.

Stručne sa venujme problému všeobecných regresných lineárnych modelov a zmienenému testovaniu submodelov. Nasledujúce tvrdenia je možné nájsť riadne dokázané v knihe [1].

Definícia 1.1 Máme náhodný vektor $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{'}$, ktorého rozdelenie závisí na parametri $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^{'}$. Nech $\boldsymbol{X}_{n\times k}$ je daná matica. Hovoríme, že $\boldsymbol{Y}$ sa riadi *lineárnym modelom* ak platí:
a) $E\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}$
b) $var\boldsymbol{Y} = \boldsymbol{V}$ existuje a nezávisí na $\boldsymbol{\beta}$.

Definícia 1.2 Hovoríme, že $\boldsymbol{b}$ je *najlepším nestranným lineárnym odhadom* (NNLO) vektoru $\boldsymbol{\beta}$ ak platí:
a) $\boldsymbol{b}$ je nestranným odhadom parametru $\boldsymbol{\beta}$, čo znamená že $\boldsymbol{b} = \boldsymbol{U}\boldsymbol{Y}$,
kde $\boldsymbol{U}_{k\times n}$ je matica a $E\boldsymbol{b} = \boldsymbol{\beta}$
b) ak je $\boldsymbol{b}^{'}$ iný nestranný lineárny odhad $\boldsymbol{\beta}$, potom platí $var\,\boldsymbol{b}^{'} - var\,\boldsymbol{b} \geq 0$.

Pre naše účely budeme ďalej predpokladať nasledujúce: $k < n$, $h(\boldsymbol{X}) < k$, $var\,\boldsymbol{Y} = \sigma^2\boldsymbol{I}$, kde $\sigma^2 > 0$ je neznámy parameter. V takomto prípade hovoríme o modeli s neúplnou hodnosťou. Parameter $\boldsymbol{\beta}$ sa odhaduje metódou najmenších štvorcov.

Veta 1.3 *Výraz* $(\boldsymbol{Y} - \boldsymbol{Xb})'(\boldsymbol{Y} - \boldsymbol{Xb})$ *nadobúda vzhľadom k* $\boldsymbol{b}$ *najmenšie hodnoty, ak je* $\boldsymbol{b}$ *riešením sústavy takzvaných normálnych rovníc, tj. sústavy* $\boldsymbol{X}'\boldsymbol{Xb} = \boldsymbol{X}'\boldsymbol{Y}$, *ktorá je ekvivalentná sústave* $\frac{\partial}{\partial b_j}(\boldsymbol{Y} - \boldsymbol{Xb})'(\boldsymbol{Y} - \boldsymbol{Xb}) = 0$, $j = 1 \ldots k$. *Pre všetky* $\boldsymbol{b}$, *ktoré sú riešením sústavy je hodnota výrazu* $(\boldsymbol{Y} - \boldsymbol{Xb})'(\boldsymbol{Y} - \boldsymbol{Xb})$ *rovnaká.*

Definícia 1.4 Veličinu $S_e = (\boldsymbol{Y} - \boldsymbol{Xb})'(\boldsymbol{Y} - \boldsymbol{Xb})$, kde $\boldsymbol{b}$ je nejaké riešenie normálnych rovníc, nazývame *reziduálny súčet štvorcov.*

Veta 1.5 *a) Vektor* $\boldsymbol{\theta} = E\boldsymbol{Y}$ *je vždy odhadnuteľný (existuje aspoň jeden lineárny nestranný odhad) a NNLO pre* $\boldsymbol{\theta}$ *je* $\hat{\boldsymbol{\theta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{Y}$, *kde* $(\boldsymbol{X}'\boldsymbol{X})^{-}$ *je pseudoinverzná matica.*
b) Platí, že parameter $\phi = \boldsymbol{c}'\boldsymbol{\beta}$ *je odhadnuteľný práve vtedy, keď je lineárnou kombináciou zložiek* $E\boldsymbol{Y}$. *NNLO pre* ϕ *je potom* $\hat{\phi} = \boldsymbol{c}'\boldsymbol{b}$, *kde* $\boldsymbol{b}$ *je ľubovolné riešenie normálnych rovníc.*

Výsledkom teórie lineárnych modelov sú nasledujúce dve dôležité tvrdenia.

Veta 1.6 *Nech* $\boldsymbol{Y} \sim N[\boldsymbol{X\beta}, \sigma^2\boldsymbol{I}]$ *a* $r_0 = h(\boldsymbol{X})$. *Potom platí*

$$S_e/\sigma^2 \sim \chi^2_{n-r_0}$$

$$s^2 = S_e/(n - r_0) \textit{je nestranným odhadom parametru } \sigma^2.$$

Veta 1.7 *Ak má* $\boldsymbol{Y}$ *normálne rozdelenie, potom vektor* $\boldsymbol{b}$ *a veličina* s^2 *sú nezávislé.*

Ďalej sa budeme venovať testovaniu submodelov. Celý mechanizmus ukážeme na jednom submodeli pôvodného modelu. Nastávajú avšak aj situácie, kedy je potrebný väčší reťazec submodelov. V takomto prípade sa postupuje analogicky. Vo všeobecnom prípade ANOVA bude pôvodný model ilustrovať rozdielne vplyvy niekoľkých znakov. V prvom submodeli bude rozdielny vplyv na náhodnú veličinu vykazovať menší počet znakov a v druhom ešte menší počet atď. V našom prípade pre jednoduchú analýzu rozptylu bude pôvodný model popisovať situáciu keď na náhodnú veličinu má významný vplyv hodnota pozorovaného znaku. Submodel bude naopak predpokladať, že rozdielne hodnoty znaku nebudú mať na náhodnú veličinu žiaden vplyv.

Uvažujme modely:
$$M : \boldsymbol{Y} \sim N[\boldsymbol{X\beta}, \sigma^2\boldsymbol{I}]$$
$$M_1 : \boldsymbol{Y} \sim N[\boldsymbol{U\alpha}, \sigma^2\boldsymbol{I}]$$

M je pôvodný model. $\boldsymbol{U}$ je typu $n \times k_1$, $h(\boldsymbol{U}) = r_1$, $\boldsymbol{\alpha}$ má k_1 zložiek.

Definícia 1.8 Hovoríme, že M_1 je *submodelom* modelu M, ak každý stĺpec matice $\boldsymbol{U}$ patrí do lineárneho obalu stĺpcov matice $\boldsymbol{X}$ a ak $r_1 < r_0$.
Prvá podmienka je splnená práve vtedy, ak existuje matica $\boldsymbol{K}_{k\times k_1}$ taká, že $\boldsymbol{U} = \boldsymbol{X}\boldsymbol{K}$. Ak platí model M_1, tak platí aj model M. V submodeli je vlastne prípustná hodnota parametru $\boldsymbol{\beta}$ redukovaná na $\boldsymbol{K}\boldsymbol{\alpha}$.

Chceme testovať hypotézu o platnosti modelu M_1. Myšlienka konštrukcie F-testu je nasledujúca: Za platnosti modelu M máme podľa vety 1.5 NNLO pre $E\boldsymbol{Y}$ v tvare $\hat{\boldsymbol{\mu}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{Y}$. V prípade platnosti submodelu M_1 máme podobne NNLO pre $E\boldsymbol{Y}$ v tvare $\hat{\boldsymbol{\nu}} = \boldsymbol{U}(\boldsymbol{U}'\boldsymbol{U})^{-}\boldsymbol{U}'\boldsymbol{Y}$. Ak platí model M_1, tak platí aj model M a teda $\hat{\boldsymbol{\mu}}$ a $\hat{\boldsymbol{\nu}}$ by sa nemali veľmi líšiť.

Veta 1.9 *Ak platí pre Y model M_1, potom*

$$F = \frac{(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}})'(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}})}{(r_0 - r_1)s^2} \sim F_{r_0-r_1,n-r_0} \ ,$$

kde $s^2 = S_e/(n - r_0)$, pričom S_e je reziduálny súčet štvorcov modelu M.

Pristúpme k aplikácii výsledkov na náš problém. Máme nezávislé náhodné výbery: $N[\mu_1, \sigma^2], \ldots, N[\mu_m, \sigma^2]$. Označme $n = n_1 + n_2 + \cdots + n_m$ rozsah celého výberu a

$$Y_{i.} = \sum_{p=1}^{n_i} Y_{ip} \qquad \bar{Y}_i = Y_{i.}/n_i$$

$$Y_{..} = \sum_{i=1}^{m}\sum_{p=1}^{n_i} Y_{ip} \qquad \bar{Y} = Y_{..}/n$$

$\bar{Y}_i$ je priemer v i-tom výbere a $\bar{Y}$ je priemer v rámci všetkých výberov. Predpoklady o veličinách môžme vyjadriť takto:

$$Y_{ip} = \mu + \alpha_i + e_{ip},\ p = 1, \ldots, n_i,\ i = 1, \ldots, m \ .$$

Parametre α_i modelujú rôzny vplyv na náhodnú veličinu spôsobený rozdielnou hodnotou znaku. Veličiny e_{ip} sú náhodné výkyvy tj. $e_{ip} \sim N[0, \sigma^2]$. Dostávame preparametrizovaný model, teda model s neúplnou hodnosťou. Vektory $\boldsymbol{Y}$ a $\boldsymbol{\beta}$ budú mať tvar $\boldsymbol{Y} = (Y_{1,1}, \ldots, Y_{1,n_1}, Y_{2,1} \ldots, Y_{2,n_2}, \ldots, Y_{m,n_m})'$, $\boldsymbol{\beta} = (\mu, \alpha_1, \ldots, \alpha_m)'$ a matica $\boldsymbol{X}$ má v prvom stĺpci samé jedničky, v druhom má jedničky na prvých n_1 miestach atď. Jej hodnosť je m. Všimnime si ešte, že pre strednú hodnotu platí

$$EY_{ip} = \mu + \alpha_i,\ i = 1, \ldots, m \quad p = 1, \ldots, n_i \ .$$

Za platnosti modelu M hľadáme NNLO pre parametre. Podľa vety 1.3 máme normálne rovnice v tvare

$$\frac{\partial}{\partial \mu} \sum_{i=1}^{m} \sum_{p=1}^{n_i} (Y_{ip} - \mu - \alpha_i)^2 = 0$$
$$\frac{\partial}{\partial \alpha_j} \sum_{i=1}^{m} \sum_{p=1}^{n_i} (Y_{ip} - \mu - \alpha_i)^2 = 0,\ j = 1, \ldots, m\ .$$

Sústavu zderivujeme a vydelíme -2, čím dostaneme

$$\sum_{i=1}^{m} \sum_{p=1}^{n_i} (Y_{ip} - \mu - \alpha_i) = 0$$
$$\sum_{p=1}^{n_j} (Y_{jp} - \mu - \alpha_j) = 0,\ j = 1, \ldots, m\ .$$

Upravujeme ďalej. Dostaneme

$$n\mu + \sum_{i=1}^{m} n_i \alpha_i = Y..$$
$$n_j \mu + n_j \alpha_j = Y_{j.},\ j = 1 \ldots m\ .$$

Keďže sú všetky riešenia sústavy vzhľadom k vete 1.3 rovnocenné, môžeme pridať takzvanú reparametrizačnú podmienku

$$\sum_{i=1}^{m} n_i \alpha_i = 0\ .$$

Sústavu už ľahko doriešime. Dostaneme odhady parametrov:

$$\hat{\mu} = \bar{Y} \qquad \hat{\alpha_i} = \bar{Y_i} - \bar{Y},\ i = 1, \ldots, m\ .$$

Podľa vety 1.5 je potom NNLO pre strednú hodnotu $\mu + \alpha_i$ veličina $\bar{Y_i}$ a NNLO pre $\alpha_i - \alpha_j$(rozdiel dvoch stredných hodnôt) je $\bar{Y_i} - \bar{Y_j}$. Hypotézu H_0 môžeme vyjadriť v tvare

$$H_0 : \alpha_1 = \cdots = \alpha_m = 0.$$

Za jej platnosti dostaneme submodel

$$Y_{ip} = \mu + e_{ip},\ p = 1, \ldots, n_i,\ i = 1, \ldots, m\ .$$

Hodnosť matice submodelu je teda rovná 1. Podobne ako pre pôvodný model odvodíme pre parameter μ NNLO. Dostaneme, že NNLO pre strednú hodnotu je veličina $\bar{Y}$.

Podľa vety 1.9 potom dostávame, že za platnosti hypotézy H_0 platí

$$F = \frac{\sum_{i=1}^{m} n_i(\bar{Y}_i - \bar{Y})^2}{(m-1)s^2} \sim F_{m-1,n-m} \ ,$$

$$\text{kde } s^2 = S_e/(n-m) = \sum_{i=1}^{m}\sum_{p=1}^{n_i}(Y_{ip} - \bar{Y}_i)^2/(n-m) \ . \tag{1.1}$$

.

Pripomeňme, že všeobecne je kritická hodnota $F_{a,b}(\alpha)$ rozdelenia $F_{a,b}$ definovaná vzorcom $P(Z \geq F_{a,b}(\alpha)) = \alpha$, kde $Z \sim F_{a,b}$. Test potom zostrojíme tak, že ak náhodná veličina F prekročí kritickú hodnotu $F_{m-1,n-m}(\alpha)$, zamietneme hypotézu H_0 na hladine α.
Venujme sa ešte reziduám modelov (tj. nevysvetlenej variabilite spôsobenej náhodnými výkyvmi e_{ip}). Reziduálny súčet štvorcov S_e pôvodného modelu má tvar

$$S_e = \sum_{i=1}^{m}\sum_{p=1}^{n_i}(Y_{ip} - \bar{Y}_i)^2 \ .$$

Reziduálny súčet štvorcov submodelu označme ako S_T. Má tvar

$$S_T = \sum_{i=1}^{m}\sum_{p=1}^{n_i}(Y_{ip} - \bar{Y})^2 \ .$$

Ich rozdiel označme S_A. Popisuje dodatočnú nevysvetlenú variabilitu spôsobenú prechodom od pôvodného modelu k submodelu. Má tvar

$$S_A = \sum_{i=1}^{m}\sum_{p=1}^{n_i}(\bar{Y}_i - \bar{Y})^2 \ .$$

Štatistiku F potom môžeme zapísať ako

$$F = \frac{S_A(n-m)}{S_e(m-1)} \sim F_{m-1,n-m} \ . \tag{1.2}$$

Výsledky testu sa zpravidla zapisujú do tabuľky.

Tabuľka 1.1: Jednoduchá analýza rozptylu

Zdroj menlivosti	Súčet štvorcov SS	Stupne volnosti df	Podiel MS	Testová štatistika	p-hodnota
faktor	S_A	$m-1$	$S_A/(m-1)$	$F = \frac{S_A(n-m)}{S_e(m-1)}$	p
Reziduálny	S_e	$n-m$	$s^2 = S_e/(n-m)$	-	-
Celkový	S_T	$n-1$	-	-	-

Skratky SS, df a MS sú odvodené z anglických slov sum of squares, degrees of freedom a mean squares. Pri testovej štatistike F prípadne pri p-hodnote môžeme jednou, dvoma či prípadne troma hviezdičkami vyjadriť, že sa kritická hodnota prekočí postupne na hladinách 0,05; 0,01; 0,001.

Kapitola 2
Metódy mnohonásobného porovnávania

2.1 Úvod k problematike

Predpokladajme, že je hypotéza $H_0 : \mu_1 = \cdots = \mu_m$ v prvom kroku zamietnutá. Je teda veľmi pravdepodobné, že existujú rozdiely medzi niektorými strednými hodnotami. Pre $m > 2$ ostáva ale otvorené, medzi ktorými základnými súbormi sa tieto rozdiely vyskytujú. Našou prvou myšlienkou môže byť viacnásobné použitie dvojvýberového t-testu. Pre m výberov máme $g = \frac{m(m-1)}{2}$ porovnaní a teda g rôznych čiastkových hypotéz, ktoré nazývajme lokálnymi hypotézami a značme

$$H_{0,k} : \mu_i = \mu_j,\ i < j \quad \text{proti} \quad H_{1,k} : \mu_i \neq \mu_j, \qquad k = 1 \ldots, g$$

$H_{0,k}$ sú navzájom rôzne. Môžeme napríklad značiť

$$H_{0,1} : \mu_1 = \mu_2, \ldots,\ H_{0,m-1} : \mu_1 = \mu_m,\ H_{0,m} : \mu_2 = \mu_3 \ldots,\ H_{0,g} : \mu_{m-1} = \mu_m$$

Pre konkrétnu lokálnu hypotézu $H_{0,k} : \mu_i = \mu_j$ môžme za jej platnosti uvážiť testovú štatistiku

$$T = \frac{\bar{Y}_i - \bar{Y}_j}{S\sqrt{1/n_i + 1/n_j}} \sim t_{n_i+n_j-2}\ \text{, kde} \quad S^2 = \frac{(n_i-1)S_i^2 + (n_j-1)S_j^2}{n_i + n_j - 2}$$

$$\text{a } S_i^2 = \frac{1}{n_i - 1}\sum_{p=1}^{n_i}(Y_{ip} - \bar{Y}_i)^2 \text{ je výberový rozptyl v i-tom súbore.}$$

Pripomeňme, že kritickú hodnotu t rozdelenia o k stupňoch voľnosti definujeme pre dané α ako také číslo $t_k(\alpha)$, pre ktoré platí $P(|T| \geq t_k(\alpha)) = \alpha$. Takže ak $|T| \geq t_{n_i+n_j-2}(\alpha)$, zamietame rovnosť $\mu_i = \mu_j$.

V ďalšom texte označujme hypotézu $H_0 : \mu_1 = \cdots = \mu_m$ ako $H_{0,G}$ a nazývajme ju globálnou. Chybu 1.druhu globálnej hypotézy značme α_G. Skúmajme, čo sa stane s hladinou globálnej hypotézy, ak lokálne hypotézy testujeme na hladine α. Využijme pri tom fakt, že sú výbery navzájom nezávislé.

$$1 - \alpha_G = P(\text{nezamietneme } H_{0,G}|\text{plaí } H_{0,G}) =$$

$$= P(\bigcap_{k=1}^{g}(\text{nezamietneme } H_{0,k}|\text{plaí } H_{0,k})) = \tag{2.1}$$

$$= \prod_{k=1}^{g} P(\text{nezamietneme } H_{0,k}|\text{platí } H_{0,k}) = (1-\alpha)^g \Rightarrow \alpha_G = 1 - (1-\alpha)^g$$

Už pri malom počte vzájomných pozorovaní môže globálna chyba veľmi vzrásť. Pre $m = 3,\ \alpha = 0,05$ dostávame $g = 3$ a $\alpha_G = 0,143$. Ak bude $m = 4$, bude $g = 6$ a chyba vzrastie na $\alpha_G = 0,265$.

2.2 LSD test a Bonferroniho korektúra

Podobne k problému pristupuje LSD(Least Significant Difference) test. Rozdiel oproti dvojvýberovému t-testu bude v tom, že odhad rozptylu σ^2 nestanovíme len z ohľadom na dva testované výbery, ale na základe všetkych m výberov. Použijeme veličinu s^2 danú vzorcom (1.1). Predpokladom je homogenita rozptylov vo všetkých základných súboroch, a teda takto stanovený odhad je presnejší ako odhad v prípade mnohonásobného použitia t-testu. LSD test je teda akési vylepšenie predchádzajúcej metódy.
Za platnosti lokálnej hypotézy o rovnosti $\mu_i = \mu_j$ platí

$$(\bar{Y}_i - \bar{Y}_j) \sim N\left[0, \sigma^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)\right] \Rightarrow \frac{(\bar{Y}_i - \bar{Y}_j)}{\sigma\sqrt{1/n_i + 1/n_j}} \sim N[0,1] \ .$$

Ďalej podľa vety 1.6

$$\frac{s^2(n-m)}{\sigma^2} \sim \chi^2_{n-m} \ .$$

A teda podľa definície t-rozdelenia s využitím faktu z vety 1.7 je

$$T_{i,j} = \frac{\frac{(\bar{Y}_i - \bar{Y}_j)}{\sigma\sqrt{1/n_i + 1/n_j}}}{\sqrt{\frac{s^2(n-m)}{\sigma^2(n-m)}}} = \frac{\bar{Y}_i - \bar{Y}_j}{s\sqrt{1/n_i + 1/n_j}} \sim t_{n-m} \ . \tag{2.2}$$

Takže všetky veličiny $T_{i,j},\ i \neq j,\ i < j,\ i = 1, \ldots, m-1,\ j = 2, \ldots, m$ majú t rozdelenie s $n - m$ stupňami voľnosti. Ak $|T_{i,j}| > t_{n-m}(\alpha)$ kde $t_{n-m}(\alpha)$ je kritická hodnota t-rozdelenia zamietneme lokálnu hypotézu $H_{0,k} : \mu_i = \mu_j$.
Ekvivalentne môžeme písať, že túto hypotézu zamietneme ak

$$\left|\bar{Y}_i - \bar{Y}_j\right| > t_{n-m}(\alpha) s\sqrt{1/n_i + 1/n_j} =: LSD \ . \tag{2.3}$$

Výraz na pravej strane nerovnosti je označovaný ako LSD a udáva medznú hodnotu, pri ktorej je rozdiel uznaný ako signifikantný. Veľkou výhodou testu je jeho jednoduchosť. Test využíva kritické hodnoty t-rozdelenia (prípadne kvantily), ktoré nájdeme v každých štatistických tabuľkách. Pripomeňme, že nutnou súčasťou tohto testu je F-test založený na vzorci (1.2), viď poznámka 2.4 v kapitole 2.6. Pri použití LSD testu sa neuskutočňuje žiadna korektúra hladiny lokálneho testu a tak sa udržuje vysoká hladina α_G. Túto hladinu však nevieme určiť podľa vzorca (2.1) ako to bolo v prípade mnohonásobného použitia dvojvýberového t-testu. V tomto prípade je totiž

$$P(\bigcap_{k=1}^{g}(\text{nezamietneme } H_{0,k}|\text{plaí } H_{0,k})) \neq \prod_{k=1}^{g} P(\text{nezamietneme } H_{0,k}|\text{plaí } H_{0,k}).$$

Čo môžeme rozpísať na

$$P\left(\bigcap_{i<j}\left[\frac{\left|\bar{Y}_i - \bar{Y}_j\right|}{s\sqrt{1/n_i + 1/n_j}} < t_{n-m}(\alpha)\middle|\mu_i = \mu_j\right]\right) \neq$$

$$\neq \prod_{i<j} P\left(\frac{\left|\bar{Y}_i - \bar{Y}_j\right|}{s\sqrt{1/n_i + 1/n_j}} < t_{n-m}(\alpha) \middle| \mu_i = \mu_j \right) .$$

Nerovnosť dostávame, pretože odhad rozptylu s^2 je rovnaký pre všetky veličiny $\frac{\left|\bar{Y}_i - \bar{Y}_j\right|}{s\sqrt{1/n_i+1/n_j}}$, $i < j$, čo spôsobuje závislosť týchto veličín. Hľadáme alternatívny spôsob ako upraviť hladinu lokálnych testov, aby sme aspoň čiastočne uspokojili náš požiadavok na hodnotu globálnej hladiny α_G.

Za platnosti $H_{0,G}$

$$\alpha_G = P(\text{ neoprávnene zamietneme } H_{0,G}) =$$

$$= P(\text{neoprávnene zamietneme aspoň jednu lokálnu hypotézu}) =$$

$$= P(\bigcup_{k=1}^{g} (\text{neoprávnene zamietneme } H_{0,k})) =$$

$$\sum_{k=1}^{g} P(\text{neopr. zamietneme } H_{0,k}) - \sum_{k<l} P(\text{ neopr. zamietneme } H_{0,k} \text{ aj } H_{0,l}) +$$

$$+ \cdots + (-1)^{g-1} P(\text{neoprávnene zamietneme všetky lokálne hypotézy})$$

Tieto pravdepodobnosti sa ale väčšinou vyčísľujú velmi ťažko. Určite ale platí:

$$P(\bigcup_{k=1}^{g} (\text{neopr. zamietneme } H_{0,k})) \leq \sum_{k=1}^{g} P(\text{neopr. zamietneme } H_{0,k}) . \quad (2.4)$$

Vidíme, že globálnu chybu možme kontrolovať prostredníctvom hladín lokálnych hypotéz. Ak zvolíme chybu 1.druhu pre všetky lokálne hypotézy rovnú α, dostaneme, že $\alpha_G \leq g\alpha$. Teda ak požadujeme, aby globálna chyba nepresiahla nejakú hranicu α_G zvolíme $\alpha = \alpha_G/g$. Tento postup nazývame Bonferroniho korektúra hladiny. Bonferroniho test potom prebehne podľa vzorca (2.3) pričom za α volíme korigovanú hladinu.

Príklad: Nech m=4 a teda g=6. Predpokladajme, že jednoduchá ANOVA prebehla na hladine 0,05 a hypotéza o rovnosti priemerov bola zamietnutá. V ďalšom postupe chceme použiť LSD test s korektúrou hladiny tak, aby globálna chyba neprekročila 0,05. Zvolíme teda $\alpha = 0,05/6 = 0,00834$. Vzájomné porovnávania teda vykonáme na hladine 0,00834.

Pravdepodobnostnú štruktúru Bonferroniho testu udáva nasledujúci vzorec.

$$P\left(\bigcap_{i<j} \left[\frac{\left|(\bar{Y}_i - \bar{Y}_j) - (\mu_i - \mu_j)\right|}{s\sqrt{1/n_i + 1/n_j}} \leq t_{n-m}(\alpha) \right] \right) \geq 1 - g * \frac{\alpha_G}{g} = 1 - \alpha_G \quad (2.5)$$

Odvodíme ho rovnako ako sme odvodili vzorec (2.2), pričom použijeme fakt daný vzorcom (2.4).

2.3 Tukeyho metóda

Definícia 2.1 Nech $Y_1, \ldots, Y_m$ je náhodný výber z rozdelenia $N[\mu, \sigma^2]$, kde $\sigma > 0$. Ďalej nech s^2 je takzvaný nezávislý odhad rozptylu σ^2 s ν stupňami voľnosti. To znamená, že $\nu s^2/\sigma^2 \sim \chi^2_\nu$ a že veličiny s^2 a $\boldsymbol{Y} = (Y_1 \ldots, Y_m)'$ sú nezávislé. Potom náhodnú veličinu

$$Q(m, \nu) = \max_{i,j=1,\ldots,m} \frac{|Y_i - Y_j|}{s}$$

nazývame *studentizované rozpätie*. Kritickú hodnotu tohto rozdelenia $q_{m,\nu}(\alpha)$ definujeme ako hodnotu, ktorú nahodná veličina $Q(m, \nu)$ prekročí s pravdepodobnosťou α, tj.$P[Q(m, \nu) > q_{m,\nu}(\alpha)] = \alpha$.
Platí

$$Q(m, \nu) = \max_{i,j=1,\ldots,m} \frac{|(Y_i - \mu)/\sigma - (Y_j - \mu)/\sigma|}{s/\sigma}$$

a teda veličina $Q(m, \nu)$ závisí skutočne len na parametroch m, ν.

Veta 2.2 (Tukeyho) *Nech $Y_1, \ldots, Y_m$ sú nezávislé náhodné veličiny, pričom $Y_i \sim N[\mu_i, \sigma^2/c^2],\ i = 1, \ldots, m;\ c > 0$ je známa konštanta. Nech s^2 je nezávislý odhad pre σ^2s ν stupňami voľnosti. Potom platí*

$$P\left[|(Y_i - Y_j) - (\mu_i - \mu_j)| \le q_{m,\nu}(\alpha)\frac{s}{c}; \text{zároveň pre } i, j = 1 \ldots, m\right] = 1 - \alpha.$$

Dôkaz: Veličiny $Y_i - \mu_i\ i = 1, \ldots, m$ sú náhodným výberom z rozdelenia $N[0, \sigma^2/c^2]$ a veličina s^2/c^2 je nezávislý odhad rozptylu s ν stupňami volnosti. Podľa definície 2.1 je veličina $\max_{i,j=1,\ldots,m} \frac{|(Y_i-\mu_i)-(Y_j-\mu_j)|}{s/c}$ studentizované rozpätie. A teda platí

$$P\left[\max_{i,j=1,\ldots,m} \frac{|(Y_i - \mu_i) - (Y_j - \mu_j)|}{s/c} \le q_{m,\nu}(\alpha)\right] = 1 - \alpha\ ,$$

čo je ekvivalentné s tvrdením vety. □

Predpokladajme rovnosť rozsahov všetkých výberov (balancovaný prípad). Označme $r = n_1 = n_2 = \cdots = n_m$. Pre naše účely máme veličiny $\bar{Y}_1, \ldots, \bar{Y}_m$, kde $\bar{Y}_i \sim N[\mu_i, \sigma^2/r]$ a veličinu

$$s^2 = S_e/(n - m) = \frac{1}{(n - m)} \sum_{i=1}^{m} \sum_{p=1}^{r} (Y_{ip} - \bar{Y}_i)^2,$$

ktorá je nezávislým odhadom rozptylu σ^2 s $n-m$ stupňami voľnosti. Nájdeme ju v tabuľke ANOVA.
Studentizované rozpätie je tvaru:

$$Q(m, n - m) = \max_{i,j=1,\ldots,m} \frac{\left|\bar{Y}_i - \bar{Y}_j\right|}{s}\sqrt{r}\ .$$

Podľa Tukeyho vety teda platí:

$$P[\left|(\bar{Y}_i - \bar{Y}_j) - (\mu_i - \mu_j)\right| \leq q_{m,n-m}(\alpha)\frac{s}{\sqrt{r}}; \text{zároveň pre } i,j = 1\ldots,m] = 1-\alpha. \tag{2.6}$$

To znamená, že neoprávnené zamietnutie globálnej nulovej hypotézy $H_{0,G}$ nastáva s pravdepodobnosťou α, ktorú môžeme nastaviť tak, aby odpovedala hladine F- testu ANOVA. V prípade rovnosti stredných hodnôt je výraz $\mu_i - \mu_j = 0$ a teda test prevedieme tak, že ak nastane

$$\left|\bar{Y}_i - \bar{Y}_j\right| > q_{m,n-m}(\alpha)\frac{s}{\sqrt{r}} =: HSD\ , \tag{2.7}$$

zamietneme lokálnu nulovú hypotézu $H_0 : \mu_i = \mu_j$. Takže Tukeyho test nám ukáže dvojicu (i, j), pre ktorú je rozdiel signifikantný. Rovnako môžeme okamžite zamietnuť hypotézy, pre ktoré platí, že ich príslušná diferencia $\left|\bar{Y}_i - \bar{Y}_j\right|$ je väčšia ako práve testovaná.

Tukeyho metódu môžeme použiť aj v prípade nevybalancovaného modelu. Bolo dokázané (viď [2]), že za platnosti nulovej globálnej hypotézy pre prípad jednoduchej ANOVA dostaneme

$$P\left(\left|\bar{Y}_i - \bar{Y}_j\right| \leq q_{m,n-m}(\alpha)s\sqrt{1/2(1/n_i + 1/n_j)} \text{ pre všetky i,j } \right) \geq 1 - \alpha\ .$$

Ak nastane

$$\left|\bar{Y}_i - \bar{Y}_j\right| > q_{m,n-m}(\alpha)s\sqrt{1/2(1/n_i + 1/n_j)}\ , \tag{2.8}$$

zamietneme hypotézu o rovnosti $\mu_i = \mu_j$. Odhad s^2 je tentokrát samozrejme myslený v jeho všeobecnom tvare pre nie nutne balancované modely tj. vzorec (1.1). Táto modifikácia sa používa v počítačových programoch, medzi ktoré patrí aj program R [5], ktorý využijeme na vyhodnocovaní dát v tretej kapitole. Zvykne sa označovať Tukey HSD. HSD pochádza z anglického výrazu Honest significant difference.

2.4 Student-Newman-Keulsov a Duncanov test

Rovnako ako v prípade Tukeyho metódy predpokladajme, že je model vybalancovaný. Máme priemery $\bar{Y}_1, \ldots, \bar{Y}_m$. Student-Newman-Keulsov test (v ďalšom len SNK) a Duncanov test slúžia na nájdenie homogénnej podskupiny priemerov, pričom homogénnou skupinou rozumieme takú podmnožinu množiny $\{\bar{Y}_1, \ldots, \bar{Y}_m\}$, pre ktorú či už prostredníctvom SNK alebo Duncanovho rozhodovacieho postupu nezamietneme hypotézu o rovnosti strednych hodnôt rozdelení prislúchajúcim k priemerom z podmnožiny. Celá metóda je založená na rovnakom tvrdení ako Tukeyho test (východiskom je Tukeyho veta).

Predpokladajme, že máme daných p priemerov $\bar{Y}_1, \ldots, \bar{Y}_p$ a chceme testovať hypotézu

$$H : \mu_1 = \cdots = \mu_p$$

na hladine α_p . Použijeme odhad s^2 daný vzorcom (1.1), ktorý má $n - m$ stupňov voľnosti, a tak podľa Tukeyovej vety dostaneme :

$$P\left[\left|(\bar{Y}_i - \bar{Y}_j) - (\mu_i - \mu_j)\right| \leq q_{p,n-m}(\alpha_p)\frac{s}{\sqrt{r}}; \text{zároveň pre } i,j = 1,..,p\right] = 1 - \alpha_p.$$

Čo je ekvivalentné so vzorcom

$$P\left[\max_{i=1,\ldots,p}(\bar{Y}_i - \mu_i) - \min_{i=1,\ldots,p}(\bar{Y}_i - \mu_i) \leq q_{p,n-m}(\alpha_p)\frac{s}{\sqrt{r}}\right] = 1 - \alpha_p \ .$$

Za platnosti hypotézy H môžeme písať

$$P\left[\max_{i=1,\ldots,p}\bar{Y}_i - \min_{i=1,\ldots,p}\bar{Y}_i \leq q_{p,n-m}(\alpha_p)\frac{s}{\sqrt{r}}\right] = 1 - \alpha_p \ .$$

K testovaniu homogenity skupiny p priemerov teda stačí porovnať rozdiel najväčšieho a najmenšieho priemeru s kritickou hranicou na pravej strane výrazu.

Uvážme situáciu, že hypotézu H zamietneme. Zaujíma nás, či nejaká podmnožina pôvodnej skupiny priemerov môže byť prehlásená za homogénnu. Nech je v podmnožine k priemerov. Platí $Q(k, n-m) \leq Q(p, n-m)$, z čoho dostaneme $q_{k,n-m}(\alpha) \leq q_{p,n-m}(\alpha)$. Navyše kritická hodnota studentizovaného rozpätia z rastúcim α klesá, a teda pre dve rôzne $\alpha_p \leq \alpha_k$ platí

$$q_{k,n-m}(\alpha_k) \leq q_{p,n-m}(\alpha_p) \ .$$

Z toho vyplýva, že ak pre $\alpha_p \leq \alpha_k$ do podmnožiny zahrnieme najmenší aj najväčší prvok s pôvodnej množiny, nemôžeme dostať homogenitu podmnožiny. Tento záver nás vedie k myšlienke postupne vynechávať najmenšie alebo najväčšie priemery.

SNK a Duncanov test môžeme zapísať vo forme nasledujúceho algoritmu. Usporiadajme všetkých m priemerov od najmenšieho po najväčší. Značíme

$\bar{Y}_{(1)}, \bar{Y}_{(2)}, \ldots, \bar{Y}_{(m)}$. Je teda $\min_{i=1,\ldots,m} \bar{Y}_i = \bar{Y}_{(1)}$ a $\max_{i=1,\ldots,m} \bar{Y}_i = \bar{Y}_{(m)}$. Symbolom $\mu_{(i)}$ značme strednú hodnotu rozdelenia, ktorému odpovedá priemer $\bar{Y}_{(i)}$, $i = 1, \ldots, m$.
1.krok: Ak platí

$$\left|\bar{Y}_{(1)} - \bar{Y}_{(m)}\right| > q_{m,n-m}(\alpha_m)\frac{s}{\sqrt{r}} \ ,$$

zamietneme hypotézu $H_{0,G}$ na hladine α_m a pokračujeme ďalším krokom. V opačnom prípade prehlásime skupinu všetkých m priemerov za homogénnu. Aby podskupina priemerov nedávala signifikantný výsledok nesmie obsahovať oba priemery $\bar{Y}_{(1)}$ a $\bar{Y}_{(m)}$.

2.krok: Skúmame množinu o $p = m - 1$ priemerov. Jediné prípustné podskupiny sú teda: $\bar{Y}_{(1)}, \ldots, \bar{Y}_{(m-1)}$ alebo $\bar{Y}_{(2)}, \ldots, \bar{Y}_{(m)}$. Aplikujeme na ne rovnaký postup. Ak platí

$$\left|\bar{Y}_{(1)} - \bar{Y}_{(m-1)}\right| > q_{m-1,n-m}(\alpha_{m-1})\frac{s}{\sqrt{r}} \ ,$$

zamietame na hladine α_{m-1} hypotézu o rovnosti príslušných stredných hodnôt $(H : \mu_{(1)} = \cdots = \mu_{(m-1)})$ a pokračujeme ďalším krokom, v ktorom skúmame skupiny: $\bar{Y}_{(1)}, \ldots, \bar{Y}_{(m-2)}$ a $\bar{Y}_{(2)}, \ldots, \bar{Y}_{(m-1)}$. Ak nerovnosť neplatí, algoritmus pre túto vetvu končí. V tomto prípade prehlásime skupinu priemerov za homogénnu.
Podobne ak platí

$$\left|\bar{Y}_{(2)} - \bar{Y}_{(m)}\right| > q_{m-1,n-m}(\alpha_{m-1})\frac{s}{\sqrt{r}} \ ,$$

zamietneme na hladine α_{m-1} hypotézu $H : \mu_{(2)} = \cdots = \mu_{(m)}$ a pokračujeme ďalším krokom.V opačnom prípade sme našli homogénnu podskupinu.

Všeobecne: Predpokladajme, že sa v nejakej vetve algoritmu dostaneme ku kroku, v ktorom skúmame skupinu o p priemeroch $p = 2, \ldots, m$. Ak je táto skupina podmnožinou nejakej skupiny, ktorá bola prehlásená za homogénnu, tak ju už ďalej neskúmame a jednoducho ju preskočíme. Ak priemery opäť preusporiadame od najmenšieho po najväčší (iba zmeníme označenie), budeme mať: $\bar{Y}_{(1)}, \ldots, \bar{Y}_{(p)}$ a môžme písať, že ak platí:

$$\left|\bar{Y}_{(1)} - \bar{Y}_{(p)}\right| > q_{p,n-m}(\alpha_p)\frac{s}{\sqrt{r}} \ , \tag{2.9}$$

zamietneme hypotézu o rovnosti p príslušných stredných hodnôt na hladine α_p a ak $p \neq 2$, pokračujeme ďalším krokom, v ktorom skúmame podmnožiny o $p-1$ prvkoch. Ak nerovnosť neplatí, nezamieta sa hypotéza o rovnosti stredných hodnôt a postup sa pre danú množinu zastaví. Po ukončení algoritmu dostaneme niekoľko skupín priemerov, ktoré vykazujú homogenitu.

Rozdiel medzi SNK testom a Duncanovým testom je rozdielna voľba hladiny α_p pre $p = m, \ldots, 2$, ktorú ďalej nazývajme *hladinou pre p priemerov.* Pre SNK je $\alpha_p = \alpha;\ p = m, \ldots, 2$, zatiaľ čo pre Duncanov test volíme

$$\alpha_p = 1 - (1 - \alpha)^{p-1}, \tag{2.10}$$

kde α je nejaká predom zvolená hladina. Napriklad pre $m = 5$ a $\alpha = 0,05$ dostaneme nasledujúce hladiny pre p priemerov.

Tabuľka 2.1: Hladiny pre p priemerov v prípade SNK a Duncanovho testu

p	5	4	3	2
α_p SNK	0,0500	0,0500	0,0500	0,0500
α_p Dunc	0,1855	0,1426	0,0975	0,0500

Hladina pre p priemerov bude v prípade Duncanovej metódy s rastúcim počtom priemerom v skupine narastať, hoci nie tak rýchlo ako napr. v prípade použitia separovaných t-testov. Pre $\alpha = 0,05$ a $p = 10$ dostaneme $\alpha_p = 0,37$, takže nárast je vcelku významný. Voľba α_p podľa vzorca (2.10) súvisí so vzorcom (2.1). Dôvodom takejto voľby je zvýšenie sily testu, pričom dosiahneme väčšiu ochranu globálnej hypotézy ako v prípade separovaných t testov.

2.5 Scheffého metóda

Uveďme znenie Scheffého vety vo všeobecnom tvare pre ľubovolný lineárny model.

Veta 2.3(Scheffého) *Nech* $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ *je vektor nezávislých normálne rozdelených náhodných veličín s rovnakým rozptylom* σ^2*. Nech sa tento vektor riadi lineárnym modelom s maticou* $\boldsymbol{X}_{n\times k}$*, pričom* $h(\boldsymbol{X}) = k$ *(teda* $E\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}$*). Ďalej nech* $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ *je odhad parametru* $\boldsymbol{\beta}$ *(ide o odhad metodóu najmenších štvorcov, ktorý je nestranný) a* s^2 *je nezávislý odhad rozptylu s* $n-k$ *stupňami voľnosti. Nech* Λ *je d-dimenzionálny podpriestor k-dimenzionálneho euklidovského priestoru. Potom*

$$P\left[\left|\boldsymbol{l}'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right| \leq (dF_{d,n-k}(\alpha))^{1/2} s(\boldsymbol{l}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{l})^{1/2}, \forall \boldsymbol{l} \in \Lambda\right] = 1 - \alpha \ ,$$

alebo ekvivalentne

$$P\left[\boldsymbol{l}'\boldsymbol{\beta} \in \boldsymbol{l}'\hat{\boldsymbol{\beta}} \pm (dF_{d,n-k}(\alpha))^{1/2} s(\boldsymbol{l}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{l})^{1/2}, \forall \boldsymbol{l} \in \Lambda\right] = 1 - \alpha \ .$$

Dôkaz: Ukáže sa platnosť ekvivalencie:

pre parameter $\boldsymbol{\beta}$ platí: $(\boldsymbol{L}\boldsymbol{\beta} - \boldsymbol{L}\hat{\boldsymbol{\beta}})'(\boldsymbol{L}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{L}')^{-1}(\boldsymbol{L}\boldsymbol{\beta} - \boldsymbol{L}\hat{\boldsymbol{\beta}}) \leq dF_{d,n-k}(\alpha)s^2$

$$\Leftrightarrow$$

pre parameter $\boldsymbol{\beta}$ platí: $\boldsymbol{l}'\boldsymbol{\beta} \in \boldsymbol{l}'\hat{\boldsymbol{\beta}} \pm (dF_{d,n-k}(\alpha))^{1/2} s(\boldsymbol{l}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{l})^{1/2}, \forall \boldsymbol{l} \in \Lambda$

$\boldsymbol{L}$ je matica typu $d \times k$, ktorej riadkové vektory tvoria bázu priestoru Λ. Navyše sa ukáže že veličina $(\boldsymbol{L}\boldsymbol{\beta} - \boldsymbol{L}\hat{\boldsymbol{\beta}})'(\boldsymbol{L}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{L}')^{-1}(\boldsymbol{L}\boldsymbol{\beta} - \boldsymbol{L}\hat{\boldsymbol{\beta}})/ds^2$ má F rozdelenie s $d, n-k$ stupňami voľnosti. Tvrdenie vety je už potom zrejmé. Dôkaz ekvivalencie je možné nájsť napr. v [3] $\square$

Poznamenajme ešte, že $F_{d,n-k}$ je kritická hodnota F rozdelenia o $d, n-k$ stupňoch voľnosti. Scheffého metóda dáva interval spoľahlivosti pre všetky lineárne kombinácie $\boldsymbol{l}'\boldsymbol{\beta} = \sum_{i=1}^{k} l_i\beta_i$. Možnosť voľby priestoru Λ a regresný charakter dáva vete široké použitie.

Máme veličiny $Y_{i,j} \sim N[\mu_i, \sigma^2],\ i = 1, \ldots, m,\ j = 1, \ldots, n_i$ ako v 1. kapitole. Uvážme tentokrát nepreparametrizovaný model. Za vektor $\boldsymbol{\beta}$ zvolme vektor stredných hodnôt. Vektor $\boldsymbol{Y} = (Y_{1,1}, \ldots, Y_{1,n_1}, Y_{2,1}, \ldots, Y_{2,n_2}, \ldots, Y_{m,n_m})'$ sa riadi lineárnym modelom s maticou $\boldsymbol{X}$, ktorá má m stĺpcov, pričom v i-tom stĺpci je na príslušných n_i miestach 1 a inde 0. Napríklad pre $m = 3$, $n_1 = 2,\ n_2 = 2,\ n_3 = 3$ máme:

$$\boldsymbol{Y} = \begin{pmatrix} Y_{1,1} \\ Y_{1,2} \\ Y_{2,1} \\ Y_{2,2} \\ Y_{3,1} \\ Y_{3,2} \\ Y_{3,3} \end{pmatrix} \qquad \boldsymbol{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}$$

Metódou najmenenších štvorcov ľahko odvodíme odhad $\hat{\boldsymbol{\beta}}$. Pre jednotlivé stredné hodnoty dostaneme odhady $\hat{\mu}_i = \bar{Y}_i$. Z teórie regresných modelov vieme, že vhodný nezávislý odhad rozptylu bude $s^2 = S_e/(n-m)$. Ide o totožný výsledok, ktorý sa dostane pre model s neúplnou hodnosťou ako je uvedený v kapitole 1 vzorcom (1.1).

Hypotéza $H_0 : \mu_1 = \mu_2 = \cdots = \mu_m$ platí práve vtedy, keď $\sum_{i=1}^{m} c_i\mu_i = 0$ pre všetky $\boldsymbol{c} = (c_1, c_2, \ldots, c_m)'$ také, že $\sum_{i=1}^{m} c_i = 0$. Priestor takýchto $\boldsymbol{c}$ tvorí lineárny pristor dimenzie $m-1$, a preto s pravdepodobnosťou $1-\alpha$ platí:

$$\sum_{i=1}^{m} c_i\mu_i \in \sum_{i=1}^{m} c_i\bar{Y}_i \pm ((m-1)F_{m-1,n-m}(\alpha))^{1/2} s \left(\sum_{i=1}^{m} \frac{c_i^2}{n_i}\right)^{1/2} \tag{2.11}$$

$$\forall(c_1, \ldots, c_m), \text{ pre ktoré } \sum_{i=1}^{m} c_i = 0 \ .$$

Pre všetky dvojice (i,j) , $i \neq j$ chceme overiť, že $\mu_i = \mu_j$. Označme $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)'$ a $\boldsymbol{c}_{ij} = (0, \ldots, 0, 1, 0, \ldots, 0, -1, 0, \ldots, 0)'$, kde 1 je na i-tom a -1 na j-tom mieste. Potom $\mu_i = \mu_j$ platí práve vtedy, keď $\boldsymbol{c}'_{ij}\boldsymbol{\mu} = \mu_i - \mu_j = 0$. Po dosadení dostaneme, že s pravdepodobnosťou najmenej $1-\alpha$ platí:

$$0 \in \bar{Y}_i - \bar{Y}_j \pm ((m-1)F_{m-1,n-m}(\alpha))^{1/2} s(1/n_i + 1/n_j)^{1/2} \ ,$$

$$\forall(i,j),\ i \neq j,\ i,j = 1, \ldots, m \ .$$

Postupne testujeme všetky intervaly na prítomnosť nuly. Ak pre nejakú dvojicu (i,j) interval neprekryje 0 zamietneme lokálnu hypotézu $\mu_i = \mu_j$ na hladine aspoň $1-\alpha$. Ekvivalentne môžeme písať, že hypotézu zamietneme, ak

$$\left|\bar{Y}_i - \bar{Y}_j\right| > \sqrt{(m-1)F_{m-1,n-m}(\alpha)s^2(1/n_i + 1/n_j)} \ . \tag{2.12}$$

2.6 Porovnanie metód

Základné techniky mnohonásobného porovnávania môžeme rozdeliť do dvoch skupín. Prvú skupinu tvoria metódy, ktoré poskytujú simultánne intervaly spoľahlivosti. V našom texte ide o Tukeyho, Scheffého a Bonferroniho metódu. Simultánne intervaly spoľahlivosti sú uvedené vzorcami (2.5), (2.6) a (2.11). Druhú skupinu tvoria viackrokové metódy. Sem patria LSD, SNK a Duncanov test.

Poznámka 2.4 *Súčasťou LSD testu je nutne ANOVA F-test založený na vzorci (1.2). Väčšinou sú metódy mnohonásobného porovnávania skonštruované tak, aby chránili celkovú chybu prvého druhu, prípadne ako uvidíme ďalej chybu rodiny(Tukey, Scheffé, Bonferroni...), takže ANOVA F-test je len dodatočnou ochranou a teoreticky ho nie je nutné vykonať. V týchto situáciách sa môže stať aj to, že prostredníctvom ANOVA F-testu a napríklad Tukeyho testu dospejeme k rozdielnym záverom. Prakticky je však zaužívaný postup, že najprv otestujeme globálnu nulovú hypotézu (všetky stredné hodnoty sa rovnajú) ANOVA F-testom a iba v prípade ak zamietneme, pokračujeme niektorou metódou mnohonásobného porovnávania.*

Pri metódach mnohonásobného porovnávania nás zaujíma chyba rodiny (v angličtine sa táto chyba bežne označuje ako experimentwise alebo familywise error rate). Táto chyba sa rovná pravdepodobnosti, že prostredníctvom daného testu nesprávne zamietnem aspoň jednu nulovú lokálnu hypotézu tj.nesprávne zamietneme nejakú rovnosť $\mu_i = \mu_j$. Pripomeňme, že týchto dvojíc je $g = m(m-1)/2$. Prehľad možných situácií ilustruje nasledujúca tabuľka 2.2.

Tabuľka 2.2: Tabuľka chýb

	# neprehlásené za signif.	# prehlásené za signif.	Spolu
# pravdivé lokálne H_0	A	B	f
# nepravdivé lokálne H_0	C	D	g-f
Spolu	$g-E$	E	g

A, B , C, D sú náhodné večiny, ktoré niesú pozorovateľné. Ani počet pravdivých hypotéz f nám nie je známy. Pozorovať môžeme len veličinu E. V duchu tejto tabuľky potom môžeme chybu rodiny vyjadriť ako pravdepodobnoť $P(B \geq 1)$. Za predpokladu, že platí globálna nulová hypotéza ide vlastne o chybu 1.druhu pre $H_{0,G}$. Pre Tukeyho test je podľa vzorca (2.6) zrejmé, že táto chyba je rovná α. Pre Scheffého a Bonferroni test bude zas podľa (2.11) a (2.5) menšia ako α. LSD test konroluje túto chybu za platnosti globálnej nulovej hypotézy prostredníctvom F-testu v prvom kroku. Ak v prvom kroku globálnu hypotézu zamietmene, dostávame sa k druhému kroku, kde už ale nemáme žiadnu ochranu k tej často globálnej nulovej hypotéze, ktorá platí(p stredných hodnôt je rovnakých). Hladina pre p priemerov je pre $p < m$ veľká. Pre SNK a Duncanov test táto chyba v dôsledku viacstupňového charakteru

testov rastie. SNK test ju kontroluje len za predpokladu, že platí globálna nulová hypotéza, pretože $\alpha_m = \alpha$.

Stretávame sa teda so zámerom viac alebo menej ochrániť chybu rodiny. Popritom sa ale skúmalo, čo sa stane zo silou testu. Za platnosti konkrétnej alternatívy je sila testu definovaná ako pravdepodobnosť, že zamietneme globálnu nulovú hypotézu, teda $P(\text{zamietnem } H_{0,G}|\text{platí alternatíva })$. Ako je tvrdené v [3], štatistik sa pri riešení daného problému musí rozhodnúť medzi väčšou silou testu a zvýšenou ochranou pre chybu rodiny. Nižšia chyba rodiny bude znamenať, že test bude slabší a naopak.

Venujme sa na záver ešte jednotlivým testom. LSD test sa v druhom kroku už nijako nesnaží chrániť nulovú hypotézu. Sila testu je vysoká. Môžeme očakávať väčšie množstvo signifikantných výsledkov ako pri použití Tukeyho alebo Scheffého testu. Použitím Bonferroniho LSD testu dostaneme väčšinou podstatne menšiu korigovanú hladinu, čo spôsobí že v porovnaní z LSD testom sú jednotlivé intervaly spoľahlivosti širšie. Dostaneme teda menej signifikantných výsledkov. Ochrana nastavená vzorcom (2.5) spôsobí zníženie sily testu. Čo sa týka Tukeyho a Scheffého testu, predstavuje Tukeyho test o niečo citlivejší prístup, pretože Scheffého test nevyužíva celú pravdepodobnosť 1.druhu.

Pozrime sa na rozdiel medzi Tukeyho, SNK a Duncanovým testom. Skúmajme pritom čo platí pre hladinu p priemerov. V kapitole 2.4 sme ukázali, že pre SNK sa zvolí $\alpha_p = \alpha$, ktorá je rovnaká pre všetky skupiny. Pre Duncanov test hladina pre p priemerov z pôvodnej $\alpha_2 = \alpha$ narastá zo zvyšujúcim sa počtom priemerov v skupine. Podľa vzorca

$$P(Q_{p,n-m} > q_{m,n-m})$$

môžme α_p vypočítať aj pre Tukeyho test. Zistíme, že táto hladina z pôvodnej $\alpha_m = \alpha$ postupne klesá. Tak napríklad pre m=5 a klasickej voľbe $\alpha = 0,05$ dostaneme $\alpha_5 = 0,05; \alpha_4 = 0,033; \alpha_3 = 0,019$ a $\alpha_2 = 0,007$. Pre porovnanie si všimnime hodnoty v tabuľke (2.1). Výsleky Tukeyho testu môžme interpretovať taktiež zostrojením homogénnych skupín. V takomto prípade nás predchádzajúce zistenia vedú k záveru, že sa pôvodné homogénne skupiny dané Tukeyho testom pri použití SNK testu rozpadajú na menšie a pri použití Duncanovho testu na ešte menšie skupiny. Tento jav môžme pozorovať v príklade 2 v kapitole 3 viď. tabuľka 3.15. Ako už bolo spomenuté v kapitole 2.4 následkom zvyšovania hladín pre p priemerov je Duncanov test spomedzi týchto testov najsilnejší avšak poskytuje najnižšiu ochanu pre chybu rodiny. Naopak, Tukeyho test poskytuje dobrú ochranu, za čo zaplatíme zníženou silou.

Kapitola 3

Testovanie dát

3.1 Úloha o hladine séra

V tejto kapitole aplikujeme popísané metódy na reálne dáta. K výpočtom použijeme štatistický program R. Uvažujeme nasledujúci problém: Výskumný pracovník chce porovnať hladinu séra t3 u pacientov, ktorí sa podrobili piatim rôznym liečbam. Zozbierané dáta sú uvedené v nasledujúcej tabuľke 3.1.

Tabuľka 3.1: Zozbierané dáta

Liečba	Sérum t3	Liečba	Sérum t3	Liečba	Sérum t3
1	94,090	3	197,180	5	82,940
1	90,450	3	207,310	5	83,140
1	99,380	3	177,500	5	89,590
1	73,560	3	226,050	5	96,430
1	74,390	3	222,740	5	96,430
2	98,810	4	102,930		
2	103,550	4	117,510		
2	115,230	4	119,920		
2	129,060	4	112,010		
2	117,610	4	101,100		

Všetci pacienti sú teda rozdelení do $m = 5$ skupín podľa toho, akú liečbu podstúpili (faktor). V každej skupine je 5 pacientov a teda ide o vybalancovaný model. Máme $n_1 = n_2 = n_3 = n_4 = n_5 = r = 5$. Celkový počet pacientov je $n = 25$. Sformulujme hypotézu:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

$$H_1 : \mu_i \neq \mu_j \text{ pre aspoň jednu } (i, j),\ i \neq j,\ i, j = 1, \ldots, 5\ .$$

K ďalšiemu potrebujeme ešte overiť, či základné súbory vykazujú predpoklad normality a homogenity rozptylu. Na test normality nám program R ponúka napríklad Shapirov-Wilkov test. Tento test jednoznačne preukáže, že výber môže byť pokladaný ako za výber z normálneho rozdelenia (dostávame vysoké p-hodnoty). K overeniu homogenity rozptylov použijeme Barlettov test. Pre tento test dostaneme p-hodnotu $0,1942$, čo teda nevyvracia predpoklad homogenity.

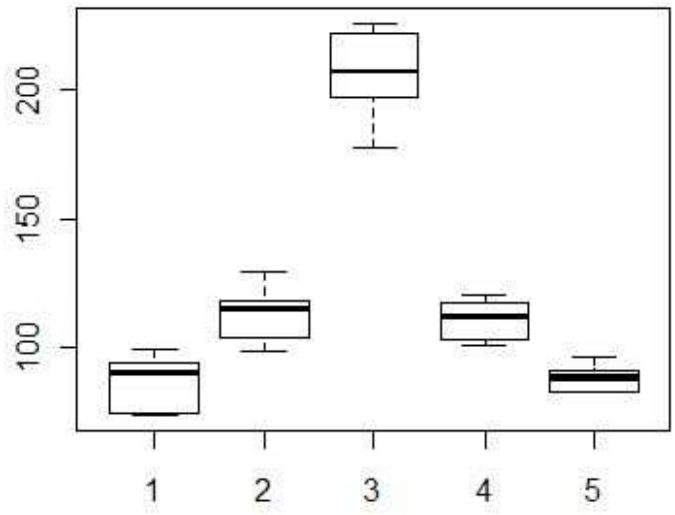
Teraz nám už nič nebráni k použitiu testovej štatistiky (1.2) a vyhodnoteniu testu. Volíme hladinu $\alpha = 0,05$. Program R má jednoduchú analýzu rozptylu priamo implementovanú. Výsledky testu dostaneme vo forme tabuľky ANOVA.

Tabuľka 3.2: Analýza rozptylu dát

Zdroj menlivosti	SS	df	MS	hodnota F	p-hodnota
Faktor(liečba)	48569	4	12142	78,084	6,479e-12 ***
Reziduálny	3110	20	156		
Celkový	51679	24			

Keďže p-hodnota vychádza blízka 0, môžeme na hladine $\alpha = 0,05$ zamietnuť hypotézu H_0. Týmto sme dokončili prvý krok a prechádzame ku kroku druhému. Zisťujeme, medzi ktorými skupinami sú významné rozdiely. Poznamenajme ešte, že $s^2 = 155,5031$. P-hodnotu definujeme ako najmenšiu hladinu α pri ktorej zamietame hypotézu. Skôr ako pristúpime k jednotlivým metódam uveďme pre ilustráciu krabicový graf dát, pomocou ktorého si môžeme vytvoriť základnú predstavu o rozdieloh medzi jednotlivými súbormi. Napríklad z neho jasne vidíme, že môžeme očakávať rozdiely medzi tretím a ostatnými súbormi.

Graf 3.1: Krabicový graf dát z tabuľky 3.1



LSD test

Jednotlivé párové porovnávania budeme vykonávať na hladine $\alpha = 0,05$. Globálna hladina testu bude samozrejme väčšia. Ďalej máme $t_{n-m}(\alpha) = qt_{n-m}(1 - \alpha/2)$, kde qt je kvantil t rozdelenia a teda

$$t_{20}(0,05) = qt_{20}(0,975) = 2,085963.$$

Kvantily uvádzame, pretože s nimi pracuje program R. Nakoniec po dosadení do (2.3) dostávame, že

$$t_{20}(0,05)s\sqrt{1/5 + 1/5} = 16,45153\ .$$

Ak je absolútna hodnota rozdielu priemerov väčšia ako toto číslo, zamietneme rovnosť príslušných stredných hodnôt. Na rýchle vyhodnotenie dát v programe R môžeme použiť funkciu LSDCI [6], ktorá dáva nasledujúcu tabuľku.

Tabuľka 3.3: Výsledky LSD testu

Skupiny	Rozdiely priemerov	Dolná hranica 95% int. spoľahlivosti	Horná hranica 95% int. spoľahlivosti	p-hodnota
1-2	-26,4726	-42,9241	-10,0210	*0,0031
1-3	-119,7820	-136,2335	-103,3304	*0,0000
1-4	-24,3200	-40,7715	-7,8684	*0,0059
1-5	-1,5980	-18,0495	14,8535	0,8415
2-3	-93,3094	-109,7609	-76,8579	*0,0000
2-4	2,1526	-14,2989	18,6041	0,7877
2-5	24,8746	8,4231	41,3261	*0,0050
3-4	95,4620	79,0105	111,9135	*0,0000
3-5	118,1840	101,7325	134,6355	*0,0000
4-5	22,7220	6,2705	39,1735	*0,0092

Z tejto tabuľky ľahko zistíme, ktoré rozdiely sú signifikantné. Platí totiž, že rozdiel priemerov je signifikantný, ak 95% interval spoľahlivosti neprekryje 0, alebo ak je p-hodnota menšia ako 0,05 (v tabuľke vyznačíme * pri p-hodnote).

BONFERRONI-LSD test

Tentokrát požadujeme, aby globálna hladina α_G bola menšia alebo rovná 0,05. Takže z Bonferroniho korektúry dostaneme, že jednotlivé párove porovnávania majú prebehnúť na hladine $\alpha = 0,05/g = 0,005$. Ďalej

$$t_{20}(0,005) = qt_{20}(0,9975) = 3,153401 \ ,$$

takže po dosadaní do vzorca (2.3) s využitím korektúry je

$$t_{20}(0,005)s\sqrt{1/5 + 1/5} = 24,87016 \ .$$

Ak je absolútna hodnota rozdielu priemerov väčšia ako toto číslo, zamietneme rovnosť príslušných stredných hodnôt. V R môžeme použiť funkciu BonferroniCI [6], ktorá dáva nasledujúcu tabuľku:

Tabuľka 3.4: Výsledky Bonferroniho testu

Skupiny	Rozdiely priemerov	Dolná hranica 95% int. spoľahlivosti	Horná hranica 95% int. spoľahlivosti	p-hodnota
1-2	-26,4726	-51,3428	-1,6024	*0,0314
1-3	-119,7820	-144,6522	-94,9118	*0,0000
1-4	-24,3200	-49,1902	0,5502	0,0586
1-5	-1,5980	-26,4682	23,2722	1,0000
2-3	-93,3094	-118,1796	-68,4392	*0,0000
2-4	2,1526	-22,7176	27,0228	1,0000
2-5	24,8746	0,0044	49,7448	*0,0499
3-4	95,4620	70,5918	120,3322	*0,0000
3-5	118,1840	93,3138	143,0542	*0,0000
4-5	22,7220	-2,1482	47,5922	0,0924

Je potrebné si uvedomiť, že v tomto prípade odpovedá interval spoľahlivosti a p-hodnota hladine α_G a nie α, ako to bolo v prípade LSD testu. Ak je p-hodnota menšia ako 0,05 je rozdiel signifikantný.

TUKEYHO test

Opať volíme $\alpha = 0,05$. Globálna hladina testu je rovná presne tejto hodnote. R dáva

$$q_{5,20}(0,05) = qtukey_{5,20}(0,95) = 4,231857 \ ,$$

kde *qtukey* je kvantil studentizovaného rozpätia. Po dosadení do (2.7) máme

$$q_{5,20}(0,05)s/\sqrt{5} = 23,60019 \ .$$

Ostáva porovnať s rozdielmi priemerov. V R použime funkciu TukeyCI [6] a obdržíme tabuľku 3.5.

Tabuľka 3.5: Výsledky Tukeyho testu

Skupiny	Rozdiely priemerov	Dolná hranica 95% int. spoľahlivosti	Horná hranica 95% int. spoľahlivosti	p-hodnota
1-2	-26,4726	-50,0728	-2,8724	*0,0233
1-3	-119,7820	-143,3822	-96,1818	*0,0000
1-4	-24,3200	-47,9202	0,7198	*0,0414
1-5	-1,5980	-25,1982	22,0022	0,9996
2-3	-93,3094	-116,9096	-69,7092	*0,0000
2-4	2,1526	-21,4476	25,7528	0,9987
2-5	24,8746	1,2744	48,4748	*0,0358
3-4	95,4620	71,8618	119,0622	*0,0000
3-5	118,1840	94,5838	141,7842	*0,0000
4-5	22,7220	-0,8782	46,3222	0,0627

Opäť ak je p-hodnota menšia ako 0,05, sú rozdiely signifikantné.

SCHEFFÉHO test

Pri $\alpha = 0,05$ je globálna hladina menšia ako 0,05. Potrebujeme vypočítať kritickú hodnotu $F_{m-1,n-m}(\alpha)$. Z R dostaneme, že

$$F_{4,20}(0,05) = qF_{4,20}0,95 = 2,866081 \ ,$$

pričom qF je kvantil F rozdelenia. Nakoniec vyčíslime pravú stranu vzorca (2.12). Dostaneme

$$\sqrt{4F_{4,20}(0,05)s^2(1/5+1/5)} = 26,703841 \ .$$

Funkcia ScheffeCI [6] dáva tabuľu 3.6.

Tabuľka 3.6: Výsledky Scheffeho testu

Skupiny	Rozdiely priemerov	Dolná hranica 95% int. spoľahlivosti	Horná hranica 95% int. spoľahlivosti	p-hodnota
1-2	-26,4726	-53,1764	0,2312	0,0528
1-3	-119,7820	-146,4858	-93,0782	*0,0000
1-4	-24,3200	-51,0238	2,3838	0,0864
1-5	-1,5980	-28,3018	25,1058	0,9998
2-3	-93,3094	-120,0132	-66,6056	*0,0000
2-4	2,1526	-24,5512	28,8564	0,9993
2-5	24,8746	-1,8292	51,5784	0,0763
3-4	95,4620	68,7582	122,1658	*0,0000
3-5	118,1840	91,4802	144,8878	*0,0000
4-5	22,7220	-3,9818	49,4258	0,1222

SNK a DUNCANOV test

Hľadáme homogénne podskupiny priemerov. Usporiadame priemery od najmenšieho po najväčší. Viď nasledujúca tabuľka.

Tabuľka 3.7: Usporiadanie priemerov

Odpovedajúca liečba	1	5	4	2	3
Usporiad. priemery	86,3740	87,9750	110,6940	112,8466	206,1560

Opäť volíme $\alpha = 0,05$. Pomocou programu R ľahko realizujeme potrebné kalkulácie. Potrebné hodnoty zapíšeme do tabuľky 3.8.

Tabuľka 3.8: Hladiny a kritické hodnoty pre SNK a Duncanov test

p	5	4	3	2
α_p SNK	0,0500	0,0500	0,0500	0,0500
α_p Dunc	0,1855	0,1426	0,0975	0,0500
$q_{p,n-m}(\alpha_p)$ SNK	4,2319	3,9583	3,5779	2,9500
$q_{p,n-m}(\alpha_p)$ Dunc	3,2546	3,1896	3,0965	2,9500
$q_{p,n-m}(\alpha_p)s/\sqrt{r}$ SNK	23,6002	22,0746	19,9534	16,4515
$q_{p,n-m}(\alpha_p)s/\sqrt{r}$ Dunc	18,1504	17,7878	17,2686	16,4515

V tabuľke 3.8 môžeme taktiež pozorovať, že kritické hodnoty pre Duncanov test sú menšie ako hodnoty prislúchajúce SNK testu, čo spôsobuje, že Duncanov test dáva viac alebo rovnaký počet signifikantných výsledkov.

Pristúpme k testovému algoritmu. Všetky potrebné hodnoty prečítame z tabuliek 3.7 a 3.8.

1. krok: Zaujíma nás rozdiel $\left|\bar{Y}_1 - \bar{Y}_3\right| = 119,7820$. Ten je väčší ako 23,6002 či 18,1504, a preto zamietame homogenitu skupiny všetkých priemerov pre oba testy.

2. krok: Máme skupiny štyroch priemerov pričom $\left|\bar{Y}_1 - \bar{Y}_2\right| = 26,4726$ a $\left|\bar{Y}_5 - \bar{Y}_3\right| = 118,1840$. Tentokrát z tabuľky berieme hodnoty pre p=4. Opäť zamietneme homogenitu v prípade oboch testov.

3. krok: Máme $\left|\bar{Y}_1 - \bar{Y}_4\right| = 24,320$, $\left|\bar{Y}_5 - \bar{Y}_2\right| = 24,8716$, $\left|\bar{Y}_4 - \bar{Y}_3\right| = 95,4620$, čo sú hodnoty väčšie ako príslušné hodnoty v tabuľke. Takže aj v tomto kroku zamietneme homogenitu skupín pre oba testy.

4. krok: Ostávajú už len skupiny o dvoch priemeroch. Postupne spočítame $\left|\bar{Y}_1 - \bar{Y}_5\right| = 1,5980$, $\left|\bar{Y}_5 - \bar{Y}_4\right| = 22,7220$, $\left|\bar{Y}_4 - \bar{Y}_2\right| = 2,1526$ a $\left|\bar{Y}_2 - \bar{Y}_3\right| = 93,3094$.

Pre oba testy dostávame totožný výsledok: Prvá homogénna skupina je tvorená liečbami 1 a 5. Druhá pozostáva z liečby 2 a 4. Pre tento príklad teda nedostaneme žiadne väčšie homogénne skupiny priemerov a navyše výsledky Duncanovho a SNK testu sú rovnaké. V ďalšej kapitole preto uveďme ešte jednu úlohu. Predtým ako k tomu pristúpime zapíšme pre prehľadnosť výsledky LSD, LSD-Bonferroniho, Tukeyho a Scheffého testu do tabuľky 3.9. Ak je rozdiel priemerov pre daný test signifikantný značíme to v tabuľke príslušným symbolom(L - LSD, B - Bonferroni, T - Tukey, S - Scheffé) inak značíme symbolom -.

Tabuľka 3.9: Zhrnutie výsledkov

Skupiny	Signifikantné pre				Skupiny	Signifikantné pre			
1-2	L	B	T	-	2-4	-	-	-	-
1-3	L	B	T	S	2-5	L	B	T	-
1-4	L	-	T	-	3-4	L	B	T	S
1-5	-	-	-	-	3-5	L	B	T	S
2-3	L	B	T	S	4-5	L	-	-	-

V tabuľke 3.9 teda môžeme pozorovať rozdielne výsledky jednotlivých metód. V zmysle kapitoly 2.6 si štatistik vyberie tú metódu, ktorá preňho predstavuje prijateľnú chybu rodiny a silu. Pripadne zmení hladinu α.

3.2 Úloha o IQ

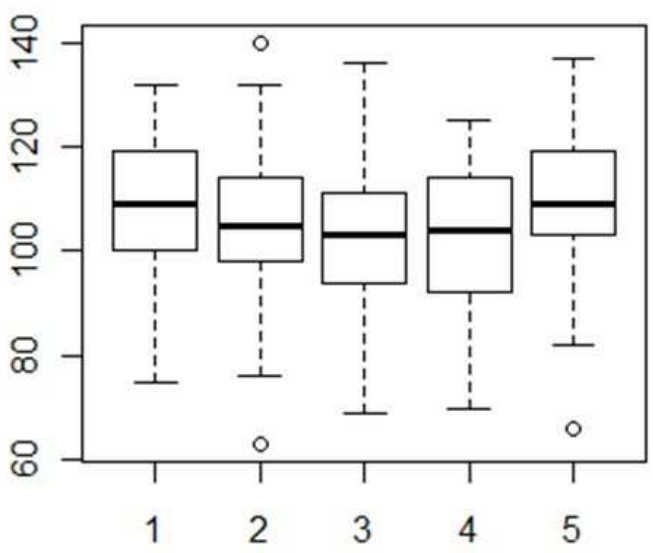
V meste je 5 škôl s rovnakým oborom štúdia. Chceme určiť, či sa na každú školu hlásia rovnako inteligentní študenti a ak nie, tak určiť ktoré školy vykazujú v tomto zmysle homogenitu. Inteligenciu študentov meriame pomocou IQ koeficientu. Na každú školu sa hlási zhruba rovnaký počet študentov. Vyberme náhodne 66 študentov z každej školy a zmerajme ich IQ. Výsledky meraní uveďme pre stručnosť vo forme priemerov v tabuľke 3.10.

Tabuľka 3.10: Priemery zozbieraných dát

Skupina(škola)	1	2	3	4	5
Priemer	107,6515	105,3030	102,9697	102,3485	109,1515

Krabicový graf dát ukáže, že tentokrát nie sú rozdiely až tak zrejmé.

Graf 3.2: Krabicový graf dát k 2.príkladu



Pre túto situáciu máme $n_1 = \cdots = n_5 = r = 66,\ n = 330,\ n - m = 325,$ $\alpha = 0,05$. Presná hodnota veličiny s^2 je $187,7733$. Výbery spĺňajú predpoklad normality aj homogenity rozptylov. Výsledky analýzy rozptylu sú uvedené v nasledujúcej tabuľke 3.11.

Tabuľka 3.11: Analýza rozptylu dát

Zdroj menlivosti	SS	df	MS	hodnota F	p-hodnota
Faktor(Škola)	2266	4	567	3,0171	0,01822 *
Reziduálny	61026	325	188		
Celkový	63292	329			

Keďže je p-hodnota menšia ako 0,05, zamietneme hypotézu, že na všetky školy sa hlásia rovnako inteligentní študenti. V ďalšom sa obmedzme na Tukeyho

Duncenovu a SNK metódu. V R zavolajme na naše dáta funkciu TukeyCI [6], čím dostaneme tabuľku 3.12.

Tabuľka 3.12: Výsledky Tukeyho testu

Skupiny	Rozdiely priemerov	Dolná hranica 95% int. spoľahlivosti	Horná hranica 95% int. spoľahlivosti	p-hodnota
1-2	2,3485	-4,1950	8,8920	0,8622
1-3	4,6818	-1,8617	11,2253	0,2867
1-4	5,3030	-1,2407	11,8465	0,1738
1-5	-1,5000	-8,0435	5,0435	0,9703
2-3	2,3333	-4,2102	8,8768	0,8650
2-4	2,9545	-3,5890	9,4980	0,7287
2-5	-3,8485	-10,3920	2,6950	0,4899
3-4	0,6212	-5,9223	7,1647	0,9990
3-5	-6,1818	-12,7253	0,3617	0,0743
4-5	-6,8030	-13,3465	-0,2595	*0,0370

P-hodnota je menšia ako 0,05 len v prípade 4-5. Výsledkom testu je teda tvrdenie, že signifikantné rozdiely vykazuje len štvrtá a piata škola. Pozrime sa ďalej na SNK a Duncanov test. Tabuľka 3.13 udáva podobne ako v minulom prípade kritické hodnoty pre rôzne p a rôzne metódy.

Tabuľka 3.13: Kritické hodnoty pre SNK a Duncanov test

p	5	4	3	2
$q_{p,n-m}(\alpha_p)s/\sqrt{r}$ SNK	6,5435	6,1600	5,6164	4,6928
$q_{p,n-m}(\alpha_p)s/\sqrt{r}$ Dunc	5,2276	5,1056	4,9402	4,6928

Tabuľka 3.14: Usporiadanie priemerov

Odpovedajúca škola	4	3	2	1	5
Usporiad. priemery	102,3485	102,9697	105,3030	107,6415	109,1515

1. krok: $\left|\bar{Y}_4 - \bar{Y}_5\right| = 6,8030 > 6,5435 > 5,2276$ ⇒nehomogenita pre obe metódy.

2. krok: $\left|\bar{Y}_3 - \bar{Y}_5\right| = 6,1818 > 6,1600 > 5,1056$ ⇒nehomogenita pre obe metódy.
   $\left|\bar{Y}_4 - \bar{Y}_1\right| = 5,3030 < 6,1600$ ale $5,3030 > 5,1056$ ⇒ nehomogenita pre Duncena, homogenita pre SNK

3. krok: $\left|\bar{Y}_2 - \bar{Y}_5\right| = 3,8485 < 4,9402 < 5,6164$ ⇒homogenita pre obe metódy.
   Skupina obsahujúca priemery zo škôl 1, 2 a 3 je obsiahnutá v nadskupine, ktorá je podľa SNK metódy homogénna a preto ju už neskúmame. Týmto algoritmus pre SNK test končí. Pokračujeme pre Duncana.

$\left|\bar{Y}_3 - \bar{Y}_1\right| = 4,6818 < 4,9402 \Rightarrow$homogenita skupiny
$\left|\bar{Y}_4 - \bar{Y}_2\right| = 2,9545 < 4,9402 \Rightarrow$homogenita skupiny
Algoritmus končí.

Výsledky testov sa niekedy zvyknú zapisovať nasledújucim spôsobom:

Tabuľka 3.15: Výsledky testov

Tukey

4	3	2	1	5
102,3485	102,9697	105,3030	107,6515	109,1515

SNK

4	3	2	1	5
102,3485	102,9697	105,3030	107,6515	109,1515

Duncan

4	3	2	1	5
102,3485	102,9697	105,3030	107,6515	109,1515

Priemery podčiarknuté rovnakou čiarou patria do rovnakej homogénnej skupiny. Štatistik zvolí tú metódu, ktorá najlepšie odpovedá jeho požiadavkam kladených na silu testu a ochranu chyby rodiny.

Literatúra

[1] Anděl, J.: Základy matematické statistiky.
Matfyzpress 2007

[2] Hayter, A.J.:A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative.
Ann. Statist. 45, 1984

[3] Müller, R.G. Jr.: Simultaneous Statistical Inference (2nd ed.).
Springer, 1981

[4] Rönz, B.: Computergestützte Statistik 1, Skript.
Humboldt-Universität zu Berlin, 2001

[5] R Development Core Team: R:A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0, URL http://www.R-project.org.

[6] Wyseure, G.: URL http://www.agr.kuleuven.ac.be/vakken/statisticsbyr/ANOVAbyRr/multiplecompJIMRC.htm

Prílohy

K práci je priložené CD s elektronickou formou bakalárskej práce. V adresári prílohy sa nachádzajú testované dáta a kópia pracovneho prostredia z programu R. Je možné nájsť tu funkcie LSDCI, BonferroniCI, TukeyCI a ScheffeCI, ktoré boli použité k rýchlemu vyhodnoteniu dát.