UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

VIACROZMERNÉ RANKOVÉ TESTY ZOHĽADŇUJÚCE CLUSTEROVANIE

DIPLOMOVÁ PRÁCA

2020 Bc. Dávid JABLONICKÝ

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

VIACROZMERNÉ RANKOVÉ TESTY ZOHĽADŇUJÚCE CLUSTEROVANIE

DIPLOMOVÁ PRÁCA

Študijný program:	Ekonomicko-finančná matematika a modelovanie
Študijný odbor:	1114 Aplikovaná matematika
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Vedúci práce:	Mgr. Ján Somorčík, PhD.

Bratislava 2020 **Bc. Dávid JABLONICKÝ**

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Dávid Jablonický
Študijný program: ekonomicko-finančná matematika a modelovanie (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: aplikovaná matematika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Viacrozmerné rankové testy zohľadňujúce clusterovanie
Multivariate cluster-adjusted rank-based tests

Anotácia: Zmyslom práce by malo byť vylepšenie či zovšeobecnenie už jestvujúcich štatistických testov a následné porovnanie s konkurenciou. Podkladové vedecké články sú z nedávnych rokov. Trochu zjednodušene povedané, ide o veľmi pokročilé testovanie odlišnosti medzi skupinami napr. pacientov, pričom u každého pacienta sa meria viacero zdravotných ukazovateľov. Pod 'zohľadnením clusterovania' sa myslí prihliadnutie na skutočnosť, že niektoré ukazovatele merané na pacientovi tvoria prirodzené skupiny (tzv. 'clusters') na základe silnej vzájomnej podobnosti.

Vedúci: Mgr. Ján Somorčík, PhD.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Daniel Ševčovič, DrSc.

Dátum zadania: 07.01.2019

Dátum schválenia: 08.01.2019 prof. RNDr. Daniel Ševčovič, DrSc.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie Chcem sa poďakovať svojmu vedúcemu diplomovej práce Mgr. Jánovi Somorčíkovi, PhD. za pomoc, ochotu a cenné rady, ktoré mi pri písaní práce poskytol.

Abstrakt v štátnom jazyku

JABLONICKÝ, Dávid: Viacrozmerné rankové testy zohľadňujúce clusterovanie [Diplomová práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: Mgr. Ján Somorčík, PhD., Bratislava, 2020, 63s.

V našej práci sa venujeme niekoľkým viacrozmerným rankovým testom. Sú to metódy z oblasti neparametrickej štatistiky. Slúžia na testovanie odlišnosti medzi súbormi pri uvažovaní zovšeobecnenej Behrens-Fisher hypotézy. Skúmané metódy majú využitie najmä v klinických štúdiách, kde je bežná clusterová (zhluková) štruktúra meraných charakteristík. V rámci týchto zhlukov preukazujú charakteristiky väčšie vzájomné podobnosti, ako charakteristiky z rôznych zhlukov. Jednou zo skúmaných metód je nedávna ranková metóda zohľadňujúca takého zhlukovanie. Táto metóda bola vyvinutá na testovanie vzájomnej odlišnosti iba dvoch súborov. Preto navrhujeme jej zovšeobecnenie pre prípad viacerých súborov. Odvádzame asymptotické vlastnosti navrhnutej testovej štatistiky a taktiež odhady potrebných parametrov. Na základe toho uvádzame postup výpočtu p-value navrhnutého testu. Pravdepodobnosť chyby 1. druhu a sila testu je skúmaná pomocou počítačových simulácií v prostredí R.

Kľúčové slová: Klinické štúdie, Neparametrická štatistika, Rankové testy, Viacrozmerné testy, Viacsúborové porovnania, Zovšeobecnená Behrens-Fisher hypotéza

Abstract

JABLONICKÝ, Dávid: Multivariate cluster-adjusted rank-based tests [Master Thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: Mgr. Ján Somorčík, PhD., Bratislava, 2020, 63p.

In our work we deal with several multivariate rank-based tests. These are non-parametric methods. They serve for testing differences among samples while considering the generalized Behrens-Fisher hypothesis. Investigated methods are used especially in clinical trials where cluster structure of measured characteristics is common. The characteristics within these clusters share more similarities than the characteristics from different clusters. One of the examined methods is a recent rank-based method that takes such clustering into account. This method was developed for testing differences between two samples only. Therefore, we propose its generalization for the case of multiple samples. We derive asymptotic properties of the proposed test statistic as well as estimates of necessary parameters. Based on this, we present a procedure to calculate the p-value of the proposed test. The type 1 error rate and the power of the test is investigated by means of computer simulations in the R environment.

Keywords: Clinical trials, Generalized Behrens-Fisher hypothesis, Multi-sample comparison, Multivariate tests, Nonparametric statistics, Rank-based tests

Obsah

Úvod

Predtým, ako je nejaký nový liek, vakcína, či iná forma liečby schválená na použitie, musí prejsť tzv. klinickými štúdiami. Klinické štúdie sú experimenty, či pozorovania robené v klinickom výskume, ktoré majú dva ciele. Prvým cieľom je zistiť účinnosť (efficacy) danej liečby, teda či je daná liečba dostatočne dobrá a druhým je zistiť jej bezpečnosť (safety). Účinnosť aj bezpečnosť liečby sú vyhodnocované relatívne k tomu, aké iné spôsoby liečby sú k dispozícií alebo aká je vážnosť danej choroby či stavu, pričom jej výhody musia vždy prevážiť jej možné riziká. Štatistika hrá veľmi významnú rolu pri vyhodnocovaní spomínanej účinnosti liečby. Využitie štatistických metód v klinických štúdiách umožňuje výskumníkom vyvodzovať zo zozbieraných dát zmysluplné závery na základe ktorých sa môžu rozhodnúť v prípade nejednoznačnosti.

Účinnosť sa v klinických štúdiách meria zvyčajne na základe väčšieho počtu charakteristík. Štatistické metódy používané v tejto oblasti preto možno rozdeliť na dve skupiny; na jednorozmerné testy a na viacrozmerné (globálne) testy. V prípade jednorozmerných testov sa posúdi účinnosť pre každú jednu charakteristiku zvlášť a o celkovej účinnosti liečby sa rozhodne napr. pomocou použitia Bonferroni korekcie. V prípade viacrozmerných testov možno vyhodnotiť celkovú účinnosť liečby okamžite. Medzi bežne používané viacrozmerné metódy v tejto oblasti patria parametrické metódy odvodené za predpokladu normality ako viacrozmerná ANOVA (MANOVA) alebo Hotellingov T^2 test. My sa budeme v našej práci zaoberať výhradne neparametrickými viacrozmernými metódami. Pôjde o rankové testy, ktoré o účinnosti liečby rozhodujú na základe testovania zovšeobecnenej Behrens-Fisher hypotézy. Konkrétne sa budeme venovať tzv. rank-sum-type testom (O'Brien [4], Huang et al. [1]), T_{max} testu (Liu et al. [3]) a T_{cMAX} testu (Zhang et al. [6]). Všetky tieto metódy vyhodnocujú účinnosť na základe porovnania nameraných charakteristík výhradne dvoch spôsobov liečby (súborov). V klinických štúdiách je bežná clusterová (zhluková) štruktúra meraných charakteristík, kde takéto zhluky predstavujú skupiny vzájomne podobných charakteristík. Posledným z vyššie uvedených testov je aktuálny test z minulého roka, ktorý vhodným spôsobom prihliada na takéto clusterovanie. Teoretické poznatky týkajúce sa týchto metód budeme čerpať najmä z pôvodných článkov [1], [3], [4] a [6].

Cieľom našej práce je skúmať výhody a nevýhody vyššie uvedených viacrozmerných

rankových testov a na základe získaných poznatkov navrhnúť zovšeobecnenie spomínaného T_{cMAX} testu pre prípad porovnania viacerých súborov. Chceme odvodiť asymptotické vlastnosti takto zovšeobecneného testu, na základe ktorých budeme môcť skúmať jeho kvalitu pomocou počítačových simulácií v prostredí R [5].

1 Metódy na porovnanie dvoch súborov

V tejto kapitole sa budeme venovať metódam slúžiacim na testovanie odlišnosti polohy dvoch súborov z viacrozmerných rozdelení. Pôjde o testy z oblasti neparametrickej štatistiky, ktoré namiesto pôvodných dát pracujú len s ich usporiadaním - rankami. Najskôr opíšeme konkrétne akým testovaním sa budeme zaoberať, uvedieme základné označenia a testovú hypotézu, za ktorej platnosti boli všetky sledované metódy odvodené. Konkrétne pôjde o nasledovné metódy: tzv. rank-sum-type testy (O'Brien [4], Huang et al. [1]), T_{max} test (Liu et al. [3]) a napokon T_{cMAX} test (Zhang et al. [6]). Autori týchto metód na seba navzájom nadväzovali, pričom každá z novších metód bola vždy v istom zmysle vylepšenou verziou tej predošlej. Preto budú spomínané testy uvedené v chronologickom poradí tak, ako sa vyvíjali, pričom na tento vývoj neskôr nadviažeme v praktickej časti našej práce. Pri každej zo spomínaných metód uvedieme motiváciu jej vzniku, definíciu príslušnej testovej štatistiky, odvodené teoretické asymptotické vlastnosti a praktický výpočet p-value, pričom čerpať budeme najmä z príslušných článkov [1], [3], [4] a [6]. Tieto informácie doplníme o vlastné dodatky, vysvetlenia a ilustrácie fungovania resp. nefungovania každej metódy pomocou počítačových simulácií v prostredí R [5].

1.1 Testovanie odlišnosti polohy dvoch súborov, zovšeobecnená Behrens-Fisher hypotéza

Ako sa aj v [4] spomína, v praxi sa často vedú klinické štúdie za účelom ohodnotenia relatívnych účinností dvoch alebo viacerých spôsobov liečby. Táto účinnosť sa zvyčajne meria pomocou viacerých charakteristík. Jedným z možných štatistických prístupov je použiť jednorozmerné metódy, ktorými sa posúdi každá jedna charakteristika zvlášť. Takýto prístup môže byť užitočný, no často je potrebné urobiť aj jedno celkové štatistické vyhodnotenie toho, či je daná testovaná liečba účinná, alebo nie. Dôvodom potreby takéhoto testovania môže byť napríklad nízky počet dát, ktoré máme k dispozícii. Má teda praktický význam sa takýmto simultánnym viacrozmerným porovnaním zaoberať. V ďalšom budeme čerpať z [3] a [6], pričom zjednotíme použité značenie.

Uvažujme teda porovnanie m spojite meraných endpointov (endpoints, outcomes). Majme dve ošetrenia (treatments), pričom nech $X = (X_1, \ldots, X_m)^T$ označuje endpointy

prvého ošetrenia a nech $Y = (Y_1, \ldots, Y_m)^T$ označuje endpointy druhého ošetrenia. Nech sa X a Y riadia pravdepodobnostnými rozdeleniami s distribučnými funkciami $F : \mathbb{R}^m \to \mathbb{R}$ a $G : \mathbb{R}^m \to \mathbb{R}$ v tomto poradí. Platí teda

$$F(t_1, \ldots, t_m) = P(X_1 \leq t_1, \ldots, X_m \leq t_m) \quad \text{a}$$
$$G(t_1, \ldots, t_m) = P(Y_1 \leq t_1, \ldots, Y_m \leq t_m).$$

Ďalej nech k jednotlivým endpointom X_k a Y_k prislúchajú marginálne distribučné funkcie $F_k : \mathbb{R} \to \mathbb{R}$, resp. $G_k : \mathbb{R} \to \mathbb{R}$, kde $k = 1, \ldots, m$. Platí teda

$$F_k(t) = P(X_k \leq t) \quad \text{a}$$
$$G_k(t) = P(Y_k \leq t), \quad k = 1, \ldots, m.$$

Pre každý endpoint teraz definujme

$$\theta_k = P(X_k < Y_k) - P(X_k > Y_k), \quad k = 1, \ldots, m. \tag{1}$$

Takto definovaný parameter θ_k sa nazýva relatívny efekt ošetrenia Y vzhľadom k ošetreniu X pre k-ty endpoint. Nulovú hypotézu, ktorú budeme ďalej v tejto kapitole uvažovať, potom zvolíme nasledovne:

$$H_0 : \theta_k = 0, \quad k = 1, \ldots, m. \tag{2}$$

Táto hypotéza sa často označuje ako zovšeobecnená alebo neparametrická Behrens-Fisher hypotéza. V prípade spojite meraných endpointov sa nulová hypotéza z (2) dá pomocou definície (1) parametra θ_k jednoducho prepísať na tvar

$$H_0 : P(X_k < Y_k) = P(X_k > Y_k) = \frac{1}{2}, \quad k = 1, \ldots, m. \tag{3}$$

Najmä z tohto tvaru je vidno, že zamietnutie nulovej hypotézy H_0 naznačuje odlišnosť medzi ošetreniami v aspoň jednom endpointe. A naopak jej nezamietnutie značí, že medzi ošetreniami v istom zmysle žiaden signifikantný rozdiel nie je. Treba ale poznamenať, že aj za platnosti tejto hypotézy, môže stále platiť odlišnosť distribúcií $F \neq G$.

Uvažujme teraz už vyššie načrtnutý problém klinických štúdií. Nech X predstavuje skupinu pacientov s istým ochorením liečeným danou metódou (treatment group 1). Nech Y predstavuje skupinu iných pacientov s tým istým ochorením, ktorí sa liečia za pomoci odlišnej metódy (treatment group 2). Jednotlivé endpointy nech vyjadrujú charakteristiky,

na základe ktorých možno účinnosť danej liečby kvantifikovať. Pri testovaní rovnosti, resp. odlišnosti účinností dvoch skúmaných spôsobov liečby možno použiť práve zovšeobecnenú Behrens-Fisher hypotézu (2) (resp. (3)). Zamietnutie nulovej hypotézy H_0 hovorí o odlišnosti spomínaných účinností, keďže sa skupiny v aspoň jednej charakteristike signifikantne líšia. Inak sú obe metódy liečby rovnako účinné.

Iným príkladom testovania, pri ktorom je vhodné uvažovať hypotézu (2), sú tzv. case-control štúdie. Nech X teraz predstavuje skupinu pacientov s daným ochorením (treatment group) a Y nech predstavuje skupinu ľudí bez tohto ochorenia (control group). Endpointy nech teraz vyjadrujú charakteristiky, ktorých hodnoty by sa mohli z dôvodu ochorenia meniť. Zaujíma nás, či sa v daných charakteristikách skupina pacientov X a skupina zdravých ľudí Y skutočne líšia. Zamietnutie hypotézy (2) (resp. (3)) nám opäť naznačuje odlišnosť skupín, keď hodnota aspoň jednej charakteristiky X_k je v istom zmysle väčšia/menšia ako príslušná hodnota charakteristiky Y_k. Naopak pri nezamietnutí nulovej hypotézy H_0 rozdiely medzi pacientami a zdravými ľuďmi nerozpoznávame, a teda dané charakteristiky so sledovaným ochorením pravdepodobne nesúvisia.

Teraz uvedieme základné označenia, ktoré budeme ďalej používať v rámci celej kapitoly, pričom vychádzať budeme z článkov [3] a [6]. Naďalej uvažujme porovnanie dvoch ošetrení X a Y s m endpointami, distribučnými funkciami F a G, a marginálnymi distribučnými funkciami F_k a G_k, $k = 1, \ldots, m$. Nech n_x značí počet subjektov s ošetrením X a nech n_y značí počet subjektov s ošetrením Y. Taktiež označme N celkový počet subjektov, ktoré uvažujeme, teda $N = n_x + n_y$. Označme teraz x_{ik} pozorovanie k-teho endpointu na i-tom subjekte s ošetrením X, $i = 1, \ldots, n_x$, $k = 1, \ldots, m$. Podobne označme y_{lk} pozorovanie k-teho endpointu na l-tom subjekte s ošetrením Y, $l = 1, \ldots, n_y$, $k = 1, \ldots, m$. Predpokladajme, že vektory $(x_{i1}, \ldots, x_{im})^T$, $i = 1, \ldots, n_x$, sú nezávislé a taktiež, že vektory $(y_{l1}, \ldots, y_{lm})^T$, $l = 1, \ldots, n_y$, sú nezávislé. V rámci k-teho endpointu spojíme obe ošetrenia X a Y, a všetkých N dostupných pozorovaní $x_{1k}, \ldots, x_{n_x k}, y_{1k}, \ldots, y_{n_y k}$ zoradíme. Označme R_{xik} a R_{ylk} ranky pozorovaní x_{ik} a y_{lk} v tomto poradí. V prípade rovnosti viacerých pozorovaní sa za ich rank R_{xik}, resp. R_{ylk} zoberie priemer rankov ktoré by dostali, keby medzi nimi „veľmi tesne“ rovnosť nenastala. Pre obe ošetrenia X a Y ešte označme priemerné ranky v k-tom endpointe cez všetky subjekty s daným ošetrením ako

$\bar{R}_{x\cdot k}$, resp. $\bar{R}_{y\cdot k}$, teda

$$\begin{aligned} \bar{R}_{x\cdot k} &= \frac{1}{n_x}\sum_{i=1}^{n_x} R_{xik} \quad \text{a} \\ \bar{R}_{y\cdot k} &= \frac{1}{n_y}\sum_{l=1}^{n_y} R_{ylk}, \quad k = 1, \ldots, m. \end{aligned} \tag{4}$$

Pomocou takto zvolených priemerných rankov $\bar{R}_{x\cdot k}$ a $\bar{R}_{y\cdot k}$ budeme vedieť zostrojiť konzistentný odhad parametrov θ_k z (1). Teraz budú nasledovať dve pomocné lemy, na základe ktorých potom tento fakt dokážeme vo vete 1.3. Prvú lemu, ktorá popisuje známy vzťah platiaci pre ranky, uvádzame bez dôkazu.

Lema 1.1. *Majme k dispozícii náhodný výber $U_1, \ldots, U_n$ rozsahu n pochádzajúci z ľubovoľného jednorozmerného pravdepodobnostného rozdelenia. Nech R_i označuje rank prislúchajúci i-tej realizácii U_i, $i = 1, \ldots, n$. Označme R_i^C tzv. centrovaný rank prislúchajúci i-tej realizácii U_i, definovaný vzťahom*

$$R_i^C = R_i - \frac{n+1}{2}, \quad i = 1, \ldots, n.$$

Pre centrovaný rank R_i^C potom platí vzťah

$$R_i^C = \frac{1}{2}\sum_{j=1}^{n} \operatorname{sgn}(U_i - U_j), \quad i = 1, \ldots, n.$$

Prípadne pre pôvodný rank R_i potom platí nasledovné:

$$R_i = \frac{1}{2}\sum_{j=1}^{n} \operatorname{sgn}(U_i - U_j) + \frac{n+1}{2}, \quad i = 1, \ldots, n. \tag{5}$$

Rovnosť (5) nám dáva návod, ako možno ranky spätne vyjadriť pomocou pôvodných dát. Tento prechodný vzťah sa ukazuje byť všeobecne pri práci s rankami veľmi užitočný. My ho teraz použijeme vo vlastnom dôkaze nasledujúcej pomocnej lemy, ktorá popisuje vzťah uvedený v článku [3].

Lema 1.2. *Uvažujme porovnanie dvoch ošetrení X a Y s m endpointami a príslušným značením zavedeným v tejto kapitole. Pre rozdiel priemerných rankov $\bar{R}_{y\cdot k}$ a $\bar{R}_{x\cdot k}$ v k-tom endpointe ošetrení X a Y potom platí vzťah*

$$\bar{R}_{y\cdot k} - \bar{R}_{x\cdot k} = \frac{N}{2n_x n_y}\sum_{i=1}^{n_x}\sum_{l=1}^{n_y}\left(I_{\{x_{ik}<y_{lk}\}} - I_{\{x_{ik}>y_{lk}\}}\right), \quad k = 1, \ldots, m, \tag{6}$$

kde $I_{\{x_{ik}<y_{lk}\}}$ a $I_{\{x_{ik}>y_{lk}\}}$ sú indikátory, teda

$$I_{\{x_{ik}<y_{lk}\}} = \begin{cases} 1 & ak \quad x_{ik} < y_{lk}, \\ 0 & inak, \end{cases} \qquad a \qquad I_{\{x_{ik}>y_{lk}\}} = \begin{cases} 1 & ak \quad x_{ik} > y_{lk}, \\ 0 & inak. \end{cases}$$

Dôkaz. Nech $k \in \{1, \ldots, m\}$ je pevné. Ranky $R_{x1k}, \ldots, R_{xn_xk}$ prislúchajúce pozorovaniam $x_{1k}, \ldots, x_{n_xk}$ k-teho endpointu v ošetrení X môžeme podľa vzťahu (5) z lemy 1.1 vyjadriť nasledovne:

$$R_{xik} = \frac{1}{2}\sum_{i'=1}^{n_x} \operatorname{sgn}(x_{ik} - x_{i'k}) + \frac{1}{2}\sum_{l=1}^{n_y} \operatorname{sgn}(x_{ik} - y_{lk}) + \frac{N+1}{2}, \quad i = 1, \ldots, n_x.$$

Dosadením tohto tvaru R_{xik} do rovnosti v (4) môžeme pre priemerný rank v k-tom endpointe $\bar{R}_{x\cdot k}$ písať

$$\bar{R}_{x\cdot k} = \frac{1}{2n_x}\sum_{i=1}^{n_x}\sum_{i'=1}^{n_x} \operatorname{sgn}(x_{ik} - x_{i'k}) + \frac{1}{2n_x}\sum_{i=1}^{n_x}\sum_{l=1}^{n_y} \operatorname{sgn}(x_{ik} - y_{lk}) + \frac{1}{n_x}\sum_{i=1}^{n_x}\frac{N+1}{2}.$$

Z dôvodu nepárnosti funkcie signum je prvá dvojitá suma rovná 0, a teda dostávame

$$\bar{R}_{x\cdot k} = \frac{1}{2n_x}\sum_{i=1}^{n_x}\sum_{l=1}^{n_y} \operatorname{sgn}(x_{ik} - y_{lk}) + \frac{N+1}{2}.$$

Analogicky pre priemerný rank v k-tom endpointe $\bar{R}_{y\cdot k}$ ošetrenia Y dostávame

$$\bar{R}_{y\cdot k} = \frac{1}{2n_y}\sum_{i=1}^{n_x}\sum_{l=1}^{n_y} \operatorname{sgn}(y_{lk} - x_{ik}) + \frac{N+1}{2}.$$

Teraz už iba vyjadríme rozdiel dvoch získaných výrazov pre $\bar{R}_{y\cdot k}$ a $\bar{R}_{x\cdot k}$, teda počítame

$$\begin{aligned}
\bar{R}_{y\cdot k} - \bar{R}_{x\cdot k} &= \frac{1}{2n_y}\sum_{i=1}^{n_x}\sum_{l=1}^{n_y} \operatorname{sgn}(y_{lk} - x_{ik}) - \frac{1}{2n_x}\sum_{i=1}^{n_x}\sum_{l=1}^{n_y} \operatorname{sgn}(x_{ik} - y_{lk}) = \\
&= \frac{1}{2n_y}\sum_{i=1}^{n_x}\sum_{l=1}^{n_y} \operatorname{sgn}(y_{lk} - x_{ik}) + \frac{1}{2n_x}\sum_{i=1}^{n_x}\sum_{l=1}^{n_y} \operatorname{sgn}(y_{lk} - x_{ik}) = \\
&= \frac{n_x + n_y}{2n_xn_y}\sum_{i=1}^{n_x}\sum_{l=1}^{n_y} \operatorname{sgn}(y_{lk} - x_{ik}) = \\
&= \frac{N}{2n_xn_y}\sum_{i=1}^{n_x}\sum_{l=1}^{n_y} \left(I_{\{x_{ik} < y_{lk}\}} - I_{\{x_{ik} > y_{lk}\}}\right),
\end{aligned}$$

pričom druhá rovnosť opäť vyplýva z nepárnosti funkcie signum. □

Teraz už bude nasledovať veta popisujúca konzistentný odhad parametrov θ_k definovaných vzťahom (1). Tento odhad je uvedený napr. v článku [6], pričom dopĺňame vlastný dôkaz.

Veta 1.3. *Uvažujme porovnanie dvoch ošetrení X a Y s m endpointami a príslušným značením zavedeným v tejto kapitole. Potom odhad $\widehat{\theta}_k$ definovaný vzťahom*

$$\widehat{\theta}_k = \frac{2}{N}\left(\bar{R}_{y\cdot k} - \bar{R}_{x\cdot k}\right), \quad k = 1, \ldots, m, \tag{7}$$

je nevychýlený odhad parametra θ_k, teda platí

$$E(\widehat{\theta}_k) = \theta_k, \quad k = 1, \ldots, m.$$

Dôkaz. Nech $k \in \{1, \ldots, m\}$ je pevné. Za rozdiel $\bar{R}_{y\cdot k} - \bar{R}_{x\cdot k}$ v (7) dosadíme podľa odvodeného vzťahu (6) z lemy 1.2 a počítame strednú hodnotu nasledovne.

$$\begin{aligned}
E(\widehat{\theta}_k) &= \frac{2}{N} E\left(\bar{R}_{y\cdot k} - \bar{R}_{x\cdot k}\right) = \\
&= \frac{2}{N} E\left(\frac{N}{2n_x n_y} \sum_{i=1}^{n_x} \sum_{l=1}^{n_y} \left(I_{\{x_{ik} < y_{lk}\}} - I_{\{x_{ik} > y_{lk}\}}\right)\right) = \\
&= E\left(\frac{1}{n_x n_y} \sum_{i=1}^{n_x} \sum_{l=1}^{n_y} I_{\{x_{ik} < y_{lk}\}}\right) - E\left(\frac{1}{n_x n_y} \sum_{i=1}^{n_x} \sum_{l=1}^{n_y} I_{\{x_{ik} > y_{lk}\}}\right) = \\
&= P\left(X_k < Y_k\right) - P\left(X_k > Y_k\right) = \\
&= \theta_k
\end{aligned}$$

□

V dôsledku vety 1.3 možno teda o platnosti nulovej hypotézy H_0 z (2) zmysluplne rozhodnúť na základe odhadov $\widehat{\theta}_k$ podľa (7), resp. samotných rozdielov $\bar{R}_{y\cdot k} - \bar{R}_{x\cdot k}$. Hodnoty odhadov $\widehat{\theta}_k$ blízke 1 alebo -1 (resp. výrazne nenulové hodnoty rozdielov $\bar{R}_{y\cdot k} - \bar{R}_{x\cdot k}$) naznačujú neplatnosť hypotézy H_0, keďže skutočné hodnoty θ_k v takom prípade nulové pravdepodobne nebudú. Naopak odhady $\widehat{\theta}_k$ (resp. rozdiely $\bar{R}_{y\cdot k} - \bar{R}_{x\cdot k}$) blízke 0 hovoria v prospech platnosti hypotézy H_0, kedy by skutočné hodnoty θ_k mohli byť rovné 0. Pri testovaní odlišnosti dvoch ošetrení, pokiaľ uvažujeme zovšeobecnenú Behrens-Fisher hypotézu, budú mať teda odhady $\widehat{\theta}_k$ významnú úlohu. Pri popisovaní konštrukcie metód v nasledujúcich podkapitolách nám bude stačiť uvažovať samotné rozdiely $\bar{R}_{y\cdot k} - \bar{R}_{x\cdot k}$, keďže obsahujú rovnaké množstvo informácie. Preto si pre budúce potreby označme vektor rozdielov $\bar{R}_{y\cdot k} - \bar{R}_{x\cdot k}$, $k = 1, \ldots, m$, nasledovne.

$$Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_m \end{pmatrix} = \begin{pmatrix} \bar{R}_{y\cdot 1} - \bar{R}_{x\cdot 1} \\ \vdots \\ \bar{R}_{y\cdot m} - \bar{R}_{x\cdot m} \end{pmatrix} \tag{8}$$

Takto zvolený vektor Z bude tvoriť jadro každej metódy v rámci tejto kapitoly, pričom na každú z metód sa bude dať pozerať ako na funkciu zložiek tohto vektora. Úlohou jednotlivých metód bude v istom zmysle efektívne využiť informáciu obsiahnutú vo vektore Z a v konečnom dôsledku poskytnúť informáciu o platnosti/neplatnosti testovanej hypotézy vo forme výslednej p-value.

1.2 Rank-sum-type testy

V tejto časti sa pozrieme na historicky najstaršie zo sledovaných metód; tzv. rank-sum-type testy. Tieto testy navrhol O'Brien v roku 1984 vo svojom článku [4]. Na tento článok neskôr nadviazali Huang et al. v článku [1] z roku 2005, kde uverejnili všeobecnejšie použiteľnú modifikáciu týchto testov. Poznatky budeme čerpať najmä z týchto pôvodných článkov [1] a [4]. Pridáme vlastné vysvetlenia, postrehy a výsledky simulácií, ktorými ilustrujeme fungovanie spomínaných testov na jednoduchých dátach.

O'Brien sa vo svojom článku venoval niekoľkým metódam slúžiacich na porovnávanie skupín s viacerými endpointami, s využitím najmä v už spomínaných klinických štúdiách. Špeciálne pre prípad dvoch ošetrení navrhol nasledovnú neparametrickú metódu. Uvažujme porovnanie dvoch ošetrení X a Y s m endpointami a príslušným značením zavedeným v tejto kapitole. Ďalej definujme celkový rank i-teho subjektu s ošetrením X a celkový rank l-tého subjektu s ošetrením Y ako sumu ich rankov cez všetkých m endpointov, a označme ich $R_{xi\cdot}$, resp. $R_{yl\cdot}$, teda

$$R_{xi\cdot} = \sum_{k=1}^{m} R_{xik}, \quad i = 1, \ldots, n_x, \quad \text{a}$$
$$R_{yl\cdot} = \sum_{k=1}^{m} R_{ylk}, \quad l = 1, \ldots, n_y.$$

Na takto definované celkové ranky $R_{x1\cdot}, \ldots, R_{xn_x\cdot}$ a $R_{y1\cdot}, \ldots, R_{yn_y\cdot}$ teraz aplikujme jednorozmerný dvojvýberový t-test. V prípade predpokladu rovnosti disperzií výberov použijeme bežný Studentov t-test a dostávame jednu testovú štatistiku, označme ju T_1. Pokiaľ tento predpoklad neuvažujeme a použijeme Welchov t-test, dostávame druhú štatistiku, označme ju T_2. Tvary výsledných štatistík sú uvedené napr. v článku [1]. Testovú štatistiku T_1 možno zapísať ako

$$T_1 = \frac{\bar{R}_{y\cdot\cdot} - \bar{R}_{x\cdot\cdot}}{\hat{\sigma}\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \tag{9}$$

a obdobne pre testovú štatistiku T_2:

$$T_2 = \frac{\bar{R}_{y\cdot\cdot} - \bar{R}_{x\cdot\cdot}}{\sqrt{\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}}}, \tag{10}$$

kde

$$\bar{R}_{x\cdot\cdot} = \frac{1}{n_x}\sum_{i=1}^{n_x} R_{xi\cdot},$$
$$\bar{R}_{y\cdot\cdot} = \frac{1}{n_y}\sum_{l=1}^{n_y} R_{yl\cdot},$$
$$\hat{\sigma}_x^2 = \frac{1}{n_x - 1}\sum_{i=1}^{n_x}\left(R_{xi\cdot} - \bar{R}_{x\cdot\cdot}\right)^2,$$
$$\hat{\sigma}_y^2 = \frac{1}{n_y - 1}\sum_{l=1}^{n_y}\left(R_{yl\cdot} - \bar{R}_{y\cdot\cdot}\right)^2, \quad \text{a}$$
$$\hat{\sigma}^2 = \frac{(n_x - 1)\,\hat{\sigma}_x^2 + (n_y - 1)\,\hat{\sigma}_y^2}{N - 2}.$$

Pokiaľ pri testovaní uvažujeme viac obmedzujúcu hypotézu H_0' o rovnosti distribúcií ošetrení X a Y, teda

$$H_0' : F = G,$$

čo je špeciálny prípad zovšeobecnenej Behrens-Fisher hypotézy H_0 z (2), tak sa obe štatistiky T_1 a T_2 riadia približne Studentovým t-rozdelením. Nech df_1 a df_2 označujú počty stupňov voľnosti t-rozdelenia štatistiky T_1 resp. T_2. Potom, ako sa aj napr. v [1] uvádza, pre df_1 platí

$$df_1 = N - 2 \tag{11}$$

a df_2 možno zapísať ako

$$df_2 = \left(\frac{\zeta^2}{n_x - 1} + \frac{(1-\zeta)^2}{n_y - 1}\right)^{-1}, \tag{12}$$

kde

$$\zeta = \frac{\frac{\hat{\sigma}_x^2}{n_x}}{\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}}.$$

Teda pri testovaní pomocou štatistiky T_j, $j = 1, 2$, definovanej podľa (9), resp. (10), hypotézu H_0' zamietame na hladine významnosti α pokiaľ $|T_j| > t_{df_j,\alpha/2}$, kde $t_{df_j,\alpha/2}$ je $\left(1 - \frac{\alpha}{2}\right)$ kvantil t-rozdelenia s df_j stupňami voľnosti. Teraz bude nasledovať veta opisujúca asymptotické rozdelenie testových štatistík T_1 a T_2 za platnosti všeobecnejšej nulovej hypotézy H_0 z (2). Toto rozdelenie odvodili Huang et al. až neskôr v ich článku [1]. Vetu uvádzame bez dôkazu. Ten je možné nájsť aj so samotnou vetou v [1].

Veta 1.4. *Nech* $(n_x/n_y) \to \lambda$ *keď* $N \to \infty$ *pre nejakú konečnú konštantu* $0 < \lambda < +\infty$. *Potom, za platnosti hypotézy* (2), *obe štatistiky* T_1 *a* T_2 *definované podľa* (9) *a* (10)

konvergujú v distribúcii k normálnemu rozdeleniu so strednou hodnotou 0 *a disperziami* h_1*, resp.* h_2*, keď* $N \to \infty$*. Disperzie* h_1*,* h_2 *sú dané vzťahmi*

$$h_1 = \frac{\sum_{k_1=1}^{m}\sum_{k_2=1}^{m}(1+\lambda)^2(a_{k_1k_2}+b_{k_1k_2}\lambda)}{\sum_{k_1=1}^{m}\sum_{k_2=1}^{m}\left[e_{k_1k_2}\lambda^3+(b_{k_1k_2}+2f_{k_1k_2})\lambda^2+(a_{k_1k_2}+2q_{k_1k_2})\lambda+p_{k_1k_2}\right]},$$
$$h_2 = \frac{\sum_{k_1=1}^{m}\sum_{k_2=1}^{m}(1+\lambda)^2(a_{k_1k_2}+b_{k_1k_2}\lambda)}{\sum_{k_1=1}^{m}\sum_{k_2=1}^{m}\left[b_{k_1k_2}\lambda^3+(e_{k_1k_2}+2q_{k_1k_2})\lambda^2+(p_{k_1k_2}+2f_{k_1k_2})\lambda+a_{k_1k_2}\right]}, \tag{13}$$

kde

$$\begin{aligned}
a_{k_1k_2} &= \operatorname{cov}\left(G_{k_1}(X_{k_1}), G_{k_2}(X_{k_2})\right),\\
b_{k_1k_2} &= \operatorname{cov}\left(F_{k_1}(Y_{k_1}), F_{k_2}(Y_{k_2})\right),\\
e_{k_1k_2} &= \operatorname{cov}\left(F_{k_1}(X_{k_1}), F_{k_2}(X_{k_2})\right),\\
f_{k_1k_2} &= \operatorname{cov}\left(F_{k_1}(X_{k_1}), G_{k_2}(X_{k_2})\right),\\
p_{k_1k_2} &= \operatorname{cov}\left(G_{k_1}(Y_{k_1}), G_{k_2}(Y_{k_2})\right) \quad a\\
q_{k_1k_2} &= \operatorname{cov}\left(G_{k_1}(Y_{k_1}), F_{k_2}(Y_{k_2})\right).
\end{aligned}$$

Výsledkami tejto vety poukázali Huang et al. na nedokonalosť navrhnutých testov O'Briena. Dá sa ukázať, že disperzie h_1 a h_2 dané (13) sú obe rovné 1, pokiaľ platí rovnosť distribúcií $F = G$. Pokiaľ teda uvažujeme prísnejšiu hypotézu $H_0' : F = G$, obe štatistiky T_1, T_2 sa asymptoticky riadia štandardným normálnym rozdelením $N(0,1)$. Dôsledkom toho dokážu príslušné testy udržať pravdepodobnosť chyby 1. druhu na požadovanej úrovni, ako je bližšie odôvodnené v [1]. Avšak pre všeobecné distribúcie $F \neq G$ máme $h_1 \neq 1$ a $h_2 \neq 1$, a teda štatistiky T_1, T_2 budú mať síce asymptoticky normálne rozdelenie, ale s nekonštantnými disperziami. A keďže za platnosti zovšeobecnenej Behrens-Fisher hypotézy H_0 z (2) rovnosť distribúcií všeobecne neplatí, použitie testových štatistík T_1 a T_2 na testovanie tejto hypotézy môže mať za následok výrazné zväčšenie pravdepodobnosti chyby 1. druhu výsledných testov.

Toto pozorovanie motivovalo Huanga et al. navrhnúť úpravy pôvodných štatistík T_1, T_2 tak, aby aj v prípade rôznych distribúcií $F \neq G$ mali testy pravdepodobnosť chyby 1. druhu pod kontrolou, a teda boli vhodné aj na testovanie nulovej hypotézy H_0 z (2). Pôvodné štatistiky navrhli vo svojom článku [1] upraviť priamo na základe výsledkov vety 1.4 doplnením konzistentných odhadov disperzií h_1 a h_2 z (13) do ich menovateľov.

Označme T_{h1} a T_{h2} upravené verzie štatistík T_1 a T_2. Testovú štatistiku T_{h1} potom môžeme zapísať ako

$$T_{h1} = \frac{\bar{R}_{y\cdot\cdot} - \bar{R}_{x\cdot\cdot}}{\hat{\sigma}\sqrt{\widehat{h}_1\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}} \tag{14}$$

a analogicky pre testovú štatistiku T_{h2}

$$T_{h2} = \frac{\bar{R}_{y\cdot\cdot} - \bar{R}_{x\cdot\cdot}}{\sqrt{\widehat{h}_2\left(\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}\right)}}, \tag{15}$$

kde $\widehat{h}_j$ je konzistentný odhad disperzie h_j, $j = 1, 2$, danej vzťahom (13) pri všeobecných F a G. Treba poznamenať, že pri testovaní pomocou týchto upravených štatistík zamietame nulovú hypotézu H_0 z (2) rovnako ako v prípade pôvodných štatistík, teda za použitia kvantilov t-rozdelenia s rovnakými počtami stupňov voľnosti df_1 a df_2 danými (11) a (12). Konzistentné odhady $\widehat{h}_1$ a $\widehat{h}_2$ použité v (14) a (15) sú v [1] odvodené pomocou empirických odhadov pre F a G nasledovne. Označme

- $R_y(x_{ik})$ rank x_{ik} v rámci $\{x_{ik}, y_{1k}, \ldots, y_{n_y k}\}$,
- $R_x(x_{ik})$ rank x_{ik} v rámci $\{x_{1k}, \ldots, x_{n_x k}\}$,
- $R_x(y_{lk})$ rank y_{lk} v rámci $\{y_{lk}, x_{1k}, \ldots, x_{n_x k}\}$ a
- $R_y(y_{lk})$ rank y_{lk} v rámci $\{y_{1k}, \ldots, y_{n_y k}\}$.

Ďalej nech A_1, A_2 sú dve $n_x \times m$ matice a nech B_1, B_2 sú dve $n_y \times m$ matice dané

$$A_1 = \left(2R_y(x_{ik}) - 2 - n_y + n_y\widehat{\theta}_k\right)_{\substack{i=1,\ldots,n_x \\ k=1,\ldots,m}}, \tag{16}$$

$$A_2 = \left(2R_x(x_{ik}) - 1 - n_x\right)_{\substack{i=1,\ldots,n_x \\ k=1,\ldots,m}},$$

$$B_1 = \left(2R_x(y_{lk}) - 2 - n_x - n_x\widehat{\theta}_k\right)_{\substack{l=1,\ldots,n_y \\ k=1,\ldots,m}}, \tag{17}$$

$$B_2 = \left(2R_y(y_{lk}) - 1 - n_y\right)_{\substack{l=1,\ldots,n_y \\ k=1,\ldots,m}},$$

kde $\widehat{\theta}_k$ je konzistentný odhad parametra θ_k definovaný vzťahom (7). Potom $\widehat{h}_1$ a $\widehat{h}_2$ možno vyjadriť ako

$$\begin{aligned} \widehat{h}_1 &= \frac{N^2 J^T \left(A_1^T A_1 + B_1^T B_1\right) J}{n_x n_y J^T \left(\left(A_1 + A_2\right)^T \left(A_1 + A_2\right) + \left(B_1 + B_2\right)^T \left(B_1 + B_2\right)\right) J} \quad \text{a} \\ \widehat{h}_2 &= \frac{N^2 J^T \left(A_1^T A_1 + B_1^T B_1\right) J}{J^T \left(n_y^2 \left(A_1 + A_2\right)^T \left(A_1 + A_2\right) + n_x^2 \left(B_1 + B_2\right)^T \left(B_1 + B_2\right)\right) J}, \end{aligned} \tag{18}$$

kde J značí m-rozmerný vektor jednotiek. Asymptotické rozdelenie testových štatistík T_{h1} a T_{h2} pri voľbe takto odvodených konzistentných odhadov $\widehat{h}_1$ a $\widehat{h}_2$ popisuje nasledujúca veta vyslovená v [1]. Uvádzame ju bez dôkazu, ten je opäť možné nájsť v spomenutom článku.

Veta 1.5. *Nech platia predpoklady vety 1.4. Potom štatistiky T_{h1} a T_{h2} definované podľa* (14) *a* (15) *s $\widehat{h}_1$ a $\widehat{h}_2$ danými vzťahom* (18) *konvergujú v distribúcii k štandardnému normálnemu rozdeleniu, keď $(n_x/n_y) \to \lambda$, a $N \to \infty$.*

Pri testovaní pomocou takto upravených štatistík T_{h1}, T_{h2} máme teda kontrolu nad pravdepodobnosťou chyby 1. druhu bez ohľadu na to, či rovnosť distribúcií $F = G$ platí, alebo nie. Túto situáciu ilustrujeme v nasledujúcich simuláciách.

Uvažujme porovnanie dvoch ošetrení X a Y s m endpointami robené pomocou testovania zovšeobecnenej Behrens-Fisher hypotézy H_0 z (2). Špeciálne majme 5-rozmerné vektory pozorovaní, teda $m = 5$ a normálne rozdelené dáta, teda $X \sim N_5(\mu_x, \Sigma_x)$, a $Y \sim N_5(\mu_y, \Sigma_y)$. Z dôvodu symetrie normálneho rozdelenia stačí na platnosť nulovej hypotézy H_0 voľba rovnakých stredných hodnôt μ_x a μ_y, pre jednoduchosť zvoľme $\mu_x = \mu_y = 0_5$. Kvôli poukázaniu na odlišnosti štatistík O'Briena a Huanga et al. z pohľadu udržateľnosti chyby 1. druhu sme uvažovali prípad $F = G$ aj $F \neq G$. Tie sme jednoducho dosiahli voľbou rovnakých resp. rôznych kovariančných matíc, konkrétne sme zvolili

- $\Sigma_x = \Sigma_y = I_5$, alebo
- $\Sigma_x = I_5$, $\Sigma_y = 4I_5$,

kde I_5 je 5×5 identická matica. Pre každý z týchto dvoch prípadov rozdelení X a Y sme v prostredí R [5] náhodne generovali 5-rozmerné dáta pre nasledovné tri voľby veľkostí ošetrení n_x a n_y:

- $n_x = 20$, $n_y = 30$,
- $n_x = 80$, $n_y = 120$ a
- $n_x = 320$, $n_y = 480$.

Na takto generované dáta sme aplikovali testy pomocou testových štatistík T_1, T_2, T_{h1}, T_{h2}, definovaných v poradí podľa (9), (10), (14) a (15), pričom sme testovali na hladine významnosti $\alpha = 0{,}05$ a vykonaných bolo 10 000 simulácií. Získané odhady pravdepodobností chyby 1. druhu zaokrúhlené na 3 desatinné miesta sú vypísané v tabuľke 1.

Tabuľka 1: Pravdepodobnosť chyby 1. druhu rank-sum-type testov

	$\Sigma_x = \Sigma_y$			$\Sigma_x \neq \Sigma_y$		
	$n_x = 20$	$n_x = 80$	$n_x = 320$	$n_x = 20$	$n_x = 80$	$n_x = 320$
	$n_y = 30$	$n_y = 120$	$n_y = 480$	$n_y = 30$	$n_y = 120$	$n_y = 480$
T_1	0,049	0,052	0,051	0,046	0,042	0,041
T_2	0,050	0,052	0,051	0,061	0,059	0,059
T_{h1}	0,049	0,052	0,051	0,051	0,050	0,048
T_{h2}	0,048	0,051	0,051	0,051	0,050	0,048

Získané výsledky sú v súlade s asymptotickou teóriou, ktorú sme v predošlých častiach popisovali. V prípade rovnosti distribúcií $F = G$ testové štatistiky T_1 a T_2 držia pravdepodobnosť chyby 1. druhu približne na nominálnej hodnote 5%. To ale neplatí pri rozdielnych distribúciách $F \neq G$, ako možno vidieť pri výsledkoch štatistiky T_2 v tabuľke 1, kde pp. chyby 1. druhu pre všetky voľby n_x a n_y vrástla rádovo na 6%. Pri upravených štatistikách T_{h1} a T_{h2} už ale tento problém nenastáva a pp. chyby 1. druhu sú na úrovni 5% bez ohľadu na tvar rozdelenia druhého súboru.

Teraz sa budeme venovať sile vyššie definovaných rank-sum-type testov. Bude nás teda zaujímať ako dobre tieto metódy zamietajú testovanú nulovú hypotézu H_0 z (2), pokiaľ neplatí. Zo vzťahov (9), (10), (14) a (15) definujúcich testové štatistiky T_1, T_2, T_{h1} a T_{h2} je zrejmé, že všetky štatistiky sú principiálne rovnaké. Jadrom každej z nich je výraz $\left(\bar{R}_{y\cdot\cdot} - \bar{R}_{x\cdot\cdot}\right)$, teda rozdiel priemerov celkových rankov subjektov s ošetrením X a Y, a odlišujú sa iba v použitom odhade disperzie tohto rozdielu. Je ľahké ukázať, že výraz $\left(\bar{R}_{y\cdot\cdot} - \bar{R}_{x\cdot\cdot}\right)$ možno prepísať na tvar

$$\bar{R}_{y\cdot\cdot} - \bar{R}_{x\cdot\cdot} = \sum_{k=1}^{m} \left(\bar{R}_{y\cdot k} - \bar{R}_{x\cdot k}\right) = \sum_{k=1}^{m} Z_k,$$

teda na sumu zložiek vektora Z podľa (8). Pri každej z metód nulovú hypotézu H_0 zamietame na základe kvantilov symetrického t-rozdelenia. Chceme teda, aby v prípade

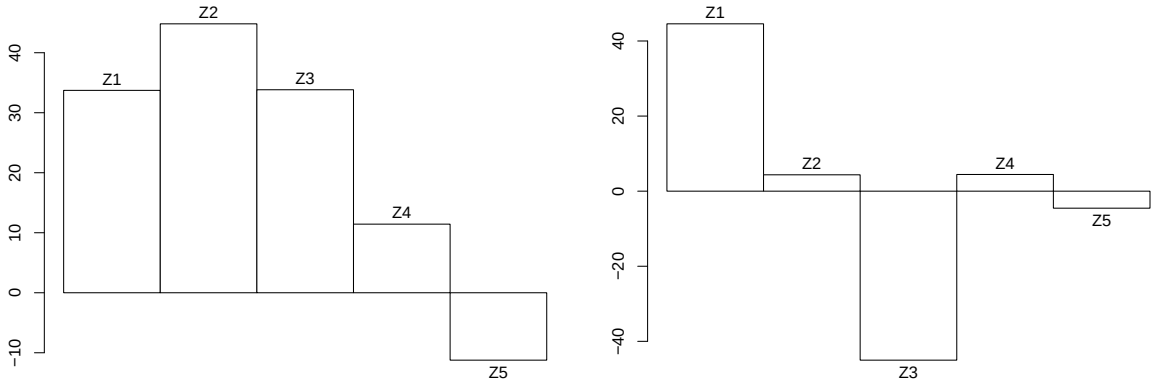
neplatnosti H_0 boli štatistiky výrazne nenulové, na ich znamienku nezáleží. Ak H_0 platí, tak všetky Z_k, $k = 1, \ldots, m$, sú blízke 0, teda aj ich suma je pravdepodobne blízka 0 a H_0 sa pravdepodobne nezamietne. Ak H_0 neplatí tak aspoň niektoré Z_k by mali byť výrazne nenulové, avšak zamietnutie alebo nezamietnutie nulovej hypotézy bude veľmi závisieť od ich znamienok. Ak majú výrazne nenulové Z_k (resp. aspoň ich väčšina) rovnaké znamienka, ich suma (a teda aj výsledná štatistika) bude opäť výrazne nenulová a H_0 sa pravdepodobne zamietne. Výsledkom by potom mali byť silné testy. V tomto prípade na znamienkach výrazne nenulových Z_k z dôvodu symetrie t-rozdelenia samozrejme nezáleží. Avšak v prípade keď sa znamienka výrazne nenulových Z_k striedajú (teda niektoré $Z_k \ll 0$, iné $Z_k \gg 0$), z dôvodu ich sčitovania môžeme dostať výsledné testové štatistiky blízke 0. Inými slovami, aj za neplatnosti H_0 z (2) môže stále platiť $\sum_{k=1}^{m} \theta_k = 0$, a teda podľa vzťahu (7) môže platiť to isté aj pre jadro týchto testových štatistík $\sum_{k=1}^{m} Z_k$. Nulovú hypotézu H_0 môžu teda testy nezamietnuť aj napriek tomu, že jasne neplatí. V tomto prípade sa môže použitím štatistík T_1, T_2, T_{h1} a T_{h2} stratiť veľké množstvo informácie a výsledné testy môžu byť veľmi slabé. Sila rank-sum-type testov bude teda veľmi závislá od charakteru neplatnosti H_0, ako to ukážeme v nasledujúcich simuláciách.

Uvažujme porovnanie dvoch ošetrení X a Y s m endpointami robené pomocou testovania zovšeobecnenej Behrens-Fisher hypotézy H_0 z (2). Rovnako ako pri predošlých simuláciách zvoľme $m = 5$ endpointov a uvažujme normálne rozdelené ošetrenia X, Y, teda $X \sim N_5(\mu_x, \Sigma_x)$, a $Y \sim N_5(\mu_y, \Sigma_y)$. Zaujíma nás sila testov, dáta teda budeme generovať pri nesplnenej nulovej hypotéze H_0. Na to nám stačí zvoliť rôzne stredné hodnoty μ_x a μ_y, dôvodom je opäť symetria normálneho rozdelenia. Motivovaní predchádzajúcim odsekom sme sa rozhodli uvažovať dva rôzne prípady dát. Strednú hodnotu μ_x ošetrenia X sme v oboch prípadoch zvolili rovnakú, za strednú hodnotu μ_y ošetrenia Y sme volili rôzne vektory μ_y^1 a μ_y^2. Konkrétne sme ich zvolili nasledovne:

$$\mu_x = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \qquad \mu_y^1 = \begin{pmatrix} 0{,}75 \\ 1 \\ 0{,}75 \\ 0{,}25 \\ -0{,}25 \end{pmatrix}, \qquad \mu_y^2 = \begin{pmatrix} 1 \\ 0{,}1 \\ -1 \\ 0{,}1 \\ -0{,}1 \end{pmatrix}.$$

Kovariančné matice Σ_x, Σ_y sme pre jednoduchosť zvolili v oboch prípadoch rovnaké,

konkrétne $\Sigma_x = \Sigma_y = 25I_5$, kde I_5 opäť značí 5×5 identickú maticu. Pre ilustráciu charakteru týchto dvoch typov dát sme na obrázku 1 znázornili priemerné hodnoty zložiek vektora Z z (8) získané z 10 000 simulácií pre voľby veľkostí ošetrení $n_x = 320$ a $n_y = 480$.



(a) Rozdelenie Y so strednou hodnotou μ_y^1 **(b)** Rozdelenie Y so strednou hodnotou μ_y^2

Obr. 1: Priemerné hodnoty zložiek vektora Z získané z 10 000 simulácií pre dva rôzne prípady normálnych rozdelení ošetrení X, Y a pri voľbe $n_x = 320$, $n_y = 480$.

Pre každý z týchto dvoch prípadov rozdelení X a Y sme v prostredí R [5] náhodne generovali 5-rozmerné dáta pre rovnaké tri voľby veľkostí ošetrení n_x a n_y ako pri predošlých simuláciách. Na takto generované dáta sme aplikovali testy pomocou testových štatistík T_1, T_2, T_{h1}, T_{h2}, definovaných v poradí podľa (9), (10), (14) a (15), pričom sme opäť testovali na hladine významnosti $\alpha = 0{,}05$ a vykonaných bolo 10 000 simulácií. Získané sily testov zaokrúhlené na 3 desatinné miesta sú vypísané v tabuľke 2.

Tabuľka 2: Sila rank-sum-type testov

	μ_y^1			μ_y^2		
	$n_x = 20$ $n_y = 30$	$n_x = 80$ $n_y = 120$	$n_x = 320$ $n_y = 480$	$n_x = 20$ $n_y = 30$	$n_x = 80$ $n_y = 120$	$n_x = 320$ $n_y = 480$
T_1	0,112	0,326	0,851	0,049	0,049	0,048
T_2	0,112	0,327	0,851	0,049	0,050	0,048
T_{h1}	0,112	0,327	0,851	0,050	0,050	0,048
T_{h2}	0,110	0,327	0,851	0,049	0,050	0,048

Z výsledkov si ako prvé môžeme všimnúť takmer totožné získané sily pre každý zo sledovaných testov. To je v súlade s teóriou, pretože štatistiky O'Briena T_1, T_2 a štatistiky

Huanga et al. T_{h1}, T_{h2} sú asymptoticky rovnaké, keďže platí $\widehat{h}_1 \approx h_1 = 1$ a $\widehat{h}_2 \approx h_2 = 1$. Čo sa týka sily testov pre jednotlivé prípady dát sú výsledky presne také ako by sme očakávali. Z obrázku 1 vidíme, že voľba μ_y^1 ako strednej hodnoty ošetrenia Y má za následok dáta vyhovujúce rank-sum-type testom. Výrazne nenulové priemery zložiek Z_1, Z_2 a Z_3 majú všetky kladné znamienko. Výsledné štatistiky boli teda často výrazne nenulové a testy nulovú hypotézu H_0 relatívne často zamietali. S rastúcim počtom dát n_x, n_y výsledné sily testov primerane rástli, pričom pri počtoch $n_x = 320$, $n_y = 480$ dosiahli sily testov úroveň až 85%. V tomto prípade teda rank-sum-type testy dokázali v dátach identifikovať signifikantné odlišnosti ošetrení X a Y. V prípade voľby μ_y^2 ako strednej hodnoty Y tomu už tak ale nebolo. Z obrázku 1 vidíme, že jediné výrazne nenulové Z_1 a Z_3 majú opačné znamienka a navyše v absolútnej hodnote veľmi podobné hodnoty. Výsledné štatistiky boli teda často blízke 0 a testy nulovú hypotézu zamietali iba v približne 5% prípadov, a to pre všetky počty n_x, n_y. V tomto prípade rank-sum-type testy nedokázali identifikovať rozdiely medzi ošetreniami X a Y, aj napriek tomu, že boli v dátach prítomné. Tento výrazný nedostatok rank-sum-type testov viedol k vzniku metódy, ktorou sa budeme zaoberať v nasledujúcej podkapitole.

1.3 T_{max} test

V tejto časti sa budeme venovať T_{max} testu, ktorý navrhli Liu et al. v roku 2010 v článku [3]. Uvedieme definíciu samotnej testovej štatistiky T_{max}, spôsob, akým sa pri tejto metóde prakticky počíta p-value, a taktiež dokázané asymptotické vlastnosti. K týmto poznatkom, ktoré budeme čerpať najmä zo samotného článku [3], opäť pridávame vlastné dodatky a ilustráciu praktického fungovania testu na základe výsledkov vlastných simulácií.

Motivovaní nedostatkom rank-sum-type testov popísaným na konci predchádzajúcej podkapitoly, Liu et al. vo svojom článku navrhli jednoduchý test, ktorý dáva uspokojivé výsledky bez ohľadu na znamienka rozdielov $\bar{R}_{y\cdot k} - \bar{R}_{x\cdot k}$. Uvažujme porovnanie dvoch ošetrení X a Y s m endpointami robené pomocou testovania zovšeobecnenej Behrens-Fisher hypotézy H_0 z (2) a príslušné značenie zavedené v tejto kapitole. Testová štatistika T_{max}, uvedená v článku [3], má potom tvar

$$T_{max} = \max_{k \in \{1,\dots,m\}} \left(\left| \bar{R}_{y\cdot k} - \bar{R}_{x\cdot k} \right| \right). \tag{19}$$

Na základe označenia (8) môžeme túto testovú štatistiku zapísať priamo ako funkciu vektora Z nasledovne:

$$T_{max} = \max_{k\in\{1,\dots,m\}} (|Z_k|) = \|Z\|_\infty, \tag{20}$$

kde $\|\cdot\|_\infty$ značí maximovú normu vektora. Takto definovaná testová štatistika T_{max} očividne dosahuje iba nezáporné hodnoty. Za platnosti nulovej hypotézy H_0 z (2) očakávame odhady pre θ_k, $k = 1, \dots, m$, blízke 0. Na základe vzťahu (7) potom platí to isté aj pre samotné zložky Z_k a výsledná testová štatistika T_{max} by teda podľa (20) mala byť blízka 0 sprava. Pokiaľ hypotéza H_0 neplatí, tak aspoň niektorý z odhadov pre θ_k očakávame výrazne nenulový. V takom prípade bude aspoň jedna zo zložiek Z_k vektora Z výrazne nenulová (či už $Z_k \ll 0$, alebo $Z_k \gg 0$) a výsledná testová štatistika T_{max} by teda podľa (20) mala byť výrazne kladná. Preto pôjde o test, ktorý nulovú hypotézu H_0 z (2) zamieta, ak platí $T_{max} > c_{max}$, pre nejakú kritickú hodnotu $c_{max} > 0$. Na rozdiel od rank-sum-type testov z predošlej podkapitoly, T_{max} test pracuje s absolútnymi hodnotami zložiek vektora Z. Pokiaľ dáta vykazujú neplatnosť nulovej hypotézy H_0, T_{max} test tento fakt zaregistruje, a to bez ohľadu na znamienka zložiek Z_k (teda rozdielov $\bar{R}_{y\cdot k} - \bar{R}_{x\cdot k}$). Dá sa teda očakávať uspokojivá sila testu pre ľubovoľný charakter neplatnosti H_0.

Prejdime teraz k spôsobu výpočtu p-value pri T_{max} teste. Nech $t \geq 0$ značí hodnotu nameranej testovej štatistiky. Na základe vzťahu (20) platí

$$P_{H_0}(T_{max} \geq t) = 1 - P_{H_0}(T_{max} < t) = 1 - P_{H_0}(\|Z\|_\infty < t),$$

čo môžeme prepísať na nasledovný vzťah uvedený v článku [3]

$$P_{H_0}(T_{max} \geq t) = 1 - P_{H_0}(|Z_1| < t, \dots, |Z_m| < t). \tag{21}$$

Využitím faktu, že testová štatistika T_{max} je vlastne vektorová norma, môžeme v prípade tohto testu p-value počítať ako mnohorozmerný integrál hustoty vektora Z z (8). Keďže ide konkrétne o maximovú normu, integračnou oblasťou je m-rozmerná kocka centrovaná v bode 0_m s hranami dĺžky $2t$ a rovnobežnými so súradnicovými osami. V dôsledku vzťahu (21) nám na výpočet p-value stačí poznať rozdelenie vektora Z, a teda rozdelenie samotnej testovej štatistiky T_{max} vôbec nepotrebujeme. Rozdelenie vektora Z z (8) za platnosti nulovej hypotézy H_0 z (2) opisuje v [3] nasledujúca veta. Uvádzame ju bez dôkazu, ten možno nájsť v prílohe spomínaného článku.

Veta 1.6. *Za platnosti nulovej hypotézy* H_0 *z* (2) *sa vektor* Z *daný vzťahom* (8) *asymptoticky riadi viacrozmerným normálnym rozdelením so strednou hodnotou* 0_m *a korelačnou maticou* $\Lambda = (\rho_{k_1k_2})_{m\times m}$ *pre* $\min\{n_x, n_y\} \to \infty$, *a* $\frac{n_x}{n_y} \to \lambda$, $0 < \lambda < \infty$, *kde*

$$\rho_{k_1k_2} = \frac{a_{k_1k_2} + \lambda b_{k_1k_2}}{\sqrt{\left(a_{k_1k_1} + \lambda b_{k_1k_1}\right)\left(a_{k_2k_2} + \lambda b_{k_2k_2}\right)}},$$

a

$$a_{k_1k_2} = \operatorname{cov}\Big(G_{k_1}\left(X_{k_1}\right), G_{k_2}\left(X_{k_2}\right)\Big),$$
$$b_{k_1k_2} = \operatorname{cov}\Big(F_{k_1}\left(Y_{k_1}\right), F_{k_2}\left(Y_{k_2}\right)\Big),$$

pre všetky $k_1 = 1, \ldots, m$, $k_2 = 1, \ldots, m$.

Poznámka: V prílohe článku [3] sú v rámci dôkazu vety 1.6 za platnosti nulovej hypotézy H_0 z (2) odvodené aj disperzie zložiek Z_k vektora Z daného vzťahom (8). Preškálovaním asymptotickej korelačnej matice Λ odmocninami z týchto disperzií vieme teda pre vektor Z za platnosti H_0 získať aj asymptotickú kovariančnú maticu.

Ako sa aj v článku [3] spomína, odhady neznámych kovariancií $a_{k_1k_2}$, $b_{k_1k_2}$ odvodili už Huang et al. v ich článku [1] nasledovne. Uvažujme pomocné matice A_1 a B_1 definované vzťahmi (16), (17) v predchádzajúcej podkapitole. Konzistentné odhady $\widehat{a}_{k_1k_2}$, $\widehat{b}_{k_1k_2}$ kovariancií $a_{k_1k_2}$, $b_{k_1k_2}$ potom možno v maticovej podobe vyjadriť ako

$$\begin{aligned}\left(\widehat{a}_{k_1k_2}\right)_{\substack{k_1=1,\ldots,m\\k_2=1,\ldots,m}} &= \frac{1}{4n_xn_y^2}A_1^TA_1 \quad \text{a}\\ \left(\widehat{b}_{k_1k_2}\right)_{\substack{k_1=1,\ldots,m\\k_2=1,\ldots,m}} &= \frac{1}{4n_x^2n_y}B_1^TB_1.\end{aligned} \tag{22}$$

Na základe výsledkov vety 1.6 s dosadenými konzistentnými odhadmi $\widehat{a}_{k_1k_2}$, $\widehat{b}_{k_1k_2}$ podľa (22) a využitím vzťahu (21), budeme teraz skúmať silu T_{max} testu z (19) v nasledujúcich simuláciách.

Uvažujme porovnanie dvoch ošetrení X a Y s m endpointami robené pomocou testovania zovšeobecnenej Behrens-Fisher hypotézy H_0 z (2). Kvôli poukázaniu na odolnosť T_{max} testu voči striedajúcim sa znamienkam rozdielov $\bar{R}_{y\cdot k} - \bar{R}_{x\cdot k}$, uvažujme druhý prípad voľby 5-rozmerných normálnych rozdelení X, Y, ktorý sme uviedli na strane 23 pri simuláciách sily tam spomínaných rank-sum-type testov. Charakter týchto dát bol ilustrovaný pomocou priemerných hodnôt zložiek vektora Z z (8) na obrázku 1 (b). Pre túto voľbu rozdelenia X a Y sme v prostredí R [5] náhodne generovali 5-rozmerné dáta pre rovnaké

tri voľby veľkostí ošetrení n_x a n_y ako pri predošlých simuláciách. Na takto generované dáta sme aplikovali test pomocou T_{max} definovanej podľa vzťahu (19) s výpočtom p-value podľa (21). Opäť sme testovali na hladine významnosti $\alpha = 0{,}05$ a vykonaných bolo 10 000 simulácií. Získané sily testu, zaokrúhlené na 3 desatinné miesta, vyšli pre jednotlivé voľby počtov dát n_x, n_y nasledovne:

- 0,088, pri voľbe $n_x = 20$, $n_y = 30$,
- 0,234, pri voľbe $n_x = 80$, $n_y = 120$ a
- 0,809, pri voľbe $n_x = 320$, $n_y = 480$.

Porovnaním týchto výsledkov s výsledkami sily rank-sum-type testov z pravej časti tabuľky 2 si môžeme hneď všimnúť výrazný rozdiel. Kým testy O'Briena a Huanga et al. nedokázali z dôvodu rôznych znamienok rozdielov $\bar{R}_{y\cdot k} - \bar{R}_{x\cdot k}$ neplatnosť nulovej hypotézy zaznamenať, T_{max} test hypotézu H_0 zamietal relatívne často. Sila T_{max} testu s rastúcim počtom dát primerane narastá a pri počtoch $n_x = 320$, $n_y = 480$ dosiahla až takmer 81%. Z dôvodu robustnejšej testovej štatistiky dokáže tento test identifikovať signifikantnú odlišnosť ošetrení X a Y bez ohľadu na charakter tejto odlišnosti. A preto sa T_{max} test javí byť vhodnou alternatívou rank-sum-type testov z predchádzajúcej podkapitoly.

1.4 T_{cMAX} test

Teraz prejdeme k poslednému testu, ktorému sa budeme v tejto kapitole venovať. Navrhli ho Zhang et al. v roku 2019 v ich článku [6]. Autori priamo nadväzujú na článok [3] autorov Liu et al., pričom navrhnutý test je v istom zmysle zovšeobecnením T_{max} testu, ktorému sme sa venovali v predchádzajúcej časti. Najskôr uvedieme motiváciu k vzniku tohto testu, definíciu samotnej testovej štatistiky T_{cMAX} a odvodené asymptotické vlastnosti, pričom čerpať budeme najmä z článku [6]. Opäť dopĺňame vlastné postrehy, vysvetlenia a výsledky našich simulácií.

Ako sa v [6] píše, v mnohých aplikáciách majú endpointy prirodzenú clusterovú štruktúru. Tieto clustery (clusters, zhluky) predstavujú skupiny výrazne kladne korelovaných endpointov, v rámci ktorých endpointy preukazujú väčšie podobnosti, ako tie spomedzi rôznych clusterov. Pre endpointy v rámci jedného clustera očakávame relatívne efekty ošetrení s rovnakými znamienkami a podobnými hodnotami, naopak medzi clusterami sa

budú relatívne efekty skôr líšiť. Preto ak test berie túto clusterovú štruktúru do úvahy, tak má k dispozícii istú informáciu navyše a dá sa očakávať jeho väčšia sila. Uvažujúc spomínanú clusterovú štruktúru endpointov teraz upravíme doposiaľ používané značenie.

Naďalej uvažujme dve ošetrenia s m endpointami. Predpokladajme, že tieto endpointy možno prirodzene zadeliť do J clusterov. Nech $X = \left(X_1^T, \ldots, X_J^T\right)^T$ označuje sadu clusterov prvého ošetrenia a nech $Y = \left(Y_1^T, \ldots, Y_J^T\right)^T$ označuje sadu clusterov druhého ošetrenia. Označme m_j počet endpointov patriacich do j-teho clustera, $j = 1, \ldots, J$, pričom pre celkový počet endpointov m platí $m = \sum_{j=1}^{J} m_j$. Potom vektory $X_j = \left(X_{j1}, \ldots, X_{jm_j}\right)^T$, resp. $Y_j = \left(Y_{j1}, \ldots, Y_{jm_j}\right)^T$, označujú j-ty cluster, $j = 1, \ldots, J$. Nech sa X a Y riadia pravdepodobnostnými rozdeleniami s distribučnými funkciami $F : \mathbb{R}^m \to \mathbb{R}$ a $G : \mathbb{R}^m \to \mathbb{R}$ v tomto poradí. Ďalej nech k jednotlivým endpointom X_{jk} a Y_{jk} prislúchajú marginálne distribučné funkcie $F_{jk} : \mathbb{R} \to \mathbb{R}$, resp. $G_{jk} : \mathbb{R} \to \mathbb{R}$, pre $j = 1, \ldots, J$, $k = 1, \ldots, m_j$. Podobne ako predtým, pre každý endpoint definujme parameter θ_{jk}, teda relatívny efekt ošetrenia Y vzhľadom k ošetreniu X pre k-ty endpoint v j-tom clusteri nasledovne:

$$\theta_{jk} = P(X_{jk} < Y_{jk}) - P(X_{jk} > Y_{jk}), \quad j = 1, \ldots, J, \ k = 1, \ldots, m_j. \tag{23}$$

Opäť budeme uvažovať zovšeobecnenú Behrens-Fisher hypotézu, teda nulovú hypotézu H_0 definovanú ako

$$H_0 : \theta_{jk} = 0, \quad j = 1, \ldots, J, \ k = 1, \ldots, m_j. \tag{24}$$

Naďalej nech n_x, n_y značia počty subjektov s ošetreniami X, Y a taktiež nech $N = n_x + n_y$ značí celkový počet subjektov. Podobne ako v predchádzajúcich podkapitolách, nech x_{ijk} značí pozorovanie k-teho endpointu v j-tom clusteri na i-tom subjekte s ošetrením X, $i = 1, \ldots, n_x$, $j = 1, \ldots, J$, $k = 1, \ldots, m_j$. Taktiež nech y_{ljk} značí pozorovanie k-teho endpointu v j-tom clusteri na l-tom subjekte s ošetrením Y, $l = 1, \ldots, n_y$, $j = 1, \ldots, J$, $k = 1, \ldots, m_j$. Naďalej uvažujme predpoklad nezávislosti vektorov pozorovaní, teda nech vektory $(x_{i11}, \ldots, x_{iJm_J})^T$, $i = 1, \ldots, n_x$, sú nezávislé a taktiež nech vektory $(y_{l11}, \ldots, y_{lJm_J})^T$, $l = 1, \ldots, n_y$, sú nezávislé. V rámci k-teho endpointu v j-tom clusteri analogicky spojíme obe ošetrenia X a Y, a všetkých N dostupných pozorovaní $x_{1jk}, \ldots, x_{n_x jk}, y_{1jk}, \ldots, y_{n_y jk}$ zoradíme. Na záver označme ranky pozorovaní x_{ijk} a y_{ljk} ako R_{xijk} a R_{yljk} v tomto poradí.

Uvažujme teda porovnanie dvoch ošetrení X a Y s m endpointami zadelenými do J

clusterov robené pomocou testovania zovšeobecnenej Behrens-Fisher hypotézy H_0 z (24) a príslušné značenie zavedené vyššie. Definujme priemerné ranky v k-tom endpointe v rámci j-teho clustera ošetrenia X a Y ako

$$\begin{aligned} \bar{R}_{x\cdot jk} &= \frac{1}{n_x}\sum_{i=1}^{n_x} R_{xijk} \quad \text{a} \\ \bar{R}_{y\cdot jk} &= \frac{1}{n_y}\sum_{l=1}^{n_y} R_{yljk}, \quad j = 1,\ldots,J,\ k = 1,\ldots,m_j. \end{aligned} \tag{25}$$

Pomocou takto definovaných priemerných rankov vieme podľa výsledkov vety 1.3 rovnako zostrojiť nevychýlený odhad $\widehat{\theta}_{jk}$ parametrov θ_{jk} z (23) nasledovne:

$$\widehat{\theta}_{jk} = \frac{2}{N}\left(\bar{R}_{y\cdot jk} - \bar{R}_{x\cdot jk}\right), \quad j = 1,\ldots,J,\ k = 1,\ldots,m_j. \tag{26}$$

Ako sa aj v [6] spomína, takto definovaný odhad $\widehat{\theta}_{jk}$ je navyše aj konzistentným odhadom neznámeho parametra θ_{jk}. Rozdiely priemerných rankov $\bar{R}_{y\cdot jk} - \bar{R}_{x\cdot jk}$ teda naďalej hrajú pri testovaní nulovej hypotézy H_0 z (24) významnú úlohu. Preto si tak, ako v úvode tejto tejto kapitoly, označme m-rozmerný vektor týchto rozdielov nasledovne.

$$Z = \begin{pmatrix} Z_{11} \\ \vdots \\ Z_{1m_1} \\ \vdots \\ Z_{J1} \\ \vdots \\ Z_{Jm_J} \end{pmatrix} = \begin{pmatrix} \bar{R}_{y\cdot 11} - \bar{R}_{x\cdot 11} \\ \vdots \\ \bar{R}_{y\cdot 1m_1} - \bar{R}_{x\cdot 1m_1} \\ \vdots \\ \bar{R}_{y\cdot J1} - \bar{R}_{x\cdot J1} \\ \vdots \\ \bar{R}_{y\cdot Jm_J} - \bar{R}_{x\cdot Jm_J} \end{pmatrix} \tag{27}$$

Pokiaľ endpointy vykazujú clusterovú štruktúru, tak v rámci daného clustera $j \in \{1,\ldots,J\}$ by mali mať relatívne efekty θ_{jk}, $k = 1,\ldots,m_j$, definované vzťahom (23), všetky rovnaké znamienka a podobné hodnoty. Na nevychýlený a konzistentný odhad týchto parametrov máme teda rovnaké očakávania, preto podľa vzťahu (26) by malo platiť to isté aj pre samotné rozdiely $\bar{R}_{y\cdot jk} - \bar{R}_{x\cdot jk}$. Sčitovaním, resp. priemerovaním zložiek vektora Z daného vzťahom (27) v rámci jednotlivých clusterov sa teda informácia kumuluje. Z pohľadu sily testu sa preto ukáže byť výhodné, ak test v prípade clusterovej štruktúry endpointov pracuje priamo s takýmito priemermi zložiek Z_{jk}. V rámci j-teho clustera preto ešte definujme priemerné ranky cez všetky endpointy daného clustera a cez všetky subjekty s

daným ošetrením X, resp. Y, ako

$$\begin{aligned} \bar{R}_{x\cdot j\cdot} &= \frac{1}{n_x m_j} \sum_{k=1}^{m_j} \sum_{i=1}^{n_x} R_{xijk} \quad \text{a} \\ \bar{R}_{y\cdot j\cdot} &= \frac{1}{n_y m_j} \sum_{k=1}^{m_j} \sum_{l=1}^{n_y} R_{yljk}, \quad j = 1, \ldots, J. \end{aligned} \tag{28}$$

Podľa definície (25) priemerných rankov $\bar{R}_{x\cdot jk}$ a $\bar{R}_{x\cdot jk}$ potom triviálne platí vzťah

$$\begin{aligned} \bar{R}_{x\cdot j\cdot} &= \frac{1}{m_j} \sum_{k=1}^{m_j} \bar{R}_{x\cdot jk} \quad \text{a} \\ \bar{R}_{y\cdot j\cdot} &= \frac{1}{m_j} \sum_{k=1}^{m_j} \bar{R}_{y\cdot jk}, \quad j = 1, \ldots, J. \end{aligned} \tag{29}$$

Na základe tohto prepisu a podľa vzťahu (26) potom platí, že $\widehat{\theta}_{j\cdot} = \frac{2}{N}\left(\bar{R}_{y\cdot j\cdot} - \bar{R}_{x\cdot j\cdot}\right)$ je nevychýleným a konzistentným odhadom pre $\theta_{j\cdot} = \frac{1}{m_j} \sum\limits_{k=1}^{m_j} \theta_{jk}$, $j = 1, \ldots, J$. Pre budúce potreby si napokon ešte označme vektor rozdielov priemerných rankov $\bar{R}_{y\cdot j\cdot}$ a $\bar{R}_{x\cdot j\cdot}$, $j = 1, \ldots, J$, definovaných vzťahom (28) nasledovne.

$$Z^c = \begin{pmatrix} Z_1^c \\ \vdots \\ Z_J^c \end{pmatrix} = \begin{pmatrix} \bar{R}_{y\cdot 1\cdot} - \bar{R}_{x\cdot 1\cdot} \\ \vdots \\ \bar{R}_{y\cdot J\cdot} - \bar{R}_{x\cdot J\cdot} \end{pmatrix} \tag{30}$$

Motivovaní dôsledkami clusterovej štruktúry endpointov popísanými vyššie, navrhli Zhang et al. v ich článku [6] robustnú testovú štatistiku T_{cMAX} definovanú vzťahom

$$T_{cMAX} = \max_{j \in \{1,\ldots,J\}} \left(\left| \frac{\bar{R}_{y\cdot j\cdot} - \bar{R}_{x\cdot j\cdot}}{\sqrt{\widehat{\text{var}}\left(\bar{R}_{y\cdot j\cdot} - \bar{R}_{x\cdot j\cdot}\right)}} \right| \right), \tag{31}$$

kde $\widehat{\text{var}}\left(\bar{R}_{y\cdot j\cdot} - \bar{R}_{x\cdot j\cdot}\right)$ je konzistentný odhad disperzie rozdielu $\bar{R}_{y\cdot j\cdot} - \bar{R}_{x\cdot j\cdot}$, $j = 1, \ldots, J$. Využitím označenia (30) môžeme pre testovú štatistiku T_{cMAX} potom písať

$$T_{cMAX} = \max_{j \in \{1,\ldots,J\}} \left(\left| \frac{Z_j^c}{\sqrt{\widehat{\text{var}}\left(Z_j^c\right)}} \right| \right). \tag{32}$$

Tak, ako v prípade T_{max} testu z predchádzajúcej podkapitoly, aj testová štatistika T_{cMAX} dosahuje iba nezáporné hodnoty. Ako sa v [6] spomína, tento test je ekvivalentný rank-sum-type testom T_{h1} z (14) resp. T_{h2} z (15), ak uvažujeme iba jeden cluster, teda voľbu $J = 1$. A pokiaľ uvažujeme iba jeden endpoint v rámci každého clustera, teda voľby

$m_j = 1$, $j = 1, \ldots, J$, tak je tento test ekvivalentný T_{max} testu z (19). Zo vzťahu (32) vidíme, že testová štatistika T_{cMAX} vzniká výberom v absolútnej hodnote najväčšej zložky štandardizovaného vektora Z^c z (30). A keďže tento vektor na základe prepisu (29) vzniká priemerovaním zložiek Z_{jk} vektora Z z (27) v rámci jednotlivých clusterov, tak sa na T_{cMAX} test dá aj všeobecne pozerať ako na istú kombináciu spomínaných rank-sum-type testov T_{h1}, resp. T_{h2}, a T_{max} testu. Ako sme už spomínali, v prípade clusterovej štruktúry endpointov budú mať rozdiely ošetrení v endpointoch v rámci daného clustera skôr rovnaké znamienka a podobné hodnoty, čo je situácia, kedy sa dá naplno využiť sila rank-sum-type testov plynúca zo sčitovania dôkazov. Preto je výhodné, že v rámci clusterov počíta testová štatistika T_{cMAX} priemery zložiek vektora Z, čoho výsledkom je vektor Z^c. Naopak, medzi rôznymi clusterami sa môžu v prípade clusterovej štruktúry endpointov rozdiely ošetrení veľmi líšiť. To je situácia, kedy je výhodné použiť robustný test akým je T_{max}, aby sa sčítaním „protichodných“ dôkazov o neplatnosti H_0 z (24) táto informácia nakoniec nestratila. Preto T_{cMAX} test vhodne volí maximum z absolútnych hodnôt (štandardizovaných) zložiek vektora Z^c ako finálnu testovú štatistiku. Dá sa teda očakávať, že v prípade clusterovej štruktúry endpointov bude mať použitie testovej štatistiky T_{cMAX} danej vzťahom (31), resp. (32), za následok veľkú silu testu.

Aj napriek tomu, že testová štatistika T_{cMAX} nie je z dôvodu prítomnej štandardizácie priamo maximovou normou vektora, vieme p-value počítať v princípe tak, ako sme to popisovali v predchádzajúcej podkapitole pre prípad T_{max} testu. Opäť ide o mnohorozmerný integrál hustoty náhodného vektora cez jednoduchú oblasť. Týmto vektorom je podľa definície testovej štatistiky (32) J-rozmerný vektor Z^c daný vzťahom (30) a integračnou oblasťou je z dôvodu štandardizácie všeobecne J-rozmerný kváder. Na výpočet p-value nám preto stačí poznať rozdelenie vektora Z^c. Toto rozdelenie spolu s potrebnými konzistentnými odhadmi popisujú v článku [6] nasledujúce dve vety. Dôkaz vety 1.7 možno nájsť vo webovej prílohe k článku [6], náčrt dôkazu vety 1.8 je uvedený v článku samotnom.

Veta 1.7. *Pre* $\min\{n_x, n_y\} \to \infty$ *a* $\frac{n_x}{n_y} \to \lambda < \infty$ *konverguje náhodný vektor* $\frac{1}{\sqrt{N}} Z^c$ *daný vzťahom* (30) *k mnohorozmernému normálnemu rozdeleniu so strednou hodnotou*

$$\mu = \frac{\sqrt{N}}{2} \left(\frac{1}{m_1} \sum_{k=1}^{m_1} \theta_{1k}, \ldots, \frac{1}{m_J} \sum_{k=1}^{m_J} \theta_{Jk} \right)^T$$

a kovariančnou maticou $\Sigma = (\sigma_{j_1 j_2})_{J\times J}$, *ktorej zložky sú dané vzťahom*

$$\sigma_{j_1 j_2} = \frac{1}{m_{j_1} m_{j_2}} \sum_{k_1=1}^{m_{j_1}} \sum_{k_2=1}^{m_{j_2}} \left(\left(1 + \frac{1}{\lambda}\right) a_{j_1 k_1, j_2 k_2} + (1+\lambda)\, b_{j_1 k_1, j_2 k_2} - \left(\sqrt{\lambda} + \frac{1}{\sqrt{\lambda}}\right)^2 \theta_{j_1 k_1} \theta_{j_2 k_2} \right),$$

kde

$$a_{j_1 k_1, j_2 k_2} = \operatorname{cov}\left(G_{j_1 k_1}\left(X_{j_1 k_1}\right), G_{j_2 k_2}\left(X_{j_2 k_2}\right)\right), \quad a$$
$$b_{j_1 k_1, j_2 k_2} = \operatorname{cov}\left(F_{j_1 k_1}\left(Y_{j_1 k_1}\right), F_{j_2 k_2}\left(Y_{j_2 k_2}\right)\right).$$

Veta 1.8. *Uvažujme predpoklady vety 1.7. Konzistentné odhady pre* $a_{j_1 k_1, j_2 k_2}$ *a* $b_{j_1 k_1, j_2 k_2}$ *sú dané v poradí vzťahmi*

$$\widehat{a}_{j_1 k_1, j_2 k_2} = \frac{1}{n_x} \sum_{i=1}^{n_x} \left[\left(\frac{1}{n_y} \sum_{l=1}^{n_y} I_{\left\{y_{l j_1 k_1} < x_{i j_1 k_1}\right\}} - \frac{1 - \widehat{\theta}_{j_1 k_1}}{2} \right) \times \left(\frac{1}{n_y} \sum_{l=1}^{n_y} I_{\left\{y_{l j_2 k_2} < x_{i j_2 k_2}\right\}} - \frac{1 - \widehat{\theta}_{j_2 k_2}}{2} \right) \right]$$

a

$$\widehat{b}_{j_1 k_1, j_2 k_2} = \frac{1}{n_y} \sum_{l=1}^{n_y} \left[\left(\frac{1}{n_x} \sum_{i=1}^{n_x} I_{\left\{x_{i j_1 k_1} < y_{l j_1 k_1}\right\}} - \frac{1 + \widehat{\theta}_{j_1 k_1}}{2} \right) \times \left(\frac{1}{n_x} \sum_{i=1}^{n_x} I_{\left\{x_{i j_2 k_2} < y_{l j_2 k_2}\right\}} - \frac{1 + \widehat{\theta}_{j_2 k_2}}{2} \right) \right],$$

kde odhad $\widehat{\theta}_{jk}$, $j = 1, \ldots, J$, $k = 1, \ldots, m_j$, *je daný vzťahom* (26). *Potom konzistentný odhad kovariančnej matice* Σ *je matica* $S = (s_{j_1 j_2})_{J\times J}$, *pre ktorej zložky platí*

$$s_{j_1 j_2} = \frac{1}{m_{j_1} m_{j_2}} \sum_{k_1=1}^{m_{j_1}} \sum_{k_2=1}^{m_{j_2}} \left(\left(1 + \frac{1}{\lambda}\right) \widehat{a}_{j_1 k_1, j_2 k_2} + (1+\lambda)\, \widehat{b}_{j_1 k_1, j_2 k_2} - \left(\sqrt{\lambda} + \frac{1}{\sqrt{\lambda}}\right)^2 \widehat{\theta}_{j_1 k_1} \widehat{\theta}_{j_2 k_2} \right).$$

Ako sa aj v [6] spomína, z výsledkov vety 1.8 vyplýva, že konzistentný odhad disperzie $\widehat{\operatorname{var}}\left(\bar{R}_{y\cdot j\cdot} - \bar{R}_{x\cdot j\cdot}\right)$ z (31) je daný vzťahom

$$\widehat{\operatorname{var}}\left(\bar{R}_{y\cdot j\cdot} - \bar{R}_{x\cdot j\cdot}\right) = N s_{jj}, \quad j = 1, \ldots, J. \tag{33}$$

Využitím výsledkov viet 1.7, 1.8 a ich dôsledku (33) môžeme teraz v ďalšom simulačne skúmať silu navrhnutého testu T_{cMAX} z (31).

Uvažujme porovnanie dvoch ošetrení X a Y s m endpointami robené pomocou testovania zovšeobecnenej Behrens-Fisher hypotézy H_0 z (24). Opäť zvoľme $m = 5$ endpointov

a normálne rozdelenie ošetrení X, Y, teda $X \sim N_5(\mu_x, \Sigma_x)$, a $Y \sim N_5(\mu_y, \Sigma_y)$. Uvažujme clusterovú štruktúru endpointov s počtom clusterov $J = 2$ a počtami endpointov v rámci daných clusterov $m_1 = 3$, $m_2 = 2$. Kvôli skúmaniu sily testov sme dáta generovali pri nesplnenej nulovej hypotéze H_0, čo sme opäť docielili voľbou rôznych vektorov stredných hodnôt μ_x a μ_y. Konkrétne sme zvolili

$$\mu_x = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \qquad \mu_y = \begin{pmatrix} 0{,}75 \\ 0{,}7 \\ 0{,}75 \\ -0{,}75 \\ -0{,}7 \end{pmatrix}.$$

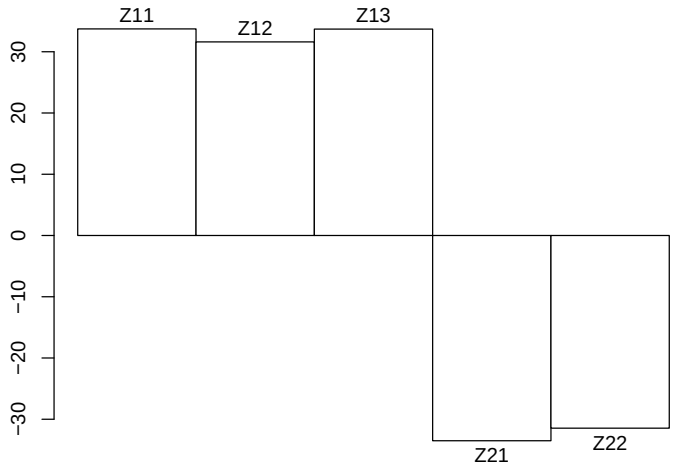
Kovariančné matice Σ_x, Σ_y sme zvolili rovnaké, konkrétne

$$\Sigma_x = \Sigma_y = \begin{pmatrix} 25 & 20 & 20 & 0 & 0 \\ 20 & 25 & 20 & 0 & 0 \\ 20 & 20 & 25 & 0 & 0 \\ 0 & 0 & 0 & 25 & 20 \\ 0 & 0 & 0 & 20 & 25 \end{pmatrix}.$$

Pre ilustráciu charakteru týchto dát sme na obrázku 2 znázornili priemerné hodnoty zložiek vektora Z z (27) získané z 10 000 simulácií pre voľby veľkostí ošetrení $n_x = 320$ a $n_y = 480$.

Pre túto voľbu rozdelenia X a Y sme v prostredí R [5] náhodne generovali 5-rozmerné dáta pre rovnaké tri voľby veľkostí ošetrení n_x a n_y ako pri predošlých simuláciách. Na takto generované dáta sme aplikovali test pomocou testovej štatistiky T_{cMAX} definovanej podľa vzťahu (31) a taktiež testy pomocou T_1, T_2, T_{h1}, T_{h2}, T_{max} z predchádzajúcich podkapitol, definovaných v poradí podľa (9), (10), (14), (15), a (19). Opäť sme testovali na hladine významnosti $\alpha = 0{,}05$ a vykonaných bolo 10 000 simulácií. Výsledné sily testov zaokrúhlené na 3 desatinné miesta sú vypísané v tabuľke 3.

Získané výsledky sú opäť v súlade s intuíciou. Z obrázka 2 vidíme, že vyššie popísaná voľba rozdelenia X a Y má za následok rovnaké znamienka a podobné hodnoty zložiek Z_{jk} vektora Z (a teda aj parametrov θ_{jk}) v rámci daného clustera, no medzi clusterami spôsobuje veľké odlišnosti. Ide teda o dáta nevyhovujúce rank-sum-type testom, čoho dôkazom je fakt, že ich sila bola pre každú voľbu počtov dát n_x, n_y výrazne nižšia, ako



Obr. 2: Priemerné hodnoty zložiek vektora Z získané z 10 000 simulácií pre normálne rozdelenie ošetrení X, Y s clusterovou štruktúrou endpointov a pri voľbe $n_x = 320$, $n_y = 480$.

Tabuľka 3: Sila rankových testov v prípade clusterovej štruktúry endpointov

	$n_x = 20$ $n_y = 30$	$n_x = 80$ $n_y = 120$	$n_x = 320$ $n_y = 480$
T_{cMAX}	0,127	0,232	0,702
T_{max}	0,100	0,217	0,674
T_1	0,051	0,060	0,095
T_2	0,050	0,060	0,096
T_{h1}	0,054	0,060	0,097
T_{h2}	0,052	0,060	0,097

sila zvyšných dvoch testov. Aj pri najvyšších počtoch $n_x = 320$, $n_y = 480$ dosiahli testy využívajúce testové štatistiky T_1, T_2, T_{h1} a T_{h2} silu iba približne 10%. Z tohto dôvodu sú v prípade clusterovej štruktúry endpointov rank-sum-type testy O'Briena a Huanga et al. nevhodné na testovanie zovšeobecnenej Behrens-Fisher hypotézy H_0 z (2). V prípade T_{max} testu sú výsledky už oveľa lepšie. Z dôvodu robustnej testovej štatistiky nie je sila tohto testu negatívne ovplyvnená striedaním znamienok Z_{jk} (resp. θ_{jk}). Preto aj napriek tomu, že neprihliada na clusterovú štruktúru, zamietal T_{max} test nulovú hypotézu H_0 relatívne často, pričom pri počtoch $n_x = 320$, $n_y = 480$ dosiahla sila testu úroveň až 67%. Najlepšie avšak obstál T_{cMAX} test. Vhodná voľba jeho testovej štatistiky má za následok najvyššiu silu testu spomedzi všetkých sledovaných testov, a to pre každú voľbu počtov

dát n_x, n_y. Pri najvyšších počtoch $n_x = 320$, $n_y = 480$ dosiahla sila T_{cMAX} testu úroveň až 70%. V prípade clusterovej štruktúry endpointov T_{cMAX} test teda výhodne kombinuje silu rank-sum-type testov O'Briena a Huanga et al. s robustnosťou T_{max} testu Liuho et al. Preto sa na testovanie zovšeobecnenej Behrens-Fisher hypotézy H_0 ukazuje byť T_{cMAX} test ako veľmi vhodná voľba.

2 Zovšeobecnenie T_{cMAX} testu pre prípad viacerých súborov

V tejto kapitole už prejdeme k vlastnému prínosu našej práce. Navrhneme metódu na testovanie odlišnosti polohy ľubovoľného počtu súborov z viacrozmerných rozdelení. Tento test bude priamym zovšeobecnením T_{cMAX} testu (Zhang et al. [6]), ktorým sme sa zaoberali v závere prvej kapitoly. Najskôr uvedieme motiváciu k zvolenému spôsobu tohto zovšeobecnenia, na základe čoho upravíme značenie z predchádzajúcej kapitoly. Ďalej opíšeme voľbu samotnej testovej štatistiky a vyjadríme sa k očakávaným vlastnostiam výsledného testu. Pre zvolenú testovú štatistiku odvodíme v ďalšej časti jej teoretické asymptotické vlastnosti a taktiež odhady potrebných parametrov využívajúce iba dostupné dáta. Vychádzať budeme z už dokázaných vzťahov a vlastností uvedených v článkoch [1], [3] a [6]. Na základe týchto výsledkov budeme potom schopní prakticky počítať p-value nášho testu, čo využijeme v poslednej časti na skúmanie jeho kvality pomocou počítačových simulácií v prostredí R [5]. Získané odhady pravdepodobnosti chyby 1. druhu a sily navrhnutého testu pre rôzne dáta budeme porovnávať s výsledkami pre známy parametrický prístup - MANOVA test.

2.1 Motivácia

Zhang et al. v ich článku [6] odvodili silný a robustný test T_{cMAX}, ktorý avšak rovnako ako aj ostatné metódy z predchádzajúcej kapitoly slúži výhradne na porovnanie dvoch súborov. No ako aj samotní autori v článku [6] píšu, malo by byť jednoduché ho rozšíriť na porovnanie ľubovoľného počtu súborov. Ako príklad takéhoto rozšírenia spomínajú autori zostrojenie rankovej testovej štatistiky na spôsob testovej štatistiky Kruskal-Wallis testu (Kruskal a Wallis [2]) v každom endpointe a následné spojenie týchto jednorozmerných štatistík pomocou nimi navrhnutého postupu, teda spriemerovaním v rámci jednotlivých clusterov, a výberom maximálnej (štandardizovanej) absolútnej hodnoty. Bez uvádzania detailov, Kruskal-Wallis test (alebo jednofaktorová ANOVA na rankoch) porovnáva ľubovoľný počet súborov na základe testovania nulovej hypotézy o tom, či dané súbory pochádzajú z rovnakého jednorozmerného rozdelenia (špeciálny prípad zovšeobecnenej Behrens-Fisher hypotézy). Testová štatistika tohto testu pracuje s priemernými rankami

pre pooled sample všetkých súborov dohromady a dosahuje iba nezáporné hodnoty (viac v [2]). Keďže relatívne efekty vystupujúce v Behrens-Fisher hypotéze sa z princípu veci viažu vždy iba k danej dvojici súborov (ošetrení), dá sa ukázať, že ich nie je možné na základe týchto celkových rankov odhadovať. Pri voľbe testovej štatistiky tohto typu by teda bolo potrebné prejsť k spomínanej hypotéze o rovnosti distribúcií. Navyše navrhnutý postup priemerovania jednorozmerných testových štatistík v rámci jednotlivých clusterov a výberu maximálnej (štandardizovanej) absolútnej hodnoty má význam len v prípade, keď môžu mať tieto testové štatistiky ľubovoľné znamienko. Nemá teda zmysel tento postup použiť v prípade nezáporných testových štatistík, akou je testová štatistika Kruskal-Wallis testu. Preto sme sa rozhodli T_{cMAX} test zovšeobecniť iným spôsobom, pri ktorom budeme môcť naďalej uvažovať zovšeobecnenú Behrens-Fisher hypotézu, a teda využiť teoretické výsledky z predchádzajúcej kapitoly.

2.2 Značenie

Teraz uvedieme označenia a základné definície, ktoré budeme v rámci tejto kapitoly uvažovať. Motivovaní predchádzajúcou časťou zvolíme značenie zo záveru predchádzajúcej kapitoly rozšírené pre prípad ľubovoľného počtu ošetrení. Všeobecne teda uvažujme g ošetrení s m endpointami. Predpokladajme, že tieto endpointy možno prirodzene zadeliť do J clusterov. Nech $X_a = \left(X_{a1}^T, \ldots, X_{aJ}^T\right)^T$ označuje sadu clusterov a-teho ošetrenia, $a = 1, \ldots, g$. Označme m_j počet endpointov patriacich do j-teho clustera, $j = 1, \ldots, J$, pričom pre celkový počet endpointov platí $m = \sum_{j=1}^{J} m_j$. Potom vektor $X_{aj} = \left(X_{aj1}, \ldots, X_{ajm_j}\right)^T$ označuje j-ty cluster a-teho ošetrenia, $a = 1, \ldots, g$, $j = 1, \ldots, J$. Nech sa X_a riadi pravdepodobnostným rozdelením s distribučnou funkciou $F_a : \mathbb{R}^m \to \mathbb{R}$, $a = 1, \ldots, g$. Ďalej nech k jednotlivým endpointom X_{ajk} prislúchajú marginálne distribučné funkcie $F_{ajk} : \mathbb{R} \to \mathbb{R}$, pre $a = 1, \ldots, g$, $j = 1, \ldots, J$, $k = 1, \ldots, m_j$. Ako sme spomínali v predošlej časti, naďalej budeme chcieť pracovať so zovšeobecnenou Behrens-Fisher hypotézou, resp. s jej rozšírením pre prípad ľubovoľného počtu ošetrení. Preto podobne ako v prvej kapitole definujeme pre každý endpoint relatívne efekty ošetrení. Avšak tentokrát bude z dôvodu všeobecného počtu ošetrení g potrebné definovať m takýchto parametrov pre každú dvojicu ošetrení zvlášť. Označme teda $\theta_{jk}^{(a,b)}$ relatívny efekt ošetrenia X_a vzhľadom k ošetreniu X_b pre k-ty

endpoint j-teho clustera definovaný vzťahom

$$\theta_{jk}^{(a,b)} = P\left(X_{ajk} < X_{bjk}\right) - P\left(X_{ajk} > X_{bjk}\right), \quad a, b = 1, \ldots, g,\ a \neq b,\ j = 1, \ldots, J,\ k = 1, \ldots, m_j. \tag{34}$$

V ďalšom budeme teda uvažovať rozšírenú verziu zovšeobecnenej Behrens-Fisher hypotézy H_0 danú vzťahom

$$H_0 : \theta_{jk}^{(a,b)} = 0, \quad a, b = 1, \ldots, g,\ a < b,\ j = 1, \ldots, J,\ k = 1, \ldots, m_j. \tag{35}$$

Ďalej označme n_a počet subjektov s ošetrením X_a, $a = 1, \ldots, g$. Potom nech x_{aijk} značí pozorovanie k-teho endpointu v j-tom clusteri na i-tom subjekte s ošetrením X_a, $a = 1, \ldots, g$, $i = 1, \ldots, n_a$, $j = 1, \ldots, J$, $k = 1, \ldots, m_j$. Uvažujme predpoklad nezávislosti vektorov pozorovaní, teda nech pre ľubovoľné $a \in \{1, \ldots, g\}$ sú vektory $(x_{ai11}, \ldots, x_{aiJm_J})^T$, $i = 1, \ldots, n_a$, nezávislé. Opäť budeme pracovať iba s usporiadaním týchto pozorovaní. Avšak, ako sme v predchádzajúcej podkapitole načrtli, bude potrebné uvažovať ranky pre pooled sample iba jednotlivých dvojíc ošetrení. Nech teda $a, b \in \{1, \ldots, g\}$, $a \neq b$. V rámci k-teho endpointu v j-tom clusteri spojíme ošetrenia X_a a X_b, a všetkých $n_a + n_b$ dostupných pozorovaní $x_{a1jk}, \ldots, x_{an_ajk}, x_{b1jk}, \ldots, x_{bn_bjk}$ zoradíme. Takto získané ranky pozorovaní x_{aijk} a x_{bljk} označme ako $R_{aijk}^{(a,b)}$ a $R_{bljk}^{(a,b)}$ v tomto poradí, $i = 1, \ldots, n_a$, $l = 1, \ldots, n_b$, $j = 1, \ldots, J$, $k = 1, \ldots, m_j$. Definujme priemerné ranky pre pooled sample X_a a X_b v k-tom endpointe v rámci j-teho clustera týchto ošetrení ako

$$\begin{aligned} \bar{R}_{a\cdot jk}^{(a,b)} &= \frac{1}{n_a} \sum_{i=1}^{n_a} R_{aijk}^{(a,b)} \quad \text{a} \\ \bar{R}_{b\cdot jk}^{(a,b)} &= \frac{1}{n_b} \sum_{l=1}^{n_b} R_{bljk}^{(a,b)}, \quad a, b = 1, \ldots, g,\ a \neq b,\ j = 1, \ldots, J,\ k = 1, \ldots, m_j. \end{aligned} \tag{36}$$

Pomocou takto definovaných priemerných rankov vieme podľa výsledkov vety 1.3 zostrojiť nevychýlený odhad $\widehat{\theta}_{jk}^{(a,b)}$ parametrov $\theta_{jk}^{(a,b)}$ z (34) nasledovne:

$$\widehat{\theta}_{jk}^{(a,b)} = \frac{2}{n_a + n_b} \left(\bar{R}_{b\cdot jk}^{(a,b)} - \bar{R}_{a\cdot jk}^{(a,b)}\right), \quad a, b = 1, \ldots, g,\ a \neq b,\ j = 1, \ldots, J,\ k = 1, \ldots, m_j. \tag{37}$$

Vychádzajúc z článku [6], takto definovaný odhad $\widehat{\theta}_{jk}^{(a,b)}$ je taktiež aj konzistentným odhadom neznámeho parametra $\theta_{jk}^{(a,b)}$. Využitím označenia (36) ešte definujme v rámci j-teho clustera priemerné ranky pre pooled sample X_a a X_b cez všetky endpointy daného clustera

a cez všetky subjekty s daným ošetrením ako

$$\begin{aligned} \bar{R}^{(a,b)}_{a\cdot j\cdot} &= \frac{1}{m_j}\sum_{k=1}^{m_j}\bar{R}^{(a,b)}_{a\cdot jk} \quad \text{a} \\ \bar{R}^{(a,b)}_{b\cdot j\cdot} &= \frac{1}{m_j}\sum_{k=1}^{m_j}\bar{R}^{(a,b)}_{b\cdot jk}, \quad a,b=1,\ldots,g,\ a\neq b,\ j=1,\ldots,J. \end{aligned} \tag{38}$$

Na základe tejto definície a podľa vzťahu (37) potom platí, že $\widehat{\theta}^{(a,b)}_{j\cdot} = \frac{2}{n_a+n_b}\left(\bar{R}^{(a,b)}_{b\cdot j\cdot} - \bar{R}^{(a,b)}_{a\cdot j\cdot}\right)$ je nevychýleným a konzistentným odhadom pre $\theta^{(a,b)}_{j\cdot} = \frac{1}{m_j}\sum_{k=1}^{m_j}\theta^{(a,b)}_{jk}$, $a,b=1,\ldots,g$, $a\neq b$, $j=1,\ldots,J$. Pre budúce potreby si napokon ešte pre každú dvojicu ošetrení X_a, X_b, $a,b=1,\ldots,g$, $a\neq b$, označme vektory rozdielov priemerných rankov $\bar{R}^{(a,b)}_{b\cdot j\cdot}$ a $\bar{R}^{(a,b)}_{a\cdot j\cdot}$, $j=1,\ldots,J$, definovaných vzťahom (38) ako

$$Z^{(a,b)} = \begin{pmatrix} Z^{(a,b)}_1 \\ \vdots \\ Z^{(a,b)}_J \end{pmatrix} = \begin{pmatrix} \bar{R}^{(a,b)}_{b\cdot 1\cdot} - \bar{R}^{(a,b)}_{a\cdot 1\cdot} \\ \vdots \\ \bar{R}^{(a,b)}_{b\cdot J\cdot} - \bar{R}^{(a,b)}_{a\cdot J\cdot} \end{pmatrix}, \quad a,b=1,\ldots,g,\ a\neq b. \tag{39}$$

2.3 Voľba testovej štatistiky T^g_{cMAX}

Uvažujme porovnanie g ošetrení X_a, $a=1,\ldots,g$, s m endpointami zadelenými do J clusterov robené pomocou testovania rozšírenej verzie zovšeobecnenej Behrens-Fisher hypotézy H_0 z (35) a príslušné značenie zavedené v predchádzajúcej časti. Podobne, ako v predchádzajúcej kapitole, aj teraz budeme o platnosti nulovej hypotézy H_0 rozhodovať na základe hodnôt odhadov $\widehat{\theta}^{(a,b)}_{jk}$ z (37), resp. samotných rozdielov priemerných rankov $\bar{R}^{(a,b)}_{b\cdot jk} - \bar{R}^{(a,b)}_{a\cdot jk}$. Keďže uvažujeme clusterovú štruktúru endpointov, z dôvodov opísaných v závere prvej kapitoly bude vhodnejšie pracovať priamo s priemermi týchto rozdielov cez všetky endpointy daného clustera, teda s rozdielmi $\bar{R}^{(a,b)}_{b\cdot j\cdot} - \bar{R}^{(a,b)}_{a\cdot j\cdot}$ definovanými podľa vzťahu (38). Tieto rozdiely sú podľa označenia (39) zložkami $Z^{(a,b)}_j$ vektorov $Z^{(a,b)}$. Ako sme aj vo vzťahu (35) uviedli, v predpise nulovej hypotézy H_0 nám stačí uvažovať relatívne efekty pre rôzne dvojice ošetrení X_a, X_b, teda pre voľbu indexov $a<b$. Dôvodom je fakt, že z nulovej hodnoty parametra $\theta^{(a,b)}_{jk}$ podľa definície (34) triviálne vyplýva aj nulová hodnota parametra $\theta^{(b,a)}_{jk} = -\theta^{(a,b)}_{jk}$. Preto nám aj pri voľbe testovej štatistiky bude stačiť uvažovať vektory rozdielov priemerných rankov $Z^{(a,b)}$ z (39) rovnako pri voľbe $a<b$. Za platnosti nulovej hypotézy očakávame všetky zložky $Z^{(a,b)}_j$ týchto vektorov blízke 0, inak by aspoň niektoré $Z^{(a,b)}_j$ mali byť výrazne nenulové (či už $Z^{(a,b)}_j \ll 0$, alebo $Z^{(a,b)}_j \gg 0$), čo

je rovnaká situácia, akú sme opisovali pri T_{cMAX} teste. Preto vychádzajúc z testovej štatistiky T_{cMAX} testu pre prípad dvoch ošetrení definovanej vzťahom (31), označme T^g_{cMAX} jej zovšeobecnenie pre prípad všeobecného počtu g ošetrení a definujme ju vzťahom

$$T^g_{cMAX} = \max_{\substack{j\in\{1,\dots,J\}\\ a,b\in\{1,\dots,g\}\\ a<b}} \left(\left| \frac{\bar{R}^{(a,b)}_{b\cdot j\cdot} - \bar{R}^{(a,b)}_{a\cdot j\cdot}}{\sqrt{\widehat{\mathrm{var}}\left(\bar{R}^{(a,b)}_{b\cdot j\cdot} - \bar{R}^{(a,b)}_{a\cdot j\cdot}\right)}} \right| \right), \tag{40}$$

kde $\widehat{\mathrm{var}}\left(\bar{R}^{(a,b)}_{b\cdot j\cdot} - \bar{R}^{(a,b)}_{a\cdot j\cdot}\right)$ je konzistentný odhad disperzie rozdielu $\bar{R}^{(a,b)}_{b\cdot j\cdot} - \bar{R}^{(a,b)}_{a\cdot j\cdot}$, $a, b = 1,\dots,g$, $a < b$, $j = 1,\dots,J$. Tento odhad, ktorý odvodili Zhang et al. v ich článku [6], sme už uvádzali vo vzťahu (33) v predchádzajúcom značení pre prípad dvoch ošetrení. V značení používanom v rámci tejto kapitoly ho uvedieme neskôr vo vzťahu (66). Využitím označenia (39) môžeme testovú štatistiku T^g_{cMAX} prepísať na tvar

$$T^g_{cMAX} = \max_{\substack{j\in\{1,\dots,J\}\\ a,b\in\{1,\dots,g\}\\ a<b}} \left(\left| \frac{Z^{(a,b)}_j}{\sqrt{\widehat{\mathrm{var}}\left(Z^{(a,b)}_j\right)}} \right| \right). \tag{41}$$

Z dôvodu zvýraznenia vzťahu tejto rozšírenej testovej štatistiky voči tej pôvodnej pre prípad dvoch ošetrení môžeme testovú štatistiku T^g_{cMAX} ešte zapísať ako

$$T^g_{cMAX} = \max_{\substack{a,b\in\{1,\dots,g\}\\ a<b}} \left(T^{(a,b)}_{cMAX} \right), \tag{42}$$

kde $T^{(a,b)}_{cMAX}$ je testová štatistika T_{cMAX} testu z (31) aplikovaná na dvojicu ošetrení X_a a X_b, teda

$$T^{(a,b)}_{cMAX} = \max_{j\in\{1,\dots,J\}} \left(\left| \frac{\bar{R}^{(a,b)}_{b\cdot j\cdot} - \bar{R}^{(a,b)}_{a\cdot j\cdot}}{\sqrt{\widehat{\mathrm{var}}\left(\bar{R}^{(a,b)}_{b\cdot j\cdot} - \bar{R}^{(a,b)}_{a\cdot j\cdot}\right)}} \right| \right).$$

Najmä z tohto vzťahu (42) je vidno, že testová štatistika T^g_{cMAX} z (40) je skutočne zovšeobecnením testovej štatistiky T_{cMAX} z (31). Pri uvažovaní iba dvoch ošetrení, teda pri voľbe $g = 2$, sú tieto dve testové štatistiky identické. Pre všeobecný počet ošetrení g vznikne T^g_{cMAX} výpočtom testovej štatistiky T_{cMAX} pre každú rôznu dvojicu ošetrení X_a a X_b, $a, b = 1,\dots,g$, $a < b$, a následným výberom tej najväčšej z takto získaných nezáporných hodnôt. Výsledkom je teda opäť nezáporná štatistika. Takto zovšeobecnený test naďalej využíva robustnosť testovej štatistiky T_{cMAX} a jej vhodnú konštrukciu pre prípad clusterovej štruktúry endpointov, preto možno očakávať uspokojivú silu testu, najmä v prípade nižšieho počtu ošetrení g. Asymptotickým vlastnostiam testovej štatistiky T^g_{cMAX} z (40) sa budeme venovať v nasledujúcej časti.

2.4 Asymptotické rozdelenie

Podobne ako pri T_{cMAX} teste z (31), aj pri tejto zovšeobecnenej verzii testovej štatistiky nám na výpočet p-value bude stačiť poznať rozdelenie istého vektora. Z dôvodu voľby testovej štatistiky T^g_{cMAX} podľa (40) pôjde o blokový vektor tvorený násobkami J-rozmerných vektorov rozdielov priemerných rankov $Z^{(a,b)}$, $a, b = 1, \ldots, g$, $a < b$, definovaných podľa vzťahu (39). V prípade uvažovania všeobecného počtu g ošetrení je počet týchto vektorov $Z^{(a,b)}$ rovný $\tau = \frac{g(g-1)}{2}$. Výsledkom je teda τJ-rozmerný vektor, označme ho Z a definujme ho nasledovne:

$$Z = \begin{pmatrix} \frac{1}{\sqrt{n_1+n_2}} Z^{(1,2)} \\ \vdots \\ \frac{1}{\sqrt{n_{g-1}+n_g}} Z^{(g-1,g)} \end{pmatrix}. \tag{43}$$

Asymptotické normálne rozdelenie blokov tohto vektora bolo odôvodnené napr. v článku [6], my sme tento fakt uvádzali vo vete 1.7. Keďže zvolený blokový vektor nie je singulárny, dá sa očakávať asymptotické normálne rozdelenie aj celého vektora Z z (43). Dôkaz tejto normality je nad rámec našej práce, avšak z výsledkov simulácií uvedených na konci tejto kapitoly sa τJ-rozmerné normálne rozdelenia vektora Z javí ako správna hypotéza. V nasledujúcej časti odvodíme strednú hodnotu a asymptotickú kovariančnú maticu blokového vektora Z z (43) za platnosti nulovej hypotézy H_0 z (35). Postupovať budeme po zložkách, teda odvodíme stredné hodnoty resp. vzájomné kovariancie všetkých jednotlivých zložiek $\frac{1}{\sqrt{n_a+n_b}} Z_j^{(a,b)}$, $a, b = 1, \ldots, g$, $a < b$, $j = 1, \ldots, J$, v dôsledku čoho bude známe aj rozdelenie samotného vektora Z. Ako sme spomínali v úvode tejto kapitoly, vychádzať budeme z už dokázaných vzťahov a vlastností uvedených v článkoch [1], [3] a [6].

Nech $a, b \in \{1, \ldots, g\}$, $a < b$ a $j \in \{1, \ldots, J\}$. Uvažujme výraz $\frac{1}{\sqrt{n_a+n_b}} Z_j^{(a,b)}$ daný označením (39) a upravme ho využitím vzťahu (38) nasledovne:

$$\frac{1}{\sqrt{n_a+n_b}} Z_j^{(a,b)} = \frac{1}{\sqrt{n_a+n_b}} \left(\bar{R}^{(a,b)}_{b\cdot j\cdot} - \bar{R}^{(a,b)}_{a\cdot j\cdot} \right) = \frac{1}{m_j \sqrt{n_a+n_b}} \sum_{k=1}^{m_j} \left(\bar{R}^{(a,b)}_{b\cdot jk} - \bar{R}^{(a,b)}_{a\cdot jk} \right).$$

Podľa dokázaného vzťahu (6) z lemy 1.2 platí rovnosť

$$\bar{R}^{(a,b)}_{b\cdot jk} - \bar{R}^{(a,b)}_{a\cdot jk} = \frac{n_a+n_b}{2n_a n_b} \sum_{i=1}^{n_a} \sum_{l=1}^{n_b} \left(I_{\{x_{aijk} < x_{bljk}\}} - I_{\{x_{aijk} > x_{bljk}\}} \right),$$

na základe ktorej dostávame

$$\frac{1}{\sqrt{n_a+n_b}}Z_j^{(a,b)} = \frac{1}{m_j\sqrt{n_a+n_b}}\sum_{k=1}^{m_j}\left(\frac{n_a+n_b}{2n_an_b}\sum_{i=1}^{n_a}\sum_{l=1}^{n_b}\left(I_{\left\{x_{aijk}<x_{bljk}\right\}} - I_{\left\{x_{aijk}>x_{bljk}\right\}}\right)\right) =$$
$$= \frac{\sqrt{n_a+n_b}}{2m_jn_an_b}\sum_{k=1}^{m_j}\sum_{i=1}^{n_a}\sum_{l=1}^{n_b}\left(I_{\left\{x_{aijk}<x_{bljk}\right\}} - I_{\left\{x_{aijk}>x_{bljk}\right\}}\right).$$

Kvôli jednoduchosti ešte zaveďme pre rozdiel indikátorov v predchádzajúcom výraze označenie

$$\psi_{iljk}^{(a,b)} = I_{\left\{x_{aijk}<x_{bljk}\right\}} - I_{\left\{x_{aijk}>x_{bljk}\right\}}$$

pre všetky $a, b = 1, \ldots, g$, $a < b$, $i = 1, \ldots, n_a$, $l = 1, \ldots, n_b$, $j = 1, \ldots, J$, $k = 1, \ldots, m_j$. Pre zložky vektorov $\frac{1}{\sqrt{n_a+n_b}}Z^{(a,b)}$ sme teda získali vyjadrenie

$$\frac{1}{\sqrt{n_a+n_b}}Z_j^{(a,b)} = \frac{\sqrt{n_a+n_b}}{2m_jn_an_b}\sum_{k=1}^{m_j}\sum_{i=1}^{n_a}\sum_{l=1}^{n_b}\psi_{iljk}^{(a,b)}. \tag{44}$$

Tento tvar sa ukazuje byť vhodný na manipuláciu s priemernými rankami, preto s ním budeme ďalej pracovať. Vychádzajúc z definície (34) je zjavné, že platí

$$E\left(\psi_{iljk}^{(a,b)}\right) = \theta_{jk}^{(a,b)}. \tag{45}$$

Pre strednú hodnotu výrazu $\frac{1}{\sqrt{n_a+n_b}}Z_j^{(a,b)}$ teda podľa (44) dostávame

$$E\left(\frac{1}{\sqrt{n_a+n_b}}Z_j^{(a,b)}\right) = E\left(\frac{\sqrt{n_a+n_b}}{2m_jn_an_b}\sum_{k=1}^{m_j}\sum_{i=1}^{n_a}\sum_{l=1}^{n_b}\psi_{iljk}^{(a,b)}\right) = \frac{\sqrt{n_a+n_b}}{2m_j}\sum_{k=1}^{m_j}\theta_{jk}^{(a,b)}.$$

Špeciálne za platnosti nulovej hypotézy H_0 z (35) teda platí

$$E_{H_0}\left(\frac{1}{\sqrt{n_a+n_b}}Z_j^{(a,b)}\right) = 0, \quad a, b = 1, \ldots, g,\ a < b,\ j = 1, \ldots, J. \tag{46}$$

Nech teraz $a_1, a_2, b_1, b_2 \in \{1, \ldots, g\}$, $a_1 < b_1$, $a_2 < b_2$ a $j_1, j_2 \in \{1, \ldots, J\}$. Označme $\Sigma_{j_1j_2}^{(a_1,b_1),(a_2,b_2)}$ kovarianciu výrazov $\frac{1}{\sqrt{n_{a_1}+n_{b_1}}}Z_{j_1}^{(a_1,b_1)}$ a $\frac{1}{\sqrt{n_{a_2}+n_{b_2}}}Z_{j_2}^{(a_2,b_2)}$. Podľa (44) môžeme písať nasledovné:

$$\Sigma_{j_1j_2}^{(a_1,b_1),(a_2,b_2)} = \operatorname{cov}\left(\frac{1}{\sqrt{n_{a_1}+n_{b_1}}}Z_{j_1}^{(a_1,b_1)}, \frac{1}{\sqrt{n_{a_2}+n_{b_2}}}Z_{j_2}^{(a_2,b_2)}\right) =$$
$$= \operatorname{cov}\left(\frac{\sqrt{n_{a_1}+n_{b_1}}}{2m_{j_1}n_{a_1}n_{b_1}}\sum_{k_1=1}^{m_{j_1}}\sum_{i_1=1}^{n_{a_1}}\sum_{l_1=1}^{n_{b_1}}\psi_{i_1l_1j_1k_1}^{(a_1,b_1)}, \frac{\sqrt{n_{a_2}+n_{b_2}}}{2m_{j_2}n_{a_2}n_{b_2}}\sum_{k_2=1}^{m_{j_2}}\sum_{i_2=1}^{n_{a_2}}\sum_{l_2=1}^{n_{b_2}}\psi_{i_2l_2j_2k_2}^{(a_2,b_2)}\right).$$

Využitím vlastností kovariancie potom dostávame vzťah

$$\Sigma_{j_1j_2}^{(a_1,b_1),(a_2,b_2)} = \frac{\sqrt{(n_{a_1}+n_{b_1})(n_{a_2}+n_{b_2})}}{4m_{j_1}m_{j_2}n_{a_1}n_{a_2}n_{b_1}n_{b_2}}\sum_{k_1=1}^{m_{j_1}}\sum_{k_2=1}^{m_{j_2}}\sum_{i_1=1}^{n_{a_1}}\sum_{i_2=1}^{n_{a_2}}\sum_{l_1=1}^{n_{b_1}}\sum_{l_2=1}^{n_{b_2}}\operatorname{cov}\left(\psi_{i_1l_1j_1k_1}^{(a_1,b_1)}, \psi_{i_2l_2j_2k_2}^{(a_2,b_2)}\right). \tag{47}$$

Pomocou tohto všeobecného vzťahu budeme teraz kovariancie $\Sigma_{j_1 j_2}^{(a_1,b_1),(a_2,b_2)}$ vyjadrovať konkrétnejšie za platnosti nulovej hypotézy H_0 (35). Ich hodnoty budú pre rôzne prípady dvojíc párov a_1, b_1 a a_2, b_2 principiálne odlišné, preto ich budeme odvádzať pre každý taký prípad osobitne. Zakaždým budeme postupovať tak, že si najskôr vyjadríme hodnoty kovariancií $\mathrm{cov}\left(\psi_{i_1 l_1 j_1 k_1}^{(a_1,b_1)}, \psi_{i_2 l_2 j_2 k_2}^{(a_2,b_2)}\right)$ pre všetky možnosti indexov i_1, i_2, l_1, l_2 a príslušnú $\Sigma_{j_1 j_2}^{(a_1,b_1),(a_2,b_2)}$ potom dostaneme dosadením týchto kovariancií do vzťahu (47). V celej nasledujúcej časti uvažujeme ľubovoľné clustery $j_1, j_2 \in \{1, \ldots, J\}$.

I. $a_1 = a_2 = a$, $b_1 = b_2 = b$

Ako prvému sa budeme venovať prípadu dvoch rovnakých párov ošetrení $a, b \in \{1, \ldots, g\}$, $a < b$. Toto je jediný prípad ktorým sa v článkoch [3] a [6] zaoberajú, keďže triviálne pri uvažovaní dvoch ošetrení ($g = 2$) iná možnosť nenastáva. Aproximáciu kovariancie typu $\Sigma_{j_1 j_2}^{(a,b),(a,b)}$ možno nájsť v spomínaných článkoch aj s dôkazom (aj keď v článku [3] bez uvažovania clusterov). Pre úplnosť jej odvodenie ale uvádzame aj tu, a to vo vlastnej verzii s našim značením a detailnejšími výpočtami. Počítajme teda hodnoty kovariancií $\mathrm{cov}\left(\psi_{i_1 l_1 j_1 k_1}^{(a,b)}, \psi_{i_2 l_2 j_2 k_2}^{(a,b)}\right)$.

(a) $i_1 \neq i_2$, $l_1 \neq l_2$

Ako prvú uvažujme situáciu rôznych vektorov pozorovaní v oboch ošetreniach a, b. Kvôli nezávislosti vektorov pozorovaní sú nezávislé aj náhodné premenné $\psi_{i_1 l_1 j_1 k_1}^{(a,b)}$ a $\psi_{i_2 l_2 j_2 k_2}^{(a,b)}$, teda platí

$$\mathrm{cov}\left(\psi_{i_1 l_1 j_1 k_1}^{(a,b)}, \psi_{i_2 l_2 j_2 k_2}^{(a,b)}\right) = 0, \quad i_1 \neq i_2, l_1 \neq l_2. \tag{48}$$

(b) $i_1 = i_2 = i$, $l_1 \neq l_2$

Teraz už z dôvodu prepojenia náhodných premenných $\psi_{i l_1 j_1 k_1}^{(a,b)}$ a $\psi_{i l_2 j_2 k_2}^{(a,b)}$ voľbou rovnakého vektoru pozorovaní z a-teho ošetrenia nezávislosť neplatí, preto ich kovarianciu vyjadrujme ďalej. Podľa vzťahu (45) za platnosti nulovej hypotézy (35) platí

$$\begin{aligned}\mathrm{cov}\left(\psi_{i l_1 j_1 k_1}^{(a,b)}, \psi_{i l_2 j_2 k_2}^{(a,b)}\right) &= E\left(\psi_{i l_1 j_1 k_1}^{(a,b)} \psi_{i l_2 j_2 k_2}^{(a,b)}\right) - E\left(\psi_{i l_1 j_1 k_1}^{(a,b)}\right) E\left(\psi_{i l_2 j_2 k_2}^{(a,b)}\right) = \\ &= E\left(\psi_{i l_1 j_1 k_1}^{(a,b)} \psi_{i l_2 j_2 k_2}^{(a,b)}\right)\end{aligned}$$

a na základe vlastností podmienenej strednej hodnoty môžeme písať

$$\begin{aligned}\mathrm{cov}\left(\psi_{i l_1 j_1 k_1}^{(a,b)}, \psi_{i l_2 j_2 k_2}^{(a,b)}\right) &= E\left[E\left(\psi_{i l_1 j_1 k_1}^{(a,b)} \psi_{i l_2 j_2 k_2}^{(a,b)} \middle| X_{a j_1 k_1} = x_{a i j_1 k_1}, X_{a j_2 k_2} = x_{a i j_2 k_2}\right)\right] = \\ &= E\left[E\left(\psi_{i l_1 j_1 k_1}^{(a,b)} \middle| X_{a j_1 k_1} = x_{a i j_1 k_1}\right) E\left(\psi_{i l_2 j_2 k_2}^{(a,b)} \middle| X_{a j_2 k_2} = x_{a i j_2 k_2}\right)\right].\end{aligned}$$

Samostatne teraz vyjadríme obe podmienené stredné hodnoty, ktoré potom spätne dosadíme do tohto vzťahu. Počítajme teda bokom:

$$E\left(\psi^{(a,b)}_{il_1j_1k_1}\middle|X_{aj_1k_1}=x_{aij_1k_1}\right)=E\left(I_{\left\{x_{aij_1k_1}<x_{bl_1j_1k_1}\right\}}-I_{\left\{x_{aij_1k_1}>x_{bl_1j_1k_1}\right\}}\middle|X_{aj_1k_1}=x_{aij_1k_1}\right)=$$
$$=P\left(x_{aij_1k_1}<x_{bl_1j_1k_1}\middle|X_{aj_1k_1}=x_{aij_1k_1}\right)-P\left(x_{aij_1k_1}>x_{bl_1j_1k_1}\middle|X_{aj_1k_1}=x_{aij_1k_1}\right)=$$
$$=1-F_{bj_1k_1}\left(x_{aij_1k_1}\middle|X_{aj_1k_1}=x_{aij_1k_1}\right)-F_{bj_1k_1}\left(x_{aij_1k_1}\middle|X_{aj_1k_1}=x_{aij_1k_1}\right)=$$
$$=1-2F_{bj_1k_1}\left(X_{aj_1k_1}\right).$$

Rovnakým postupom by sme získali analogicky druhú podmienenú strednú hodnotu

$$E\left(\psi^{(a,b)}_{il_2j_2k_2}\middle|X_{aj_2k_2}=x_{aij_2k_2}\right)=1-2F_{bj_2k_2}\left(X_{aj_2k_2}\right).$$

Ich spätným dosadením potom dostávame

$$\begin{aligned}\operatorname{cov}\left(\psi^{(a,b)}_{il_1j_1k_1},\psi^{(a,b)}_{il_2j_2k_2}\right)&=E\Big[\Big(1-2F_{bj_1k_1}\left(X_{aj_1k_1}\right)\Big)\Big(1-2F_{bj_2k_2}\left(X_{aj_2k_2}\right)\Big)\Big]=\\&=4E\left[\left(\frac{1}{2}-F_{bj_1k_1}\left(X_{aj_1k_1}\right)\right)\left(\frac{1}{2}-F_{bj_2k_2}\left(X_{aj_2k_2}\right)\right)\right].\end{aligned}$$

Ako sa v [3] píše, za platnosti hypotézy H_0 (35) platí

$$E\Big(F_{bj_1k_1}\left(X_{aj_1k_1}\right)\Big)=E\Big(F_{bj_2k_2}\left(X_{aj_2k_2}\right)\Big)=\frac{1}{2}.$$

Využitím tejto vlastnosti v predchádzajúcom vzťahu dostávame finálne vyjadrenie

$$\operatorname{cov}\left(\psi^{(a,b)}_{i_1l_1j_1k_1},\psi^{(a,b)}_{i_2l_2j_2k_2}\right)=4\operatorname{cov}\left(F_{bj_1k_1}\left(X_{aj_1k_1}\right),F_{bj_2k_2}\left(X_{aj_2k_2}\right)\right),\quad i_1=i_2=i,l_1\neq l_2. \tag{49}$$

(c) $i_1\neq i_2$, $l_1=l_2=l$

Tu nastáva v princípe rovnaká situácia ako v bode (b), preto kovarianciu $\operatorname{cov}\left(\psi^{(a,b)}_{i_1lj_1k_1},\psi^{(a,b)}_{i_2lj_2k_2}\right)$ analogicky upravujme. Podľa vzťahu (45) za platnosti nulovej hypotézy (35) a na základe vlastností podmienenej strednej hodnoty obdobne platí

$$\operatorname{cov}\left(\psi^{(a,b)}_{i_1lj_1k_1},\psi^{(a,b)}_{i_2lj_2k_2}\right)=E\left[E\left(\psi^{(a,b)}_{i_1lj_1k_1}\middle|X_{bj_1k_1}=x_{blj_1k_1}\right)E\left(\psi^{(a,b)}_{i_2lj_2k_2}\middle|X_{bj_2k_2}=x_{blj_2k_2}\right)\right].$$

Opäť počítajme bokom:

$$E\left(\psi^{(a,b)}_{i_1lj_1k_1}\middle|X_{bj_1k_1}=x_{blj_1k_1}\right)=E\left(I_{\left\{x_{ai_1j_1k_1}<x_{blj_1k_1}\right\}}-I_{\left\{x_{ai_1j_1k_1}>x_{blj_1k_1}\right\}}\middle|X_{bj_1k_1}=x_{blj_1k_1}\right)=$$
$$=P\left(x_{ai_1j_1k_1}<x_{blj_1k_1}\middle|X_{bj_1k_1}=x_{blj_1k_1}\right)-P\left(x_{ai_1j_1k_1}>x_{blj_1k_1}\middle|X_{bj_1k_1}=x_{blj_1k_1}\right)=$$
$$=F_{aj_1k_1}\left(x_{blj_1k_1}\middle|X_{bj_1k_1}=x_{blj_1k_1}\right)-\left(1-F_{aj_1k_1}\left(x_{blj_1k_1}\middle|X_{bj_1k_1}=x_{blj_1k_1}\right)\right)=$$
$$=2F_{aj_1k_1}\left(X_{bj_1k_1}\right)-1.$$

Analogicky pre druhú podmienenú strednú hodnotu potom platí

$$E\left(\psi^{(a,b)}_{i_2lj_2k_2}\middle|X_{bj_2k_2}=x_{blj_2k_2}\right)=2F_{aj_2k_2}\left(X_{bj_2k_2}\right)-1$$

a po ich spätnom dosadení dostávame

$$\begin{aligned}\operatorname{cov}\left(\psi^{(a,b)}_{i_1lj_1k_1},\psi^{(a,b)}_{i_2lj_2k_2}\right)&=E\Big[\Big(2F_{aj_1k_1}\left(X_{bj_1k_1}\right)-1\Big)\Big(2F_{aj_2k_2}\left(X_{bj_2k_2}\right)-1\Big)\Big]=\\&=4E\left[\left(F_{aj_1k_1}\left(X_{bj_1k_1}\right)-\frac{1}{2}\right)\left(F_{aj_2k_2}\left(X_{bj_2k_2}\right)-\frac{1}{2}\right)\right].\end{aligned}$$

Opäť sa odvoláme na článok [3], podľa ktorého za platnosti hypotézy H_0 (35) platí

$$E\Big(F_{aj_1k_1}\left(X_{bj_1k_1}\right)\Big)=E\Big(F_{aj_2k_2}\left(X_{bj_2k_2}\right)\Big)=\frac{1}{2},$$

a využitím tejto vlastnosti získavame vyjadrenie

$$\operatorname{cov}\left(\psi^{(a,b)}_{i_1l_1j_1k_1},\psi^{(a,b)}_{i_2l_2j_2k_2}\right)=4\operatorname{cov}\left(F_{aj_1k_1}\left(X_{bj_1k_1}\right),F_{aj_2k_2}\left(X_{bj_2k_2}\right)\right),\quad i_1\neq i_2,l_1=l_2=l. \tag{50}$$

(d) $i_1=i_2=i$, $l_1=l_2=l$

Ako posledná nám ostala voľba rovnakých vektorov pozorovaní v rámci oboch ošetrení a aj b. Pre neznáme kovariancie vo vyjadreniach (49) a (50) z častí (b) a (c) boli odvodené konzistentné odhady, ktoré aj neskôr v práci uvedieme. V tomto prípade sa avšak k podobnému vyjadreniu nie je možné dopracovať, hodnoty kovariancií $\operatorname{cov}\left(\psi^{(a,b)}_{ilj_1k_1},\psi^{(a,b)}_{ilj_2k_2}\right)$ sa teda nedajú odhadovať na základe samotných dát, a preto sa ich budeme musieť „zbaviť“ využitím asymptotiky pre rastúce počty dát n_a, n_b. V tomto bode preto ukážeme, že sú skúmané kovariancie ohraničené konštantami nezávislými od počtu dát n_a, n_b, keďže sa v takom prípade ich hodnoty neskôr ukážu byť asymptoticky zanedbateľné. Náhodné premenné $\psi^{(a,b)}_{ilj_1k_1}$ a $\psi^{(a,b)}_{ilj_2k_2}$ môžeme v prípade spojitých endpointov zapísať nasledovne:

$$\psi^{(a,b)}_{ilj_1k_1}=\begin{cases}1 & \text{ak}\quad x_{aij_1k_1}<x_{blj_1k_1},\\ -1 & \text{ak}\quad x_{aij_1k_1}>x_{blj_1k_1},\end{cases}\qquad\text{a}\qquad \psi^{(a,b)}_{ilj_2k_2}=\begin{cases}1 & \text{ak}\quad x_{aij_2k_2}<x_{blj_2k_2},\\ -1 & \text{ak}\quad x_{aij_2k_2}>x_{blj_2k_2}.\end{cases}$$

Ich druhé mocniny $\left(\psi^{(a,b)}_{ilj_1k_1}\right)^2$ a $\left(\psi^{(a,b)}_{ilj_2k_2}\right)^2$ sú potom zjavne konštantné náhodné premenné, ktorých hodnota je 1 s pravdepodobnosťou 1. Na základe tohto a podľa vzťahu (45) za platnosti nulovej hypotézy (35) môžeme počítať disperzie $\psi^{(a,b)}_{ilj_1k_1}$ a $\psi^{(a,b)}_{ilj_2k_2}$, teda

$$\begin{aligned}\operatorname{var}\left(\psi^{(a,b)}_{ilj_1k_1}\right)&=E\left[\left(\psi^{(a,b)}_{ilj_1k_1}\right)^2\right]-E\left[\psi^{(a,b)}_{ilj_1k_1}\right]^2=1,\\ \operatorname{var}\left(\psi^{(a,b)}_{ilj_2k_2}\right)&=E\left[\left(\psi^{(a,b)}_{ilj_2k_2}\right)^2\right]-E\left[\psi^{(a,b)}_{ilj_2k_2}\right]^2=1.\end{aligned}$$

Ako priamy dôsledok dostávame rovnosť

$$\operatorname{cov}\left(\psi_{ilj_1k_1}^{(a,b)}, \psi_{ilj_2k_2}^{(a,b)}\right) = \operatorname{cor}\left(\psi_{ilj_1k_1}^{(a,b)}, \psi_{ilj_2k_2}^{(a,b)}\right)$$

a teda pre skúmanú kovarianciu triviálne platí ohraničenie

$$\operatorname{cov}\left(\psi_{i_1l_1j_1k_1}^{(a,b)}, \psi_{i_2l_2j_2k_2}^{(a,b)}\right) \in [-1,1], \quad i_1 = i_2 = i, l_1 = l_2 = l. \tag{51}$$

Tým sme doriešili všetky jednotlivé prípady kovariancií $\psi_{i_1l_1j_1k_1}^{(a,b)}$ a $\psi_{i_2l_2j_2k_2}^{(a,b)}$, a môžeme sa vrátiť k samotnej $\Sigma_{j_1j_2}^{(a,b),(a,b)}$. Uvažujme vyjadrenie (47) pri voľbe $a_1 = a_2 = a$, $b_1 = b_2 = b$, dosaďme doň podľa odvodených vzťahov (48), (49) a (50), a ďalej upravujme nasledovne:

$$\begin{aligned}
\Sigma_{j_1j_2}^{(a,b),(a,b)} &= \frac{n_a + n_b}{4m_{j_1}m_{j_2}n_a^2n_b^2} \sum_{k_1=1}^{m_{j_1}} \sum_{k_2=1}^{m_{j_2}} \sum_{i_1=1}^{n_a} \sum_{i_2=1}^{n_a} \sum_{l_1=1}^{n_b} \sum_{l_2=1}^{n_b} \operatorname{cov}\left(\psi_{i_1l_1j_1k_1}^{(a,b)}, \psi_{i_2l_2j_2k_2}^{(a,b)}\right) = \\
&= \frac{n_a + n_b}{4m_{j_1}m_{j_2}n_a^2n_b^2} \sum_{k_1=1}^{m_{j_1}} \sum_{k_2=1}^{m_{j_2}} \left(\sum_{\substack{i_1=1 \\ i_1=i_2=i}}^{n_a} \sum_{\substack{i_2=1 \\ l_1=l_2=l}}^{n_a} \sum_{l_1=1}^{n_b} \sum_{l_2=1}^{n_b} \operatorname{cov}\left(\psi_{ilj_1k_1}^{(a,b)}, \psi_{ilj_2k_2}^{(a,b)}\right) \right. \\
&\quad + \sum_{\substack{i_1=1 \\ i_1\neq i_2}}^{n_a} \sum_{i_2=1}^{n_a} \sum_{\substack{l_1=1 \\ l_1=l_2=l}}^{n_b} \sum_{l_2=1}^{n_b} 4 \operatorname{cov}\left(F_{aj_1k_1}\left(X_{bj_1k_1}\right), F_{aj_2k_2}\left(X_{bj_2k_2}\right)\right) \\
&\quad + \sum_{\substack{i_1=1 \\ i_1=i_2=i}}^{n_a} \sum_{i_2=1}^{n_a} \sum_{\substack{l_1=1 \\ l_1\neq l_2}}^{n_b} \sum_{l_2=1}^{n_b} 4 \operatorname{cov}\left(F_{bj_1k_1}\left(X_{aj_1k_1}\right), F_{bj_2k_2}\left(X_{aj_2k_2}\right)\right) \\
&\quad \left. + \sum_{\substack{i_1=1 \\ i_1\neq i_2}}^{n_a} \sum_{i_2=1}^{n_a} \sum_{\substack{l_1=1 \\ l_1\neq l_2}}^{n_b} \sum_{l_2=1}^{n_b} 0 \right) = \\
&= \frac{n_a + n_b}{4m_{j_1}m_{j_2}n_a^2n_b^2} \sum_{k_1=1}^{m_{j_1}} \sum_{k_2=1}^{m_{j_2}} \left(\sum_{i=1}^{n_a} \sum_{l=1}^{n_b} \operatorname{cov}\left(\psi_{ilj_1k_1}^{(a,b)}, \psi_{ilj_2k_2}^{(a,b)}\right) \right. \\
&\quad + 4n_a(n_a-1)n_b \operatorname{cov}\left(F_{aj_1k_1}\left(X_{bj_1k_1}\right), F_{aj_2k_2}\left(X_{bj_2k_2}\right)\right) \\
&\quad \left. + 4n_an_b(n_b-1) \operatorname{cov}\left(F_{bj_1k_1}\left(X_{aj_1k_1}\right), F_{bj_2k_2}\left(X_{aj_2k_2}\right)\right) \right) = \\
&= \frac{1}{m_{j_1}m_{j_2}} \sum_{k_1=1}^{m_{j_1}} \sum_{k_2=1}^{m_{j_2}} \left(\frac{n_a + n_b}{4n_a^2n_b^2} \sum_{i=1}^{n_a} \sum_{l=1}^{n_b} \operatorname{cov}\left(\psi_{ilj_1k_1}^{(a,b)}, \psi_{ilj_2k_2}^{(a,b)}\right) \right. \\
&\quad + \frac{(n_a+n_b)(n_a-1)}{n_an_b} \operatorname{cov}\left(F_{aj_1k_1}\left(X_{bj_1k_1}\right), F_{aj_2k_2}\left(X_{bj_2k_2}\right)\right) \\
&\quad \left. + \frac{(n_a+n_b)(n_b-1)}{n_an_b} \operatorname{cov}\left(F_{bj_1k_1}\left(X_{aj_1k_1}\right), F_{bj_2k_2}\left(X_{aj_2k_2}\right)\right) \right).
\end{aligned}$$

Kvôli neznámym hodnotám kovariancií $\operatorname{cov}\left(\psi_{ilj_1k_1}^{(a,b)}, \psi_{ilj_2k_2}^{(a,b)}\right)$ sa pozrime na asymptotické správanie posledného výrazu. Nech $\frac{n_a}{n_b} \to \lambda^{(a,b)} < \infty$ pre $\min\{n_a, n_b\} \to \infty$. Počítajme

teraz limity jednotlivých členov obsahujúcich počty dát n_a, n_b pre $\min\{n_a, n_b\} \to \infty$. Pre prvý člen na základe ohraničenia (51) môžeme písať

$$\frac{n_a+n_b}{4n_a^2n_b^2}\sum_{i=1}^{n_a}\sum_{l=1}^{n_b}-1 \leq \frac{n_a+n_b}{4n_a^2n_b^2}\sum_{i=1}^{n_a}\sum_{l=1}^{n_b}\operatorname{cov}\left(\psi^{(a,b)}_{ilj_1k_1},\psi^{(a,b)}_{ilj_2k_2}\right) \leq \frac{n_a+n_b}{4n_a^2n_b^2}\sum_{i=1}^{n_a}\sum_{l=1}^{n_b}1,$$

teda

$$-\frac{n_a+n_b}{4n_an_b} \leq \frac{n_a+n_b}{4n_a^2n_b^2}\sum_{i=1}^{n_a}\sum_{l=1}^{n_b}\operatorname{cov}\left(\psi^{(a,b)}_{ilj_1k_1},\psi^{(a,b)}_{ilj_2k_2}\right) \leq \frac{n_a+n_b}{4n_an_b}.$$

Keďže zjavne platí

$$\lim_{\min\{n_a,n_b\}\to\infty} \pm\frac{n_a+n_b}{4n_an_b} = \lim_{\min\{n_a,n_b\}\to\infty} \pm\left(\frac{1}{4n_a}+\frac{1}{4n_b}\right) = 0,$$

pre ohraničený člen rovnako dostávame

$$\lim_{\min\{n_a,n_b\}\to\infty} \frac{n_a+n_b}{4n_a^2n_b^2}\sum_{i=1}^{n_a}\sum_{l=1}^{n_b}\operatorname{cov}\left(\psi^{(a,b)}_{ilj_1k_1},\psi^{(a,b)}_{ilj_2k_2}\right) = 0,$$

a teda hodnoty kovariancií $\operatorname{cov}\left(\psi^{(a,b)}_{i_1l_1j_1k_1},\psi^{(a,b)}_{i_2l_2j_2k_2}\right)$ pre $i_1 = i_2 = i$, $l_1 = l_2 = l$ sú skutočne asymptoticky nesignifikantné, ako bolo už avizované. Dopočítajme ešte limity zvyšných dvoch výrazov, teda

$$\begin{aligned}
\lim_{\min\{n_a,n_b\}\to\infty} \frac{(n_a+n_b)(n_a-1)}{n_an_b} &= \lim_{\min\{n_a,n_b\}\to\infty} \frac{n_an_b+n_a^2-n_a-n_b}{n_an_b} = \\
&= \lim_{\min\{n_a,n_b\}\to\infty} \left(1+\frac{n_a}{n_b}-\frac{1}{n_b}-\frac{1}{n_a}\right) = 1+\lambda^{(a,b)}, \\
\lim_{\min\{n_a,n_b\}\to\infty} \frac{(n_a+n_b)(n_b-1)}{n_an_b} &= \lim_{\min\{n_a,n_b\}\to\infty} \frac{n_an_b+n_b^2-n_a-n_b}{n_an_b} = \\
&= \lim_{\min\{n_a,n_b\}\to\infty} \left(1+\frac{n_b}{n_a}-\frac{1}{n_b}-\frac{1}{n_a}\right) = 1+\frac{1}{\lambda^{(a,b)}}.
\end{aligned}$$

Dosadením získaných limít do vyjadrenia pre kovarianciu $\Sigma^{(a,b),(a,b)}_{j_1j_2}$ dostávame za platnosti nulovej hypotézy H_0 (35) a dostatočne veľké počty dát n_a, n_b aproximáciu

$$\begin{aligned}
\Sigma^{(a,b),(a,b)}_{j_1j_2} \approx \frac{1}{m_{j_1}m_{j_2}}\sum_{k_1=1}^{m_{j_1}}\sum_{k_2=1}^{m_{j_2}}\Bigg(&\left(1+\lambda^{(a,b)}\right)\operatorname{cov}\left(F_{aj_1k_1}\left(X_{bj_1k_1}\right),F_{aj_2k_2}\left(X_{bj_2k_2}\right)\right) \\
&+\left(1+\frac{1}{\lambda^{(a,b)}}\right)\operatorname{cov}\left(F_{bj_1k_1}\left(X_{aj_1k_1}\right),F_{bj_2k_2}\left(X_{aj_2k_2}\right)\right)\Bigg).
\end{aligned} \tag{52}$$

II. $a_1 = a_2 = a$, $b_1 \neq b_2$

Teraz sa už dostávame k vlastným odvodeniam kovariancií $\Sigma^{(a_1,b_1),(a_2,b_2)}_{j_1j_2}$, pri ktorých sa uvažuje viacero ošetrení ($g > 2$). Postupy budú ale analogické tým z článkov [3] a [6]. V tomto prípade bude závislosť/nezávislosť náhodných premenných $\psi^{(a,b_1)}_{i_1l_1j_1k_1}$ a $\psi^{(a,b_2)}_{i_2l_2j_2k_2}$

závisieť iba od voľby indexov i_1, i_2 pozorovaní a-teho ošetrenia. Pozorovania ošetrení b_1, b_2 v ďalšom preto uvažujeme ľubovoľné, teda $l_1 \in \{1, \ldots, n_{b_1}\}$, $l_2 \in \{1, \ldots, n_{b_2}\}$.

(a) $i_1 \neq i_2$

Podobne ako v bode (a) prípadu I. tu máme situáciu štyroch rôznych vektorov pozorovaní, a teda z dôvodu ich nezávislosti platí

$$\operatorname{cov}\left(\psi_{i_1 l_1 j_1 k_1}^{(a,b_1)}, \psi_{i_2 l_2 j_2 k_2}^{(a,b_2)}\right) = 0, \quad i_1 \neq i_2. \tag{53}$$

(b) $i_1 = i_2 = i$

Tu je situácia principiálne totožná tej z bodu (b) prípadu I. a pri upravovaní kovariancie $\operatorname{cov}\left(\psi_{i_1 l_1 j_1 k_1}^{(a,b_1)}, \psi_{i_2 l_2 j_2 k_2}^{(a,b_2)}\right)$ sa dá postupovať presne rovnakým spôsobom. Výsledné vyjadrenie možno preto získať zo vzťahu (49) zámenou indexu b za index b_1, resp. b_2 na príslušných miestach. Platí teda

$$\operatorname{cov}\left(\psi_{i_1 l_1 j_1 k_1}^{(a,b_1)}, \psi_{i_2 l_2 j_2 k_2}^{(a,b_2)}\right) = 4 \operatorname{cov}\left(F_{b_1 j_1 k_1}\left(X_{a j_1 k_1}\right), F_{b_2 j_2 k_2}\left(X_{a j_2 k_2}\right)\right), \quad i_1 = i_2 = i. \tag{54}$$

Už máme k dispozícii všetko potrebné a môžeme vyjadriť výslednú kovarianciu $\Sigma_{j_1 j_2}^{(a,b_1),(a,b_2)}$. Do vzťahu (47) pri voľbe $a_1 = a_2 = a$, $b_1 \neq b_2$ dosaďme podľa (53) a (54), a upravujme:

$$\begin{aligned}
\Sigma_{j_1 j_2}^{(a,b_1),(a,b_2)} &= \frac{\sqrt{(n_a + n_{b_1})(n_a + n_{b_2})}}{4 m_{j_1} m_{j_2} n_a^2 n_{b_1} n_{b_2}} \sum_{k_1=1}^{m_{j_1}} \sum_{k_2=1}^{m_{j_2}} \sum_{i_1=1}^{n_a} \sum_{i_2=1}^{n_a} \sum_{l_1=1}^{n_{b_1}} \sum_{l_2=1}^{n_{b_2}} \operatorname{cov}\left(\psi_{i_1 l_1 j_1 k_1}^{(a,b_1)}, \psi_{i_2 l_2 j_2 k_2}^{(a,b_2)}\right) = \\
&= \frac{\sqrt{(n_a + n_{b_1})(n_a + n_{b_2})}}{4 m_{j_1} m_{j_2} n_a^2 n_{b_1} n_{b_2}} \sum_{k_1=1}^{m_{j_1}} \sum_{k_2=1}^{m_{j_2}} \left(\sum_{\substack{i_1=1 \\ i_1 \neq i_2}}^{n_a} \sum_{i_2=1}^{n_a} \sum_{l_1=1}^{n_{b_1}} \sum_{l_2=1}^{n_{b_2}} 0 \right. \\
&\quad \left. + \sum_{\substack{i_1=1 \\ i_1=i_2=i}}^{n_a} \sum_{i_2=1}^{n_a} \sum_{l_1=1}^{n_{b_1}} \sum_{l_2=1}^{n_{b_2}} 4 \operatorname{cov}\left(F_{b_1 j_1 k_1}\left(X_{a j_1 k_1}\right), F_{b_2 j_2 k_2}\left(X_{a j_2 k_2}\right)\right) \right) = \\
&= \frac{\sqrt{(n_a + n_{b_1})(n_a + n_{b_2})}}{4 m_{j_1} m_{j_2} n_a^2 n_{b_1} n_{b_2}} \sum_{k_1=1}^{m_{j_1}} \sum_{k_2=1}^{m_{j_2}} 4 n_a n_{b_1} n_{b_2} \operatorname{cov}\left(F_{b_1 j_1 k_1}\left(X_{a j_1 k_1}\right), F_{b_2 j_2 k_2}\left(X_{a j_2 k_2}\right)\right) = \\
&= \frac{1}{m_{j_1} m_{j_2}} \frac{\sqrt{(n_a + n_{b_1})(n_a + n_{b_2})}}{n_a} \sum_{k_1=1}^{m_{j_1}} \sum_{k_2=1}^{m_{j_2}} \operatorname{cov}\left(F_{b_1 j_1 k_1}\left(X_{a j_1 k_1}\right), F_{b_2 j_2 k_2}\left(X_{a j_2 k_2}\right)\right).
\end{aligned}$$

A napokon algebraickou úpravou druhého zlomku dostávame pre kovarianciu $\Sigma_{j_1 j_2}^{(a,b_1),(a,b_2)}$, $b_1 \neq b_2$, za platnosti nulovej hypotézy H_0 (35) vyjadrenie

$$\begin{aligned}
&\Sigma_{j_1 j_2}^{(a,b_1),(a,b_2)} = \\
&\frac{1}{m_{j_1} m_{j_2}} \sqrt{\left(1 + \frac{n_{b_1}}{n_a}\right)\left(1 + \frac{n_{b_2}}{n_a}\right)} \sum_{k_1=1}^{m_{j_1}} \sum_{k_2=1}^{m_{j_2}} \operatorname{cov}\left(F_{b_1 j_1 k_1}\left(X_{a j_1 k_1}\right), F_{b_2 j_2 k_2}\left(X_{a j_2 k_2}\right)\right).
\end{aligned} \tag{55}$$

III. $a_1 \neq a_2$, $b_1 = b_2 = b$

Teraz bude o závislosti/nezávislosti náhodných premenných $\psi^{(a_1,b)}_{i_1 l_1 j_1 k_1}$ a $\psi^{(a_2,b)}_{i_2 l_2 j_2 k_2}$ rozhodovať iba voľba indexov l_1, l_2 pozorovaní b-teho ošetrenia. Pozorovania ošetrení a_1, a_2 v ďalšom preto uvažujeme ľubovoľné, teda $i_1 \in \{1, \ldots, n_{a_1}\}$, $i_2 \in \{1, \ldots, n_{a_2}\}$.

(a) $l_1 \neq l_2$

Rovnako ako v predchádzajúcich prípadoch, pre štyri rôzne nezávislé vektory pozorovaní platí

$$\operatorname{cov}\left(\psi^{(a_1,b)}_{i_1 l_1 j_1 k_1}, \psi^{(a_2,b)}_{i_2 l_2 j_2 k_2}\right) = 0, \quad l_1 \neq l_2. \tag{56}$$

(b) $l_1 = l_2 = l$

Tu je situácia principiálne totožná tentokrát tej z bodu (c) prípadu I., preto možno výsledné vyjadrenie získať zo vzťahu (50) zámenou indexu a za index a_1, resp. a_2 na príslušných miestach. Platí teda

$$\operatorname{cov}\left(\psi^{(a_1,b)}_{i_1 l_1 j_1 k_1}, \psi^{(a_2,b)}_{i_2 l_2 j_2 k_2}\right) = 4 \operatorname{cov}\left(F_{a_1 j_1 k_1}\left(X_{b j_1 k_1}\right), F_{a_2 j_2 k_2}\left(X_{b j_2 k_2}\right)\right), \quad l_1 = l_2 = l. \tag{57}$$

Teraz môžeme pristúpiť k vyjadreniu samotnej $\Sigma^{(a_1,b),(a_2,b)}_{j_1 j_2}$. Dalo by sa postupovať tak, ako v predošlých prípadoch dosadením získaných kovariancií do vyjadrenia (47). Na základe porovnania vzťahov (56) a (57) so vzťahmi (53) a (54) si ale môžeme uvedomiť, že by bol tento postup úplne rovnaký tomu v predchádzajúcom prípade II., iba pri vymenenej úlohe ošetrení a_1, a_2 a b_1, b_2. Preto môžeme výsledné vyjadrenie kovariancie získať priamo zo vzťahu (55) zámenou indexov a, b_1, b_2 za indexy b, a_1, a_2 v tomto poradí. Za platnosti nulovej hypotézy H_0 (35) dostávame teda pre kovarianciu $\Sigma^{(a_1,b),(a_2,b)}_{j_1 j_2}$, $a_1 \neq a_2$, vzťah

$$\begin{aligned}&\Sigma^{(a_1,b),(a_2,b)}_{j_1 j_2} = \\ &\frac{1}{m_{j_1} m_{j_2}} \sqrt{\left(1 + \frac{n_{a_1}}{n_b}\right)\left(1 + \frac{n_{a_2}}{n_b}\right)} \sum_{k_1=1}^{m_{j_1}} \sum_{k_2=1}^{m_{j_2}} \operatorname{cov}\left(F_{a_1 j_1 k_1}\left(X_{b j_1 k_1}\right), F_{a_2 j_2 k_2}\left(X_{b j_2 k_2}\right)\right).\end{aligned} \tag{58}$$

IV. $a_1 \neq a_2$, $b_1 \neq b_2$, $a_1 = b_2 = s$

V tomto prípade spravíme rozdelenie na základe voľby indexov i_1, l_2 pozorovaní ošetrenia s, pozorovania ostatných ošetrení uvažujeme ľubovoľné, teda $i_2 \in \{1, \ldots, n_{a_2}\}$, $l_1 \in \{1, \ldots, n_{b_1}\}$.

(a) $i_1 \neq l_2$

Tu máme opäť situáciu štyroch rôznych vektorov pozorovaní, preto z ich nezávislosti vyplýva

$$\operatorname{cov}\left(\psi^{(s,b_1)}_{i_1 l_1 j_1 k_1}, \psi^{(a_2,s)}_{i_2 l_2 j_2 k_2}\right) = 0, \quad i_1 \neq l_2. \tag{59}$$

(b) $i_1 = l_2 = p$

Z dôvodu závislosti kovarianciu $\operatorname{cov}\left(\psi^{(s,b_1)}_{pl_1j_1k_1}, \psi^{(a_2,s)}_{i_2pj_2k_2}\right)$ ďalej upravujme. Na základe vzťahu (45) za platnosti nulovej hypotézy (35) a využitím vlastností podmienenej strednej hodnoty dostávame

$$\operatorname{cov}\left(\psi^{(s,b_1)}_{pl_1j_1k_1}, \psi^{(a_2,s)}_{i_2pj_2k_2}\right) = E\left[E\left(\psi^{(s,b_1)}_{pl_1j_1k_1}\middle|X_{sj_1k_1} = x_{spj_1k_1}\right) E\left(\psi^{(a_2,s)}_{i_2pj_2k_2}\middle|X_{sj_2k_2} = x_{spj_2k_2}\right)\right].$$

Obe tieto podmienené stredné hodnoty sme už počítali v časti I. v bode (b), resp. (c). Využitím týchto výsledkov teda môžeme po prispôsobení indexácie priamo písať

$$E\left(\psi^{(s,b_1)}_{pl_1j_1k_1}\middle|X_{sj_1k_1} = x_{spj_1k_1}\right) = 1 - 2F_{b_1j_1k_1}\left(X_{sj_1k_1}\right)$$

a

$$E\left(\psi^{(a_2,s)}_{i_2pj_2k_2}\middle|X_{sj_2k_2} = x_{spj_2k_2}\right) = 2F_{a_2j_2k_2}\left(X_{sj_2k_2}\right) - 1.$$

Ich spätným dosadením dostávame

$$\begin{aligned}\operatorname{cov}\left(\psi^{(s,b_1)}_{pl_1j_1k_1}, \psi^{(a_2,s)}_{i_2pj_2k_2}\right) &= E\Big[\Big(1 - 2F_{b_1j_1k_1}\left(X_{sj_1k_1}\right)\Big)\Big(2F_{a_2j_2k_2}\left(X_{sj_2k_2}\right) - 1\Big)\Big] = \\ &= -4E\left[\left(\frac{1}{2} - F_{b_1j_1k_1}\left(X_{sj_1k_1}\right)\right)\left(\frac{1}{2} - F_{a_2j_2k_2}\left(X_{sj_2k_2}\right)\right)\right]\end{aligned}$$

a opäť využitím vlastnosti

$$E\Big(F_{b_1j_1k_1}\left(X_{sj_1k_1}\right)\Big) = E\Big(F_{a_2j_2k_2}\left(X_{sj_2k_2}\right)\Big) = \frac{1}{2}$$

spomínanej v článku [3] dostávame za platnosti nulovej hypotézy H_0 (35) vyjadrenie

$$\operatorname{cov}\left(\psi^{(s,b_1)}_{i_1l_1j_1k_1}, \psi^{(a_2,s)}_{i_2l_2j_2k_2}\right) = -4\operatorname{cov}\left(F_{b_1j_1k_1}\left(X_{sj_1k_1}\right), F_{a_2j_2k_2}\left(X_{sj_2k_2}\right)\right), \quad i_1 = l_2 = p. \quad (60)$$

Vyjadrujme teraz výslednú kovarianciu $\Sigma^{(s,b_1),(a_2,s)}_{j_1j_2}$. Do vzťahu (47) pri voľbe $a_1 \neq a_2$,

$b_1 \neq b_2$, $a_1 = b_2 = s$ dosaďme podľa (59) a (60), a upravujme:

$$\Sigma_{j_1 j_2}^{(s,b_1),(a_2,s)} = \frac{\sqrt{(n_s + n_{b_1})(n_{a_2} + n_s)}}{4 m_{j_1} m_{j_2} n_s^2 n_{a_2} n_{b_1}} \sum_{k_1=1}^{m_{j_1}} \sum_{k_2=1}^{m_{j_2}} \sum_{i_1=1}^{n_s} \sum_{i_2=1}^{n_{a_2}} \sum_{l_1=1}^{n_{b_1}} \sum_{l_2=1}^{n_s} \operatorname{cov}\left(\psi_{i_1 l_1 j_1 k_1}^{(s,b_1)}, \psi_{i_2 l_2 j_2 k_2}^{(a_2,s)}\right) =$$

$$= \frac{\sqrt{(n_s + n_{b_1})(n_{a_2} + n_s)}}{4 m_{j_1} m_{j_2} n_s^2 n_{a_2} n_{b_1}} \sum_{k_1=1}^{m_{j_1}} \sum_{k_2=1}^{m_{j_2}} \left(\sum_{\substack{i_1=1 \\ i_1 \neq l_2}}^{n_s} \sum_{l_2=1}^{n_s} \sum_{i_2=1}^{n_{a_2}} \sum_{l_1=1}^{n_{b_1}} 0 \right.$$

$$\left. + \sum_{\substack{i_1=1 \\ i_1=l_2=p}}^{n_s} \sum_{l_2=1}^{n_s} \sum_{i_2=1}^{n_{a_2}} \sum_{l_1=1}^{n_{b_1}} -4 \operatorname{cov}\left(F_{b_1 j_1 k_1}\left(X_{s j_1 k_1}\right), F_{a_2 j_2 k_2}\left(X_{s j_2 k_2}\right)\right) \right) =$$

$$= \frac{\sqrt{(n_s + n_{b_1})(n_{a_2} + n_s)}}{4 m_{j_1} m_{j_2} n_s^2 n_{a_2} n_{b_1}} \sum_{k_1=1}^{m_{j_1}} \sum_{k_2=1}^{m_{j_2}} -4 n_s n_{a_2} n_{b_1} \operatorname{cov}\left(F_{b_1 j_1 k_1}\left(X_{s j_1 k_1}\right), F_{a_2 j_2 k_2}\left(X_{s j_2 k_2}\right)\right) =$$

$$= -\frac{1}{m_{j_1} m_{j_2}} \frac{\sqrt{(n_s + n_{b_1})(n_{a_2} + n_s)}}{n_s} \sum_{k_1=1}^{m_{j_1}} \sum_{k_2=1}^{m_{j_2}} \operatorname{cov}\left(F_{b_1 j_1 k_1}\left(X_{s j_1 k_1}\right), F_{a_2 j_2 k_2}\left(X_{s j_2 k_2}\right)\right).$$

Pre kovarianciu $\Sigma_{j_1 j_2}^{(s,b_1),(a_2,s)}$ za platnosti nulovej hypotézy H_0 (35) teda platí vzťah

$$\Sigma_{j_1 j_2}^{(s,b_1),(a_2,s)} =$$
$$- \frac{1}{m_{j_1} m_{j_2}} \sqrt{\left(1 + \frac{n_{b_1}}{n_s}\right)\left(1 + \frac{n_{a_2}}{n_s}\right)} \sum_{k_1=1}^{m_{j_1}} \sum_{k_2=1}^{m_{j_2}} \operatorname{cov}\left(F_{b_1 j_1 k_1}\left(X_{s j_1 k_1}\right), F_{a_2 j_2 k_2}\left(X_{s j_2 k_2}\right)\right). \tag{61}$$

V. $a_1 \neq a_2$, $b_1 \neq b_2$, $a_2 = b_1 = s$

Pre úplnosť uvádzame aj túto možnosť voľby indexov ošetrení. Z dôvodu symetrie kovariancie avšak platí rovnosť $\Sigma_{j_1 j_2}^{(a_1,s),(s,b_2)} = \Sigma_{j_2 j_1}^{(s,b_2),(a_1,s)}$, tým pádom je tento prípad totožný prípadu IV., a teda môžeme použiť už odvodený vzťah (61).

VI. $a_1 \neq a_2, b_1 \neq b_2, a_1 \neq b_2, a_2 \neq b_1$

Ako posledná nám ostala voľba štyroch odlišných ošetrení a_1, a_2, b_1, b_2. V tomto prípade iná ako voľba štyroch rôznych vektorov pozorovaní nastať nemôže, a teda z dôvodu ich nezávislosti budú všetky kovariancie náhodných premenných $\psi_{i_1 l_1 j_1 k_1}^{(a_1,b_1)}$ a $\psi_{i_2 l_2 j_2 k_2}^{(a_2,b_2)}$ nulové. Pre kovarianciu $\Sigma_{j_1 j_2}^{(a_1,b_1),(a_2,b_2)}$, $a_1 \neq a_2, b_1 \neq b_2, a_1 \neq b_2, a_2 \neq b_1$, teda podľa vzťahu (47) na záver dostávame

$$\Sigma_{j_1 j_2}^{(a_1,b_1),(a_2,b_2)} = 0. \tag{62}$$

Týmto sme vyčerpali všetky možnosti voľby štvorice indexov ošetrení a kovarianciu $\Sigma_{j_1 j_2}^{(a_1,b_1),(a_2,b_2)}$ máme teda vyjadrenú pre všetky $a_1, a_2, b_1, b_2 \in \{1, \ldots, g\}$, $a_1 < b_1$, $a_2 < b_2$, $j_1, j_2 \in \{1, \ldots, J\}$. Keď sa teraz spätne pozrieme na výsledné vyjadrenia v jednotlivých

prípadoch I.-VI., tak si môžeme všimnúť, že sme dostali iba 3 principiálne odlišné výsledky. Prvým je vzťah (52) z bodu I. (dve spoločné ošetrenia), druhým sú vzťahy (55), (58) a (61) z bodov II.-V. (jedno spoločné ošetrenie), a tretím je vzťah (62) z posledného bodu VI. (žiadne spoločné ošetrenie). Uvažujme teraz prípady II.-V. a pokúsme sa pre nich získané vyjadrenia kovarincií $\Sigma_{j_1 j_2}^{(a_1,b_1),(a_2,b_2)}$ zjednotiť. Nech teda $a_1, a_2, b_1, b_2 \in \{1, \ldots, g\}$, $a_1 < b_1$, $a_2 < b_2$ a nech platí práve jedna z rovností $a_1 = a_2, b_1 = b_2, a_1 = b_2, a_2 = b_1$. Tieto rovnajúce sa indexy označme s, zvyšný index z dvojice a_1, b_1 označme r_1 a podobne označme r_2 zvyšný index z dvojice a_2, b_2. V takomto značení možno odmocninové výrazy vystupujúce vo vzťahoch (55), (58) a (61) jednotne zapísať ako $\sqrt{\left(1 + \frac{n_{r_1}}{n_s}\right)\left(1 + \frac{n_{r_2}}{n_s}\right)}$. A podobne pre kovariancie vystupujúce v dvojitých sumách v každom z týchto vzťahov, tie môžeme jednotne zapísať ako $\operatorname{cov}\left(F_{r_1 j_1 k_1}\left(X_{s j_1 k_1}\right), F_{r_2 j_2 k_2}\left(X_{s j_2 k_2}\right)\right)$. Vyjadrenia (55) a (58) sa od vyjadrenia (61) avšak ešte líšia v znamienku. Môžeme si všimnúť, že v prípadoch II. a III. platí rovnaké usporiadanie dvojíc indexov ošetrení s, r_1 a s, r_2, konkrétne $s < r_1, s < r_2$ v prípade II. a $s > r_1, s > r_2$ v prípade III. To ale neplatí pre prípad IV., kde majú dvojice s, r_1 a s, r_2 opačné usporiadania, teda $s < r_1, s > r_2$. Na základe tohto pozorovania môžeme toto znamienko jednotne zapísať pomocou funkcie signum, napr. ako súčin $\operatorname{sgn}(s - r_1)\operatorname{sgn}(s - r_2)$. A teda ak uvažujeme voľbu indexov a_1, a_2, b_1, b_2, resp. r_1, r_2, s, popísanú vyššie, dostávame pre kovarianciu $\Sigma_{j_1 j_2}^{(a_1,b_1),(a_2,b_2)}$ spoločný zápis

$$\begin{aligned}\Sigma_{j_1 j_2}^{(a_1,b_1),(a_2,b_2)} = \operatorname{sgn}(s - r_1)\operatorname{sgn}(s - r_2)\frac{1}{m_{j_1} m_{j_2}}\sqrt{\left(1 + \frac{n_{r_1}}{n_s}\right)\left(1 + \frac{n_{r_2}}{n_s}\right)} \\ \times \sum_{k_1=1}^{m_{j_1}} \sum_{k_2=1}^{m_{j_2}} \operatorname{cov}\left(F_{r_1 j_1 k_1}\left(X_{s j_1 k_1}\right), F_{r_2 j_2 k_2}\left(X_{s j_2 k_2}\right)\right).\end{aligned} \tag{63}$$

Kvôli praktickému využitiu odvodených kovariancií $\Sigma_{j_1 j_2}^{(a_1,b_1),(a_2,b_2)}$ je na záver ešte potrebné odhadnúť neznáme kovariancie typu $\operatorname{cov}\left(F_{r_1 j_1 k_1}\left(X_{s j_1 k_1}\right), F_{r_2 j_2 k_2}\left(X_{s j_2 k_2}\right)\right)$ vystupujúce vo vzťahoch (52) a (63). Ako sme aj spomínali, pre kovariancie tohto typu už boli odvodené ich konzistentné odhady. Je ich možné nájsť napr. v článku [6] alebo v inej forme v článku [1] aj s dôkazom. Nech teraz $r_1, r_2, s \in \{1, \ldots, g\}$, $r_1 \neq s, r_2 \neq s$, $j_1, j_2 \in \{1, \ldots, J\}$, $k_1 \in \{1, \ldots, m_{j_1}\}$, $k_2 \in \{1, \ldots, m_{j_2}\}$ a nech $\widehat{\operatorname{cov}}\left(F_{r_1 j_1 k_1}\left(X_{s j_1 k_1}\right), F_{r_2 j_2 k_2}\left(X_{s j_2 k_2}\right)\right)$ označuje konzistentný odhad kovariancie $\operatorname{cov}\left(F_{r_1 j_1 k_1}\left(X_{s j_1 k_1}\right), F_{r_2 j_2 k_2}\left(X_{s j_2 k_2}\right)\right)$. Podľa článkov [1] a

[6] potom platí vzťah

$$\begin{aligned}
&\widehat{\mathrm{cov}}\Big(F_{r_1j_1k_1}\left(X_{sj_1k_1}\right), F_{r_2j_2k_2}\left(X_{sj_2k_2}\right)\Big) = \\
&= \frac{1}{n_s}\sum_{i=1}^{n_s}\left[\left(\frac{1}{n_{r_1}}\sum_{l_1=1}^{n_{r_1}} I_{\left\{x_{r_1l_1j_1k_1}<x_{sij_1k_1}\right\}} - \frac{1+\widehat{\theta}_{j_1k_1}^{(r_1,s)}}{2}\right)\right. \\
&\qquad\left.\times\left(\frac{1}{n_{r_2}}\sum_{l_2=1}^{n_{r_2}} I_{\left\{x_{r_2l_2j_2k_2}<x_{sij_2k_2}\right\}} - \frac{1+\widehat{\theta}_{j_2k_2}^{(r_2,s)}}{2}\right)\right],
\end{aligned} \tag{64}$$

kde $\widehat{\theta}_{jk}^{(a,b)}$ je nevychýlený odhad parametra $\theta_{jk}^{(a,b)}$ z (34), definovaný podľa vzťahu (37). Teraz už máme k dispozícii všetko potrebné k tomu, aby sme skúmané kovariancie odhadli iba za pomoci dát. Nech $a_1, a_2, b_1, b_2 \in \{1, \ldots, g\}$, $a_1 < b_1$, $a_2 < b_2$ a nech $j_1, j_2 \in \{1, \ldots, J\}$. Označme $S_{j_1j_2}^{(a_1,b_1),(a_2,b_2)}$ konzistentný odhad kovariancie $\Sigma_{j_1j_2}^{(a_1,b_1),(a_2,b_2)}$. Uvažujme odvodené vyjadrenia (52), (62) a (63). Využitím odhadov zo vzťahu (64) a nahradením $\lambda^{(a,b)}$ podielom počtov dát n_a a n_b dostávame pre konzistentný odhad $S_{j_1j_2}^{(a_1,b_1),(a_2,b_2)}$ vyjadrenie

$$S_{j_1j_2}^{(a_1,b_1),(a_2,b_2)} = \begin{cases}
\dfrac{1}{m_{j_1}m_{j_2}}\displaystyle\sum_{k_1=1}^{m_{j_1}}\sum_{k_2=1}^{m_{j_2}}\left(\left(1+\frac{n_a}{n_b}\right)\widehat{\mathrm{cov}}\Big(F_{aj_1k_1}\left(X_{bj_1k_1}\right), F_{aj_2k_2}\left(X_{bj_2k_2}\right)\Big)\right. \\
\qquad\qquad\left. +\ \left(1+\dfrac{n_b}{n_a}\right)\widehat{\mathrm{cov}}\Big(F_{bj_1k_1}\left(X_{aj_1k_1}\right), F_{bj_2k_2}\left(X_{aj_2k_2}\right)\Big)\right) \\
\quad \text{ak } a_1 = a_2 = a, b_1 = b_2 = b, \\
\operatorname{sgn}(s-r_1)\operatorname{sgn}(s-r_2)\dfrac{1}{m_{j_1}m_{j_2}}\sqrt{\left(1+\dfrac{n_{r_1}}{n_s}\right)\left(1+\dfrac{n_{r_2}}{n_s}\right)} \\
\times\displaystyle\sum_{k_1=1}^{m_{j_1}}\sum_{k_2=1}^{m_{j_2}}\widehat{\mathrm{cov}}\Big(F_{r_1j_1k_1}\left(X_{sj_1k_1}\right), F_{r_2j_2k_2}\left(X_{sj_2k_2}\right)\Big) \\
\quad \text{ak platí práve jedna z rovností } a_1 = a_2 = s, b_1 = b_2 = s, \\
\quad a_1 = b_2 = s, a_2 = b_1 = s \text{ a } r_1, r_2 \text{ označujú zostávajúce indexy} \\
\quad \text{tak, aby k nim prislúchali indexy } j_1, j_2 \text{ v tomto poradí}, \\
0 \quad \text{inak.}
\end{cases} \tag{65}$$

V dôsledku vyjadrenia (65) dostávame pre konzistentný odhad $\widehat{\mathrm{var}}\left(\bar{R}_{b\cdot j\cdot}^{(a,b)} - \bar{R}_{a\cdot j\cdot}^{(a,b)}\right)$ vystupujúci v definícii (40) testovej štatistiky T_{cMAX}^{g} vzťah

$$\widehat{\mathrm{var}}\left(\bar{R}_{b\cdot j\cdot}^{(a,b)} - \bar{R}_{a\cdot j\cdot}^{(a,b)}\right) = (n_a + n_b)\, S_{jj}^{(a,b),(a,b)}, \quad a, b = 1, \ldots, g,\ a < b,\ j = 1, \ldots, J. \tag{66}$$

2.5 Výpočet p-value

Vychádzajúc zo vzťahov odvodených v predchádzajúcej časti teraz uvedieme praktický spôsob výpočtu p-value zovšeobecneného T^g_{cMAX} testu s testovou štatistikou definovanou podľa vzťahu (40). Tento postup bude v princípe rovnaký tomu pre prípad T_{cMAX} testu pri uvažovaní dvoch ošetrení ($g = 2$). Nech $t \geq 0$ značí hodnotu nameranej testovej štatistiky. Podľa tvaru testovej štatistiky T^g_{cMAX} uvedeného v (41) pre hodnotu p-value platí

$$P_{H_0}\left(T^g_{cMAX} \geq t\right) = 1 - P_{H_0}\left(T^g_{cMAX} < t\right) = \\ = 1 - P_{H_0}\left(\max_{\substack{j\in\{1,\ldots,J\}\\ a,b\in\{1,\ldots,g\}\\ a<b}} \left(\left|\frac{Z_j^{(a,b)}}{\sqrt{\widehat{\mathrm{var}}\left(Z_j^{(a,b)}\right)}}\right|\right) < t\right).$$

Po úprave poslednej pravdepodobnosti môžeme písať

$$P_{H_0}\left(T^g_{cMAX} \geq t\right) = \\ = 1 - P_{H_0}\left(\left|\frac{Z_j^{(a,b)}}{\sqrt{\widehat{\mathrm{var}}\left(Z_j^{(a,b)}\right)}}\right| < t, \quad j = 1,\ldots,J,\ a,b = 1,\ldots,g,\ a < b\right)$$

a dosadením za odhad $\widehat{\mathrm{var}}\left(Z_j^{(a,b)}\right)$ podľa (66) dostávame na záver vzťah

$$P_{H_0}\left(T^g_{cMAX} \geq t\right) = \\ = 1 - P_{H_0}\left(\left|\frac{Z_j^{(a,b)}}{\sqrt{(n_a + n_b)\, S_{jj}^{(a,b),(a,b)}}}\right| < t, \quad j = 1,\ldots,J,\ a,b = 1,\ldots,g,\ a < b\right). \tag{67}$$

Výrazy vystupujúce v absolútnych hodnotách vo vzťahu (67) sú jednotlivé štandardizované zložky vektora Z z (43). Preto môžeme skúmanú p-value počítať pomocou integrálu τJ-rozmernej hustoty vektora Z, kde $\tau = \frac{g(g-1)}{2}$. Túto hustotu vieme za platnosti nulovej hypotézy H_0 z (35) aproximovať na základe predpokladanej normality vektora Z a využitím jeho odvodenej strednej hodnoty podľa (46), a odhadu kovariančnej matice podľa (65), ktorý preškálujeme na korelačnú maticu. Integračnou oblasťou je podľa (67) τJ-rozmerná kocka centrovaná v bode $0_{\tau J}$ s hranami dĺžky $2t$ a rovnobežnými so súradnicovými osami. Využitím asymptotickej normality vektora Z a na základe vzťahov (46), (65), a (67), budeme v nasledujúcej časti simulačne skúmať kvalitu zovšeobecneného testu T^g_{cMAX}.

2.6 Simulácie

Uvažujme porovnanie g ošetrení X_a, $a = 1, \ldots, g$, s m endpointami zadelenými do J clusterov robené pomocou testovania rozšírenej verzie zovšeobecnenej Behrens-Fisher hypotézy H_0 z (35) a príslušné značenie zavedené v rámci tejto kapitoly. V nasledujúcich simuláciách sa bližšie pozrieme na kvalitu navrhnutého zovšeobecnenia T_{cMAX} testu. Pomocou odhadov pravdepodobnosti chyby 1. druhu overíme správnosť vyššie odvodených asymptotických vlastností a odhadov a taktiež správnosť spôsobu výpočtu p-value, ktorý sme uviedli v predchádzajúcej časti. Taktiež nás budú zaujímať odhady sily navrhnutého testu, ktoré nám zase niečo povedia o vhodnosti voľby samotnej zovšeobecnenej testovej štatistiky T^g_{cMAX} z (40). Dáta budeme generovať pre 3 rôzne scenáre v prostredí R [5]. V každom z nich budeme uvažovať $m = 6$ endpointov s clusterovou štruktúrou, pričom vždy zvolíme počet clusterov $J = 3$ a počty endpointov v rámci daných clusterov $m_1 = 3$, $m_2 = 2, m_3 = 1$. Kvôli porovnaniu budeme pri každom scenári uvádzať aj výsledky získané pre jednofaktorový MANOVA test.

V rámci prvého scenára uvažujme $g = 4$, teda testovanie odlišnosti štyroch ošetrení X_a, $a = 1, \ldots, 4$. Budeme uvažovať dve rôzne voľby rozdelení týchto ošetrení. Prvým bude 6-rozmerné normálne rozdelenie, teda $X_a \sim N_6(\mu_a, \Sigma_a)$, $a = 1, \ldots, 4$. Druhým bude 6-rozmerné zovšeobecnené Studentovo t-rozdelenie s počtom stupňov voľnosti $df = 3$, pričom k ošetreniu X_a prislúcha vektor stredných hodnôt μ_a a škálovacia matica Σ_a, $a = 1, \ldots, 4$. V prípade počítania odhadov pravd. chyby 1. druhu je potrebné dáta generovať za platnosti nulovej hypotézy H_0 z (35). To pre oba prípady rozdelenia ošetrení docielime voľbou rovnakých vektorov stredných hodnôt $\mu_a = 0_6$, $a = 1, \ldots, 4$. Pri skúmaní sily testu sme stredné hodnoty zvolili rôzne, konkrétne

$$\mu_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \qquad \mu_2 = \begin{pmatrix} 1{,}5 \\ 0{,}5 \\ 0{,}5 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \qquad \mu_3 = \begin{pmatrix} -2 \\ -0{,}5 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \qquad \mu_4 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Kovariančné matice v prípade normálneho rozdelenia, resp. škálovacie matice v prípade

zovšeobecneného t-rozdelenia, sme zvolili rovnaké, konkrétne

$$\Sigma_a = \begin{pmatrix} 100 & 36 & 16 & 0 & 0 & 0 \\ 36 & 25 & 4 & 0 & 0 & 0 \\ 16 & 4 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0{,}0225 & 0{,}01 & 0 \\ 0 & 0 & 0 & 0{,}01 & 0{,}01 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0{,}25 \end{pmatrix}, \quad a = 1, \ldots, 4.$$

Pre každý z týchto dvoch prípadov rozdelení X_a, $a = 1, \ldots, 4$, sme v prostredí R [5] náhodne generovali 6-rozmerné dáta pre nasledovné tri voľby veľkostí ošetrení n_a:

- $n_1 = 50$, $n_2 = 60$, $n_3 = 50$, $n_4 = 40$,
- $n_1 = 100$, $n_2 = 120$, $n_3 = 100$, $n_4 = 80$ a
- $n_1 = 200$, $n_2 = 240$, $n_3 = 200$, $n_4 = 160$.

Na takto generované dáta sme aplikovali test pomocou T^g_{cMAX} definovanej podľa vzťahu (40). Testovali sme na hladine významnosti $\alpha = 0{,}05$ a vykonaných bolo vždy 10 000 simulácií. Získané odhady pravdepodobností chyby 1. druhu a sily testu zaokrúhlené na 3 desatinné miesta sú vypísané v tabuľke 4. Pre porovnanie uvádzame aj výsledky pre MANOVA test.

Tabuľka 4: Pravdepodobnosť chyby 1. druhu a sila T^g_{cMAX} testu, 4 ošetrenia, homoskedasticita

	Pravd. chyby 1. druhu			Sila testu		
	$n_a \sim 50$	$n_a \sim 100$	$n_a \sim 200$	$n_a \sim 50$	$n_a \sim 100$	$n_a \sim 200$
6-rozmerné normálne rozdelenie						
T^g_{cMAX}	0,080	0,068	0,056	0,199	0,353	0,654
MANOVA	0,050	0,052	0,049	0,291	0,621	0,953
6-rozmerné zovšeobecnené Studentovo t-rozdelenie						
T^g_{cMAX}	0,085	0,067	0,053	0,160	0,226	0,443
MANOVA	0,039	0,046	0,046	0,124	0,235	0,467

Z výsledkov uvedených v tabuľke 4 vidíme, že odhady pravdepodobnosti chyby 1. druhu navrhnutého testu sú pre obe voľby rozdelení nad úrovňou 5%. Tento nedostatok

je najvýraznejší pri nižších počtoch dát n_a. Jeho dôvodom môže byť fakt, že pre zvolený počet ošetrení $g = 4$ má kovariančná matica vektora Z veľké množstvo parametrov na odhadovanie, čoho dôsledkom môže byť menšia presnosť výsledného odhadu. S rastúcim počtom dát sa avšak situácia zlepšuje, čo môže indikovať správnosť uvedených asymptotických vlastností a odvodených asymptotických odhadov. Čo sa týka sily rozšíreného testu, tam sú výsledky v princípe v poriadku. Test daný testovou štatistikou T^g_{cMAX} z (40) v dátach identifikuje signifikantné odlišnosti a nulovú hypotézu H_0 zamieta relatívne často. V porovnaní s výsledkami pre MANOVA test je ale navrhnutý test pre obe voľby rozdelení slabší. A to aj napriek tomu, že z dôvodu vyššej pravd. chyby 1. druhu môže byť sila nášho testu mierne zväčšená. Kvôli presnejšiemu porovnaniu sme sa preto v nasledujúcom scenári pozreli na prípad nižšieho počtu ošetrení a vyššieho počtu dát.

Zvoľme teraz $g = 3$, čo je prípad troch ošetrení X_a, $a = 1, \ldots, 3$. Uvažujme rovnaké dva typy rozdelení, ako v predchádzajúcom scenári, teda 6-rozmerné normálne rozdelenie a 6-rozmerné zovšeobecnené Studentovo t-rozdelenie s počtom stupňov voľnosti $df = 3$, pričom zachovajme označenia parametrov μ_a, Σ_a, $a = 1, \ldots, 3$. V prípade počítania odhadov pravd. chyby 1. druhu opäť zvolíme rovnaké nulové vektory stredných hodnôt $\mu_a = 0_6$, $a = 1, \ldots, 3$. Pri skúmaní sily testu sme stredné hodnoty zvolili nasledovne:

$$\mu_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \qquad \mu_2 = \begin{pmatrix} 1 \\ 0{,}5 \\ 0{,}25 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \qquad \mu_3 = \begin{pmatrix} -1{,}5 \\ -0{,}25 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Kovariančné (resp. škálovacie) matice Σ_a, $a = 1, \ldots, 3$, sme zvolili rovnako, ako v predchádzajúcom prípade pre $g = 4$. Pre každý z týchto dvoch prípadov rozdelení X_a, $a = 1, \ldots, 3$, sme v prostredí R [5] náhodne generovali 6-rozmerné dáta pre nasledovné tri voľby veľkostí ošetrení n_a:

- $n_1 = 100$, $n_2 = 125$, $n_3 = 140$,
- $n_1 = 200$, $n_2 = 250$, $n_3 = 280$ a
- $n_1 = 400$, $n_2 = 500$, $n_3 = 560$.

Na takto generované dáta sme opäť aplikovali test pomocou T^g_{cMAX} z (40), pričom sme testovali rovnako na hladine významnosti $\alpha = 0{,}05$ a vykonaných bolo 10 000 simulácií. Získané odhady pravdepodobností chyby 1. druhu a sily testu zaokrúhlené na 3 desatinné miesta sú vypísané v tabuľke 5. Pre porovnanie opäť uvádzame aj výsledky pre MANOVA test.

Tabuľka 5: Pravdepodobnosť chyby 1. druhu a sila T^g_{cMAX} testu, 3 ošetrenia, homoskedasticita

	Pravd. chyby 1. druhu			Sila testu		
	$n_a \sim 125$	$n_a \sim 250$	$n_a \sim 500$	$n_a \sim 125$	$n_a \sim 250$	$n_a \sim 500$
6-rozmerné normálne rozdelenie						
T^g_{cMAX}	0,057	0,054	0,051	0,203	0,376	0,719
MANOVA	0,046	0,051	0,048	0,434	0,800	0,991
6-rozmerné zovšeobecnené Studentovo t-rozdelenie						
T^g_{cMAX}	0,062	0,052	0,055	0,148	0,258	0,511
MANOVA	0,045	0,047	0,047	0,167	0,324	0,626

Získané odhady pravdepodobnosti chyby 1. druhu uvedené v tabuľke 5 sú v súlade s našim očakávaním. Navýšenie počtov dát n_a malo za následok, že výsledné odhady sú opäť o niečo bližšie k hodnote 5%. Pre obe voľby rozdelení a všetky počty dát sú tieto odhady pravd. chyby 1. druhu maximálne na úrovni 6%. Dá sa teda očakávať, že asymptoticky drží zovšeobecnený test daný testovou štatistikou T^g_{cMAX} z (40) pravd. chyby 1. druhu na požadovanej úrovni α. Odhadované sily tohto testu sú ale stále menšie, ako tie získané pre MANOVA test. Tento rozdiel je najvýraznejší pri normálnom rozdelení ošetrení. To je avšak očakávateľné, keďže pri zvolenom normálnom rozdelení sú z dôvodu voľby rovnakých kovariančných matíc (teda homoskedasticity) splnené všetky predpoklady MANOVA testu. Preto sme sa v nasledujúcom scenári pozreli na prípad heteroskedastických dát.

Ako posledný zvoľme prípad rozdelení z predchádzajúceho scenára pri narušenej homoskedasticite dát. Majme teda $g = 3$ ošetrenia X_a, $a = 1, \ldots, 3$, a naďalej uvažujme 6-rozmerné normálne rozdelenie a 6-rozmerné zovšeobecnené Studentovo t-rozdelenie s počtom stupňov voľnosti $df = 3$, pričom zachovajme označenia parametrov μ_a, Σ_a, $a = 1, \ldots, 3$. Vektory stredných hodnôt μ_a, $a = 1, \ldots, 3$, zvoľme aj za platnosti, aj za neplatnosti H_0 z (35) rovnako, ako v predchádzajúcom scenári. Z dôvodu uvažovania heteros-

kedasticity ale zmeníme voľby kovariančných (resp. škálovacích) matíc Σ_a, $a = 1, \ldots, 3$. Maticu Σ_1 zvolíme tak, ako v predchádzajúcich simuláciách. Ostatné matice Σ_2, Σ_3 zvolíme iné, konkrétne

$$\Sigma_2 = \begin{pmatrix} 36 & 10 & 4 & 0 & 0 & 0 \\ 10 & 4 & 1{,}5 & 0 & 0 & 0 \\ 4 & 1{,}5 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0{,}04 & 0{,}02 & 0 \\ 0 & 0 & 0 & 0{,}02 & 0{,}0225 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0{,}64 \end{pmatrix}, \Sigma_3 = \begin{pmatrix} 64 & 25 & 10 & 0 & 0 & 0 \\ 25 & 16 & 5 & 0 & 0 & 0 \\ 10 & 5 & 2{,}25 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0{,}09 & 0{,}05 & 0 \\ 0 & 0 & 0 & 0{,}05 & 0{,}04 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0{,}04 \end{pmatrix}.$$

Pre každý z týchto dvoch prípadov rozdelení X_a, $a = 1, \ldots, 3$, sme v prostredí R [5] náhodne generovali 6-rozmerné dáta pre rovnaké tri voľby veľkostí ošetrení n_a, $a = 1, \ldots, 3$, ako v predchádzajúcom scenári. Na takto generované dáta sme aplikovali test pomocou T^g_{cMAX} z (40). Testovali sme rovnako na hladine významnosti $\alpha = 0{,}05$ a vykonaných bolo 10 000 simulácií. Získané odhady pravdepodobností chyby 1. druhu a sily testu zaokrúhlené na 3 desatinné miesta sú vypísané v tabuľke 6. Pre porovnanie opäť uvádzame aj výsledky pre MANOVA test.

Tabuľka 6: Pravdepodobnosť chyby 1. druhu a sila T^g_{cMAX} testu, 3 ošetrenia, heteroskedasticita

	Pravd. chyby 1. druhu			Sila testu		
	$n_a \sim 125$	$n_a \sim 250$	$n_a \sim 500$	$n_a \sim 125$	$n_a \sim 250$	$n_a \sim 500$
6-rozmerné normálne rozdelenie						
T^g_{cMAX}	0,057	0,055	0,050	0,397	0,721	0,968
MANOVA	0,070	0,075	0,072	0,529	0,895	0,999
6-rozmerné zovšeobecnené Studentovo t-rozdelenie						
T^g_{cMAX}	0,058	0,054	0,058	0,263	0,515	0,870
MANOVA	0,063	0,065	0,064	0,215	0,401	0,750

Porovnaním získaných výsledkov s výsledkami uvedenými v tabuľke 5 si môžeme všimnúť, že zmena kovariančných matíc neovplyvnila pravd. chyby 1. druhu nášho testu. V prípade MANOVA testu mala avšak táto zmena z dôvodu porušenia predpokladov za následok ich výraznejší nárast. Čo sa týka sily testu pomocou T^g_{cMAX} z (40), tak mal

prechod k heteroskedasticite dokonca priaznivé účinky. V porovnaní s výsledkami pre homoskedasticitu uvedenými v tabuľke 5 sila testu výrazne vzrástla. V prípade normálneho rozdelenia sú odhady sily MANOVA testu stále o niečo vyššie, treba avšak prihliadať na výrazný nárast pravd. chyby 1. druhu tohto testu. Pre prípad t-rozdelenia sú odhady sily testu pomocou T^g_{cMAX} už pre všetky voľby počtov dát n_a vyššie ako pre MANOVA test, a to aj napriek zvýšenej pravd. chyby 1. druhu MANOVA testu. Podobne ako test pomocou T_{cMAX} z (31) pre prípad dvoch ošetrení, aj pri použití T^g_{cMAX} podľa (40) je výsledkom robustná neparametrická metóda, pre ktorú sa dajú očakávať výborné výsledky pre heteroskedastické dáta. Preto sa testová štatistika T^g_{cMAX} z (40) javí byť jej vhodným zovšeobecnením.

Záver

Cieľom tejto diplomovej práce bolo nadviazať na vývoj rankových testov z podkladových článkov, navrhnúť zovšeobecnenie najaktuálnejšej z týchto metód pre prípad ľubovoľného počtu súborov a po odvodení asymptotických vlastností a odhadov skúmať kvalitu navrhnutého testu pomocou počítačových simulácií v prostredí R [5]. Poznatky týkajúce sa sledovaných viacrozmerných metód na porovnanie dvoch súborov sme čerpali najmä z pôvodných článkov [1], [3], [4] a [6].

V úvode prvej kapitoly sme stručne opísali problematiku klinických štúdií, čím sme poukázali na dôležitosť viacrozmerného testovania. Po zadefinovaní zovšeobecnenej Behrens-Fisher hypotézy sme vysvetlili jej prepojenie so spomínanými klinickými štúdiami. Uviedli sme pomocnú lemu a vetu s vlastnými dôkazmi a na základe ich výsledkov sme vysvetlili, ako možno túto nulovú hypotézu testovať za pomoci rankových testov.

Ako prvým sme sa venovali tzv. rank-sum-type testom. Vysvetlili sme, ako môže mať použitie testových štatistík O'Briena z [4] na testovanie zovšeobecnenej Behrens-Fisher hypotézy za následok zväčšenie pravdepodobnosti chyby 1. druhu príslušných testov. Uviedli sme úpravy týchto testových štatistík navrhnuté Huangom et al. v [1], pri ktorých majú výsledné testy aj pre prípad tejto nulovej hypotézy pravd. chyby 1. druhu pod kontrolou. Tento fakt sme potvrdili na základe výsledkov vlastných simulácií. Vysvetlili sme závislosť sily týchto testov od charakteru neplatnosti zovšeobecnenej Behrens-Fisher hypotézy. V prípade rovnakých znamienok relatívnych efektov θ_k testy správne kumulujú dôkazy o neplatnosti nulovej hypotézy, čo má za následok veľkú silu testu. Inak sa ale z dôvodu sčitovania hodnôt s opačnými znamienkami, teda spájania dôkazov nerovnakého typu, môže stratiť veľké množstvo informácie a výsledné testy môžu byť veľmi slabé. Tento jav sme pre oba prípady ilustrovali vo vlastných simuláciách.

V ďalšej časti sme sa v krátkosti zaoberali T_{max} testom, ktorý navrhli Liu et. v [3]. Na základe výsledkov vlastných simulácií sme ukázali, že táto voľba jednoduchej, no robustnej testovej štatistiky má za následok uspokojivú silu testu pre ľubovoľný charakter neplatnosti nulovej hypotézy.

V ďalšom sme spomenuli, že v mnohých aplikáciách tvoria endpointy prirodzenú clusterovú štruktúru. Vysvetlili sme, ako bol tento fakt využití pri konštrukcii testovej štatistiky T_{cMAX}, ktorú navrhli Zhang et al. v [6]. Naznačili sme, že tento test vhodne kombinuje

silu rank-sum-type testov a robustnosť testovej štatistiky T_{max}, a teda že preň možno v prípade clusterovej štruktúry endpointov očakávať veľkú silu testu. Tieto očakávania potvrdili výsledky našich simulácií, kde sme na základe porovnania odhadov sily všetkých sledovaných rankových testov demonštrovali výhodnosť použitia tejto metódy na testovanie zovšeobecnenej Behrens-Fisher hypotézy.

V rámci druhej kapitoly sme navrhli vlastné zovšeobecnenie testovej štatistiky T_{cMAX} pre prípad viacerých súborov. Vysvetlili sme, prečo sme sa rozhodli túto testovú štatistiku zovšeobecniť iným spôsobom, ako tým, ktorý navrhli samotní autori v článku [6]. Uviedli sme definíciu navrhnutej testovej štatistiky T^g_{cMAX} a vyjadrili sme sa k očakávaným vlastnostiam výsledného testu. Čiastočne sme odôvodnili asymptotickú normalitu pomocného vektora Z. Vychádzajúc z článkov [1], [3] a [6] sme za platnosti zovšeobecnenej Behrens-Fisher hypotézy detailne odvodili jeho strednú hodnotu a asymptotickú kovariančnú maticu, uviedli sme odhady potrebných parametrov a taktiež sme opísali praktický spôsob výpočtu p-value nášho testu. Využitím týchto poznatkov sme simulačne skúmali kvalitu navrhnutého zovšeobecnenia T^g_{cMAX} v porovnaní so známou parametrickou metódou - jednofaktorovým MANOVA testom. V prostredí R [5] sme generovali dáta s clusterovou štruktúrou endpointov pre 3 rôzne scenáre. Získané odhady pravd. chyby 1. druhu nášho testu sú s rastúcim počtom dát stále bližšie k požadovanej úrovni α, čo naznačuje správnosť odvodených asymptotických vlastností a odhadov. V prípade neplatnosti nulovej hypotézy ju navrhnutý test zamietal relatívne často, čoho dôkazom sú všeobecne uspokojivé odhady sily testu. Najlepšie výsledky náš test vykazoval pre prípad heteroskedastických dát z rozdelenia s ťažkými chvostami, kde dosiahol vyššiu silu testu ako MANOVA test a to pri nižších úrovniach pravd. chyby 1. druhu. Celkovo sa teda navrhnutý test daný testovou štatistikou T^g_{cMAX} javí byť výhodnou neparametrickou metódou na testovanie zovšeobecnenej Behrens-Fisher hypotézy pre prípad viacerých súborov.

Zoznam použitej literatúry

[1] Huang, P., Tilley, B.C., Woolson, R.F., Lipsitz, S.: *Adjusting O'Brien's test to control type I error for the generalized nonparametric Behrens–Fisher problem*, Biometrics 61 (2005), 532-539

[2] Kruskal, W.H., Wallis, W.A.: *Use of ranks in one-criterion variance analysis*, Journal of the American statistical Association 47 (1952), 583-621

[3] Liu, A., Li, Q., Liu, C., Yu, K., Yu, K.F.: *A Rank-Based Test for Comparison of Multidimensional Outcomes*, Journal of the American Statistical Association 105 (2010), 578-587

[4] O'Brien, P.C.: *Procedures for comparing samples with multiple endpoints*, Biometrics (1984), 1079-1087

[5] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[6] Zhang, W., Liu, A., Tang, L.L., Li, Q.: *A cluster-adjusted rank-based test for a clinical trial concerning multiple endpoints with application to dietary intervention assessment*, Biometrics 75 (2019), 821–830