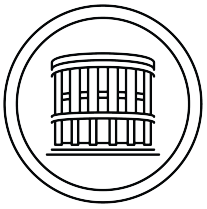
UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY



# VIACROZMERNÉ NEPARAMETRICKÉ TESTY POLOHY ZOHĽADŇUJÚCE CLUSTEROVANIE

DIPLOMOVÁ PRÁCA

2024 Bc. Marek ŠPITALSKÝ

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

# VIACROZMERNÉ NEPARAMETRICKÉ TESTY POLOHY 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 2024 **Bc. Marek ŠPITALSKÝ**






Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

# ZADANIE ZÁVEREČNEJ PRÁCE

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

**Názov:** Viacrozmerné neparametrické testy polohy zohľadňujúce clusterovanie
*Multivariate Non-Parametric Cluster-Adjusted Location Tests*

**Anotácia:** Trochu zjednodušene povedané, ide o pomerne 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 korelácie. Pri skúmaní vlastností týchto testov z teoretickej i praktickej stránky sa ukazuje byť dôležité dbať na splenie istých predpokladov o povahe vstupných dát.

**Vedúci:** Mgr. Ján Somorčík, PhD.
**Katedra:** FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
**Vedúci katedry:** doc. Mgr. Igor Melicherčík, PhD.

**Dátum zadania:** 10.01.2023

**Dátum schválenia:** 11.01.2023 prof. RNDr. Daniel Ševčovič, DrSc.
garant študijného programu

.....
študent

.....
vedúci práce

**Poďakovanie** Na tomto mieste by som sa chcel poďakovať svojmu vedúcemu diplomovej práce Mgr. Jánovi Somorčíkovi, PhD. za odborné rady, ochotu, čas a podnetné pripomienky, ktoré mi veľmi pomohli pri písaní tejto práce. Takisto by som sa chcel poďakovať môjmu otcovi Vladimírovi Špitalskému za jeho cenné rady k matematike v práci.

# Abstrakt v štátnom jazyku

ŠPITALSKÝ, Marek: Viacrozmerné neparametrické testy polohy zohľadňujúce clusterovanie [Diplomová práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: Mgr. Ján Somorčík, PhD., Bratislava, 2024, 106s.

V diplomovej práci sa zaoberáme nedávno publikovaným viacrozmerným neparametrickým testom využívajúcim clusterovú štruktúru čŕt dát s označením $T_{cMAX}$. Uvádzame stručný úvod do zovšeobecneného Behrensovho-Fisherovho problému a teórie $U$-štatistík. Poukazujeme na chybu pri dôkaze vety o asymptotickom rozdelení testovej štatistiky $T_{cMAX}$ testu, ktorá v niektorých prípadoch spôsobuje indefinitnosť odhadu kovariančnej matice. Túto chybu v práci opravujeme a predstavujeme niekoľko návrhov na novú verziu $T_{cMAX}$ testu. Následne simulačne porovnávame sily týchto verzií $T_{cMAX}$ testu spolu s pravdepodobnosťami chýb prvého druhu. Ďalej sa zameriavame na všeobecné rozdelenia dát, teda aj na prípady, kedy majú dáta diskrétne rozdelenia s prekrývajúcimi sa supportami. Nakoniec predstavíme, ako sa zmení asymptotická kovariančná matica $T_{cMAX}$ testu a jej odhad kvôli tomuto zovšeobecneniu.

**Kľúčové slová:** Viacrozmerné neparametrické testy, Clusterová štruktúra čŕt dát, Zovšeobecnený Behrensov-Fisherov problém, $U$-štatistiky

# Abstract

ŠPITALSKÝ, Marek: Multivariate Non-Parametric Cluster-Adjusted Location 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, 2024, 106p.

In the master thesis we discuss recently published multivariate non-parametric test that considers also the cluster structure of endpoints. This test is known as the $T_{cMAX}$ test. We present a brief introduction to the generalized Behrens-Fisher problem and theory of $U$-statistics. We point out a flaw in the proof of the theorem concerning the asymptotic distribution of the $T_{cMAX}$ test statistics, which causes the proposed covariance matrix to be indefinite in some cases. We fix this issue and present several candidates for an improved $T_{cMAX}$ test. Then we compare by simulations powers of these tests and their probabilities of type I error. Furthermore, we focus on general type of underlying distributions, even on discrete distributions with overlapping supports. At the end we give formulas for the asymptotic covariance matrix of the $T_{cMAX}$ test and its estimators that are valid in this general case.

**Keywords:** Multivariate non-parametric tests, Cluster structure of data, Generalized Behrens-Fisher problem, $U$-statistics

# Obsah

**Úvod** **9**

**1 $T_{cMAX}$ test** **11**

1.1 Zavedenie označenia a predpokladov . . . . . 11

1.2 História . . . . . 12

1.2.1 Rank-sum-type test - O'Brien (1984) . . . . . 12

1.2.2 Rank-sum-type test - Huang et al. (2005) . . . . . 13

1.2.3 $T_{MAX}$ test (2005) . . . . . 13

1.2.4 $T_{cMAX}$ test (2019) . . . . . 14

1.3 Pomocné tvrdenia pre $T_{cMAX}$ test . . . . . 15

1.4 $U$-štatistiky . . . . . 16

1.5 Veta 1.6 [13, Theorem 1] . . . . . 18

1.6 Veta 1.7 [13, Theorem 2] . . . . . 22

**2 Oprava $T_{cMAX}$ testu: spojitý prípad** **25**

2.1 Problém $T_{cMAX}$ testu - numerická chyba vo vete 1.6 [13, Theorem 1] . . . . 28

2.2 Problém $T_{cMAX}$ testu - chyba pri výpočte kovariancie vo vete 1.6 [13, Theorem 1] . . . . . 28

2.3 Kontrapríklad k dôkazu vety 1.6 [13, Theorem 1] . . . . . 29

2.4 Oprava vety 1.6 [13, Theorem 1] . . . . . 33

2.5 Oprava vety 1.7 [13, Theorem 2] . . . . . 34

2.6 Oprava vety 1.7 [13, Theorem 2] - neasymptotický odhad . . . . . 35

2.7 Oprava vety 1.6 [13, Theorem 1] - finálna verzia; dôkaz vety 2.1 . . . . . 36

2.8 Oprava vety 1.7 [13, Theorem 2] - finálna verzia . . . . . 40

2.9 Dosadzovanie a odhadovanie $\theta$ . . . . . 40

2.10 Simulačné štúdie - spojité dáta . . . . . 42

**3 Oprava $T_{cMAX}$ testu: všeobecný prípad** **48**

3.1 Všeobecný prípad rozdelení - správne vyjadrenie kovariancií z vety 1.6 [13, Theorem 1] . . . . . 50

3.2 Kontrapríklad k vete 1.6 [13, Theorem 1] pre všeobecný prípad . . . . . 51

3.3 Všeobecná verzia $T_{cMAX}$ testu; dôkaz vety 3.1 . . . . . 54
3.4 Nekonzistentnosť odhadov z vety 1.7 [13, Theorem 2] pre všeobecný prípad 57
3.5 Kontrapríklad k vete 1.7 [13, Theorem 2] pre všeobecný prípad . . . . . 58
3.6 Dôkaz vety 3.2 . . . . . 58
3.6.1 Asymptotická nevychýlenosť . . . . . 59
3.6.2 Limitne nulová variancia . . . . . 62
3.6.3 Záverečná časť dôkazu vety 3.2 . . . . . 64
3.7 Vyšetrenie nekonzistentnosti odhadov z vety 1.7 [13, Theorem 2] v prípade nesplnenia predpokladu (1.1) . . . . . 64
3.8 Kontrapríklad k vete 1.7 [13, Theorem 2] - dokončenie kapitoly 3.5 . . . . . 67
3.9 Simulačné štúdie - diskrétne dáta . . . . . 69

**4 Príloha A: Súvis rozdielu priemerných midrankov a priemeru rozdielov indikátorov 76**

**5 Príloha B: Výsledky simulácií z kapitoly 2.10 80**

**6 Príloha C: Pomocné tvrdenia o podmienenej strednej hodnote ku kapitole 2.7 84**

**7 Príloha D: Pomocné tvrdenia pre dôkaz konzistentnosti odhadu kovariančnej matice z kapitoly 3.6 90**

**8 Príloha E: Výsledky simulácií z kapitoly 3.9 92**

**9 Príloha F: Porovnanie výsledkov simulácií s výsledkami uvedenými v článku [13] 97**
9.1 Porovnanie výsledkov simulácií s výsledkami uvedenými v článku [13] pre spojité rozdelenia . . . . . 97
9.2 Porovnanie výsledkov simulácií s výsledkami uvedenými v článku [13] pre diskrétne rozdelenia . . . . . 100

**Záver 103**

**Zoznam použitej literatúry 105**

# Úvod

Úloha štatistiky v zdravotníctve je veľmi významná. V ľubovoľnej klinickej skúške je jej použitie nutné pri vytváraní záverov z nazbieraných dát. Častokrát sa však stáva, že klasické štatistické testy, ako napríklad Studentov t-test, nemôžeme použiť, keďže dáta nie sú normálne rozdelené. Preto sa v tomto odvetví do veľkej miery využívajú neparametrické testy, ktoré nekladú žiadne podmienky na pravdepodobnostné rozdelenia dát. Často sa v týchto dátach objavujú skupiny čŕt, ktoré medzi sebou veľmi úzko súvisia. Príkladom môže byť krvný tlak a hladina cholesterolu v krvi. Tieto skupiny čŕt nám v dátach vytvárajú akúsi clusterovú štruktúru čŕt, ktorá, keď je zobraná do úvahy, napomáha pri analýzach týchto dát.

V tejto práci sa budeme venovať istej triede neparametrických viacrozmerných 2-súborových testov polohy založených na rankovaní každej súradnice dát. Ide o testy, ktoré porovnávajú rozdelenia dvoch náhodných výberov na základe toho, ako sú voči sebe „umiestnené“. Čo sa historického vývoja týchto testov týka, tak prvý takýto test (tzv. rank-sum test) publikoval v roku 1984 O'Brien v článku [6]. Jeho vylepšenie bolo navrhnuté v roku 2005 v článku [3]. Ďalší z tejto skupiny testov bol publikovaný v roku 2010 v článku [5] pod označením $T_{MAX}$. My sa hlavne budeme zaoberať testom z roku 2019 z článku [13], ktorý je nadstavbou $T_{MAX}$ testu. Tento test oproti $T_{MAX}$ testu zohľadňuje aj spomínanú clusterovú štruktúru čŕt dát, podľa čoho ho autori označili ako $T_{cMAX}$ test.

Na začiatku diplomovej práce sa budeme venovať úvodu do problematiky neparametrických testov polohy založených na rankovaní v súradniciach a priblížime spomínané testy po matematickej stránke. Ďalej uvedieme pomocné lemy a časť teórie k $U$-štatistikám, ktoré sú využívané pri dôkaze vety [13, Theorem 1]. Ďalej uvedieme znenie hlavných viet článku [13, Theorem 1 a Theorem 2]. V druhej kapitole sa budeme venovať tomu, prečo kovariančné matice v originálnych zneniach viet [13, Theorem 1 a Theorem 2] vychádzajú niekedy indefinitné. Ukážeme, ktorý predpoklad v dôkaze vete [13, Theorem 1] nebol správny a uvedieme k tejto vete kontrapríklad. V ďalšej časti tejto kapitoly budeme prezentovať opravy tejto vety a následne v simulačných štúdiách porovnáme pravdepodobnosti chýb prvého druhu a sily navrhovaných testy. V tretej kapitole sa budeme venovať $T_{cMAX}$ testu za všeobecných predpokladov, zahŕňajúcich prípad, keď sú rozdelenia náhodných

vektorov diskrétne s prekrývajúcimi sa supportami. Ukážeme, ako sa zmení asymptotická kovariančná matica vo vete [13, Theorem 1]. Ďalej sa pozrieme na odhad tejto matice z vety [13, Theorem 2] a uvedieme, prečo za všeobecných predpokladov nie je konzistentný. Následne navrhneme aj korekciu tohto odhadu, ktorá už konzistentná bude. Nakoniec v simulačných štúdiách budeme porovnávať pravdepodobnosť chyby prvého druhu a silu jednotlivých testov pri diskrétnych rozdeleniach.

# 1 $T_{cMAX}$ test

V tejto kapitole popíšeme neparametrický test polohy $T_{cMAX}$, ktorý bol navrhnutý v článku [13].

## 1.1 Zavedenie označenia a predpokladov

Nech $X, Y$ sú nezávislé $m$-rozmerné náhodné vektory, $X$ má združenú distribučnú funkciu $F$ a $Y$ má združenú distribučnú funkciu $G$. Predpokladajme ďalej, že zložky vektorov $X$ a $Y$ sa dajú prirodzene rozdeliť do $J$ clusterov s veľkosťami $m_1, \ldots, m_J$, pričom všetky $m_j$ sú kladné a $\sum_{j=1}^{J} m_j = m$. Tejto clusterovej štruktúre prispôsobíme aj označenie zložiek vektorov. Označíme

$$X = (X_1^T, X_2^T, \ldots, X_J^T)^T \quad \text{pričom } X_j = (X_{j1}, X_{j2}, \ldots, X_{jm_j})^T$$

pre $j = 1, \ldots J$. Analogicky

$$Y = (Y_1^T, Y_2^T, \ldots, Y_J^T)^T \quad \text{pričom } Y_j = (Y_{j1}, Y_{j2}, \ldots, Y_{jm_j})^T.$$

Ďalej na distribúcie $F$, $G$ kladieme predpoklad, že pre každé $j = 1, \ldots, J$, $k = 1, \ldots, m_j$ platí[1]

$$P(X_{jk} = Y_{jk}) = 0. \tag{1.1}$$

Pre každú zložku vektorov definujeme takzvaný *relatívny efekt* pre $k-$tu zložku $j$-teho clustera nasledovne

$$\theta_{jk} = P(X_{jk} < Y_{jk}) - P(X_{jk} > Y_{jk}), \quad j = 1, \ldots, J, \ k = 1, \ldots, m_j. \tag{1.2}$$

Hypotéza, ktorú testujeme, je takzvaná *zovšeobecnená Behrensova-Fisherova hypotéza*:

$$H_0 \colon \ \theta_{jk} = 0 \quad \text{pre každé } j = 1, \ldots, J, \ k = 1, \ldots, m_j. \tag{1.3}$$

Majme dva nezávislé náhodné výbery

$$X_i \ (i = 1, \ldots, n_x) \quad \text{a} \quad Y_\ell \ (\ell = 1, \ldots, n_y)$$

z rozdelení $F$ a $G$. Označme $N = n_x + n_y$. Nech $x_i$ $(i = 1, \ldots, n_x)$ a $y_\ell$ $(\ell = 1, \ldots, n_y)$ sú realizácie týchto náhodných výberov.

[1] Poznamenajme, že v článku [13] túto podmienku autori explicitne neuvádzajú, avšak ju využívajú.

Nakoľko sa zaoberáme neparametrickými testami, tak zavedieme ranky na náhodných výberoch z $F$ a $G$. Zoberme si $j-$ty cluster a $k-$tu zložku v tomto clusteri. Spojíme hodnoty tejto zložky z oboch realizácií $x_{1jk}, x_{2jk}, \ldots, x_{n_x jk}$ a $y_{1jk}, y_{2jk}, \ldots, y_{n_y jk}$ do jedného vektora. Následne v tomto vektore spočítame midranky[2]; nech $R_{xijk}$ je midrank $x_{ijk}$ a $R_{y\ell jk}$ je midrank $y_{\ell jk}$.

Ďalej uvedieme označenia týkajúce sa rankov, ktoré sa budú vyskytovať v rámci práce:

- $R_{xi\cdot\cdot} = \sum_{j=1}^{J} \sum_{k=1}^{m_j} R_{xijk}$, $R_{y\ell\cdot\cdot} = \sum_{j=1}^{J} \sum_{k=1}^{m_j} R_{y\ell jk}$
- $\bar{R}_x = \frac{1}{n_x} \sum_{i=1}^{n_x} R_{xi\cdot\cdot}$, $\bar{R}_y = \frac{1}{n_y} \sum_{\ell=1}^{n_y} R_{y\ell\cdot\cdot}$
- $\bar{R}_{x\cdot jk} = \frac{1}{n_x} \sum_{i=1}^{n_x} R_{xijk}$, $\bar{R}_{y\cdot jk} = \frac{1}{n_y} \sum_{\ell=1}^{n_y} R_{y\ell jk}$
- $\bar{R}_{x\cdot j\cdot} = \frac{1}{n_x m_j} \sum_{i=1}^{n_x} \sum_{k=1}^{m_j} R_{xijk}$, $\bar{R}_{y\cdot j\cdot} = \frac{1}{n_y m_j} \sum_{\ell=1}^{n_y} \sum_{k=1}^{m_j} R_{y\ell jk}$

## 1.2 História

V tejto podkapitole sa budeme venovať stručnej histórii testovania zovšeobecnenej Behrensovej-Fisherovej hypotézy (1.3). Preberieme ju z článku [13, str. 823].

### 1.2.1 Rank-sum-type test - O'Brien (1984)

V roku 1984 Peter C. O'Brien navrhol vo svojom článku [6] dva neparametrické testy, pričom aplikoval klasický dvojvýberový $t-$test na zosumované orankované pozorovania $\{R_{x1\cdot\cdot}, \ldots, R_{xn_x\cdot\cdot}, R_{y1\cdot\cdot}, \ldots, R_{yn_y\cdot\cdot}\}$. Predpisy testovacích štatistík týchto testov sú nasledovné

$$T_1 = \frac{\bar{R}_y - \bar{R}_x}{\sqrt{\frac{1}{N-2}\left(\frac{1}{n_x} + \frac{1}{n_y}\right)\left\{(n_x - 1)\,\hat{\sigma}_x^2 + (n_y - 1)\,\hat{\sigma}_y^2\right\}}}$$

$$T_2 = \frac{\bar{R}_y - \bar{R}_x}{\sqrt{\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}}},$$

[2]Midranky sú definované ako klasické ranky v prípade, že všetky prvky vektora sú navzájom rôzne. V prípade zhody sa každému z týchto pozorovaní priradí ich priemerný rank. Napríklad ak máme vektor pozorovaní $(5, 9, 13, 13, 44)$ tak midranky tejto množiny by boli $(1, 2, 3.5, 3.5, 5)$.

kde $\sigma_x^2 = \frac{1}{n_x-1}\sum_{i=1}^{n_x}\left(R_{xi\cdot\cdot} - \bar{R}_x\right)^2$ a $\sigma_y^2 = \frac{1}{n_y-1}\sum_{i=1}^{n_y}\left(R_{yi\cdot\cdot} - \bar{R}_y\right)^2$. Testové štatistiky $T_1$ a $T_2$, za platnosti (klasickej) Behrensovej-Fisherovej hypotézy $H_0 : F = G$ majú asymptoticky štandardizované normálne rozdelenie. Pre prípady, kedy $F \neq G$[3], je variancia týchto testových štatistík nekonštantná, čo má za následok zvýšenie pravdepodobnosti chyby prvého druhu. Teda tieto testy sa ukazajú často ako nevhodné pri testovaní zovšeobecnenej Behrensovej-Fisherovej hypotézy.

### 1.2.2 Rank-sum-type test - Huang et al. (2005)

Tomuto problému O'Brienovych testov sa venovali Huang a kolektív v článku [3], kde navrhli modifikáciu O'Brienovych rank-sum testov. Do menovateľov štatistík $T_1$ a $T_2$ pridali špeciálne členy závislé od dát $\hat{h}_1$ a $\hat{h}_2$, ktoré zaručili asymptotickú normalitu testových štatistík aj za platnosti zovšeobecnenej Behrensovej-Fisherovej hypotézy:

$$T_{h1} = \frac{\bar{R}_y - \bar{R}_x}{\sqrt{\hat{h}_1 \frac{1}{N-2}\left(\frac{1}{n_x}+\frac{1}{n_y}\right)\left\{(n_x-1)\,\hat{\sigma}_x^2 + (n_y-1)\,\hat{\sigma}_y^2\right\}}}$$

$$T_{h2} = \frac{\bar{R}_y - \bar{R}_x}{\sqrt{\hat{h}_2\left\{\frac{\hat{\sigma}_x^2}{n_x}+\frac{\hat{\sigma}_y^2}{n_y}\right\}}}$$

Predpis parametrov $\hat{h}_1$ a $\hat{h}_2$ je uvedený v článku [3, str. 534].

### 1.2.3 $T_{MAX}$ test (2005)

Rank-sum testy sa ukázali ako dostatočne silné v prípadoch, kedy všetky relatívne efekty $\theta_{jk}$ z (1.2) mali rovnaké znamienko. Problém nastával v prípade, že niektoré relatívne efekty $\theta_{jk}$ boli kladné a niektoré boli záporné. Toto zapríčiňovalo pokles absolútnej hodnoty $\bar{R}_y - \bar{R}_x$, čo malo za následok pokles sily tohto testu. Na toto sa zamerali autori článku [5], kde navrhli nový test, ktorého testová štatistika je nasledovná

$$T_{MAX} = \max_{\substack{j\in\{1,\ldots,J\}\\ k\in\{1,\ldots,m_j\}}} \left|\bar{R}_{y\cdot jk} - \bar{R}_{x\cdot jk}\right|. \tag{1.4}$$

Táto testová štatistika sa za platnosti zovšeobecnenej Behrensovej-Fisherovej hypotézy asymptoticky riadi normálnym rozdelením s parametrami, ktoré je možné nájsť v [5,

[3]Zovšeobecnená Behrensova-Fisherova hypotéza (1.3) však môže stále platiť.

str. 580]. Takouto úpravou autori dosiahli zlepšenie sily testu oproti rank-sum testom v prípade, že znamienka relatívnych efektov sa líšia. Avšak v prípade, kedy všetky znamienka sú rovnaké, sa ukázalo, že môže dôjsť k poklesu sily tohto testu oproti spomínaným rank-sum testom.

### 1.2.4 $T_{cMAX}$ test (2019)

V tejto kapitole sa pozrieme na $T_{cMAX}$ test, ktorý navrhli autori článku [13]. Tomuto testu sa budeme venovať aj vo zvyšku práce.

Najprv sa ale vrátime ku clusterom, ktoré sme zaviedli na začiatku kapitoly 1.1. Častokrát sa stretávame v aplikáciách s prípadmi, kedy jednotlivé črty (premenné) dát spolu súvisia. Toto môže viesť k akejsi prirodzenej clusterovej štruktúre týchto čŕt. Autori článku [13] predpokladajú, že črty sa budú správať podobnejšie v rámci jedného clustera ako naprieč clusterami. Využitie tohto predpokladu by teda mohlo pomôcť pri testovaní (1.3).

Vráťme sa k samotnému $T_{cMAX}$ testu. Začneme tým, že máme danú takúto prirodzenú clusterovú štruktúru jednotlivých premenných. Narozdiel od $T_{MAX}$ testu, kde počítame priemerné ranky v každej premennej, autori článku [13] pracovali s priemernými rankami v rámci týchto clusterov, teda počítali $\bar{R}_{x\cdot j\cdot}, \bar{R}_{y\cdot j\cdot}$ pre každý cluster. Očakávali, že rozdiely relatívnych efektov $\theta_{jk}$ budú pre jednotlivé premenné v rámci clusterov podobné, čo viedlo k predpokladu väčšej sily pri testovaní zovšeobecnenej Behrensovej-Fisherovej hypotézy.

Testovú štatistiku $T_{cMAX}$ testu navrhli autori [13] nasledovne

$$T_{cMAX} = \max_{j\in\{1,\dots,J\}} \left\{ \left| \frac{\bar{R}_{y\cdot j\cdot} - \bar{R}_{x\cdot j\cdot}}{\sqrt{\widehat{\text{var}}\left(\bar{R}_{y\cdot j\cdot} - \bar{R}_{x\cdot j\cdot}\right)}} \right| \right\}, \tag{1.5}$$

kde $\widehat{\text{var}}\left(\bar{R}_{y\cdot j\cdot} - \bar{R}_{x\cdot j\cdot}\right)$ je odhad variancie $\bar{R}_{y\cdot j\cdot} - \bar{R}_{x\cdot j\cdot}$.

Môže byť zaujímavé podotknúť, že tento test je priamym zovšeobecnením rank-sum testu z roku 2005 a aj $T_{MAX}$ testu. V prípade, že máme iba jeden cluster, tak $T_{cMAX}$ test je ekvivalentný s rank-sum testom navrhnutým Huangom a kolektívom. V prípade, že všetky clustery majú len jednu zložku, tak $T_{cMAX}$ je ekvivalentný s $T_{MAX}$ testom[4].

[4]Toto uvádzajú autori v článku [13, str. 824]. Avšak test $T_{MAX}$ (1.4) hľadá maximum z neštandardizovaných rozdielov priemerných rankov jednotlivých čŕt.

## 1.3 Pomocné tvrdenia pre $T_{cMAX}$ test

Predtým, ako budeme vyšetrovať rozdelenie testovej štatistiky $T_{cMAX}$ testu, uvedieme niekoľko pomocných tvrdení. Pre $i = 1, \ldots, n_x$, $\ell = 1, \ldots, n_y$, $j = 1, \ldots, J$, $k = 1, \ldots, m_j$ označme dva odhady relatívneho efektu $\theta_{jk}$ (viď (1.2)):

$$\Psi_{i\ell jk} = I(X_{ijk} < Y_{\ell jk}) - I(X_{ijk} > Y_{\ell jk}), \tag{1.6}$$

$$\hat{\theta}_{jk} = \frac{1}{n_x n_y} \sum_{i=1}^{n_x} \sum_{\ell=1}^{n_y} I(X_{ijk} < Y_{\ell jk}) - I(X_{ijk} > Y_{\ell jk}) = \frac{1}{n_x n_y} \sum_{i=1}^{n_x} \sum_{\ell=1}^{n_y} \Psi_{i\ell jk}. \tag{1.7}$$

Začneme tvrdením, kde o týchto odhadoch ukážeme, že sú nevychýlené.

**Veta 1.1.** *Pre každé $i, \ell, j, k$, premenné $\Psi_{i\ell jk}$ a $\hat{\theta}_{jk}$ sú nevychýlenými odhadmi $\theta_{jk}$.*

*Dôkaz.* Začneme odhadom $\Psi_{i\ell jk}$. Platí nasledovná séria rovností:

$$\begin{aligned}
E\left(\Psi_{i\ell jk}\right) &= E\left[I(X_{ijk} < Y_{\ell jk}) - I(X_{ijk} > Y_{\ell jk})\right] \\
&= E\left[I(X_{ijk} < Y_{\ell jk})\right] - E\left[I(X_{ijk} > Y_{\ell jk})\right] \\
&= \left[1 \cdot P(X_{jk} < Y_{jk}) + 0 \cdot P(X_{jk} \geq Y_{jk})\right] - \left[1 \cdot P(X_{jk} > Y_{jk}) + 0 \cdot P(X_{jk} \leq Y_{jk})\right] \\
&= P(X_{ijk} < Y_{\ell jk}) - P(X_{ijk} > Y_{\ell jk}) \\
&= \theta_{jk}.
\end{aligned}$$

Ďalej ukážeme nevychýlenosť $\hat{\theta}_{jk}$:

$$\begin{aligned}
E\left(\hat{\theta}_{jk}\right) &= E\left[\frac{1}{n_x n_y} \sum_{i=1}^{n_x} \sum_{\ell=1}^{n_y} \Psi_{i\ell jk}\right] \\
&= \frac{1}{n_x n_y} \sum_{i=1}^{n_x} \sum_{\ell=1}^{n_y} E\left[\Psi_{i\ell jk}\right].
\end{aligned}$$

Z predchádzajúcej časti vieme, že $E\left[\Psi_{i\ell jk}\right] = \theta_{jk}$. Máme teda

$$E\left(\hat{\theta}_{jk}\right) = \frac{1}{n_x n_y} \sum_{i=1}^{n_x} \sum_{\ell=1}^{n_y} \theta_{jk} = \theta_{jk}.$$

□

Pokračujeme tvrdením, ktoré dáva do súvisu rozdiel priemerných rankov $k-$tej zložky $j$-teho clustera s odhadom relatívneho efektu $\hat{\theta}_{jk}$. Táto veta vyplýva z vety 4.1 v kapitole 4.

**Veta 1.2** ([14, str. 2])**.** *Nech $\hat{\theta}_{jk}$ je definované vzťahom* (1.7)*. Potom platí*

$$\hat{\theta}_{jk} = \frac{2}{N}\left(\bar{R}_{y \cdot jk} - \bar{R}_{x \cdot jk}\right).$$

## 1.4 $U$-štatistiky

Podstatný pojem, ktorý vystupuje v dôkaze hlavnej vety článku [13], je $U$-štatistika. Tento pojem zaviedol Wassily Hoeffding v roku 1948 v článku [2]. Avšak Hoeffding sa zaoberal $U-$štatistikami, ktoré sú skonštruované len z jedného náhodného výberu. V článku [8] Nariaki Sugiura tento pojem zovšeobecnil a ukázal, že podobné vlastnosti majú aj $U-$štatistiky skonštruované z viacerých náhodných výberov. Ďalej uvedieme mierne zjednodušenú verziu definície z článku [8], nakoľko my sa v práci zaoberáme len dvoma náhodnými výbermi. Označenie bude oproti pôvodnej definícii tiež pozmenené, kvôli konzistentnosti so zvyškom práce.

**Definícia 1.3** ([8, str. 386]). *Majme dva $m-$rozmerné náhodné výbery $X_i$ $(i = 1, \ldots, n_x)$ a $Y_\ell$ $(\ell = 1, \ldots, n_y)$. Nech $\omega_x, \omega_y$ sú prirodzené čísla také, že $\omega_x \leq n_x$ a $\omega_y \leq n_y$. Nech*

$$\Phi\colon\ (\mathbb{R}^m)^{\omega_x} \times (\mathbb{R}^m)^{\omega_y} \to \mathbb{R}, \qquad \Phi = \Phi\left(x_1, \ldots, x_{\omega_x}; y_1, \ldots, y_{\omega_y}\right)$$

*je borelovská funkcia, ktorá je symetrická vzhľadom na prvú skupinu premenných $x_1, \ldots, x_{\omega_x}$ a aj na druhú skupinu premenných $y_1, \ldots, y_{\omega_y}$.*

*Položme*

$$U = \binom{n_x}{\omega_x}^{-1} \binom{n_y}{\omega_y}^{-1} \sum_{\alpha} \sum_{\beta} \Phi\left(X_{\alpha_1}, \ldots, X_{\alpha_{\omega_x}}; Y_{\beta_1}, \ldots, Y_{\beta_{\omega_y}}\right), \tag{1.8}$$

*kde sumy $\sum_\alpha \sum_\beta$ idú cez všetky možné podmnožiny indexov $\{\alpha_1, \ldots, \alpha_{\omega_x}\}, \{\beta_1, \ldots, \beta_{\omega_y}\}$ také, že $1 \leq \alpha_1 < \cdots < \alpha_{\omega_x} \leq n_x, 1 \leq \beta_1 < \cdots < \beta_{\omega_y} \leq n_y$. Potom $U$ nazývame* viacrozmernou zovšeobecnenou $U$-štatistikou.

Podstatnou vlastnosťou $U-$štatistík je ich asymptotická normalita. O tom, za akých podmienok sú zovšeobecnené $U-$štatistiky asymptoticky normálne, hovorí nasledujúca veta, prebratá z [8]. Jej znenie sme tiež zjednodušili len pre dva náhodné výbery. Predtým však zavedieme pomocné označenie.

Majme štyri $m-$rozmerné nezávislé náhodné výbery: $X_i, X_i'$ $(i = 1, \ldots, n_x)$ a $Y_\ell, Y_\ell'$ $(\ell = 1, \ldots, n_y)$, kde náhodné výbery $X_i, X_i'$ pochádzajú z rozdelenia so združenou distribučnou funkciou $F$ a $Y_\ell, Y_\ell'$ pochádzajú z rozdelenia so združenou distribučnou funkciou $G$. Majme dve funkcie

$$\Phi^{(a)} = \Phi^{(a)}\left(x_1, \ldots, x_{\omega_x^a}; y_1, \ldots, y_{\omega_y^a}\right), \quad \Phi^{(b)} = \Phi^{(b)}\left(x_1, \ldots, x_{\omega_x^b}; y_1, \ldots, y_{\omega_y^b}\right)$$

spĺňajúce vlastnosti z definície 1.3. Označme $\omega_x^{(a,b)} = \min\{\omega_x^a, \omega_x^b\}$, $\omega_y^{(a,b)} = \min\{\omega_y^a, \omega_y^b\}$. Nech $d_x, d_y$ sú také prirodzené čísla, ktoré spĺňajú $1 \leq d_x \leq \omega_x^{(a,b)}$ a $1 \leq d_y \leq \omega_y^{(a,b)}$. Potom symbolom $\xi_{d_x,d_y}^{(a,b)}$ označujeme kovarianciu náhodných premenných

$$\Phi^{(a)}\left(X_1, \ldots, X_{\omega_x^a}; Y_1, \ldots, Y_{\omega_y^a}\right)$$

a

$$\Phi^{(b)}\left(X_1, \ldots, X_{d_x}, X'_{d_x+1}, \ldots, X'_{\omega_x^b};\ Y_1, \ldots, Y_{d_y}, Y'_{d_y+1}, \ldots, Y'_{\omega_y^b}\right).$$

**Veta 1.4** ([8, Theorem 2.1(i)–(ii)]). *Majme dva nezávislé $m-$rozmerné náhodné výbery $X_i$ $(i = 1, \ldots, n_x)$ a $Y_\ell$ $(\ell = 1, \ldots, n_y)$. Majme $J$ viacrozmerných zovšeobecnených $U-$štatistík $U^{(j)}$ $(j = 1, \ldots, J)$, ktoré sú určené funkciami $\Phi^{(j)}$ s parametrami $\omega_x^j, \omega_y^j$. Náhodné premenné $\Phi^{(j)}\left(X_1, \ldots, X_{\omega_x^j}; Y_1, \ldots, Y_{\omega_y^j}\right)$ označme $\tilde{\Phi}^{(j)}$.*

1. *Ak $E\left[\tilde{\Phi}^{(j)}\right] = \mu^j$, potom $E\left[U^{(j)}\right] = \mu^j$.*
2. *Ak $E\left[(\tilde{\Phi}^{(j)})^2\right] < \infty$ pre každé $j$, potom pre všetky $a, b \in \{1, \ldots, J\}$ je*

$$\begin{aligned}&\operatorname{cov}\left[U^{(a)}, U^{(b)}\right]\\ &= \binom{n_x}{\omega_x^b}^{-1}\binom{n_y}{\omega_y^b}^{-1}\sum_{d_x=0}^{\omega_x^{(a,b)}}\sum_{d_y=0}^{\omega_y^{(a,b)}}\left[\binom{\omega_x^a}{d_x}\binom{n_x-\omega_x^a}{\omega_x^b-d_x}\binom{\omega_y^a}{d_y}\binom{n_y-\omega_y^a}{\omega_y^b-d_y}\right]\xi_{d_x,d_y}^{(a,b)}.\end{aligned}$$

**Veta 1.5** ([8, Theorem 2.1(iii)]). *Nech $X_i$, $Y_\ell$, $U^j$, $\Phi^{(j)}$, $\tilde{\Phi}^{(j)}$ sú ako vo vete 1.4. Nech $E\left[\Phi^{(j)}\right] = \mu^j$ a $E\left[\left(\Phi^{(j)}\right)^2\right] < \infty$ pre každé $j$. Ďalej nech $n_x = \rho_x N$ a $n_y = \rho_y N$, kde $\rho_x, \rho_y$ sú kladné konštanty nezávislé od $N$. Potom $J$-rozmerná štatistika*

$$\sqrt{N}\left(U^{(1)} - \mu_1, \ldots, U^{(J)} - \mu_J\right) \tag{1.9}$$

*konverguje pre $N \to \infty$ v distribúcii k $J$-rozmernému normálnemu rozdeleniu so strednou hodnotou $0_{J\times 1}$ a kovariančnou maticou $\Sigma = (\sigma_{ab})_{J\times J}$, kde*

$$\sigma_{ab} = \frac{\omega_x^a\omega_x^b}{\rho_x}\xi_{1,0}^{(a,b)} + \frac{\omega_y^a\omega_y^b}{\rho_y}\xi_{0,1}^{(a,b)}.$$

*Poznámka*: Predpoklad $n_x = \rho_x N$, $n_y = \rho_y N$ v predošlej vete je obmedzujúci. Môžeme ho nahradiť predpokladom, že existuje $\lambda \in (0, \infty)$ také, že

$$n_x, n_y \to \infty, \quad \frac{n_x}{n_y} \to \lambda. \tag{1.10}$$

Pre $m = 1$ viď napr. [10, str. 165]; pre všeobecné $m$ citáciu nájsť nevieme, ale veta je v takomto tvare často používaná.

Limity v (1.10) chápeme nasledovne. Predpokladáme, že $n_x = n_x(\nu)$, $n_y = n_y(\nu)$ sú závislé od parametra $\nu \in \mathbb{N}$. Vzťah (1.10) znamená, že

$$\lim_{\nu\to\infty} n_x(\nu) = \infty, \quad \lim_{\nu\to\infty} n_y(\nu) = \infty, \quad \lim_{\nu\to\infty} \frac{n_x(\nu)}{n_y(\nu)} = \lambda.$$

## 1.5 Veta 1.6 [13, Theorem 1]

Ďalej sa už budeme venovať hlavnej vete článku [13], ktorá hovorí o rozdelení $T_{cMAX}$ štatistiky. Znenie tejto vety preberáme v mierne upravenom znení z [13, str. 824]. Dôkaz tejto vety je prebratý z prílohy tohto článku [14, str. 3-5], pričom niektoré časti dôkazu uvedieme podrobnejšie. Predpoklady a označenie boli uvedené v kapitole 1.1.

**Veta 1.6** ([13, Theorem 1]). *Nech* $\min\{n_x, n_y\} \to \infty$ *a* $n_x/n_y \to \lambda \in (0, \infty)$. *Potom náhodný vektor*

$$\left( \frac{\bar{R}_{y\cdot 1\cdot} - \bar{R}_{x\cdot 1\cdot}}{\sqrt{N}} - \frac{\sqrt{N}}{2m_1} \sum_{k=1}^{m_1} \theta_{1k}, \ldots, \frac{\bar{R}_{y\cdot J\cdot} - \bar{R}_{x\cdot J\cdot}}{\sqrt{N}} - \frac{\sqrt{N}}{2m_J} \sum_{k=1}^{m_J} \theta_{Jk} \right)^T \tag{1.11}$$

*konverguje k* $J$*-rozmernému normálnemu rozdeleniu s nulovou strednou hodnotou a kovariančnou maticou* $\Sigma = (\sigma_{ab})_{J\times J}$*, ktorej zložky sú*

$$\sigma_{ab} = \frac{1}{m_a m_b} \sum_{k_1=1}^{m_a} \sum_{k_2=1}^{m_b} \left\{ \left(1 + \frac{1}{\lambda}\right) c_{ak_1,bk_2} + (1+\lambda) d_{ak_1,bk_2} - \left(\sqrt{\lambda} + \frac{1}{\sqrt{\lambda}}\right)^2 \theta_{ak_1}\theta_{bk_2} \right\}, \tag{1.12}$$

*kde* $a, b = 1, \ldots, J$,

$$\begin{aligned} c_{ak_1,bk_2} &= \operatorname{cov}\left[G_{ak_1}(X_{ak_1}), G_{k_2}(X_{bk_2})\right], \\ d_{ak_1,bk_2} &= \operatorname{cov}\left[F_{ak_1}(Y_{ak_1}), F_{k_2}(Y_{bk_2})\right]. \end{aligned}$$

*Dôkaz.* Pre $j-$ty cluster máme

$$\begin{aligned} \frac{1}{\sqrt{N}} \left(\bar{R}_{y\cdot j\cdot} - \bar{R}_{x\cdot j\cdot}\right) &= \frac{1}{\sqrt{N} m_j} \sum_{k=1}^{m_j} \left( \frac{1}{n_y} \sum_{\ell=1}^{n_y} R_{y\ell jk} - \frac{1}{n_x} \sum_{i=1}^{n_x} R_{xijk} \right) \\ &= \frac{1}{\sqrt{N} m_j} \sum_{k=1}^{m_j} \left(\bar{R}_{y\cdot jk} - \bar{R}_{y\cdot jk}\right). \end{aligned}$$

Z vety 1.2 dostávame, že

$$\frac{1}{\sqrt{N}} \left(\bar{R}_{y\cdot j\cdot} - \bar{R}_{x\cdot j\cdot}\right) = \frac{\sqrt{N}}{2m_j} \sum_{k=1}^{m_j} \hat{\theta}_{jk} = \frac{\sqrt{N}}{2m_j} \frac{1}{n_x n_y} \sum_{k=1}^{m_j} \sum_{i=1}^{n_x} \sum_{\ell=1}^{n_y} \Psi_{i\ell jk},$$

kde

$$\Psi_{i\ell jk} = I(X_{ijk} < Y_{\ell jk}) - I(X_{ijk} > Y_{\ell jk})$$

bolo definované v (1.6).

Z predpokladu (1.1) a nezávislosti skoro isto platí

$$\Psi_{i\ell jk} = 1 - 2I(X_{ijk} > Y_{\ell jk}) = 2I(X_{ijk} < Y_{\ell jk}) - 1. \tag{1.13}$$

Pre $J-$zložkový vektor rozdielu priemerných rankov clusterov dostávame, že

$$\frac{1}{\sqrt{N}} \begin{pmatrix} \bar{R}_{y\cdot 1\cdot} - \bar{R}_{x\cdot 1\cdot} \\ \vdots \\ \bar{R}_{y\cdot J\cdot} - \bar{R}_{x\cdot J\cdot} \end{pmatrix} = \frac{\sqrt{N}}{2n_x n_y} \sum_{i=1}^{n_x} \sum_{\ell=1}^{n_y} \begin{pmatrix} \frac{1}{m_1} \sum_{k=1}^{m_1} \Psi_{i\ell 1k} \\ \vdots \\ \frac{1}{m_J} \sum_{k=1}^{m_J} \Psi_{i\ell Jk} \end{pmatrix}. \tag{1.14}$$

Vezmime si $\omega_x^j = 1$, $\omega_y^j = 1$ a $\Phi^{(j)}(X_i, Y_\ell) = \frac{1}{m_j} \sum_{k=1}^{m_j} \Psi_{i\ell jk}$. Potom môžeme vidieť, že podľa definície 1.3 je výraz $\frac{1}{n_x n_y} \sum_{i=1}^{n_x} \sum_{\ell=1}^{n_y} \left( \frac{1}{m_1} \sum_{k=1}^{m_1} \Psi_{i\ell jk} \right)$ viacrozmernou zovšeobecnenou $U-$štatistikou. Z ohraničenosti $\Phi^{(j)}$ a vety 1.1 dostávame, že predpoklady vety 1.5 sú splnené. Takže pravá strana rovnosti (1.14) a teda aj vektor

$$\frac{1}{\sqrt{N}} \left( \bar{R}_{y\cdot 1\cdot} - \bar{R}_{x\cdot 1\cdot}, \ldots, \bar{R}_{y\cdot J\cdot} - \bar{R}_{x\cdot J\cdot} \right)^T$$

majú asymptoticky normálne rozdelenie.

Strednú hodnotu tohto rozdelenia dostávame nasledovne. Fixujme $j \in 1, \ldots, J$. Potom

$$\begin{aligned} E\left[ \frac{1}{\sqrt{N}} \left( \bar{R}_{y\cdot j\cdot} - \bar{R}_{x\cdot j\cdot} \right) \right] &= E\left[ \frac{\sqrt{N}}{2n_x n_y} \sum_{i=1}^{n_x} \sum_{\ell=1}^{n_y} \left( \frac{1}{m_j} \sum_{k=1}^{m_j} \Psi_{i\ell jk} \right) \right] \\ &= \frac{\sqrt{N}}{2n_x n_y} \frac{1}{m_j} \sum_{i=1}^{n_x} \sum_{\ell=1}^{n_y} \sum_{k=1}^{m_j} E\left[ \Psi_{i\ell jk} \right]. \end{aligned}$$

Z vety 1.1 vieme, že platí

$$E\left[ \frac{1}{\sqrt{N}} \left( \bar{R}_{y\cdot j\cdot} - \bar{R}_{x\cdot j\cdot} \right) \right] = \frac{\sqrt{N}}{2n_x n_y} \frac{1}{m_j} \sum_{i=1}^{n_x} \sum_{\ell=1}^{n_y} \sum_{k=1}^{m_j} \theta_{jk},$$

teda

$$E\left[ \frac{1}{\sqrt{N}} \left( \bar{R}_{y\cdot j\cdot} - \bar{R}_{x\cdot j\cdot} \right) \right] = \frac{\sqrt{N}}{2} \frac{1}{m_j} \sum_{k=1}^{m_j} \theta_{jk}.$$

Takže stredná hodnota vektora $\frac{1}{\sqrt{N}} \left( \bar{R}_{y\cdot 1\cdot} - \bar{R}_{x\cdot 1\cdot}, \ldots, \bar{R}_{y\cdot J\cdot} - \bar{R}_{x\cdot J\cdot} \right)^T$ je

$$\frac{\sqrt{N}}{2} \left( \frac{1}{m_1} \sum_{k=1}^{m_1} \theta_{1k}, \ldots, \frac{1}{m_J} \sum_{k=1}^{m_J} \theta_{Jk} \right)^T. \tag{1.15}$$

Čím je dokázané, že vektor (1.11) má nulovú strednú hodnotu.

Ďalej sa pozrieme na jednotlivé zložky kovariančnej matice vektora (1.11), ktorá je zrejme zhodná s kovariančnou maticou vektora $\frac{1}{\sqrt{N}}\left(\bar{R}_{y\cdot 1\cdot}-\bar{R}_{x\cdot 1\cdot},\ldots,\bar{R}_{y\cdot J\cdot}-\bar{R}_{x\cdot J\cdot}\right)^T$. Pre ľubovoľné $a,b\in\{1,2,\ldots,J\}$ z (1.14) dostávame, že

$$\begin{aligned}
&\operatorname{cov}\left[\frac{1}{\sqrt{N}}\left(\bar{R}_{y\cdot a\cdot}-\bar{R}_{x\cdot a\cdot}\right),\frac{1}{\sqrt{N}}\left(\bar{R}_{y\cdot b\cdot}-\bar{R}_{x\cdot b\cdot}\right)\right]\\
&=\frac{N}{4n_x^2n_y^2}\sum_{i_1=1}^{n_x}\sum_{i_2=1}^{n_x}\sum_{\ell_1=1}^{n_y}\sum_{\ell_2=1}^{n_y}\operatorname{cov}\left[\frac{1}{m_a}\sum_{k_1=1}^{m_a}\Psi_{i_1\ell_1ak_1},\frac{1}{m_b}\sum_{k_2=1}^{m_b}\Psi_{i_2\ell_2bk_2}\right]\\
&=\frac{N}{4n_x^2n_y^2m_am_b}\sum_{i_1=1}^{n_x}\sum_{i_2=1}^{n_x}\sum_{\ell_1=1}^{n_y}\sum_{\ell_2=1}^{n_y}\sum_{k_1=1}^{m_a}\sum_{k_2=1}^{m_b}\operatorname{cov}\left[\Psi_{i_1\ell_1ak_1},\Psi_{i_2\ell_2bk_2}\right].
\end{aligned}\tag{1.16}$$

Pre ľubovoľné $a,b\in\{1,2,\ldots,J\}$ budeme pre výraz $\operatorname{cov}\left[\Psi_{i_1\ell_1ak_1},\Psi_{i_2\ell_2bk_2}\right]$ uvažovať štyri prípady podľa toho, aké sú hodnoty $i_1,i_2,\ell_1,\ell_2$.

V prvom prípade, kedy $i_1\neq i_2$ a $\ell_1\neq\ell_2$ z nezávislosti dostávame, že

$$\operatorname{cov}\left[\Psi_{i_1\ell_1ak_1},\Psi_{i_2\ell_2bk_2}\right]=0.\tag{1.17}$$

V druhom prípade predpokladajme, že $i_1=i_2=i$ a $\ell_1\neq\ell_2$. Z definície kovariancie dostávame, že

$$\operatorname{cov}\left[\Psi_{i\ell_1ak_1},\Psi_{i\ell_2bk_2}\right]=E\left[\Psi_{i\ell_1ak_1}\Psi_{i\ell_2bk_2}\right]-E\left[\Psi_{i\ell_1ak_1}\right]E\left[\Psi_{i\ell_2bk_2}\right].$$

Z vety 1.1 vieme, že $E\left(\Psi_{i\ell_1ak_1}\right)=\theta_{ak_1}$ a $E\left(\Psi_{i\ell_2bk_2}\right)=\theta_{bk_2}$. Ďalej využijeme rovnosť (1.13) a dostávame, že

$$\operatorname{cov}\left[\Psi_{i\ell_1ak_1},\Psi_{i\ell_2bk_2}\right]=E\left[\left\{1-2I(X_{iak_1}>Y_{\ell_1ak_1})\right\}\left\{1-2I(X_{ibk_2}>Y_{\ell_2bk_2})\right\}\right]-\theta_{ak_1}\theta_{bk_2}.$$

Autori v článku [13] uvádzajú nasledovnú sériu rovností, ktorú však nebudeme teraz objasňovať. Vrátime sa k nej v kapitole 2, nakoľko je v tejto sérii rovností chyba.

$$\begin{aligned}
\operatorname{cov}\left[\Psi_{i\ell_1ak_1},\Psi_{i\ell_2bk_2}\right]&=E\left[\left\{1-2G_{ak_1}(X_{ak_1})\right\}\left\{1-2G_{bk_2}(X_{bk_2})\right\}\right]-\theta_{ak_1}\theta_{bk_2}\\
&=4\operatorname{cov}\left[G_{ak_1}(X_{ak_1}),G_{bk_2}(X_{bk_2})\right]-\theta_{ak_1}\theta_{bk_2}.
\end{aligned}\tag{1.18}$$

Tretí prípad bude analogický k druhému prípadu, pričom uvažujeme $i_1\neq i_2$ a $\ell_1=\ell_2=\ell$. Z definície kovariancie dostávame, že

$$\operatorname{cov}\left[\Psi_{i_1\ell ak_1},\Psi_{i_2\ell bk_2}\right]=E\left[\Psi_{i_1\ell ak_1}\Psi_{i_2\ell bk_2}\right]-E\left[\Psi_{i_1\ell ak_1}\right]E\left[\Psi_{i_2\ell bk_2}\right].$$

Z vety 1.1 a rovnosti (1.13) máme nasledovnú rovnosť

$$\operatorname{cov}\left[\Psi_{i_1\ell ak_1}, \Psi_{i_2\ell bk_2}\right] = E\left[\{2I(X_{iak_1} < Y_{\ell_1 ak_1}) - 1\}\{2I(X_{ibk_2} < Y_{\ell_2 bk_2}) - 1\}\right] - \theta_{ak_1}\theta_{bk_2}.$$

Ako v druhom prípade, nasledovnú sériu rovností iba uvedieme a vrátime sa k nej v kapitole 2, kvôli spomínanej chybe.

$$\begin{aligned}\operatorname{cov}\left[\Psi_{i_1\ell_1 ak_1}, \Psi_{i_2\ell_2 bk_2}\right] &= E\left[\{2F_{ak_1}(Y_{ak_1}) - 1\}\{2F_{bk_2}(Y_{bk_2}) - 1\}\right] - \theta_{ak_1}\theta_{bk_2} \\ &= 4\operatorname{cov}\left[F_{ak_1}(Y_{ak_1}), F_{bk_2}(Y_{bk_2})\right] - \theta_{ak_1}\theta_{bk_2}.\end{aligned} \tag{1.19}$$

V poslednom štvrtom prípade uvažujme situáciu, keď $i_1 = i_2 = i$ a $\ell_1 = \ell_2 = \ell$. Tu využijeme definíciu kovariancie spolu s rovnosťou (1.13) a dostávame

$$\operatorname{cov}\left[\Psi_{i\ell ak_1}, \Psi_{i\ell bk_2}\right] = E\left[\Psi_{i\ell ak_1}\Psi_{i\ell bk_2}\right] - \theta_{ak_1}\theta_{bk_2}. \tag{1.20}$$

Ďalej rozpíšeme strednú hodnotu $E\left[\Psi_{i\ell ak_1}\Psi_{i\ell bk_2}\right]$

$$\begin{aligned}E\left[\Psi_{i\ell ak_1}\Psi_{i\ell bk_2}\right] = E[&(I(X_{iak_1} < Y_{\ell ak_1}) - I(X_{iak_1} > Y_{\ell ak_1})) \\ &\cdot (I(X_{ibk_2} < Y_{\ell bk_2}) - I(X_{ibk_2} > Y_{\ell bk_2}))]\end{aligned}$$

Keďže $X_i$ má rovnaké rozdelenie ako $X$ a $Y_\ell$ má rovnaké rozdelenie ako $Y$, tak môžeme písať

$$\begin{aligned}E\left[\Psi_{i\ell ak_1}\Psi_{i\ell bk_2}\right] &= E\Big[(I(X_{ak_1} < Y_{ak_1}) - I(X_{ak_1} > Y_{ak_1})) \\ &\qquad \cdot (I(X_{bk_2} < Y_{bk_2}) - I(X_{bk_2} > Y_{bk_2}))\Big] \\ &= E\left[\Psi_{ak_1}\Psi_{bk_2}\right],\end{aligned}$$

kde $\Psi_{ak_1}$, $\Psi_{bk_2}$ sú definované analogicky ako v (1.6), len namiesto $X_i$ a $Y_\ell$ sa vezme $X$ a $Y$.

Dosadením týchto výrazov do (1.16) autori článku [13] dostali nasledujúce vyjadrenie pre kovarianciu zložiek $a, b$ vektora (1.11). Avšak toto vyjadrenie obsahuje numerickú chybu, ktorej sa budeme venovať v kapitole 2.

$$\begin{aligned}&\operatorname{cov}\left[\frac{1}{\sqrt{N}}\left(\bar{R}_{y\cdot a\cdot} - \bar{R}_{x\cdot a\cdot}\right), \frac{1}{\sqrt{N}}\left(\bar{R}_{y\cdot b\cdot} - \bar{R}_{x\cdot b\cdot}\right)\right] \\ &= \frac{1}{m_a m_b}\sum_{k_1=1}^{m_a}\sum_{k_2=1}^{m_b}\left\{\frac{N(n_y-1)}{n_x n_y}\operatorname{cov}\left[G_{ak_1}(X_{ak_1}), G_{bk_2}(X_{bk_2})\right]\right. \\ &\quad + \frac{N(n_x-1)}{n_x n_y}\operatorname{cov}\left[F_{ak_1}(Y_{ak_1}), F_{bk_2}(Y_{bk_2})\right] \\ &\quad \left. + \frac{N}{4n_x n_y}E\left[\Psi_{ak_1}\Psi_{bk_2}\right] - \frac{N\left(N - \frac{7}{4}\right)}{n_x n_y}\theta_{ak_1}\theta_{bk_2}\right\}.\end{aligned} \tag{1.21}$$

Nech $\min\{n_x, n_y\} \to \infty$ a $n_x/n_y \to \lambda \in (0, \infty)$. Potom podľa (1.21) platí, že

$$\begin{aligned}
&\operatorname{cov}\left[\frac{1}{\sqrt{N}}\left(\bar{R}_{y\cdot a\cdot} - \bar{R}_{x\cdot a\cdot}\right), \frac{1}{\sqrt{N}}\left(\bar{R}_{y\cdot b\cdot} - \bar{R}_{x\cdot b\cdot}\right)\right] \\
&\to \frac{1}{m_a m_b}\sum_{k_1=1}^{m_a}\sum_{k_2=1}^{m_b}\left\{\left(1+\frac{1}{\lambda}\right)c_{ak_1,bk_2} + (1+\lambda)d_{ak_1,bk_2} - \left(\sqrt{\lambda}+\frac{1}{\sqrt{\lambda}}\right)^2\theta_{ak_1}\theta_{bk_2}\right\},
\end{aligned}$$

kde $c_{ak_1,bk_2} = \operatorname{cov}\left[G_{ak_1}(X_{ak_1}), G_{k_2}(X_{bk_2})\right]$ a $d_{ak_1,bk_2} = \operatorname{cov}\left[F_{ak_1}(Y_{ak_1}), F_{k_2}(Y_{bk_2})\right]$. □

*Poznámka*: Autori ukázali, že vektor

$$\left(\frac{\bar{R}_{y\cdot 1\cdot} - \bar{R}_{x\cdot 1\cdot}}{\sqrt{N}} - \frac{\sqrt{N}}{2m_1}\sum_{k=1}^{m_1}\theta_{1k}, \ldots, \frac{\bar{R}_{y\cdot J\cdot} - \bar{R}_{x\cdot J\cdot}}{\sqrt{N}} - \frac{\sqrt{N}}{2m_J}\sum_{k=1}^{m_J}\theta_{Jk}\right)^T$$

sa asymptoticky riadi normálnym rozdelením s nulovou strednou hodnotou a spomínanou kovariančnou maticou, bez ohľadu na platnosť zovšeobecnenej Behrensovej-Fisherovej hypotézy. Vďaka tomuto výsledku môžeme pomocou $T_{cMAX}$ testu testovať aj iné hypotézy (o parametroch $\theta_{jk}$), nielen zovšeobecnenú Behrensovu-Fisherovu hypotézu.

## 1.6 Veta 1.7 [13, Theorem 2]

Ďalej uvedieme vetu, ktorá sa venuje odhadu kovariančnej matice z vety 1.6.

**Veta 1.7** ([13, Theorem 2]). *Konzistentnými odhadmi pre $c_{ak_1,bk_2}$ a $d_{ak_1,bk_2}$ sú*

$$\begin{aligned}
\hat{c}_{ak_1,bk_2} = \frac{1}{n_x}\sum_{i=1}^{n_x}\Bigg(&\left\{\frac{1}{n_y}\sum_{\ell=1}^{n_y} I(Y_{\ell ak_1} < X_{iak_1}) - \frac{1-\hat{\theta}_{ak_1}}{2}\right\} \\
&\cdot\left\{\frac{1}{n_y}\sum_{\ell=1}^{n_y} I(Y_{\ell bk_2} < X_{ibk_2}) - \frac{1-\hat{\theta}_{bk_2}}{2}\right\}\Bigg),
\end{aligned}$$

$$\begin{aligned}
\hat{d}_{ak_1,bk_2} = \frac{1}{n_y}\sum_{\ell=1}^{n_y}\Bigg(&\left\{\frac{1}{n_x}\sum_{i=1}^{n_x} I(X_{iak_1} < Y_{\ell ak_1}) - \frac{1+\hat{\theta}_{ak_1}}{2}\right\} \\
&\cdot\left\{\frac{1}{n_x}\sum_{i=1}^{n_x} I(X_{ibk_2} < Y_{\ell bk_2}) - \frac{1+\hat{\theta}_{bk_2}}{2}\right\}\Bigg),
\end{aligned}$$

*kde $a, b = 1, \ldots, J$, $k_1 = 1, \ldots, m_a$, $k_2 = 1, \ldots, m_b$; ďalej, podľa* (1.7),

$$\hat{\theta}_{jk} = \frac{1}{n_x n_y}\sum_{i=1}^{n_x}\sum_{\ell=1}^{n_y} I(X_{ijk} < Y_{\ell jk}) - I(X_{ijk} > Y_{\ell jk}),$$

*$j = 1, \ldots, J$, $k = 1, \ldots, m_j$. Teda konzistentný odhad pre $\Sigma$ je $S = (s_{ab})_{J\times J}$, kde*

$$s_{ab} = \frac{1}{m_a m_b}\sum_{k_1=1}^{m_a}\sum_{k_2=1}^{m_b}\left\{\left(1+\frac{1}{\lambda}\right)\hat{c}_{ak_1,bk_2} + (1+\lambda)\hat{d}_{ak_1,bk_2} - \left(\sqrt{\lambda}+\frac{1}{\sqrt{\lambda}}\right)^2\hat{\theta}_{ak_1}\hat{\theta}_{bk_2}\right\}. \quad (1.22)$$

*Stručná idea dôkazu.* Nakoľko v článku [13] nie je dôkaz tejto vety, tak uvedieme stručnú ideu odvodenia odhadov $\hat{c}_{ak_1,bk_2}$ a $\hat{d}_{ak_1,bk_2}$. Začneme odhadom $\hat{c}_{ak_1,bk_2}$:

$$\begin{aligned}
\hat{c}_{ak_1,bk_2} &= \widehat{\mathrm{cov}}\left[G_{ak_1}(X_{ak_1}), G_{bk_2}(X_{bk_2})\right] \\
&= \hat{E}\left[\left\{G_{ak_1}(X_{ak_1}) - E\left[G_{ak_1}(X_{ak_1})\right]\right\}\left\{G_{bk_2}(X_{bk_2}) - E\left[G_{bk_2}(X_{bk_2})\right]\right\}\right] \\
&= \frac{1}{n_x}\sum_{i=1}^{n_x}\left(\left\{\frac{1}{n_y}\sum_{\ell=1}^{n_y} I(Y_{\ell ak_1} < X_{iak_1}) - \frac{1-\hat{\theta}_{ak_1}}{2}\right\}\right. \\
&\qquad \left.\cdot\left\{\frac{1}{n_y}\sum_{\ell=1}^{n_y} I(Y_{\ell bk_2} < X_{ibk_2}) - \frac{1-\hat{\theta}_{bk_2}}{2}\right\}\right),
\end{aligned}$$

pričom $\frac{1}{n_y}\sum_{\ell=1}^{n_y} I(Y_{\ell ak_1} < x_{iak_1})$ je empirický odhad distribučnej funkcie $G_{ak_1}$ v bode $x_{iak_1}$. Ďalej za platnosti predpokladu (1.1) a využitím (1.13) dostávame, že $(1-\hat{\theta}_{ak_1})/2$ je odhadom $E\left[G_{ak_1}(X_{ak_1})\right]$[5]. Podobne pre členy v druhom súčiniteli s indexami $bk_2$.

Analogicky pre $\hat{d}_{ak_1,bk_2}$ dostávame:

$$\begin{aligned}
\hat{d}_{ak_1,bk_2} &= \widehat{\mathrm{cov}}\left[F_{ak_1}(Y_{ak_1}), F_{bk_2}(Y_{bk_2})\right] \\
&= \hat{E}\left[\left\{F_{ak_1}(Y_{ak_1}) - E\left[F_{ak_1}(Y_{ak_1})\right]\right\}\left\{F_{bk_2}(Y_{bk_2}) - E\left[F_{bk_2}(Y_{bk_2})\right]\right\}\right] \\
&= \frac{1}{n_y}\sum_{\ell=1}^{n_y}\left(\left\{\frac{1}{n_x}\sum_{i=1}^{n_x} I(X_{iak_1} < Y_{\ell ak_1}) - \frac{1+\hat{\theta}_{ak_1}}{2}\right\}\right. \\
&\qquad \left.\cdot\left\{\frac{1}{n_x}\sum_{i=1}^{n_x} I(X_{ibk_2} < Y_{\ell bk_2}) - \frac{1+\hat{\theta}_{bk_2}}{2}\right\}\right),
\end{aligned}$$

pričom $\frac{1}{n_x}\sum_{i=1}^{n_x} I(X_{iak_1} < y_{\ell ak_1})$ je empirický odhad distribučnej funkcie $F_{ak_1}$ v bode $y_{\ell ak_1}$. Ďalej za platnosti predpokladu (1.1) a využitím (1.13) dostávame, že $(1+\hat{\theta}_{ak_1})/2$ je odhadom $E\left[F_{ak_1}(Y_{ak_1})\right]$. Podobne pre $bk_2$. □

Vrátime sa ešte k výpočtu samotnej testovej štatistiky $T_{cMAX}$ testu. Jej predpis je

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

Ostáva nám ešte dopočítať odhad variancie $\widehat{\mathrm{var}}\left(\bar{R}_{y\cdot j\cdot} - \bar{R}_{x\cdot j\cdot}\right)$. Testovú štatistiku $T_{cMAX}$ testu môžeme prepísať nasledovne

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

[5]Tomu prečo sú $(1-\hat{\theta}_{ak_1})/2$ a $(1+\hat{\theta}_{ak_1})/2$ dobrými odhadmi postupne $E\left[G_{ak_1}(X_{ak_1})\right]$ a $E\left[F_{ak_1}(Y_{ak_1})\right]$ sa budeme podrobnejšie venovať v kapitolách 3.4 a 3.6.1.

Z vety 1.7 [13, Theorem 2] vieme, že konzistentným odhadom $\widehat{\text{var}}\left((\bar{R}_{y\cdot j\cdot} - \bar{R}_{x\cdot j\cdot})/\sqrt{N}\right)$ za platnosti zovšeobecnenej Behrensovej-Fisherovej hypotézy (1.3) je $s_{jj}$. Z toho dostávame, že

$$T_{cMAX} = \max_{j\in\{1,\ldots,J\}} \left\{ \left| \frac{(\bar{R}_{y\cdot j\cdot} - \bar{R}_{x\cdot j\cdot})/\sqrt{N}}{\sqrt{s_{jj}}} \right| \right\} = \max_{j\in\{1,\ldots,J\}} \left\{ \left| \frac{\bar{R}_{y\cdot j\cdot} - \bar{R}_{x\cdot j\cdot}}{\sqrt{N s_{jj}}} \right| \right\}.$$

# 2 Oprava $T_{cMAX}$ testu: spojitý prípad

Pri implementácii testu podľa vety [13, Theorem 2] sme narazili na problém. Relatívne často vychádzali odhady kovariančnej matice $\Sigma$ indefinitné. Hľadali sme dôvody tejto indefinitnosti a možnosti, ktoré nám napadli, boli nasledovné:

1. Jednou z možností bol fakt, že vo vete 1.7 [13, Theorem 2] sa každý z členov kovariančnej matice odhadoval „nezávisle“ od ostatných, čo mohlo viesť k indefinitnosti tejto matice.

2. Ďalej sme si všimli, že s rastúcou veľkosťou náhodných výberov klesala relatívna frekvencia týchto indefinitných matíc (viď obrázok 1 čierna krivka). Tento výsledok nás viedol k domnienke, že indefinitnosť môže byť spôsobená zanedbaním členov, ktoré sú v limite nulové. Napríklad

   (a) $(n_y - 1)/n_y$ sa nahradilo jednotkou;

   (b) $n_x/n_y$ sa nahradilo $\lambda$;

   (c) $N/(n_x n_y)$ sa nahradilo nulou ...

Pri hľadaní dôvodu (možnej) indefinitnosti matice $\Sigma$ sme prišli na niekoľko chýb v dôkaze vety 1.6 [13, Theorem 1]. Prvou z chýb bol zle spočítaný koeficient pri člene $\theta_{ak_1}\theta_{bk_2}$ viď (1.21). Túto chybu popisujeme v kapitole 2.1 spolu s jej opravou. Po tejto oprave sme však zistili, že odhady $\Sigma$ sú aj naďalej relatívne často indefinitné (viď obrázok 1 červená krivka).

Ďalším problémom v dôkaze vety 1.6 [13, Theorem 1] bolo chybné odvodenie kovariancií $c_{ak_1,bk_2}$ a $d_{ak_1,bk_2}$, viď (1.18) a (1.19). Túto chybu popisujeme v kapitolách 2.2 a 2.3. Korektné znenie vety 1.6 [13, Theorem 1] uvádzame v kapitole 2.4 a novú kovariančnú maticu označíme $\Sigma^{(1)}$. Pri implementácii opravenej verzie $T_{cMAX}$ testu (viď kapitola 2.5) sme zistili, že odhady $S^{(1)}$ tejto kovariančnej matice už indefinitné nie sú (viď obrázok 1 zelená krivka).

V kapitole 2.6 sme sa pozreli na opravenú verziu $T_{cMAX}$ testu v prípade, že nepoužijeme zanedbanie popísané v bode 2 vyššie na tejto strane.

Potom sme sa pokúsili nájsť iný spôsob vyjadrenia kovariančnej matice z vety 1.6 [13, Theorem 1]. Odvodenie tvaru kovariančnej matice $\Sigma^{(2)}$ popisujeme v kapitole 2.7. Na to

nadväzuje odhad $S^{(2)}$, ktorý uvádzame v „asymptotickom“ aj „neasymptotickom“ znení v kapitole 2.8.

Hlavnými výsledkami sú teda vety 2.1 a 2.2, ktoré opravujú [13, Theorem 1] a [13, Theorem 2]. Ich dôkazy uvádzame v spomínaných kapitolách 2.7 a 2.8 [6].

**Veta 2.1.** *Majme dva nezávislé náhodné vektory $X$, $Y$, ktoré spĺňajú podmienku* (1.1). *Nech* $\min\{n_x, n_y\} \to \infty$ *a* $n_x/n_y \to \lambda \in (0, \infty)$. *Potom náhodný vektor*

$$\left( \frac{\bar{R}_{y\cdot 1\cdot} - \bar{R}_{x\cdot 1\cdot}}{\sqrt{N}} - \frac{\sqrt{N}}{2m_1} \sum_{k=1}^{m_1} \theta_{1k}, \ldots, \frac{\bar{R}_{y\cdot J\cdot} - \bar{R}_{x\cdot J\cdot}}{\sqrt{N}} - \frac{\sqrt{N}}{2m_J} \sum_{k=1}^{m_J} \theta_{Jk} \right)^T$$

*konverguje k $J$-rozmernému normálnemu rozdeleniu s nulovou strednou hodnotou a kovariančnou maticou* $\Sigma^{(2)} = (\sigma_{ab}^{(2)})_{J\times J}$, *ktorej zložky sú*

$$\sigma_{ab}^{(2)} = \frac{1}{m_a m_b} \sum_{k_1=1}^{m_a} \sum_{k_2=1}^{m_b} \left\{ \left(1 + \frac{1}{\lambda}\right) c_{ak_1,bk_2} + (1+\lambda) d_{ak_1,bk_2} - \left(\sqrt{\lambda} + \frac{1}{\sqrt{\lambda}}\right)^2 \theta_{ak_1}\theta_{bk_2} \right\}, \quad (2.1)$$

*kde*

$$c_{ak_1,bk_2} = \operatorname{cov}\left[G_{ak_1}(X_{ak_1}), G_{k_2}(X_{bk_2})\right],$$
$$d_{ak_1,bk_2} = \operatorname{cov}\left[F_{ak_1}(Y_{ak_1}), F_{k_2}(Y_{bk_2})\right],$$

$a, b = 1, \ldots, J$, $k_1 = 1, \ldots, m_a$, $k_2 = 1, \ldots, m_b$.

**Veta 2.2.** *Za platnosti* (1.1) *konzistentnými odhadmi pre* $c_{ak_1,bk_2}$ *a* $d_{ak_1,bk_2}$ *z vety 2.1 sú*

$$\hat{c}_{ak_1,bk_2} = \frac{1}{n_x} \sum_{i=1}^{n_x} \left( \left\{ \frac{1}{n_y} \sum_{\ell=1}^{n_y} I(Y_{\ell ak_1} < X_{iak_1}) - \frac{1-\hat{\theta}_{ak_1}}{2} \right\} \cdot \left\{ \frac{1}{n_y} \sum_{\ell=1}^{n_y} I(Y_{\ell bk_2} < X_{ibk_2}) - \frac{1-\hat{\theta}_{bk_2}}{2} \right\} \right),$$

$$\hat{d}_{ak_1,bk_2} = \frac{1}{n_y} \sum_{\ell=1}^{n_y} \left( \left\{ \frac{1}{n_x} \sum_{i=1}^{n_x} I(X_{iak_1} < Y_{\ell ak_1}) - \frac{1+\hat{\theta}_{ak_1}}{2} \right\} \cdot \left\{ \frac{1}{n_x} \sum_{i=1}^{n_x} I(X_{ibk_2} < Y_{\ell bk_2}) - \frac{1+\hat{\theta}_{bk_2}}{2} \right\} \right),$$

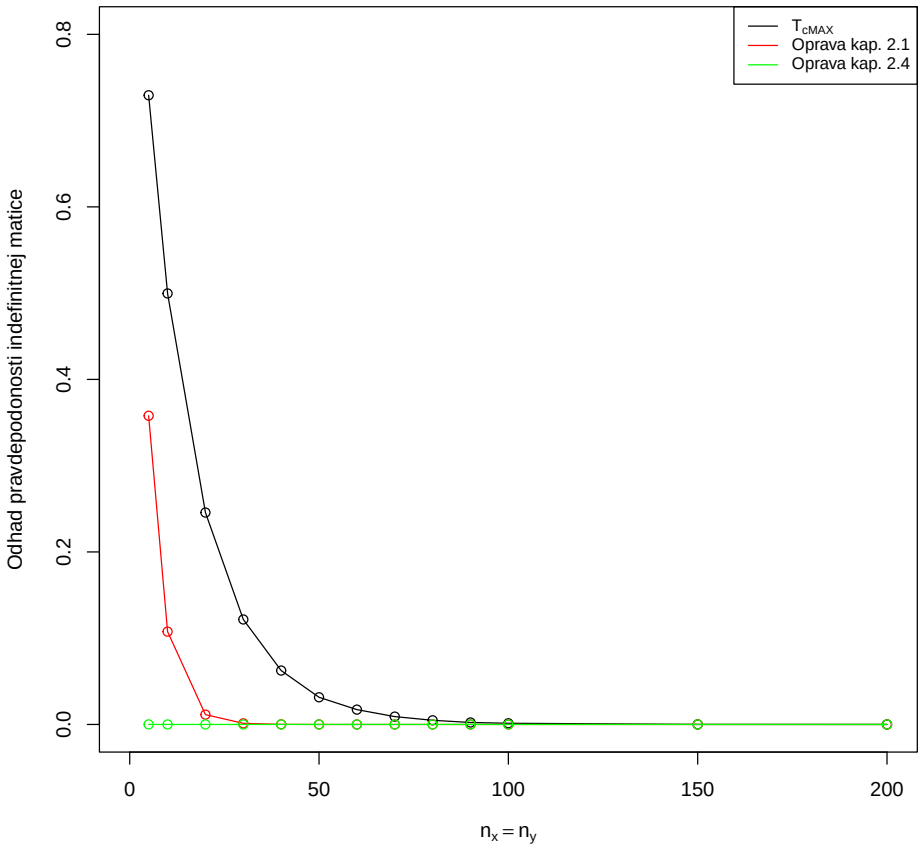
[6] Pri finalizácii diplomovej práce sme našli článok [12], ktorý sa venuje podobnému problému ako článok [13]. V článku [12, str. 6-7] sú uvedené ekvivalenty viet 1.6 [13, Theorem 1] a 1.7 [13, Theorem 2]. Treba podotknúť, že tieto vety [12, Theorem 1 a Theorem 2] sú korektné za predpokladu (1.1) a ich znenie je veľmi podobné našim vetám 2.1 a 2.2. Článok [12] napriek tomu, že cituje článok [13], explicitne neuvádza, že vety [13, Theorem 1 a Theorem 2] sú chybné a vety [12, Theorem 1 a Theorem 2] sú ich opravami.

*kde* $a, b = 1, \ldots, J$, $k_1 = 1, \ldots, m_a$, $k_2 = 1, \ldots, m_b$; *ďalej, podľa* (1.7),

$$\hat{\theta}_{jk} = \frac{1}{n_x n_y} \sum_{i=1}^{n_x} \sum_{\ell=1}^{n_y} I(X_{ijk} < Y_{\ell jk}) - I(X_{ijk} > Y_{\ell jk}),$$

$j = 1, \ldots, J$, $k = 1, \ldots, m_j$. *Teda konzistentný odhad pre* $\Sigma$ *z vety 2.1 je* $S^{(2)} = (s_{ab}^{(2)})_{J \times J}$, *kde*

$$s_{ab}^{(2)} = \frac{1}{m_a m_b} \sum_{k_1=1}^{m_a} \sum_{k_2=1}^{m_b} \left\{ \left(1 + \frac{1}{\lambda}\right) \hat{c}_{ak_1,bk_2} + (1+\lambda) \hat{d}_{ak_1,bk_2} - \left(\sqrt{\lambda} + \frac{1}{\sqrt{\lambda}}\right)^2 \hat{\theta}_{ak_1} \hat{\theta}_{bk_2} \right\}. \quad (2.2)$$




**Obr. 1:** Relatívna frekvencia indefinitných matíc podľa veľkostí náhodných výberov pre rôzne implementácie $T_{cMAX}$ testu. Simulácie boli generované zo šesť-rozmerného log-normálneho rozdelenia, pričom zovšeobecnená Behrensova-Fisherova hypotéza (1.3) bola splnená. Pre všetky simulácie platilo, že $n_x = n_y$.

## 2.1 Problém $T_{cMAX}$ testu - numerická chyba vo vete 1.6 [13, Theorem 1]

V tejto sekcii sa pozrieme na spomínanú numerickú chybu v rovnosti (1.21). Chyba je v koeficiente pri člene $\theta_{ak_1}\theta_{bk_2}$. Počítajme, akej hodnote sa rovná tento koeficient skutočne.

Pre druhý prípad vety 1.6 [13, Theorem 1] máme $i_1 = i_2 = i$ a $l_1 \neq l_2$. Počet možností, kedy tento prípad nastáva, je $n_x n_y(n_y - 1)$. Pre tretí prípad, kedy $i_1 \neq i_2$ a $\ell_1 = \ell_2 = \ell$, je počet možností rovný $n_y n_x(n_x - 1)$. Nakoniec pre štvrtý prípad, kedy $i_1 = i_2 = i$ a $\ell_1 = \ell_2 = \ell$, je týchto možností $n_x n_y$. Dostávame teda, že výraz $\theta_{ak_1}\theta_{bk_2}$ sa vyskytuje $n_x n_y(n_x + n_y - 1)$ krát. Po vynásobení koeficientom z (1.16) dostávame, že koeficient pred členom $\theta_{ak_1}\theta_{bk_2}$ je

$$\frac{N}{4n_x^2 n_y^2} n_x n_y(n_x + n_y - 1) = \frac{N(N-1)}{4n_x n_y},$$

čo je zjavne rozdielne od $[N\,(N - 7/4)]/[n_x n_y]$ z [14, str. 5]. Pre $\min\{n_x, n_y\} \to \infty$ a $n_x/n_y \to \lambda \in (0, \infty)$ tak dostávame, že

$$\sigma_{ab} = \frac{1}{m_a m_b} \sum_{k_1=1}^{m_a} \sum_{k_2=1}^{m_b} \left\{ \left(1 + \frac{1}{\lambda}\right) c_{ak_1,bk_2} + (1+\lambda) d_{ak_1,bk_2} - \mathbf{\frac{1}{4}} \left(\sqrt{\lambda} + \frac{1}{\sqrt{\lambda}}\right)^2 \theta_{ak_1}\theta_{bk_2} \right\}.$$

K analogickej zmene dôjde vo vete 1.7 [13, Theorem 2], kde

$$s_{ab} = \frac{1}{m_a m_b} \sum_{k_1=1}^{m_a} \sum_{k_2=1}^{m_b} \left\{ \left(1 + \frac{1}{\lambda}\right) \hat{c}_{ak_1,bk_2} + (1+\lambda) \hat{d}_{ak_1,bk_2} - \mathbf{\frac{1}{4}} \left(\sqrt{\lambda} + \frac{1}{\sqrt{\lambda}}\right)^2 \hat{\theta}_{ak_1}\hat{\theta}_{bk_2} \right\}.$$

Nakoľko táto numerická chyba nebola jedinou chybou pri odvodení kovariančnej matice z vety 1.6 [13, Theorem 1], tak podiel indefinitných matíc pri implementácii $T_{cMAX}$ testu bol stále relatívne veľký (viď obrázok 1 červená krivka).

## 2.2 Problém $T_{cMAX}$ testu - chyba pri výpočte kovariancie vo vete 1.6 [13, Theorem 1]

Ako sme spomínali v dôkaze vety 1.6 [13, Theorem 1], tak v druhom a treťom prípade, konkrétne v rovnostiach (1.18) a (1.19), sa nachádza chyba. Začneme druhým prípadom. Ak $i_1 = i_2 = i$ a $l_1 \neq l_2$, autori v [14, str. 5] píšu, že

$$\begin{aligned} & E\left[\{1 - 2G_{ak_1}(X_{ak_1})\}\{1 - 2G_{bk_2}(X_{bk_2})\}\right] - \theta_{ak_1}\theta_{bk_2} \\ =& 4E\left[\left\{G_{ak_1}(X_{ak_1}) - \frac{1}{2}\right\}\left\{G_{bk_2}(X_{bk_2}) - \frac{1}{2}\right\}\right] - \theta_{ak_1}\theta_{bk_2} \\ =& 4\,\mathrm{cov}\left[G_{ak_1}(X_{ak_1}), G_{bk_2}(X_{bk_2})\right] - \theta_{ak_1}\theta_{bk_2}. \end{aligned} \tag{2.3}$$

Z definície kovariancie dostávame, že

$$\begin{aligned}&\operatorname{cov}\left[G_{ak_1}(X_{ak_1}), G_{bk_2}(X_{bk_2})\right]\\&= E\left[\left\{G_{ak_1}(X_{ak_1}) - E\left[G_{ak_1}(X_{ak_1})\right]\right\}\left\{G_{bk_2}(X_{bk_2}) - E\left[G_{bk_2}(X_{bk_2})\right]\right\}\right].\end{aligned}$$

Problémom je rovnosť (2.3), kde autori v [14, str. 5] vlastne tvrdia, že

$$\begin{aligned}&E\left[\left\{G_{ak_1}(X_{ak_1}) - E\left[G_{ak_1}(X_{ak_1})\right]\right\}\left\{G_{bk_2}(X_{bk_2}) - E\left[G_{bk_2}(X_{bk_2})\right]\right\}\right]\\&= E\left[\left\{G_{ak_1}(X_{ak_1}) - \frac{1}{2}\right\}\left\{G_{bk_2}(X_{bk_2}) - \frac{1}{2}\right\}\right].\end{aligned}$$

Je možné, že autori využili rovnosti $E\left[G_{ak_1}(X_{ak_1})\right] = E\left[G_{bk_2}(X_{bk_2})\right] = 1/2$. Toto však vo všeobecnosti neplatí[7]. Analogický problém nastáva aj v rovnostiach (1.19), kedy $i_1 \neq i_2$ a $\ell_1 = \ell_2 = \ell$.

## 2.3 Kontrapríklad k dôkazu vety 1.6 [13, Theorem 1]

Nech $m = J = m_1 = 1$, $F = U[0;1]$ a $G = U[0.5;1.5]$. Teda $X = X_{11} \sim F_{11} = F$ a $Y = Y_{11} \sim G_{11} = G$. Majme dva nezávislé náhodné výbery $X_i$ $(i = 1, \ldots, n_x)$ a $Y_\ell$ $(\ell = 1, \ldots, n_y)$ z rozdelení $F$ a $G$. Nech ďalej platí, že $n_x = n_y$, teda $n_x/n_y = 1 = \lambda$. V tejto podkapitole ukážeme, že správna asymptotická kovariančná matica $\Sigma$ vo vete 1.6 [13, Theorem 1] je $5/48$, ale z vety 1.6 [13, Theorem 1] pomocou (1.12) dostávame, že $\Sigma = -103/48$.

Distribučné funkcie $X$, $Y$ sú

$$F_{11}(x) = \begin{cases} 0 & x < 0, \\ x & x \in [0;1], \\ 1 & x > 1, \end{cases} \qquad G_{11}(y) = \begin{cases} 0 & y < 0.5, \\ y - \frac{1}{2} & y \in [0.5;1.5], \\ 1 & y > 1.5 \end{cases}$$

a kvantilové funkcie sú

$$F_{11}^{-}(p) = p, \qquad G_{11}^{-}(q) = q + \frac{1}{2} \qquad p, q \in (0;1). \tag{2.4}$$

Platí, že

$$\theta_{11} = P\left(X_{11} < Y_{11}\right) - P\left(X_{11} > Y_{11}\right) = \frac{7}{8} - \frac{1}{8} = \frac{3}{4}.$$

[7] Rovnosti platia napríklad pri splnení zovšeobecnenej Behrensovej-Fisherovej hypotézy (1.3), ako uvádzajú autori v článku [5, str. 584].

Pre $i = 1, \ldots, n_x$ a $\ell = 1, \ldots, n_y$ má náhodná premenná

$$\Psi_{i\ell 11} = I(X_{i11} < Y_{\ell 11}) - I(X_{i11} > Y_{\ell 11})$$

nasledovné rozdelenie

$$\Psi_{i\ell 11} = \begin{cases} -1 & \text{pr. } \frac{1}{8}, \\ 1 & \text{pr. } \frac{7}{8}. \end{cases}$$

Z toho dostávame, že

$$E[\Psi_{i\ell 11}] = \frac{3}{4}, \qquad \text{var}(\Psi_{i\ell 11}) = \frac{7}{16}. \tag{2.5}$$

Začnime druhým prípadom z dôkazu vety 1.6 [13, Theorem 1], kde $i_1 = i_2 = i$ a $\ell_1 \neq \ell_2$. Počítame $\text{cov}\,[\Psi_{i\ell_1 11}, \Psi_{i\ell_2 11}]$. Dá sa ukázať, že

$$\Psi_{i\ell_1 11}\Psi_{i\ell_2 11} = \begin{cases} -1 & \text{pr. } \frac{1}{6}, \\ 1 & \text{pr. } \frac{5}{6}. \end{cases}$$

Z toho vyplýva, že

$$E[\Psi_{i\ell_1 11}\Psi_{i\ell_2 11}] = \frac{2}{3}.$$

Nakoniec platí, že

$$\text{cov}\,[\Psi_{i\ell_1 11}, \Psi_{i\ell_2 11}] = E[\Psi_{i\ell_1 11}\Psi_{i\ell_2 11}] - E[\Psi_{i\ell_1 11}]E[\Psi_{i\ell_2 11}] = \frac{5}{48}.$$

Analogicky pre tretí prípad z dôkazu vety 1.6 [13, Theorem 1], kde $i_1 \neq i_2$ a $\ell_1 = \ell_2 = \ell$, dostávame, že

$$\text{cov}\,[\Psi_{i_1 \ell 11}, \Psi_{i_2 \ell 11}] = \frac{5}{48}.$$

Nakoniec počítajme kovarianciu v štvrtom prípade z dôkazu vety 1.6 [13, Theorem 1], kde $i_1 = i_2 = i$ a $\ell_1 = \ell_2 = \ell$. Z (2.5) máme

$$\text{cov}\,[\Psi_{i\ell 11}, \Psi_{i\ell 11}] = \text{var}(\Psi_{i\ell 11}) = \frac{7}{16}.$$

Kovariančná matica z (1.16) je rovná

$$\begin{aligned}
&\operatorname{cov}\left[\frac{1}{\sqrt{N}}\left(\bar{R}_{y\cdot 1\cdot}-\bar{R}_{x\cdot 1\cdot}\right),\frac{1}{\sqrt{N}}\left(\bar{R}_{y\cdot 1\cdot}-\bar{R}_{x\cdot 1\cdot}\right)\right]\\
&=\frac{N}{4n_x^2n_y^2}\left[\sum_{i=1}^{n_x}\sum_{\ell_1=1}^{n_y}\sum_{\ell_2=1}^{n_y}\operatorname{cov}\left[\Psi_{i\ell_1 11},\Psi_{i\ell_2 11}\right]+\sum_{i_1=1}^{n_x}\sum_{i_2=1}^{n_x}\sum_{\ell=1}^{n_y}\operatorname{cov}\left[\Psi_{i_1\ell 11},\Psi_{i_2\ell 11}\right]\right.\\
&\qquad\left.+\sum_{i=1}^{n_x}\sum_{\ell=1}^{n_y}\operatorname{cov}\left[\Psi_{i\ell 11},\Psi_{i\ell 11}\right]\right]\\
&=\frac{N(n_y-1)}{4n_xn_y}\cdot\frac{5}{48}+\frac{N(n_x-1)}{4n_xn_y}\cdot\frac{5}{48}+\frac{N}{4n_xn_y}\cdot\frac{7}{16}\\
&=\frac{2n_x(n_x-1)}{4n_xn_x}\cdot\frac{5}{48}+\frac{2n_x(n_x-1)}{4n_xn_x}\cdot\frac{5}{48}+\frac{2n_x}{4n_xn_x}\cdot\frac{7}{16}\\
&=\frac{n_x-1}{n_x}\cdot\frac{5}{48}+\frac{1}{2n_x}\cdot\frac{7}{16}=:\Sigma_{n_x}.
\end{aligned}$$

Teda pre $n_x\to\infty$ platí, že

$$\lim_{n_x\to\infty}\Sigma_{n_x}=\lim_{n_x\to\infty}\left(\frac{n_x-1}{n_x}\cdot\frac{5}{48}+\frac{2}{n_x}\cdot\frac{7}{16}\right)=\frac{5}{48}=:\Sigma_{\text{správna}}. \tag{2.6}$$

Ďalej sa pozrieme na chybný vzorec (1.12) pre kovariančnú maticu zo znenia vety 1.6 [13, Theorem 1]. Podobne ako vyššie, začneme s $c_{11,11}$. Budeme počítať

$$c_{11,11}=\operatorname{cov}\left[G_{11}(X_{11}),G_{11}(X_{11})\right]=\operatorname{var}\left[G_{11}(X_{11})\right]. \tag{2.7}$$

Označme si náhodnú premennú $G_{11}(X_{11})$ symbolom $Z$ a počítajme jej distribučnú funkciu

$$\tilde{G}(z)=P(Z<z)=P(G_{11}(X_{11})<z).$$

Uvažujme tri prípady. Ak $z\le 0$, tak $P(G_{11}(X_{11})<z)=0$. Podobne ak $z\ge 1$, tak $P(G_{11}(X_{11})<z)=1$. Pre $z\in(0;1)$ pomocou (2.4) dostávame, že

$$P(G_{11}(X_{11})<z)=P(X_{11}<G_{11}^{-}(z))=P\left(X_{11}<z+\frac{1}{2}\right)=\begin{cases}z+\frac{1}{2}, & \text{ak } z\in\left(0,\frac{1}{2}\right]\\ 1, & \text{ak } z>\frac{1}{2}.\end{cases}$$

Teda distribučná funkcia náhodnej premennej $Z$ je

$$\tilde{G}(z)=\begin{cases}0 & z\le 0,\\ z+\frac{1}{2} & z\in\left(0;\frac{1}{2}\right],\\ 1 & z>\frac{1}{2}.\end{cases}$$

Ak si zoberieme dve náhodné premenné $W_1\equiv 0$ a $W_2\sim U[0;0.5]$, tak si môžeme všimnúť, že distribučná funkcia $\tilde{G}$ je konvexnou kombináciou distribučných funkcií náhodných premenných $W_1$ a $W_2$.

**Lema 2.3.** *Majme náhodnú premennú*

$$\tilde{Z} = B \cdot W_1 + (1 - B)W_2,$$

*pričom* $B \sim Alt(0.5)$, $W_1 \equiv 0$, $W_2 \sim U[0; 0.5]$ *a* $B, W_1, W_2$ *sú navzájom nezávislé. Potom* $\tilde{Z} \stackrel{d}{=} Z$ *t. j.* $\tilde{Z}$ *a* $Z$ *majú zhodnú distribúciu.*

*Dôkaz.* Využitím vety o iterovanej strednej hodnote máme:

$$P(\tilde{Z} < \tilde{z}) = E[I(\tilde{Z} < \tilde{z})] = E[E[I(\tilde{Z} < \tilde{z}) \mid B]].$$

Zrejme

$$E[I(\tilde{Z} < \tilde{z}) \mid B] = \begin{cases} E[I(W_1 < \tilde{z})] & \text{ak} \quad B = 1, \\ E[I(W_2 < \tilde{z})] & \text{ak} \quad B = 0. \end{cases}$$

Z toho dostávame, že

$$\begin{aligned} P(\tilde{Z} < \tilde{z}) &= E[I(W_1 < \tilde{z})]P(B = 1) + E[I(W_2 < \tilde{z})]P(B = 0) \\ &= \frac{1}{2}P(W_1 < \tilde{z}) + \frac{1}{2}P(W_2 < \tilde{z}) \\ &= G(\tilde{z}). \end{aligned}$$

Teda $P(\tilde{Z} < \tilde{z}) = P(Z < \tilde{z})$ pre ľubovoľné $\tilde{z} \in \mathbb{R}$, čiže $\tilde{Z} \stackrel{d}{=} Z$. □

Analogicky sa dá ukázať, že pre ľubovoľnú borelovskú funkciu $h : \mathbb{R} \to \mathbb{R}$ platí nasledujúca rovnosť za predpokladu, že stredné hodnoty na pravej strane sú konečné

$$E[h(\tilde{Z})] = \frac{1}{2}E[h(W_1)] + \frac{1}{2}E[h(W_2)].$$

Môžeme teda počítať varianciu $\tilde{Z}$

$$\begin{aligned} E[\tilde{Z}] &= \frac{1}{2}E[W_1] + \frac{1}{2}E[W_2] = \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{8} \\ E[\tilde{Z}^2] &= \frac{1}{2}E[W_1^2] + \frac{1}{2}E[W_2^2] = \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot \frac{1}{12} = \frac{1}{24} \\ \text{var}[\tilde{Z}] &= \frac{1}{24} - \left(\frac{1}{8}\right)^2 = \frac{5}{192}. \end{aligned} \tag{2.8}$$

Pomocou lemy 2.3 môžeme dokončiť výpočet (2.7)

$$c_{11,11} = \text{var}(Z) = \text{var}(\tilde{Z}) = \frac{5}{192}.$$

Postup pre $d_{11,11}$ je analogický. Dá sa ukázať, že distribučná funkcia náhodnej premennej $F_{11}(Y_{11})$ je

$$\tilde{F}(z) = \begin{cases} 0 & z \leq \frac{1}{2}, \\ z - \frac{1}{2} & z \in \left(\frac{1}{2}; 1\right], \\ 1 & z > \frac{1}{2}. \end{cases} \tag{2.9}$$

Podobne platí, že

$$d_{11,11} = \operatorname{cov}\left[F_{11}(Y_{11}), F_{11}(Y_{11})\right] = \operatorname{var}\left[F_{11}(Y_{11})\right] = \frac{5}{192}.$$

Z vety 1.6 [13, Theorem 1] dostávame, že

$$\begin{aligned} \Sigma &= \left(1 + \frac{1}{\lambda}\right) c_{11,11} + (1+\lambda) d_{11,11} - \left(\sqrt{\lambda} + \frac{1}{\sqrt{\lambda}}\right)^2 \theta_{11}\theta_{11} \\ &= 2 \cdot \frac{5}{192} + 2 \cdot \frac{5}{192} - 4 \cdot \left(\frac{3}{4}\right)^2 \\ &= \frac{-103}{48}. \end{aligned}$$

Vidíme, že vzorec (1.12) z vety 1.6 [13, Theorem 1] nielenže nedáva rovnaký výsledok ako (2.6) ale dokonca je záporný, čo je spor s kladnou semidefinitnosťou kovariančnej matice.

*Poznámka*: Keby sme zakomponovali opravu z kapitoly 2.1, tak kovariančná matica by bola

$$\begin{aligned} \Sigma &= \left(1 + \frac{1}{\lambda}\right) c_{11,11} + (1+\lambda) d_{11,11} - \frac{1}{4}\left(\sqrt{\lambda} + \frac{1}{\sqrt{\lambda}}\right)^2 \theta_{11}\theta_{11} \\ &= 2 \cdot \frac{5}{192} + 2 \cdot \frac{5}{192} - \frac{1}{4} \cdot 4 \cdot \left(\frac{3}{4}\right)^2 \\ &= \frac{-11}{24}. \end{aligned}$$

Teda vidíme, že aj po oprave chyby z (1.21) nie je $\Sigma$ správna a stále nie je pozitívne semidefinitná.

## 2.4 Oprava vety 1.6 [13, Theorem 1]

Na to, aby sme dostali korektné znenie vety 1.6 [13, Theorem 1], treba ešte opraviť série rovností (1.18) a (1.19) pri výpočte kovariančnej matice $\Sigma$. Toto dosiahneme jednoducho tým, že v nich vynecháme poslednú rovnosť. Dostaneme

$$\begin{aligned} \operatorname{cov}\left[\Psi_{i\ell_1 ak_1}, \Psi_{i\ell_2 bk_2}\right] &= 4E\left[\left\{G_{ak_1}(X_{ak_1}) - \frac{1}{2}\right\}\left\{G_{bk_2}(X_{bk_2}) - \frac{1}{2}\right\}\right] - \theta_{ak_1}\theta_{bk_2} \\ \operatorname{cov}\left[\Psi_{i_1 \ell ak_1}, \Psi_{i_2 \ell bk_2}\right] &= 4E\left[\left\{F_{ak_1}(Y_{ak_1}) - \frac{1}{2}\right\}\left\{F_{bk_2}(Y_{bk_2}) - \frac{1}{2}\right\}\right] - \theta_{ak_1}\theta_{bk_2}. \end{aligned}$$

Znenie opravenej vety 1.6 [13, Theorem 1] je teda podobné s jej originálnym znením. Zmena nastáva v tom, že položíme

$$c'_{ak_1,bk_2} = E\left[\left\{G_{ak_1}(X_{ak_1}) - \frac{1}{2}\right\}\left\{G_{bk_2}(X_{bk_2}) - \frac{1}{2}\right\}\right]$$
$$d'_{ak_1,bk_2} = E\left[\left\{F_{ak_1}(Y_{ak_1}) - \frac{1}{2}\right\}\left\{F_{bk_2}(Y_{bk_2}) - \frac{1}{2}\right\}\right]$$

a dostávame teda nový tvar kovariančnej matice z (1.12), ktorú označíme $\Sigma^{(1)} = (\sigma^{(1)}_{ab})_{J\times J}$, kde

$$\sigma^{(1)}_{ab} = \frac{1}{m_a m_b}\sum_{k_1=1}^{m_a}\sum_{k_2=1}^{m_b}\left\{\left(1+\frac{1}{\lambda}\right)c'_{ak_1,bk_2} + (1+\lambda)d'_{ak_1,bk_2} - \frac{1}{4}\left(\sqrt{\lambda}+\frac{1}{\sqrt{\lambda}}\right)^2\theta_{ak_1}\theta_{bk_2}\right\}. \tag{2.10}$$

*Poznámka*: Úplné znenie opravenej vety neuvádzame, nakoľko v kapitole 2.7 odvodíme ekvivalentné tvrdenie s iným tvarom kovariančnej matice; viď veta 2.1.

*Poznámka*: Po tejto oprave je asymptotická kovariančná matica $\Sigma^{(1)}$ v poriadku aj pre kontrapríklad z kapitoly 2.3. Počítajme

$$\begin{aligned} c'_{11,11} &= E\left[\left\{G_{11}(X_{11}) - \frac{1}{2}\right\}\left\{G_{11}(X_{11}) - \frac{1}{2}\right\}\right] \\ &= E\left[G_{11}(X_{11})^2\right] - E\left[G_{11}(X_{11})\right] + \frac{1}{4}. \end{aligned}$$

Pomocou výsledkov z (2.8) dostávame, že

$$\begin{aligned} c'_{11,11} &= E\left[\tilde{Z}^2\right] - E\left[\tilde{Z}\right] + \frac{1}{4} \\ &= \frac{1}{24} - \frac{1}{8} + \frac{1}{4} = \frac{1}{6}. \end{aligned}$$

Analogicky by sme vedeli ukázať, že

$$d'_{11,11} = E\left[\left\{F_{11}(Y_{11}) - \frac{1}{2}\right\}\left\{F_{11}(Y_{11}) - \frac{1}{2}\right\}\right] = \frac{1}{6}.$$

Po dosadení do (2.10) dostaneme, že $\Sigma^{(1)}$ je $5/48$, čo sa zhoduje so $\Sigma_{\text{správna}}$ vypočítanou v (2.6).

## 2.5 Oprava vety 1.7 [13, Theorem 2]

Kvôli chybe vety 1.6 [13, Theorem 1] je nutné opraviť aj vetu 1.7 [13, Theorem 2]. Podobne ako v predchádzajúcej kapitole, kde sme ukázali, že treba zmeniť predpis členov $c_{ak_1,bk_2}$

a $d_{ak_1,bk_2}$, tak aj v oprave vety 1.7 [13, Theorem 2] treba zmeniť ich odhady. Tie sú nasledovné

$$\begin{aligned}
\hat{c}'_{ak_1,bk_2} &= \hat{E}\left[\left\{G_{ak_1}(X_{ak_1}) - \frac{1}{2}\right\}\left\{G_{bk_2}(X_{bk_2}) - \frac{1}{2}\right\}\right] \\
&= \frac{1}{n_x}\sum_{i=1}^{n_x}\left(\left\{\frac{1}{n_y}\sum_{\ell=1}^{n_y} I(Y_{\ell ak_1} < X_{iak_1}) - \frac{1}{2}\right\}\left\{\frac{1}{n_y}\sum_{\ell=1}^{n_y} I(Y_{\ell bk_2} < X_{ibk_2}) - \frac{1}{2}\right\}\right), \\
\hat{d}'_{ak_1,bk_2} &= \hat{E}\left[\left\{F_{ak_1}(Y_{ak_1}) - \frac{1}{2}\right\}\left\{F_{bk_2}(Y_{bk_2}) - \frac{1}{2}\right\}\right] \\
&= \frac{1}{n_y}\sum_{\ell=1}^{n_y}\left(\left\{\frac{1}{n_x}\sum_{i=1}^{n_x} I(X_{iak_1} < Y_{\ell ak_1}) - \frac{1}{2}\right\}\left\{\frac{1}{n_x}\sum_{i=1}^{n_x} I(X_{ibk_2} < Y_{\ell bk_2}) - \frac{1}{2}\right\}\right).
\end{aligned}$$

Následne kovariančná matica $S^{(1)}$ odhadujúca opravenú maticu $\Sigma^{(1)}$ má prvky

$$s_{ab}^{(1)} = \frac{1}{m_a m_b}\sum_{k_1=1}^{m_a}\sum_{k_2=1}^{m_b}\left\{\left(1+\frac{1}{\lambda}\right)\hat{c}'_{ak_1,bk_2} + (1+\lambda)\hat{d}'_{ak_1,bk_2} - \frac{1}{4}\left(\sqrt{\lambda}+\frac{1}{\sqrt{\lambda}}\right)^2\hat{\theta}_{ak_1}\hat{\theta}_{bk_2}\right\}.$$

Zvyšok znenia vety ostáva rovnaký. Úplné znenie podobne ako v minulej kapitole neuvádzame, nakoľko v kapitole 2.8 uvedieme iný odhad kovariančnej matice.

## 2.6 Oprava vety 1.7 [13, Theorem 2] - neasymptotický odhad

Druhou možnosťou, ako odhadovať kovariančnú maticu $\Sigma$, by bolo zvoliť „neasymptotický“ odhad. To znamená nevyužiť aproximácie

$$\frac{n_x}{n_y} \approx \lambda,\ \frac{1}{n_x} \approx 0,\ \frac{1}{n_y} \approx 0, \ldots$$

V takomto prípade máme

$$\begin{aligned}
&\operatorname{cov}\left[\frac{1}{\sqrt{N}}\left(\bar{R}_{y\cdot a\cdot} - \bar{R}_{x\cdot a\cdot}\right), \frac{1}{\sqrt{N}}\left(\bar{R}_{y\cdot b\cdot} - \bar{R}_{x\cdot b\cdot}\right)\right] \\
&= \frac{1}{m_a m_b}\sum_{k_1=1}^{m_a}\sum_{k_2=1}^{m_b}\left\{\frac{N(n_y-1)}{n_x n_y}c'_{ak_1,bk_2} + \frac{N(n_x-1)}{n_x n_y}d'_{ak_1,bk_2}\right. \\
&\qquad\left. + \frac{N}{4n_x n_y}E\left(\Psi_{ak_1}\Psi_{bk_2}\right) - \frac{N\left(N-1\right)}{4n_x n_y}\theta_{ak_1}\theta_{bk_2}\right\}.
\end{aligned}$$

Odhady pre $c'_{ak_1,bk_2}$ a $d'_{ak_1,bk_2}$ môžeme použiť z predchádzajúcej kapitoly 2.5 a za odhad $\theta$ zoberieme odhad z vety 1.7 [13, Theorem 2]. Ostáva nám teda nájsť odhad výrazu $E\left(\Psi_{ak_1}\Psi_{bk_2}\right)$. Rozpíšme, čomu sa rovná táto stredná hodnota:

$$\begin{aligned}
E\left(\Psi_{ak_1}\Psi_{bk_2}\right) = E\Big[&\left\{I(X_{ak_1} < Y_{ak_1}) - I(X_{ak_1} > Y_{ak_1})\right\} \\
&\cdot\left\{I(X_{bk_2} < Y_{bk_2}) - I(X_{bk_2} > Y_{bk_2})\right\}\Big].
\end{aligned}$$

Rozumným odhadom tejto strednej hodnoty je priemer, teda odhad vyzerá nasledovne

$$\hat{E}\left(\Psi_{ak_1}\Psi_{bk_2}\right) = \frac{1}{n_x}\frac{1}{n_y}\sum_{i=1}^{n_x}\sum_{\ell=1}^{n_y}\left\{I(X_{iak_1} < Y_{\ell ak_1}) - I(X_{iak_1} > Y_{\ell ak_1})\right\} \cdot \left\{I(X_{ibk_2} < Y_{\ell bk_2}) - I(X_{ibk_2} > Y_{\ell bk_2})\right\}.$$

Označme odhad $\hat{E}\left(\Psi_{ak_1}\Psi_{bk_2}\right) = \hat{e}'_{ak_1,bk_2}$.

Finálnym neasymptotickým odhadom skutočnej kovariančnej matice by bola matica $S^{(1e)} = (s_{ab}^{(1e)})_{J\times J}$, kde

$$s_{ab}^{(1e)} = \frac{1}{m_a m_b}\sum_{k_1=1}^{m_a}\sum_{k_2=1}^{m_b}\left\{\frac{N(n_y-1)}{n_x n_y}\hat{c}'_{ak_1,bk_2} + \frac{N(n_x-1)}{n_x n_y}\hat{d}'_{ak_1,bk_2} + \frac{N}{4n_x n_y}\hat{e}'_{ak_1,bk_2} - \frac{N\left(N-1\right)}{4n_x n_y}\hat{\theta}_{ak_1}\hat{\theta}_{bk_2}\right\}.$$

## 2.7 Oprava vety 1.6 [13, Theorem 1] - finálna verzia; dôkaz vety 2.1

V tejto kapitole odvodíme iný spôsob vyjadrenia kovariančnej matice, ktorý má „elegantnejší“ zápis ako ten z kapitoly 2.4 a ktorý je využitý vo vete 2.1. Tento postup následne využijeme aj v kapitole 3.3, kde sa budeme venovať odvodeniu $T_{cMAX}$ testu pre všeobecný prípad rozdelení $X$ a $Y$ bez nutnosti splnenia predpokladu (1.1). Najprv si však uvedieme pomocnú lemu.

**Lema 2.4.** *Nech $X$, $Y$ sú $m$-rozmerné nezávislé náhodné vektory. Majme dva nezávislé náhodné výbery*

$$X_i\ (i = 1, \ldots, n_x) \quad a \quad Y_\ell\ (\ell = 1, \ldots, n_y)$$

*z rozdelení $X$, $Y$. Nech $\bigcirc$, $\square$ sú dve borelovské zobrazenia z $\mathbb{R}\times\mathbb{R}$ do $\{Pravda, Nepravda\}$. Nech $g_{ak_1}, h_{bk_2}\colon \mathbb{R}\to\mathbb{R}$ sú funkcie definované vzťahmi*

$$g_{ak_1}(x) = P(x \bigcirc Y_{ak_1}),\ h_{bk_2}(x) = P(x\ \square\ Y_{bk_2}).$$

*Potom pre všetky $a, b \in \{1, 2, \ldots, J\}$, $k_1 \in \{1, 2, \ldots, m_a\}$, $k_2 \in \{1, 2, \ldots, m_b\}$, $i \in \{1, 2, \ldots, n_x\}$ a rôzne $\ell_1, \ell_2 \in \{1, 2, \ldots, n_y\}$ platí, že*

$$E\left[I(X_{iak_1} \bigcirc Y_{\ell_1 ak_1})\right] = E\left[g_{ak_1}\left(X_{ak_1}\right)\right],$$

$$E\left[I(X_{ibk_2}\ \square\ Y_{\ell_2 bk_2})\right] = E\left[h_{bk_2}\left(X_{bk_2}\right)\right],$$

$$E\left[I(X_{iak_1} \bigcirc Y_{\ell_1 ak_1})I(X_{ibk_2}\ \square\ Y_{\ell_2 bk_2})\right] = E\left[g_{ak_1}\left(X_{ak_1}\right)h_{bk_2}\left(X_{bk_2}\right)\right],$$

$$\operatorname{cov}\left[I(X_{iak_1} \bigcirc Y_{\ell_1 ak_1}), I(X_{ibk_2}\ \square\ Y_{\ell_2 bk_2})\right] = \operatorname{cov}\left[g_{ak_1}\left(X_{ak_1}\right), h_{bk_2}\left(X_{bk_2}\right)\right].$$

*Dôkaz.* Z definície kovariancie dostávame

$$\begin{aligned}\operatorname{cov}\left[I(X_{iak_1} \bigcirc Y_{\ell_1 ak_1}), I(X_{ibk_2} \square Y_{\ell_2 bk_2})\right] &= E\left[I(X_{iak_1} \bigcirc Y_{\ell_1 ak_1}) I(X_{ibk_2} \square Y_{\ell_2 bk_2})\right] \\ &\quad - E\left[I(X_{iak_1} \bigcirc Y_{\ell_1 ak_1})\right] E\left[I(X_{ibk_2} \square Y_{\ell_2 bk_2})\right].\end{aligned}$$

Ďalej si označíme $\tilde{\alpha} = E\left[I(X_{iak_1} \bigcirc Y_{\ell_1 ak_1}) I(X_{ibk_2} \square Y_{\ell_2 bk_2})\right]$, $\tilde{\beta}_a = E\left[I(X_{iak_1} \bigcirc Y_{\ell_1 ak_1})\right]$ a $\tilde{\beta}_b = E\left[I(X_{ibk_2} \square Y_{\ell_2 bk_2})\right]$.

Pozrime sa najskôr na člen $\tilde{\alpha}$. Z vety o iterovanej strednej hodnote vieme, že

$$\tilde{\alpha} = E\Big[E\left[I(X_{iak_1} \bigcirc Y_{\ell_1 ak_1}) I(X_{ibk_2} \square Y_{\ell_2 bk_2}) \mid X_{iak_1}, X_{ibk_2}\right]\Big]$$

Keďže podmieňujeme $\sigma$-algebrou generovanou $X_{iak_1}$ a $X_{ibk_2}$ a $\ell_1 \neq \ell_2$, tak náhodné premenné $I(X_{iak_1} \bigcirc Y_{\ell_1 ak_1})$ a $I(X_{ibk_2} \square Y_{\ell_2 bk_2})$ sú (podmienene) nezávislé. Môžeme teda písať

$$\tilde{\alpha} = E\Big[E\left[I(X_{iak_1} \bigcirc Y_{\ell_1 ak_1}) \mid X_{iak_1}, X_{ibk_2}\right] \cdot E\left[I(X_{ibk_2} \square Y_{\ell_2 bk_2}) \mid X_{iak_1}, X_{ibk_2}\right]\Big].$$

Využitím lemy 6.5 na $E\left[I(X_{iak_1} \bigcirc Y_{\ell_1 ak_1}) | X_{iak_1}, X_{ibk_2}\right]$ a $E\left[I(X_{ibk_2} \square Y_{\ell_2 bk_2}) | X_{iak_1}, X_{ibk_2}\right]$ dostávame, že

$$\tilde{\alpha} = E\Big[P\left(X_{iak_1} \bigcirc Y_{\ell_1 ak_1} \mid X_{iak_1}\right) P\left(X_{ibk_2} \square Y_{\ell_2 bk_2} \mid X_{ibk_2}\right)\Big].$$

Keďže $Y_{\ell_1}, Y_{\ell_2}$ a $Y$ majú rovnaké rozdelenie, platí $g_{ak_1}(x) = P(x \bigcirc Y_{\ell_1 ak_1})$, $h_{bk_2}(x) = P(x \square Y_{\ell_2 bk_2})$. Využitím lemy 6.4 dostávame, že

$$\tilde{\alpha} = E\Big[g_{ak_1}\left(X_{iak_1}\right) h_{bk_2}\left(X_{ibk_2}\right)\Big].$$

Ďalej využijeme, že vektory $X_i$ a $X$ majú rovnaké rozdelenie a preto

$$\tilde{\alpha} = E\Big[g_{ak_1}\left(X_{ak_1}\right) h_{bk_2}\left(X_{bk_2}\right)\Big].$$

Pre člen $\tilde{\beta}_a$ použitím vety o iterovanej strednej hodnote dostávame, že

$$\begin{aligned}\tilde{\beta}_a &= E\Big[E\left[I(X_{iak_1} \bigcirc Y_{\ell_1 ak_1}) \mid X_{iak_1}\right]\Big] \\ &= E\Big[P\left(X_{iak_1} \bigcirc Y_{\ell_1 ak_1} \mid X_{iak_1}\right)\Big].\end{aligned}$$

Podobne ako pri úpravách výrazu $\tilde{\alpha}$ máme

$$\tilde{\beta}_a = E\Big[g_{ak_1}\left(X_{ak_1}\right)\Big].$$

Podobne dostaneme, že $\tilde{\beta}_b = E\Big[h_{bk_2}\left(X_{bk_2}\right)\Big]$. Máme teda, že

$$\begin{aligned}
&\operatorname{cov}\left[I(X_{iak_1} \bigcirc Y_{\ell_1 ak_1}), I(X_{ibk_2} \square Y_{\ell_2 bk_2})\right] \\
&= E\Big[g_{ak_1}\left(X_{ak_1}\right) h_{bk_2}\left(X_{bk_2}\right)\Big] - E\Big[g_{ak_1}\left(X_{ak_1}\right)\Big] E\Big[h_{bk_2}\left(X_{bk_2}\right)\Big] \\
&= \operatorname{cov}\left[g_{ak_1}\left(X_{ak_1}\right), h_{bk_2}\left(X_{bk_2}\right)\right].
\end{aligned}$$

□

Teraz prejdeme k samotnému odvodeniu korektného tvaru zložiek kovariančnej matice z vety 1.6 [13, Theorem 1]. Najprv sa pozrieme na druhý prípad, kedy $i_1 = i_2 = i$ a $l_1 \neq l_2$. Z definície (1.6) dostávame, že

$$\begin{aligned}
\operatorname{cov}\left[\Psi_{i\ell_1 ak_1}, \Psi_{i\ell_2 bk_2}\right] = \operatorname{cov}\Big[\; & I(X_{iak_1} < Y_{\ell_1 ak_1}) - I(X_{iak_1} > Y_{\ell_1 ak_1}), \\
& I(X_{ibk_2} < Y_{\ell_2 bk_2}) - I(X_{ibk_2} > Y_{\ell_2 bk_2})\Big].
\end{aligned}$$

Za platnosti (1.1) platí, že

$$\operatorname{cov}\left[\Psi_{i\ell_1 ak_1}, \Psi_{i\ell_2 bk_2}\right] = \operatorname{cov}\left[1 - 2I(X_{iak_1} > Y_{\ell_1 ak_1}), 1 - 2I(X_{ibk_2} > Y_{\ell_2 bk_2})\right].$$

Následne využijeme bilineárnosť kovariancie

$$\operatorname{cov}\left[\Psi_{i\ell_1 ak_1}, \Psi_{i\ell_2 bk_2}\right] = 4\operatorname{cov}\left[I(X_{iak_1} > Y_{\ell_1 ak_1}), I(X_{ibk_2} > Y_{\ell_2 bk_2})\right]$$

Teraz využijeme lemu 2.4, kde za $\bigcirc$ aj $\square$ dosadíme $>$. Dostaneme, že

$$\operatorname{cov}\left[\Psi_{i\ell_1 ak_1}, \Psi_{i\ell_2 bk_2}\right] = 4\operatorname{cov}\left[g_{ak_1}\left(X_{ak_1}\right), h_{bk_2}\left(X_{bk_2}\right)\right].$$

Keďže funkcia $g_{ak_1}$ je v tomto prípade definovaná vzťahom $P(x > Y_{ak_1})$, tak vidíme, že ide o distribučnú funkciu náhodnej premennej $Y_{ak_1}$ (t. j. $G_{ak_1}$). Podobne $h_{bk_2}$ je distribučná funkcia $Y_{bk_2}$ (t. j. $G_{bk_2}$). Dostávame, že

$$\operatorname{cov}\left[\Psi_{i\ell_1 ak_1}, \Psi_{i\ell_2 bk_2}\right] = 4\operatorname{cov}\left[G_{ak_1}\left(X_{ak_1}\right), G_{bk_2}\left(X_{bk_2}\right)\right].$$

Analogicky pre tretí prípad, kedy $i_1 \neq i_2$ a $\ell_1 = \ell_2 = \ell$, dostaneme opravu vzťahu (1.19):

$$\operatorname{cov}\left[\Psi_{i_1 \ell ak_1}, \Psi_{i_2 \ell bk_2}\right] = 4\operatorname{cov}\left[F_{ak_1}(Y_{ak_1}), F_{k_2}(Y_{bk_2})\right].$$

Pre štvrtý prípad, kedy $i_1 = i_2 = i$ a $\ell_1 = \ell_2 = \ell$, dostaneme

$$\begin{aligned}
&\operatorname{cov}\left[\Psi_{i\ell ak_1}, \Psi_{i\ell bk_2}\right] \\
&= \operatorname{cov}\left[I(X_{iak_1} < Y_{\ell ak_1}) - I(X_{iak_1} > Y_{\ell ak_1}), I(X_{ibk_2} < Y_{\ell bk_2}) - I(X_{ibk_2} > Y_{\ell bk_2})\right] \\
&= \operatorname{cov}\left[1 - 2I(X_{iak_1} > Y_{\ell ak_1}), 1 - 2I(X_{ibk_2} > Y_{\ell bk_2})\right] \\
&= 4\operatorname{cov}\left[I(X_{iak_1} > Y_{\ell ak_1}), I(X_{ibk_2} > Y_{\ell bk_2})\right] \\
&= 4\operatorname{cov}\left[I(X_{ak_1} > Y_{ak_1}), I(X_{bk_2} > Y_{bk_2})\right].
\end{aligned}$$

Dosadením do (1.16) dostávame, že

$$\begin{aligned}
&\operatorname{cov}\left[\frac{1}{\sqrt{N}}(\bar{R}_{x\cdot a\cdot} - \bar{R}_{y\cdot a\cdot}), \frac{1}{\sqrt{N}}(\bar{R}_{x\cdot b\cdot} - \bar{R}_{y\cdot b\cdot})\right] \\
&= \frac{1}{m_a m_b}\sum_{k_1=1}^{m_a}\sum_{k_2=1}^{m_b}\left\{\frac{N(n_y-1)}{n_x n_y}\operatorname{cov}\left[G_{ak_1}(X_{ak_1}), G_{bk_2}(X_{bk_2})\right]\right. \\
&\quad + \frac{N(n_x-1)}{n_x n_y}\operatorname{cov}\left[F_{ak_1}(Y_{ak_1}), F_{bk_2}(Y_{bk_2})\right] \\
&\quad \left. + \frac{N}{n_x n_y}\operatorname{cov}\left[I(X_{ak_1} > Y_{ak_1}), I(X_{bk_2} > Y_{bk_2})\right]\right\}.
\end{aligned}$$

Označme:

- $c_{ak_1,bk_2} = \operatorname{cov}\left[G_{ak_1}(X_{ak_1}), G_{bk_2}(X_{bk_2})\right]$,
- $d_{ak_1,bk_2} = \operatorname{cov}\left[F_{ak_1}(Y_{ak_1}), F_{bk_2}(Y_{bk_2})\right]$,
- $e_{ak_1,bk_2} = \operatorname{cov}\left[I(X_{ak_1} > Y_{ak_1}), I(X_{bk_2} > Y_{bk_2})\right]$.

Teda

$$\begin{aligned}
&\operatorname{cov}\left[\frac{1}{\sqrt{N}}(\bar{R}_{x\cdot a\cdot} - \bar{R}_{y\cdot a\cdot}), \frac{1}{\sqrt{N}}(\bar{R}_{x\cdot b\cdot} - \bar{R}_{y\cdot b\cdot})\right] \\
&= \frac{1}{m_a m_b}\sum_{k_1=1}^{m_a}\sum_{k_2=1}^{m_b}\left\{\frac{N(n_y-1)}{n_x n_y}c_{ak_1,bk_2} + \frac{N(n_x-1)}{n_x n_y}d_{ak_1,bk_2} + \frac{N}{n_x n_y}e_{ak_1,bk_2}\right\}.
\end{aligned} \tag{2.11}$$

Kovariancie $c_{ak_1,bk_2}$, $d_{ak_1,bk_2}$, $e_{ak_1,bk_2}$ nezávisia od veľkostí náhodných výberov $n_x, n_y$. Preto pre $\min\{n_x, n_y\} \to \infty$ za predpokladu, že $n_x/n_y \to \lambda \in (0, \infty)$, dostávame, že

$$\operatorname{cov}\left[\frac{1}{\sqrt{N}}(\bar{R}_{x\cdot a\cdot} - \bar{R}_{y\cdot a\cdot}), \frac{1}{\sqrt{N}}(\bar{R}_{x\cdot b\cdot} - \bar{R}_{y\cdot b\cdot})\right] \to \sigma^{(2)}_{ab},$$

kde

$$\sigma^{(2)}_{ab} = \frac{1}{m_a m_b}\sum_{k_1=1}^{m_a}\sum_{k_2=1}^{m_b}\left\{\left(1 + \frac{1}{\lambda}\right)c_{ak_1,bk_2} + (1+\lambda)\,d_{ak_1,bk_2}\right\}. \tag{2.12}$$

Tým sme dokázali vetu 2.1.

## 2.8 Oprava vety 1.7 [13, Theorem 2] - finálna verzia

Vo vete 1.7 [13, Theorem 2] dôjde k nasledovnej zmene a to, že

$$s_{ab}^{(2)} = \frac{1}{m_a m_b} \sum_{k_1=1}^{m_a} \sum_{k_2=1}^{m_b} \left\{ \left(1 + \frac{1}{\lambda}\right) \hat{c}_{ak_1,bk_2} + (1+\lambda)\, \hat{d}_{ak_1,bk_2} \right\}.$$

Dôkaz tejto vety (t. j. vety 2.2) neuvádzame, nakoľko bude uvedený v kapitole 3.6 pri všeobecnejších predpokladoch.

Mohli by sme zvoliť aj odhad založený na vzorci (2.11), ktorý by však bol menej „elegantný“. Ten by vyzeral nasledovne

$$s_{ab}^{(2e)} = \frac{1}{m_a m_b} \sum_{k_1=1}^{m_a} \sum_{k_2=1}^{m_b} \left\{ \frac{N(n_y-1)}{n_x n_y} \hat{c}_{ak_1,bk_2} + \frac{N(n_x-1)}{n_x n_y} \hat{d}_{ak_1,bk_2} + \frac{N}{n_x n_y} \hat{e}_{ak_1,bk_2} \right\}. \quad (2.13)$$

## 2.9 Dosadzovanie a odhadovanie $\theta$

V tejto kapitole sa vrátime k pôvodným zneniam vety 1.6 [13, Theorem 1] a vety 1.7 [13, Theorem 2], o ktorých sme ukázali, že sú chybné a uviedli sme ich korektné znenia vo vetách 2.1 a 2.2. V [14, str. 6] uvádzajú znenie vety [13, Theorem 1] špecificky v prípade, že testujeme zovšeobecnenú Behrensovu-Fisherovu hypotézu (1.3). Za predpokladu jej platnosti dostávame, že

$$\frac{1}{\sqrt{N}} \left( \bar{R}_{y\cdot 1\cdot} - \bar{R}_{x\cdot 1\cdot}, \ldots, \bar{R}_{y\cdot J\cdot} - \bar{R}_{x\cdot J\cdot} \right)^T \xrightarrow{d} N(0, \tilde{\Sigma}),$$

kde pre zložky kovariančnej matice $\tilde{\Sigma} = (\tilde{\sigma}_{ab})_{J\times J}$ platí, že

$$\tilde{\sigma}_{ab} = \frac{1}{m_a m_b} \sum_{k_1=1}^{m_a} \sum_{k_2=1}^{m_b} \left\{ \left(1 + \frac{1}{\lambda}\right) c_{ak_1,bk_2} + (1+\lambda) d_{ak_1,bk_2} \right\}. \quad (2.14)$$

Autori teda za výraz $\theta_{ak_1}\theta_{bk_2}$ dosadili nuly zo zovšeobecnenej Behrensovej-Fisherovej hypotézy. Keď sa následne pozrieme na [14, str. 3], tak pri výpočte matice $\tilde{S}$ (odhadu matice $\tilde{\Sigma}$) je vo výrazoch $\hat{c}_{ak_1,bk_2}$ a $\hat{d}_{ak_1,bk_2}$ parameter $\theta$ odhadovaný. Teda pri výpočte jednotlivých zložiek matice $\tilde{S}$ je parameter $\theta$ na jednom mieste dosadený podľa testovanej hypotézy a na druhom mieste je odhadovaný. Toto vo svojej podstate nie je chyba. My by sme však chceli mať parameter $\theta$ jednotný v celom odhade kovariančnej matice.

*Poznámka*: Môžeme si všimnúť, že tvar $T_{cMAX}$ testu, ktorý uvádzajú jeho autori, pri testovaní zovšeobecnenej Behrensovej-Fisherovej hypotézy, je zhodný s tvarom našej opravy z vety 2.1. Dôvodom je, že špecificky za platnosti tejto hypotézy a platnosti predpokladu (1.1) je tvrdenie vety 1.6 [13, Theorem 1] správne.

V tejto a nasledujúcej kapitole sa pozrieme na dve varianty, tentokrát už nami opraveného $T_{cMAX}$ testu. V prvej variante, ktorú označíme $T_{cMAX2}$, budeme parameter $\theta$ odhadovať vo výpočte kovariancií $\hat{c}_{ak_1,bk_2}$ (2.15) a $\hat{d}_{ak_1,bk_2}$ (2.17). Za odhad parametra $\theta$ zvolíme $\hat{\theta}$ popísaný v (1.7). Pre $T_{cMAX2}$ test bude odhadom kovariančnej matice matica $S^{(2)}$, ktorej zložky sú uvedené v (2.19). Test $T_{cMAX2}$ je testom na základ vety 2.2. V druhej variante, ktorú označíme $T_{cMAX3}$, budeme parameter $\theta$ v kovarianciách (2.16) a (2.18) dosadzovať podľa testovanej nulovej hypotézy. Odhadom kovariančnej matice $T_{cMAX3}$ testu bude matica $S^{(3)}$, ktorej zložky sú uvedené v (2.20).

Pozrieme sa aj na „neasymptotické“ verzie týchto testov založené na vzorci (2.13). Analogicky ako pri „asymptotickej“ verzii označíme $T_{cMAX2e}$ verziu, kde v odhade kovariančnej matici (2.21) sa parameter $\theta$ bude odhadovať, a $T_{cMAX3e}$ verziu, kde parameter $\theta$ budeme v odhade kovariančnej matici (2.22) dosadzovať podľa testovanej hypotézy.

$$\begin{aligned}\hat{c}_{ak_1,bk_2} = \frac{1}{n_x}\sum_{i=1}^{n_x}\Bigg(&\left\{\frac{1}{n_y}\sum_{\ell=1}^{n_y} I(Y_{\ell ak_1} < X_{iak_1}) - \frac{1-\hat{\theta}_{ak_1}}{2}\right\}\\ &\cdot\left\{\frac{1}{n_y}\sum_{\ell=1}^{n_y} I(Y_{\ell bk_2} < X_{ibk_2}) - \frac{1-\hat{\theta}_{bk_2}}{2}\right\}\Bigg)\end{aligned} \tag{2.15}$$

$$\begin{aligned}\hat{c}^{H_0}_{ak_1,bk_2} = \frac{1}{n_x}\sum_{i=1}^{n_x}\Bigg(&\left\{\frac{1}{n_y}\sum_{\ell=1}^{n_y} I(Y_{\ell ak_1} < X_{iak_1}) - \frac{1-\theta^{H_0}_{ak_1}}{2}\right\}\\ &\cdot\left\{\frac{1}{n_y}\sum_{\ell=1}^{n_y} I(Y_{\ell bk_2} < X_{ibk_2}) - \frac{1-\theta^{H_0}_{bk_2}}{2}\right\}\Bigg)\end{aligned} \tag{2.16}$$

$$\begin{aligned}\hat{d}_{ak_1,bk_2} = \frac{1}{n_y}\sum_{\ell=1}^{n_y}\Bigg(&\left\{\frac{1}{n_x}\sum_{i=1}^{n_x} I(X_{iak_1} < Y_{\ell ak_1}) - \frac{1+\hat{\theta}_{ak_1}}{2}\right\}\\ &\cdot\left\{\frac{1}{n_x}\sum_{i=1}^{n_x} I(X_{ibk_2} < Y_{\ell bk_2}) - \frac{1+\hat{\theta}_{bk_2}}{2}\right\}\Bigg)\end{aligned} \tag{2.17}$$

$$\begin{aligned}\hat{d}^{H_0}_{ak_1,bk_2} = \frac{1}{n_y}\sum_{\ell=1}^{n_y}\Bigg(&\left\{\frac{1}{n_x}\sum_{i=1}^{n_x} I(X_{iak_1} < Y_{\ell ak_1}) - \frac{1+\theta^{H_0}_{ak_1}}{2}\right\}\\ &\cdot\left\{\frac{1}{n_x}\sum_{i=1}^{n_x} I(X_{ibk_2} < Y_{\ell bk_2}) - \frac{1+\theta^{H_0}_{bk_2}}{2}\right\}\Bigg)\end{aligned} \tag{2.18}$$

$$\hat{e}_{ak_1,bk_2} = \frac{1}{n_y n_x} \sum_{i=1}^{n_x} \sum_{\ell=1}^{n_y} \Bigg( \left\{ I(Y_{\ell ak_1} < X_{iak_1}) - \frac{1-\hat{\theta}_{ak_1}}{2} \right\} \cdot \left\{ I(Y_{\ell bk_2} < X_{ibk_2}) - \frac{1-\hat{\theta}_{bk_2}}{2} \right\} \Bigg)$$

$$\hat{e}^{H_0}_{ak_1,bk_2} = \frac{1}{n_y n_x} \sum_{i=1}^{n_x} \sum_{\ell=1}^{n_y} \Bigg( \left\{ I(Y_{\ell ak_1} < X_{iak_1}) - \frac{1-\theta^{H_0}_{ak_1}}{2} \right\} \cdot \left\{ I(Y_{\ell bk_2} < X_{ibk_2}) - \frac{1-\theta^{H_0}_{bk_2}}{2} \right\} \Bigg)$$

$$s^{(2)}_{ab} = \frac{1}{m_a m_b} \sum_{k_1=1}^{m_a} \sum_{k_2=1}^{m_b} \left\{ \left(1 + \frac{1}{\lambda}\right) \hat{c}_{ak_1,bk_2} + (1+\lambda)\, \hat{d}_{ak_1,bk_2} \right\} \tag{2.19}$$

$$s^{(3)}_{ab} = \frac{1}{m_a m_b} \sum_{k_1=1}^{m_a} \sum_{k_2=1}^{m_b} \left\{ \left(1 + \frac{1}{\lambda}\right) \hat{c}^{H_0}_{ak_1,bk_2} + (1+\lambda)\, \hat{d}^{H_0}_{ak_1,bk_2} \right\} \tag{2.20}$$

$$s^{(2e)}_{ab} = \frac{1}{m_a m_b} \sum_{k_1=1}^{m_a} \sum_{k_2=1}^{m_b} \left\{ \frac{N(n_y-1)}{n_x n_y} \hat{c}_{ak_1,bk_2} + \frac{N(n_x-1)}{n_x n_y} \hat{d}_{ak_1,bk_2} + \frac{N}{n_x n_y} \hat{e}_{ak_1,bk_2} \right\} \tag{2.21}$$

$$s^{(3e)}_{ab} = \frac{1}{m_a m_b} \sum_{k_1=1}^{m_a} \sum_{k_2=1}^{m_b} \left\{ \frac{N(n_y-1)}{n_x n_y} \hat{c}^{H_0}_{ak_1,bk_2} + \frac{N(n_x-1)}{n_x n_y} \hat{d}^{H_0}_{ak_1,bk_2} + \frac{N}{n_x n_y} \hat{e}^{H_0}_{ak_1,bk_2} \right\} \tag{2.22}$$

## 2.10 Simulačné štúdie - spojité dáta

Teraz sa pozrieme na pravdepodobnosti chyby prvého druhu a sily jednotlivých testov pri testovaní zovšeobecnenej Behrensovej-Fisherovej hypotézy. Testovať ju budeme na hladine významnosti 0.05. Pravdepodobnosti chýb prvého druhu a sily testov odhadneme simulačne. Porovnávať budeme testy, ktoré sme navrhli v kapitole 2.9 spolu s originálnym $T_{cMAX}$ testom z vety 1.6 [13, Theorem 2]. V tejto časti sa zameriame na spojité rozdelenia pravdepodobnosti. Rovnako ako v článku [13, str. 825] sa pozrieme na prípad, kedy $X$ a $Y$ majú šesť-rozmerné log-normálne rozdelenie, kde prvé tri črty patria do prvého clustera a zvyšné črty patria do druhého clustera. Budeme pracovať so štyrmi rôznymi nastaveniami. Vo všetkých nastaveniach bude mať kovariančná matica $\Sigma_X$ na diagonále všetky zložky rovné 1 a mimo diagonály 0.6. Matica $\Sigma_Y$ bude mať vo všetkých prípadoch na diagonále 10 a mimo diagonály 1. Medzi nastaveniami sa budú meniť len vektory stredných hodnôt, ktoré sú pre jednotlivé nastavenia nasledovné:

1. $\mu_x = (0, 0, 0;\ 0, 0, 0)$, $\mu_y = (0, 0, 0;\ 0, 0, 0)$,

2. $\mu_x = (0.5, 0.6, 0.7;\ -0.5, -0.6, -0.7)$, $\mu_y = (-0.2, -0.1, 0;\ -0.1, -0.2, -0.3)$,

3. $\mu_x = (0.5, 0.6, 0.7;\ -0.5, -0.6, -0.7)$, $\mu_y = (-0.2, -0.1, 0;\ -0.1, -0.2, -1.1)$,

4. $\mu_x = (0.5, 0.6, 0.7;\ -0.5, -0.6, -0.7)$, $\mu_y = (0.6, 0.7, 0.8;\ 0.1, 0, -0.1)$.

Pre jednotlivé nastavenia sú príslušné vektory parametrov $\theta$ nasledovné

1. $\theta = (0, 0, 0;\ 0, 0, 0)$,

2. $\theta = (-0.167, -0.167, -0.167;\ 0.096, 0.096, 0.096)$,

3. $\theta = (-0.167, -0.167, -0.167;\ 0.096, 0.096, -0.096)$,

4. $\theta = (0.024, 0.024, 0.024;\ 0.143, 0.143, 0.143)$.

Vo všetkých prípadoch bude pre veľkosti náhodných výberov $n_x$ a $n_y$ platiť $n_x = n_y$. Konkrétne veľkosti $n_x$ a $n_y$, s ktorými sme pracovali, sú uvedené v kapitole 5.

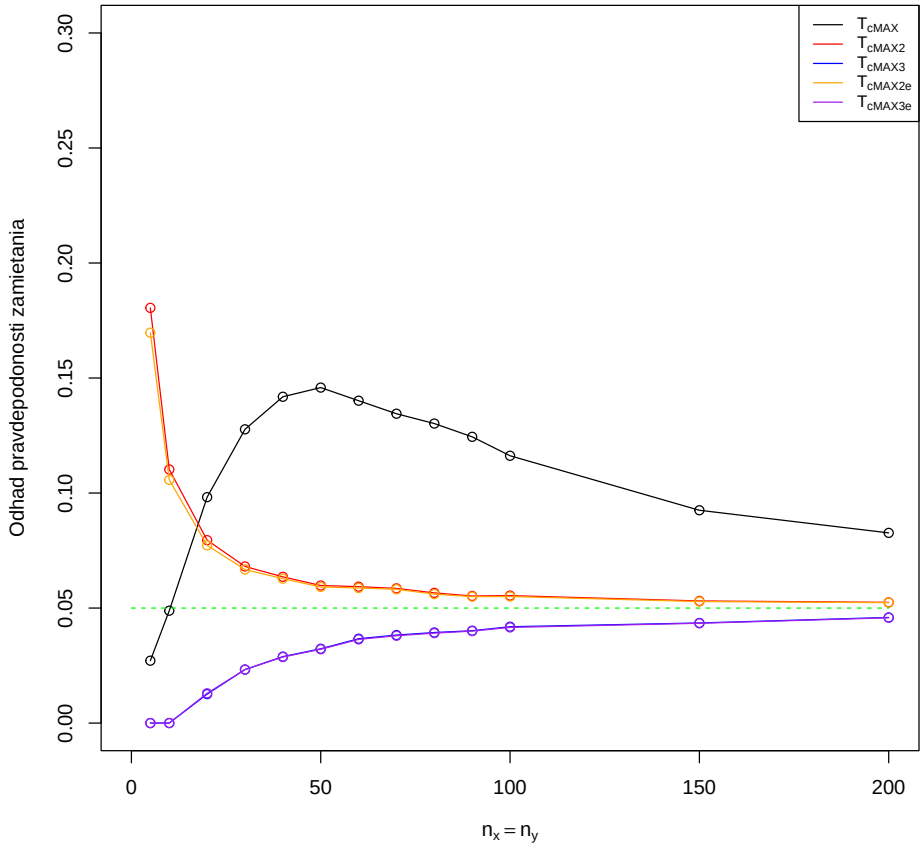
Pri prvom nastavení je naším cieľom odhadnúť pravdepodobnosť chyby prvého druhu jednotlivých testov, nakoľko pri tomto nastavení zovšeobecnená Behrensova-Fisherova hypotéza zjavne platí. Z obrázku 2 vidíme, že pravdepodobnosť chyby prvého druhu pre všetky nami navrhnuté varianty ($T_{cMAX2}$, $T_{cMAX2e}$, $T_{cMAX3}$, $T_{cMAX3e}$) konverguje k želanej hladine 0.05 pri rastúcej veľkosti náhodných výberov $n_x$ a $n_y$. Vidíme aj to, že testy, kde je parameter $\theta$ dosadený z hypotézy ($T_{cMAX3}$, $T_{cMAX3e}$), sú konzervatívnejšie ako testy, kde sa parameter $\theta$ odhaduje ($T_{cMAX2}$, $T_{cMAX2e}$).

Ďalšie tri nastavenia testujú sily spomínaných testov, nakoľko ani v jednom prípade neplatí zovšeobecnená Behrensova-Fisherova hypotéza. Podobne ako v prvom prípade môžeme na obrázkoch 3, 4 a 5 vidieť, že ochota zamietať hypotézu je v prípade testov $T_{cMAX3}$ a $T_{cMAX3e}$ nižšia ako pri $T_{cMAX2}$ a $T_{cMAX2e}$. Rozdiely v pravdepodobnosti zamietania sa ale zmenšujú s narastajúcimi veľkosťami náhodných výberov.

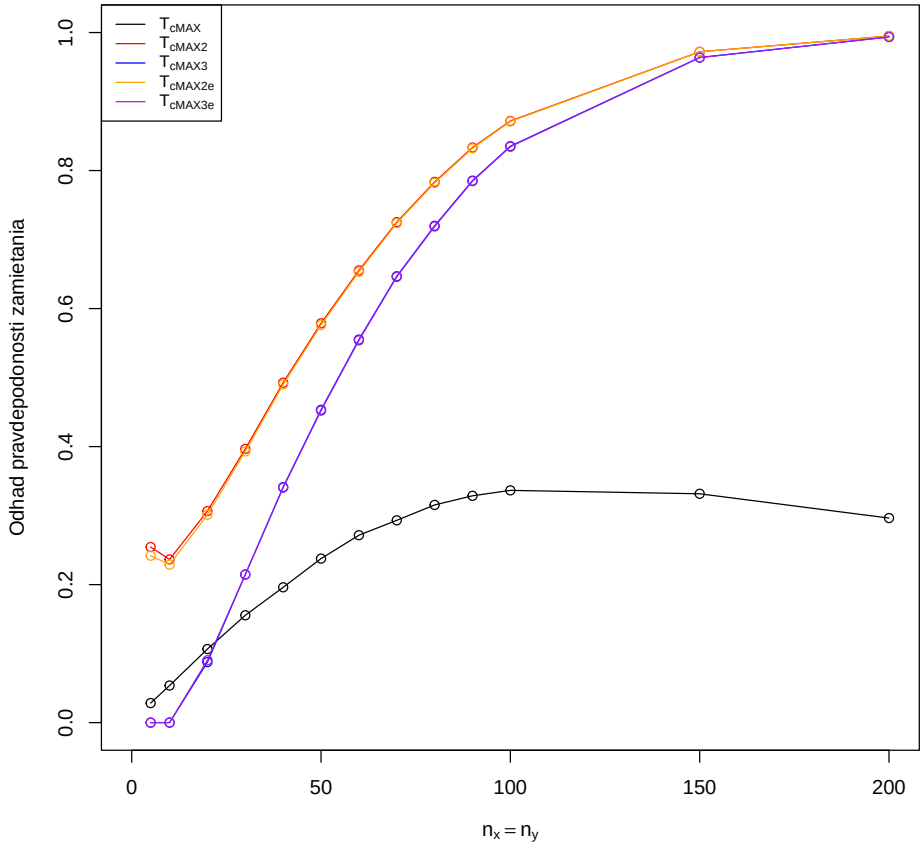
Takisto si môžeme všimnúť, že rozdiely v pravdepodobnosti chýb prvého druhu respektíve sily testov medzi verziami testov využívajúcimi presnú respektíve asymptotickú kovariančnú maticu sú zanedbateľné okrem prípadov, kedy sú veľkosti náhodných výberov veľmi malé.

*Poznámka*: Konkrétne hodnoty odhadov pravdepodobnosti zamietania pre jednotlivé testy sú vypísané v tabuľkách v kapitole 5.

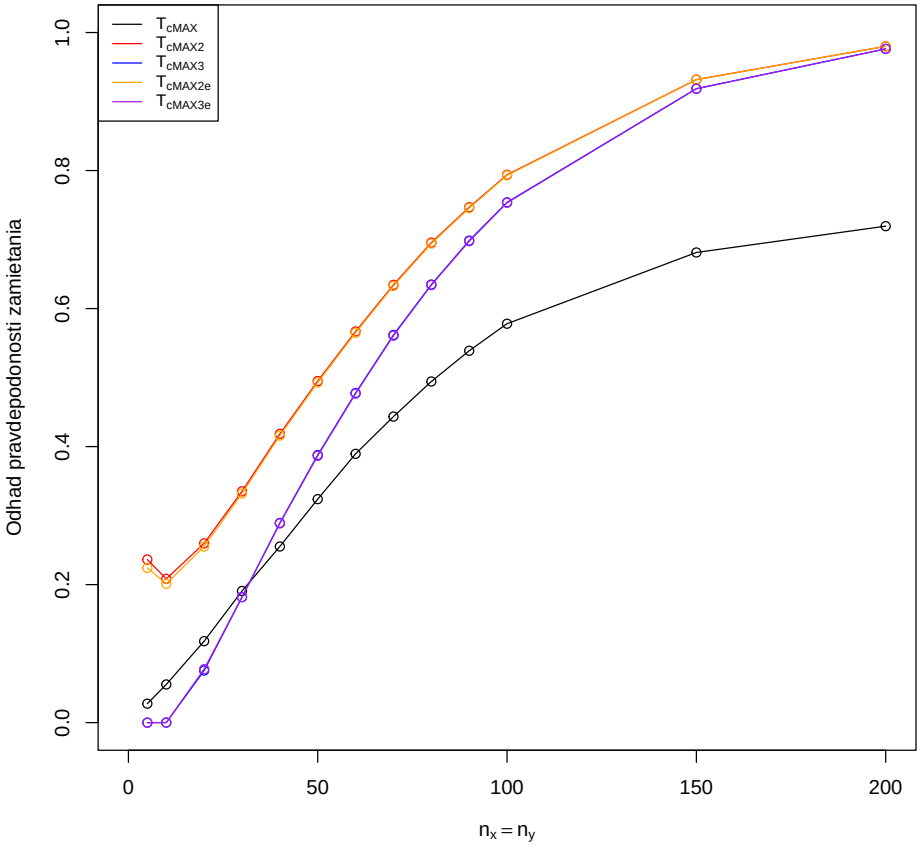
*Poznámka*: Môžeme si všimnúť, že sila testov $T_{cMAX2}$ a $T_{cMAX2e}$ je pri veľkosti náhodných výberov $n_x = n_y = 10$ nižšia ako pri $n_x = n_y = 5$. Toto je spôsobené tým, že pri menších veľkostiach náhodných výberov sa častejšie stáva, že v niektorom zo zlomkov z (1.5) je čitateľ nenulový a menovateľ nulový. V takom prípade test automaticky zovšeobecnenú Behrensovu-Fisherovu hypotézu zamieta.



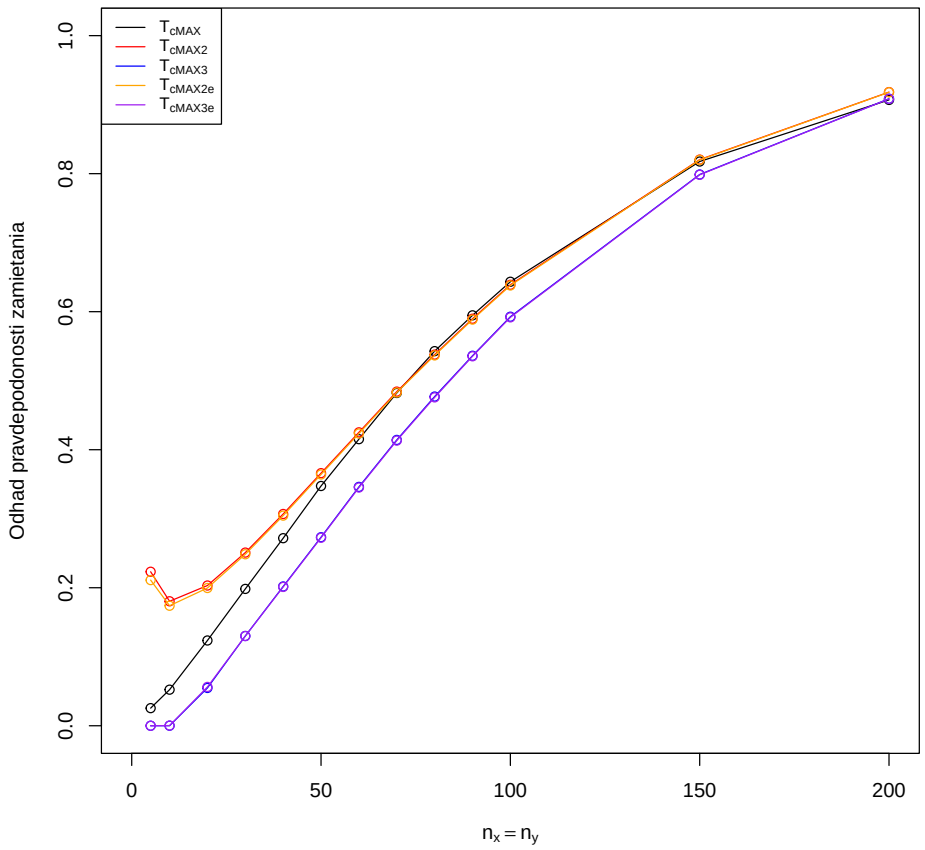

**Obr. 2:** Odhad pravdepodobnosti chyby prvého druhu pre jednotlivé testy pri rastúcej veľkosti náhodných výberov $n_x$ a $n_y$, keď $X$ a $Y$ majú šesť-rozmerné lognormálne rozdelenie a $\theta = (0, 0, 0;\ 0, 0, 0)$.




**Obr. 3:** Odhad sily pre jednotlivé testy pri rastúcej veľkosti náhodných výberov $n_x$ a $n_y$, keď $X$ a $Y$ majú šesť-rozmerné lognormálne rozdelenie a $\theta = (-0.167, -0.167, -0.167;\ 0.096, 0.096, 0.096)$.




**Obr. 4:** Odhad sily pre jednotlivé testy pri rastúcej veľkosti náhodných výberov $n_x$ a $n_y$, keď $X$ a $Y$ majú šesť-rozmerné lognormálne rozdelenie a $\theta = (-0.167, -0.167, -0.167;\ 0.096, 0.096, -0.096)$.




**Obr. 5:** Odhad sily pre jednotlivé testy pri rastúcej veľkosti náhodných výberov $n_x$ a $n_y$, keď $X$ a $Y$ majú šesť-rozmerné lognormálne rozdelenie a $\theta = (0.024, 0.024, 0.024;\ 0.143, 0.143, 0.143)$.

# 3 Oprava $T_{cMAX}$ testu: všeobecný prípad

V tejto kapitole sa budeme venovať $T_{cMAX}$ testu vo všeobecnom prípade, keď rozdelenia náhodných vektorov $X$ a $Y$ nemusia nutne spĺňať podmienku (1.1) [8].

V prvej časti sa budeme venovať teoretickému rozdeleniu testovej štatistiky $T_{cMAX}$ testu. Ukážeme, kde dôkaz vety 1.6 [13, Theorem 1] nie je korektný v prípade, že podmienka (1.1) neplatí. Ďalej v kapitole 3.2 uvedieme kontrapríklad k tejto vete. Následne v kapitole 3.3 odvodíme tvar kovariančnej matice vystupujúcej v $T_{cMAX}$ teste v tomto všeobecnom prípade. Hlavným výsledkom tejto časti je veta 3.1, ktorá je korektným znením [13, Theorem 1] pre všeobecný prípad rozdelení náhodných vektorov $X$ a $Y$. Pripomeňme, že $F_{ak}$ a $G_{ak}$ sú distribučné funkcie $X_{ak}$ a $Y_{ak}$:

$$F_{ak}(x) = P(X_{ak} < x) \qquad G_{ak}(y) = P(Y_{ak} < y).$$

Označme ešte pre každé $x, y \in \mathbb{R}$

$$\tilde{F}_{ak}(x) = P(X_{ak} = x) \qquad \tilde{G}_{ak}(y) = P(Y_{ak} = y).$$

**Veta 3.1.** *Majme dva nezávislé náhodné vektory $X$, $Y$. Nech* $\min\{n_x, n_y\} \to \infty$ *a* $n_x/n_y \to \lambda \in (0, \infty)$. *Potom náhodný vektor*

$$\left( \frac{\bar{R}_{y\cdot 1\cdot} - \bar{R}_{x\cdot 1\cdot}}{\sqrt{N}} - \frac{\sqrt{N}}{2m_1} \sum_{k=1}^{m_1} \theta_{1k}, \ldots, \frac{\bar{R}_{y\cdot J\cdot} - \bar{R}_{x\cdot J\cdot}}{\sqrt{N}} - \frac{\sqrt{N}}{2m_J} \sum_{k=1}^{m_J} \theta_{Jk} \right)^T$$

*konverguje k $J$-rozmernému normálnemu rozdeleniu s nulovou strednou hodnotou a kovariančnou maticou $\Sigma^{(4)} = (\sigma^{(4)}_{ab})_{J\times J}$, ktorej zložky sú*

$$\sigma^{(4)}_{ab} = \frac{1}{m_a m_b} \sum_{k_1=1}^{m_a} \sum_{k_2=1}^{m_b} \left\{ \left(1 + \frac{1}{\lambda}\right) \mathcal{C}_{ak_1,bk_2} + (1+\lambda)\, \mathcal{D}_{ak_1,bk_2} \right\}, \tag{3.1}$$

*kde*

$$\begin{aligned}
\mathcal{C}_{ak_1,bk_2} =& \operatorname{cov}\left[G_{ak_1}(X_{ak_1}), G_{bk_2}(X_{bk_2})\right] + \frac{1}{2}\operatorname{cov}\left[G_{ak_1}(X_{ak_1}), \tilde{G}_{bk_2}(X_{bk_2})\right] \\
&+ \frac{1}{2}\operatorname{cov}\left[\tilde{G}_{ak_1}(X_{ak_1}), G_{bk_2}(X_{bk_2})\right] + \frac{1}{4}\operatorname{cov}\left[\tilde{G}_{ak_1}(X_{ak_1}), \tilde{G}_{bk_2}(X_{bk_2})\right], \\
\mathcal{D}_{ak_1,bk_2} =& \operatorname{cov}\left[F_{ak_1}(Y_{ak_1}), F_{bk_2}(Y_{bk_2})\right] + \frac{1}{2}\operatorname{cov}\left[F_{ak_1}(Y_{ak_1}), \tilde{F}_{bk_2}(Y_{bk_2})\right] \\
&+ \frac{1}{2}\operatorname{cov}\left[\tilde{F}_{ak_1}(Y_{ak_1}), F_{bk_2}(Y_{bk_2})\right] + \frac{1}{4}\operatorname{cov}\left[\tilde{F}_{ak_1}(Y_{ak_1}), \tilde{F}_{bk_2}(Y_{bk_2})\right],
\end{aligned}$$

$a, b = 1, \ldots, J$, $k_1 = 1, \ldots, m_a$, $k_2 = 1, \ldots, m_b$.

[8] V prípade, že predpoklad (1.1) neplatí, tak ani znenia viet z článku [12, Theorem 1 a Theorem 2] nie sú korektné, na rozdiel od prípadu, keď predpoklad (1.1) platí.

Ďalej sa zameriame na vetu 1.7 [13, Theorem 2] a jej znenie pre všeobecný prípad. V kapitole 3.4 poukážeme na to, kde môžu nastať problémy pri odhadoch kovariančnej matice ak podmienka (1.1) neplatí; v kapitolách 3.5 a 3.8 následne uvádzame kontrapríklad. Hlavným výsledkom tejto časti je veta 3.2, ktorá je opravou [13, Theorem 2] a jej dôkaz uvádzame v kapitole 3.6. Pre $i = 1, \ldots, n_x$, $\ell = 1, \ldots, n_y$, $j = 1, \ldots, J$ a $k = 1, \ldots, m_j$ zaveďme nasledovné označenie

$$\hat{I}^{<}_{jk} = \frac{1}{n_x n_y} \sum_{i=1}^{n_x} \sum_{\ell=1}^{n_y} I(Y_{\ell jk} < X_{ijk}) \qquad \hat{I}^{=}_{jk} = \frac{1}{n_x n_y} \sum_{i=1}^{n_x} \sum_{\ell=1}^{n_y} I(Y_{\ell jk} = X_{ijk})$$
$$\hat{I}^{>}_{jk} = \frac{1}{n_x n_y} \sum_{i=1}^{n_x} \sum_{\ell=1}^{n_y} I(Y_{\ell jk} > X_{ijk})$$

a

$$\hat{G}_{jk}(X_{ijk}) = \frac{1}{n_y} \sum_{\ell=1}^{n_y} I(Y_{\ell jk} < X_{ijk}) \qquad \hat{\tilde{G}}_{jk}(X_{ijk}) = \frac{1}{n_y} \sum_{\ell=1}^{n_y} I(Y_{\ell jk} = X_{ijk})$$
$$\hat{F}_{jk}(Y_{\ell jk}) = \frac{1}{n_x} \sum_{i=1}^{n_x} I(Y_{\ell jk} > X_{ijk}) \qquad \hat{\tilde{F}}_{jk}(Y_{\ell jk}) = \frac{1}{n_x} \sum_{i=1}^{n_x} I(Y_{\ell jk} = X_{ijk})$$

**Veta 3.2.** *Konzistentnými odhadmi pre $\mathcal{C}_{ak_1,bk_2}$ a $\mathcal{D}_{ak_1,bk_2}$ z vety 3.1 sú*

$$\begin{aligned}
\hat{\mathcal{C}}_{ak_1,bk_2} = {} & \frac{1}{n_x} \sum_{i=1}^{n_x} (\hat{G}_{ak_1}(X_{iak_1}) - \hat{I}^{<}_{ak_1})(\hat{G}_{bk_2}(X_{ibk_2}) - \hat{I}^{<}_{bk_2}) \\
& + \frac{1}{2} \cdot \frac{1}{n_x} \sum_{i=1}^{n_x} (\hat{G}_{ak_1}(X_{iak_1}) - \hat{I}^{<}_{ak_1})(\hat{\tilde{G}}_{bk_2}(X_{ibk_2}) - \hat{I}^{=}_{bk_2}) \\
& + \frac{1}{2} \cdot \frac{1}{n_x} \sum_{i=1}^{n_x} (\hat{\tilde{G}}_{ak_1}(X_{iak_1}) - \hat{I}^{=}_{ak_1})(\hat{G}_{bk_2}(X_{ibk_2}) - \hat{I}^{<}_{bk_2}) \\
& + \frac{1}{4} \cdot \frac{1}{n_x} \sum_{i=1}^{n_x} (\hat{\tilde{G}}_{ak_1}(X_{iak_1}) - \hat{I}^{=}_{ak_1})(\hat{\tilde{G}}_{bk_2}(X_{ibk_2}) - \hat{I}^{=}_{bk_2}),
\end{aligned}$$

$$\begin{aligned}
\hat{\mathcal{D}}_{ak_1,bk_2} = {} & \frac{1}{n_y} \sum_{\ell=1}^{n_y} (\hat{F}_{ak_1}(Y_{\ell ak_1}) - \hat{I}^{>}_{ak_1})(\hat{F}_{bk_2}(Y_{\ell bk_2}) - \hat{I}^{>}_{bk_2}) \\
& + \frac{1}{2} \cdot \frac{1}{n_y} \sum_{\ell=1}^{n_y} (\hat{F}_{ak_1}(Y_{\ell ak_1}) - \hat{I}^{>}_{ak_1})(\hat{\tilde{F}}_{bk_2}(Y_{\ell bk_2}) - \hat{I}^{=}_{bk_2}) \\
& + \frac{1}{2} \cdot \frac{1}{n_y} \sum_{\ell=1}^{n_y} (\hat{\tilde{F}}_{ak_1}(Y_{\ell ak_1}) - \hat{I}^{=}_{ak_1})(\hat{F}_{bk_2}(Y_{\ell bk_2}) - \hat{I}^{>}_{bk_2}) \\
& + \frac{1}{4} \cdot \frac{1}{n_y} \sum_{\ell=1}^{n_y} (\hat{\tilde{F}}_{ak_1}(Y_{\ell ak_1}) - \hat{I}^{=}_{ak_1})(\hat{\tilde{F}}_{bk_2}(Y_{\ell bk_2}) - \hat{I}^{=}_{bk_2}),
\end{aligned}$$

*kde $a, b = 1, \ldots, J$, $k_1 = 1, \ldots, m_a$, $k_2 = 1, \ldots, m_b$.*

*Teda konzistentný odhad pre $\Sigma^{(4)}$ z vety 3.1 je $S^{(4)} = (s_{ab}^{(4)})_{J\times J}$, kde*

$$s_{ab}^{(4)} = \frac{1}{m_a m_b} \sum_{k_1=1}^{m_a} \sum_{k_2=1}^{m_b} \left\{ \left(1 + \frac{1}{\lambda}\right) \hat{\mathcal{C}}_{ak_1,bk_2} + (1+\lambda)\hat{\mathcal{D}}_{ak_1,bk_2} \right\}. \tag{3.2}$$

## 3.1 Všeobecný prípad rozdelení - správne vyjadrenie kovariancií z vety 1.6 [13, Theorem 1]

V dôkaze vety 1.6 [13, Theorem 1] a takisto vety 2.1 bol pri odvodzovaní kovariančnej matice využitý predpoklad (1.1)

$$P(X_{ak} = Y_{ak}) = 0 \quad \forall a = 1, \ldots, J, \ \forall k = 1, \ldots, m_a.$$

Tento predpoklad však autori článku [13] explicitne nespomínajú. Naopak, v [13, str. 822] autori uvádzajú, že črty $X$,$Y$ môžu mať ľubovoľné ordinálne rozdelenie; podobne, predpoklad (1.1) nie je splnený ani pri simuláciách, ktorých výsledky sú uvedené v [13, str. 827-828].

Za predpokladu (1.1) platí nasledujúca séria rovností, ktorá je kľúčová pri odvodení kovariančných matíc v dôkazoch viet 1.6 [13, Theorem 1] a 2.1:

$$\begin{aligned}
\operatorname{cov}\left[\Psi_{i\ell_1 ak_1}, \Psi_{i\ell_2 bk_2}\right] &= \operatorname{cov}\Big[I(X_{iak_1} < Y_{\ell_1 ak_1}) - I(X_{iak_1} > Y_{\ell_1 ak_1}), \\
&\qquad I(X_{ibk_2} < Y_{\ell_2 bk_2}) - I(X_{ibk_2} > Y_{\ell_2 bk_2})\Big] \\
&= \operatorname{cov}\left[1 - 2I(X_{iak_1} > Y_{\ell_1 ak_1}), 1 - 2I(X_{ibk_2} > Y_{\ell_2 bk_2})\right] \\
&= 4\operatorname{cov}\left[I(X_{iak_1} > Y_{\ell_1 ak_1}), I(X_{ibk_2} > Y_{\ell_2 bk_2})\right]
\end{aligned}$$

Ak (1.1) neplatí, tak $\operatorname{cov}\left[\Psi_{i\ell_1 ak_1}, \Psi_{i\ell_2 bk_2}\right]$ musíme počítať nasledovne:

$$\begin{aligned}
\operatorname{cov}\left[\Psi_{i\ell_1 ak_1}, \Psi_{i\ell_2 bk_2}\right] &= \operatorname{cov}\Big[I(X_{iak_1} < Y_{\ell_1 ak_1}) - I(X_{iak_1} > Y_{\ell_1 ak_1}), \\
&\qquad I(X_{ibk_2} < Y_{\ell_2 bk_2}) - I(X_{ibk_2} > Y_{\ell_2 bk_2})\Big] \\
&= \operatorname{cov}\Big[1 - 2I(X_{iak_1} > Y_{\ell_1 ak_1}) - I(X_{iak_1} = Y_{\ell_1 ak_1}), \\
&\qquad 1 - 2I(X_{ibk_2} > Y_{\ell_2 bk_2}) - I(X_{ibk_2} = Y_{\ell_2 bk_2})\Big]
\end{aligned}$$

Ďalej využitím bilineárnosti kovariancie dostávame, že

$$\begin{aligned}\operatorname{cov}\left[\Psi_{i\ell_1ak_1}, \Psi_{i\ell_2bk_2}\right] = 4\operatorname{cov}\Big[I(X_{iak_1} > Y_{\ell_1ak_1}), I(X_{ibk_2} > Y_{\ell_2bk_2})\Big] \\ +2\operatorname{cov}\left[I(X_{iak_1} > Y_{\ell_1ak_1}), I(X_{ibk_2} = Y_{\ell_2bk_2})\right] \\ +2\operatorname{cov}\left[I(X_{iak_1} = Y_{\ell_1ak_1}), I(X_{ibk_2} > Y_{\ell_2bk_2})\right] \\ +\operatorname{cov}\left[I(X_{iak_1} = Y_{\ell_1ak_1}), I(X_{ibk_2} = Y_{\ell_2bk_2})\right].\end{aligned}$$

## 3.2 Kontrapríklad k vete 1.6 [13, Theorem 1] pre všeobecný prípad

V tejto časti ukážeme, že vynechanie predpokladu (1.1) v dôkaze viet 1.6 [13, Theorem 1] a 2.1 nie je možné a vedie k chybným výsledkom.

Nech $m = J = m_1 = 1$, $F = Alt(p)$ a $G = Alt(q)$, kde $p, q \in (0,1)$ sú ľubovoľné. Teda $X = X_{11} \sim F_{11} = F$ a $Y = Y_{11} \sim G_{11} = G$. Majme dva nezávislé náhodné výbery $X_i$ $(i = 1, \ldots, n_x)$ a $Y_\ell$ $(\ell = 1, \ldots, n_y)$ z rozdelení $F$ a $G$. Nech ďalej platí, že $n_x = n_y$, teda $n_x/n_y = 1 = \lambda$. Náhodné premenné $X_{11}$ a $Y_{11}$ majú pravdepodobnostné rozdelenie

$$X_{11} = \begin{cases} 0 & \text{pr. } 1-p, \\ 1 & \text{pr. } p, \end{cases} \qquad Y_{11} = \begin{cases} 0 & \text{pr. } 1-q, \\ 1 & \text{pr. } q \end{cases} \tag{3.3}$$

a distribučné funkcie

$$F_{11}(x) = \begin{cases} 0 & x \le 0, \\ 1-p & x \in (0;1], \\ 1 & x > 1, \end{cases} \qquad G_{11}(y) = \begin{cases} 0 & y \le 0, \\ 1-q & y \in (0;1], \\ 1 & y > 1. \end{cases} \tag{3.4}$$

Zrejme rozdelenie náhodných premenných $I(X_{i11} < Y_{\ell 11})$ a $I(X_{i11} > Y_{\ell 11})$ je nasledovné

$$I(X_{i11} < Y_{\ell 11}) = \begin{cases} 1 & \text{pr. } q(1-p), \\ 0 & \text{pr. } 1-q(1-p), \end{cases}$$

$$I(X_{i11} > Y_{\ell 11}) = \begin{cases} 1 & \text{pr. } p(1-q), \\ 0 & \text{pr. } 1-p(1-q). \end{cases}$$

Z toho dostávame, že

$$E[I(X_{i11} < Y_{\ell 11})] = q(1-p),$$
$$E[I(X_{i11} > Y_{\ell 11})] = p(1-q).$$

Náhodná premenná $\Psi_{i\ell 11} = I(X_{i11} < Y_{\ell 11}) - I(X_{i11} > Y_{\ell 11})$ má rozdelenie

$$\Psi_{i\ell 11} = \begin{cases} -1 & \text{pr. } p(1-q), \\ 0 & \text{pr. } (1-p)(1-q) + pq, \\ 1 & \text{pr. } q(1-p). \end{cases}$$

Teda

$$E[\Psi_{i\ell 11}] = q - p, \quad \text{var}[\Psi_{i\ell 11}] = p^2 + q^2 + p + q.$$

Z vety 1.1 vieme, že $\Psi_{i\ell 11}$ je nevychýlený odhad $\theta$, z čoho vyplýva, že

$$\theta_{11} = q - p. \tag{3.5}$$

Začneme druhým prípadom z dôkazu vety 1.6 [13, Theorem 1] teda $i_1 = i_2 = i$ a $\ell_1 \neq \ell_2$. Vyšetrením všetkých možností jednoducho overíme, že

$$\Psi_{i\ell_1 11}\Psi_{i\ell_2 11} = \begin{cases} 0 & \text{pr. } 1 - p(1-q)^2 + (1-p)q^2, \\ 1 & \text{pr. } p(1-q)^2 + (1-p)q^2. \end{cases}$$

Teda

$$E[\Psi_{i\ell_1 11}\Psi_{i\ell_2 11}] = p(1-q)^2 + (1-p)q^2.$$

Z toho dostávame, že

$$\begin{aligned} \text{cov}\,[\Psi_{i\ell_1 11}, \Psi_{i\ell_2 11}] &= E[\Psi_{i\ell_1 11}\Psi_{i\ell_2 11}] - E[\Psi_{i\ell_1 11}]E[\Psi_{i\ell_2 11}] \\ &= p(1-q)^2 + (1-p)q^2 - (q-p)^2 = p(1-p). \end{aligned}$$

Ďalej chceme spočítať $\text{cov}\,[G_{11}(X_{11}), G_{11}(X_{11})] = \text{var}\,[G_{11}(X_{11})]$. Z (3.3) a (3.4) vieme ukázať, že platí

$$G_{11}(X) = \begin{cases} 0 & \text{pr. } 1-p, \\ 1-q & \text{pr. } p. \end{cases}$$

Z toho vyplýva, že

$$E\left[G_{11}(X)\right] = (1-q)p$$
$$E\left[G_{11}^2(X)\right] = (1-q)^2 p$$

a teda

$$\operatorname{cov}\left[G_{11}(X_{11}), G_{11}(X_{11})\right] = \operatorname{var}\left[G_{11}(X_{11})\right] = p(1-p)(1-q)^2. \tag{3.6}$$

Analogicky vieme ukázať, pre tretí prípad dôkazu vety 1.6 [13, Theorem 1], kedy $i_1 \neq i_2$ a $\ell_1 = \ell_2 = \ell$, že

$$\operatorname{cov}\left[\Psi_{i_1\ell 11}, \Psi_{i_2\ell 11}\right] = q(1-q)$$
$$\operatorname{cov}\left[F_{11}(Y_{11}), F_{11}(Y_{11})\right] = (1-p)^2 q(1-q) \tag{3.7}$$

V štvrtom prípade dôkazu vety 1.6 [13, Theorem 1], kedy $i_1 = i_2 = 1$ a $\ell_1 = \ell_2 = \ell$ dostávame, že

$$\operatorname{cov}\left[\Psi_{i\ell 11}, \Psi_{i\ell 11}\right] = \operatorname{var}[\Psi_{i\ell 11}] = p^2 + q^2 + p + q.$$

Nakoniec spočítame kovariančnú maticu. Z (1.16) dostávame, že

$$\begin{aligned}
&\operatorname{cov}\left[\frac{1}{\sqrt{N}}\left(\bar{R}_{y\cdot 1\cdot} - \bar{R}_{x\cdot 1\cdot}\right), \frac{1}{\sqrt{N}}\left(\bar{R}_{y\cdot 1\cdot} - \bar{R}_{x\cdot 1\cdot}\right)\right] \\
&= \frac{N}{4n_x^2 n_y^2}\left[\sum_{i=1}^{n_x}\sum_{\ell_1=1}^{n_y}\sum_{\ell_2=1}^{n_y}\operatorname{cov}\left[\Psi_{i\ell_1 11}, \Psi_{i\ell_2 11}\right] + \sum_{i_1=1}^{n_x}\sum_{i_2=1}^{n_x}\sum_{\ell=1}^{n_y}\operatorname{cov}\left[\Psi_{i_1\ell 11}, \Psi_{i_2\ell 11}\right]\right. \\
&\qquad \left. + \sum_{i=1}^{n_x}\sum_{\ell=1}^{n_y}\operatorname{cov}\left[\Psi_{i\ell 11}, \Psi_{i\ell 11}\right]\right] \\
&= \frac{N(n_y-1)}{4n_x n_y}\cdot p(1-p) + \frac{N(n_x-1)}{4n_x n_y}\cdot q(1-q) + \frac{N}{4n_x n_y}\cdot(p^2+q^2+p+q) \\
&= \frac{2n_x(n_x-1)}{4n_x n_x}\cdot p(1-p) + \frac{2n_x(n_x-1)}{4n_x n_x}\cdot q(1-q) + \frac{2n_x}{4n_x n_x}\cdot(p^2+q^2+p+q) \\
&= \frac{n_x-1}{2n_x}\cdot p(1-p) + \frac{n_x-1}{2n_x}\cdot q(1-q) + \frac{1}{2n_x}\cdot(p^2+q^2+p+q) =: \Sigma_{n_x}.
\end{aligned}$$

Teda pre $n_x \to \infty$ platí, že

$$\begin{aligned}
\lim_{n_x\to\infty}\Sigma_{n_x} &= \lim_{n_x\to\infty}\left(\frac{n_x-1}{2n_x}\cdot p(1-p) + \frac{n_x-1}{2n_x}\cdot q(1-q) + \frac{1}{2n_x}\cdot(p^2+q^2+p+q)\right) \\
&= \frac{p(1-p)}{2} + \frac{q(1-q)}{2} =: \Sigma_{\text{správna}}.
\end{aligned} \tag{3.8}$$

Kovariančná matica podľa vety 1.6 [13, Theorem 1] ale je

$$\begin{aligned}\Sigma &= \left(1+\frac{1}{\lambda}\right)\operatorname{cov}\left[G_{11}(X_{11}), G_{11}(X_{11})\right] + (1+\lambda)\operatorname{cov}\left[F_{11}(Y_{11}), F_{11}(Y_{11})\right] \\ &\quad - \left(\sqrt{\lambda}+\frac{1}{\sqrt{\lambda}}\right)^2 \theta_{11}\theta_{11} \\ &= 2p(1-p)(1-q)^2 + 2(1-p)^2 q(1-q) - (1+1)^2(q-p)^2 \\ &= 2(1-p)(1-q)[p(1-q)+q(1-p)] - 4(q-p)^2. \end{aligned} \tag{3.9}$$

Je zjavne vidieť, že $\Sigma_{\text{správna}}$ a $\Sigma$ sú rozdielne. Ak by sme napríklad položili $p = 1/3$ a $q = 2/3$ tak $\Sigma_{\text{správna}} = 2/9$ a $\Sigma = -16/81$. Vidíme, že pre diskrétne rozdelenia s prekrývajúcimi sa supportami veta 1.6 [13, Theorem 1] vo všeobecnosti neplatí.

*Poznámka*: V prípade, že by sme upustili od predpokladu neprekrývajúcich sa supportov (1.1) ($P(X_{jk} = Y_{jk}) = 0$) vo vete 2.1, tak

$$\begin{aligned}\Sigma^{(2)} &= (1+1/\lambda)\operatorname{cov}\left[G_{11}(X_{11}), G_{11}(X_{11})\right] + (1+\lambda)\operatorname{cov}\left[F_{11}(Y_{11}), F_{11}(Y_{11})\right] \\ &= 2p(1-p)(1-q)^2 + 2(1-p)^2 q(1-q) \\ &= 20/81. \end{aligned}$$

Teda aj veta 2.1 nedáva korektné výsledky pre diskrétne rozdelenia s prekrývajúcimi sa supportami. To ukazuje opodstatnenosť predpokladu (1.1) v tejto vete.

## 3.3 Všeobecná verzia $T_{cMAX}$ testu; dôkaz vety 3.1

V tejto kapitole sa pozrieme na to, ako sa zmení odhad kovariančnej matice $\Sigma$ v $T_{cMAX}$ teste, ak vynecháme predpoklad (1.1).

Najprv si označíme pravdepodobnostné funkcie náhodných premenných $X_{jk}$ a $Y_{jk}$ ($j = 1, \ldots, J,\ k = 1, \ldots, m_J$) nasledovne

$$\tilde{F}_{jk}(x) = P(X_{jk} = x), \qquad \tilde{G}_{jk}(y) = P(Y_{jk} = y).$$

*Poznámka*: V prípade, že $X_{jk}$ je spojitá náhodná premenná tak zrejme $\tilde{F}_{jk}(x) \equiv 0$. Analogicky, ak $Y_{jk}$ je spojitá náhodná premenná tak $\tilde{G}_{jk}(y) \equiv 0$.

Budeme teda počítať kovariančnú maticu z (1.16)

$$\begin{aligned}&\operatorname{cov}\left[\frac{1}{\sqrt{N}}\left(\bar{R}_{y\cdot a\cdot} - \bar{R}_{x\cdot a\cdot}\right), \frac{1}{\sqrt{N}}\left(\bar{R}_{y\cdot b\cdot} - \bar{R}_{x\cdot b\cdot}\right)\right] \\ &= \frac{N}{4n_x^2 n_y^2 m_a m_b} \sum_{i_1=1}^{n_x}\sum_{i_2=1}^{n_x}\sum_{\ell_1=1}^{n_y}\sum_{\ell_2=1}^{n_y}\sum_{k_1=1}^{m_a}\sum_{k_2=1}^{m_b} \operatorname{cov}\left[\Psi_{i_1\ell_1 a k_1}, \Psi_{i_2\ell_2 b k_2}\right],\end{aligned}$$

kde $a, b \in \{1, \ldots, J\}$. Analogicky ako v dôkaze vety 1.6 [13, Theorem 1] sa tento výpočet dá rozdeliť na štyri prípady. V prvom prípade, keď $i_1 \neq i_2$ a $l_1 \neq l_2$ z nezávislosti dostávame, že

$$\operatorname{cov}\left[\Psi_{i_1\ell_1 ak_1}, \Psi_{i_2\ell_2 bk_2}\right] = 0.$$

V druhom prípade, kedy $i_1 = i_2 = i$ a $l_1 \neq l_2$, z kapitoly 3.1 vieme, že

$$\begin{aligned}\operatorname{cov}\left[\Psi_{i\ell_1 ak_1}, \Psi_{i\ell_2 bk_2}\right] = {} & 4\operatorname{cov}\left[I(X_{iak_1} > Y_{\ell_1 ak_1}), I(X_{ibk_2} > Y_{\ell_2 bk_2})\right] \\ & +2\operatorname{cov}\left[I(X_{iak_1} > Y_{\ell_1 ak_1}), I(X_{ibk_2} = Y_{\ell_2 bk_2})\right] \\ & +2\operatorname{cov}\left[I(X_{iak_1} = Y_{\ell_1 ak_1}), I(X_{ibk_2} > Y_{\ell_2 bk_2})\right] \\ & +\operatorname{cov}\left[I(X_{iak_1} = Y_{\ell_1 ak_1}), I(X_{ibk_2} = Y_{\ell_2 bk_2})\right].\end{aligned}$$

Označme si

$$\begin{aligned}\alpha &= \operatorname{cov}\left[I(X_{iak_1} > Y_{\ell_1 ak_1}), I(X_{ibk_2} > Y_{\ell_2 bk_2})\right] \\ \beta &= \operatorname{cov}\left[I(X_{iak_1} > Y_{\ell_1 ak_1}), I(X_{ibk_2} = Y_{\ell_2 bk_2})\right] \\ \gamma &= \operatorname{cov}\left[I(X_{iak_1} = Y_{\ell_1 ak_1}), I(X_{ibk_2} > Y_{\ell_2 bk_2})\right] \\ \delta &= \operatorname{cov}\left[I(X_{iak_1} = Y_{\ell_1 ak_1}), I(X_{ibk_2} = Y_{\ell_2 bk_2})\right].\end{aligned}$$

Využitím lemy 2.4 na výrazy $\alpha, \beta, \gamma, \delta$ (podobne ako v kapitole 2.7) dostávame, že

$$\begin{aligned}\alpha &= \operatorname{cov}\left[G_{ak_1}(X_{ak_1}), G_{bk_2}(X_{bk_2})\right] \\ \beta &= \operatorname{cov}\left[G_{ak_1}(X_{ak_1}), \tilde{G}_{bk_2}(X_{bk_2})\right] \\ \gamma &= \operatorname{cov}\left[\tilde{G}_{ak_1}(X_{ak_1}), G_{bk_2}(X_{bk_2})\right] \\ \delta &= \operatorname{cov}\left[\tilde{G}_{ak_1}(X_{ak_1}), \tilde{G}_{bk_2}(X_{bk_2})\right]\end{aligned}$$

a teda

$$\begin{aligned}\operatorname{cov}\left[\Psi_{i\ell_1 ak_1}, \Psi_{i\ell_2 bk_2}\right] = {} & 4\operatorname{cov}\left[G_{ak_1}(X_{ak_1}), G_{bk_2}(X_{bk_2})\right] + 2\operatorname{cov}\left[G_{ak_1}(X_{ak_1}), \tilde{G}_{bk_2}(X_{bk_2})\right] \\ & +2\operatorname{cov}\left[\tilde{G}_{ak_1}(X_{ak_1}), G_{bk_2}(X_{bk_2})\right] + \operatorname{cov}\left[\tilde{G}_{ak_1}(X_{ak_1}), \tilde{G}_{bk_2}(X_{bk_2})\right] \\ = {} & 4 \cdot \mathcal{C}_{ak_1, bk_2}.\end{aligned}$$

Analogicky pre tretí prípad, keď $i_1 \neq i_2$ a $\ell_1 = \ell_2 = \ell$ dostaneme

$$\begin{aligned}\operatorname{cov}\left[\Psi_{i_1\ell ak_1}, \Psi_{i_2\ell bk_2}\right] = {} & 4\operatorname{cov}\left[F_{ak_1}(Y_{ak_1}), F_{bk_2}(Y_{bk_2})\right] + 2\operatorname{cov}\left[F_{ak_1}(Y_{ak_1}), \tilde{F}_{bk_2}(Y_{bk_2})\right] \\ & +2\operatorname{cov}\left[\tilde{F}_{ak_1}(Y_{ak_1}), F_{bk_2}(Y_{bk_2})\right] + \operatorname{cov}\left[\tilde{F}_{ak_1}(Y_{ak_1}), \tilde{F}_{bk_2}(Y_{bk_2})\right] \\ = {} & 4 \cdot \mathcal{D}_{ak_1, bk_2}.\end{aligned}$$

Pre štvrtý prípad, kedy $i_1 = i_2 = i$ a $\ell_1 = \ell_2 = \ell$ podľa kapitoly 3.1 máme

$$\begin{aligned}\operatorname{cov}\left[\Psi_{i\ell ak_1}, \Psi_{i\ell bk_2}\right] = & \, 4\operatorname{cov}\left[I(X_{iak_1} > Y_{\ell ak_1}), I(X_{ibk_2} > Y_{\ell bk_2})\right] \\ & + 2\operatorname{cov}\left[I(X_{iak_1} > Y_{\ell ak_1}), I(X_{ibk_2} = Y_{\ell bk_2})\right] \\ & + 2\operatorname{cov}\left[I(X_{iak_1} = Y_{\ell ak_1}), I(X_{ibk_2} > Y_{\ell bk_2})\right] \\ & + \operatorname{cov}\left[I(X_{iak_1} = Y_{\ell ak_1}), I(X_{ibk_2} = Y_{\ell bk_2})\right].\end{aligned}$$

Ďalej využijeme, že $X_i$ má rovnaké rozdelenie ako $X$ a $Y_\ell$ má rovnaké rozdelenie ako $Y$, teda

$$\begin{aligned}\operatorname{cov}\left[\Psi_{i\ell ak_1}, \Psi_{i\ell bk_2}\right] = & \, 4\operatorname{cov}\left[I(X_{ak_1} > Y_{ak_1}), I(X_{bk_2} > Y_{bk_2})\right] \\ & + 2\operatorname{cov}\left[I(X_{ak_1} > Y_{ak_1}), I(X_{bk_2} = Y_{bk_2})\right] \\ & + 2\operatorname{cov}\left[I(X_{ak_1} = Y_{ak_1}), I(X_{bk_2} > Y_{bk_2})\right] \\ & + \operatorname{cov}\left[I(X_{ak_1} = Y_{ak_1}), I(X_{bk_2} = Y_{bk_2})\right] \\ =: & \, 4 \cdot \mathcal{E}_{ak_1,bk_2}.\end{aligned}$$

Dosadením do (1.16) dostávame, že

$$\begin{aligned}& \operatorname{cov}\left[\frac{1}{\sqrt{N}}(\bar{R}_{x\cdot a\cdot} - \bar{R}_{y\cdot a\cdot}), \frac{1}{\sqrt{N}}(\bar{R}_{x\cdot b\cdot} - \bar{R}_{y\cdot b\cdot})\right] \\ = & \, \frac{1}{m_a m_b}\sum_{k_1=1}^{m_a}\sum_{k_2=1}^{m_b}\left\{\frac{N(n_y-1)}{n_x n_y}\mathcal{C}_{ak_1,bk_2} + \frac{N(n_x-1)}{n_x n_y}\mathcal{D}_{ak_1,bk_2} + \frac{N}{n_x n_y}\mathcal{E}_{ak_1,bk_2}\right\}.\end{aligned}$$

Kovariancie $\mathcal{C}_{ak_1,bk_2}$, $\mathcal{D}_{ak_1,bk_2}$, $\mathcal{E}_{ak_1,bk_2}$ nezávisia od veľkostí náhodných výberov $n_x, n_y$. Preto pre $\min\{n_x, n_y\} \to \infty$ za predpokladu, že $n_x/n_y \to \lambda \in (0, \infty)$, dostávame

$$\sigma_{ab}^{(4)} = \frac{1}{m_a m_b}\sum_{k_1=1}^{m_a}\sum_{k_2=1}^{m_b}\left\{\left(1 + \frac{1}{\lambda}\right)\mathcal{C}_{ak_1,bk_2} + (1+\lambda)\,\mathcal{D}_{ak_1,bk_2}\right\}. \tag{3.10}$$

Tým sme dokázali vetu 3.1.

*Poznámka*: Všimnime si, že ak je splnený predpoklad (1.1), tak

$$\mathcal{C}_{ak_1,bk_2} = \operatorname{cov}\left[G_{ak_1}(X_{ak_1}), G_{bk_2}(X_{bk_2})\right] = c_{ak_1,bk_2}$$

$$\mathcal{D}_{ak_1,bk_2} = \operatorname{cov}\left[F_{ak_1}(Y_{ak_1}), F_{bk_2}(Y_{bk_2})\right] = d_{ak_1,bk_2}$$

a teda vyjadrenie $\sigma_{ab}^{(2)}$ pomocou (3.10) a (2.12) je rovnaké.

## 3.4 Nekonzistentnosť odhadov z vety 1.7 [13, Theorem 2] pre všeobecný prípad

Vynechanie predpokladu (1.1) sa prenesie aj do odhadov vo vete 1.7 [13, Theorem 2]. Namiesto odhadov kovariancií $c_{ak_1,bk_2}$ a $d_{ak_1,bk_2}$ budeme potrebovať odhady kovariancií $\mathcal{C}_{ak_1,bk_2}$ a $\mathcal{D}_{ak_1,bk_2}$. Prejaví sa to ale aj vo výpočte samotných odhadov. Vynechaním predpokladu (1.1) výrazy $(1+\hat{\theta}_{jk})/2$ a $(1-\hat{\theta}_{jk})/2$ už nie sú rovné dobrým odhadom $\hat{I}^{>}_{jk}$ a $\hat{I}^{<}_{jk}$ pre $E\left[F_{jk}(Y_{jk})\right]$ a $E\left[G_{jk}(X_{jk})\right]$. Dôvod je nasledovný. Za platnosti predpokladu (1.1) s pravdepodobnosťou 1 platí, že $I(X_{jk}<Y_{jk})+I(X_{jk}>Y_{jk})=1$, z čoho vyplýva

$$\begin{aligned}\frac{1+\hat{\theta}_{jk}}{2} &= \frac{1+\frac{1}{n_x n_y}\sum_{i=1}^{n_x}\sum_{\ell=1}^{n_y}[I(x_{ijk}<y_{\ell jk})-I(x_{ijk}>y_{\ell jk})]}{2}\\ &= \frac{1+2\frac{1}{n_x n_y}\sum_{i=1}^{n_x}\sum_{\ell=1}^{n_y}I(x_{ijk}<y_{\ell jk})-1}{2}\\ &= \frac{1}{n_x n_y}\sum_{i=1}^{n_x}\sum_{\ell=1}^{n_y}I(x_{ijk}<y_{\ell jk})=\hat{I}^{>}_{jk}.\end{aligned}$$

Vynechaním predpokladu (1.1) máme len $I(X_{jk}<Y_{jk})+I(X_{jk}=Y_{jk})+I(X_{jk}>Y_{jk})=1$, z čoho vyplýva

$$\begin{aligned}\frac{1+\hat{\theta}_{jk}}{2} &= \frac{1+2\frac{1}{n_x n_y}\sum_{i=1}^{n_x}\sum_{\ell=1}^{n_y}I(x_{ijk}<y_{\ell jk})+\frac{1}{n_x n_y}\sum_{i=1}^{n_x}\sum_{\ell=1}^{n_y}I(x_{ijk}=y_{\ell jk})-1}{2}\\ &= \frac{1}{n_x n_y}\sum_{i=1}^{n_x}\sum_{\ell=1}^{n_y}I(x_{ijk}<y_{\ell jk})+\frac{1}{2}\frac{1}{n_x n_y}\sum_{i=1}^{n_x}\sum_{\ell=1}^{n_y}I(x_{ijk}=y_{\ell jk})\\ &= \hat{I}^{>}_{jk}+\frac{1}{2}\hat{I}^{=}_{jk}.\end{aligned}$$

Analogicky by sa ukázalo, že vynechaním predpokladu (1.1) dostávame

$$\frac{1-\hat{\theta}_{jk}}{2}=\hat{I}^{<}_{jk}+\frac{1}{2}\hat{I}^{=}_{jk}.$$

V nasledujúcej kapitole uvedieme kontrapríklad, kde ukážeme, že vynechanie predpokladu (1.1) skutočne spôsobuje nekonzistentnosť odhadov $\hat{c}_{ak_1,bk_2}$ a $\hat{d}_{ak_1,bk_2}$ vo vete 1.7 [13, Theorem 2].

Oprava tohto problému je jednoduchá. V odhadoch $\hat{c}_{ak_1,bk_2}$ a $\hat{d}_{ak_1,bk_2}$ stačí nahradiť $(1+\hat{\theta}_{jk})/2$ a $(1-\hat{\theta}_{jk})/2$ postupne výrazmi

$$\begin{aligned}\hat{I}^{>}_{jk} &= \frac{1}{n_x n_y}\sum_{i=1}^{n_x}\sum_{\ell=1}^{n_y}I(y_{\ell jk}>x_{ijk})\\ \hat{I}^{<}_{jk} &= \frac{1}{n_x n_y}\sum_{i=1}^{n_x}\sum_{\ell=1}^{n_y}I(y_{\ell jk}<x_{ijk}).\end{aligned}$$

V kapitole 3.6 ukážeme, že táto oprava je korektná.

## 3.5 Kontrapríklad k vete 1.7 [13, Theorem 2] pre všeobecný prípad

V tejto časti si uvedieme, že pre dva vyvážené hody mincou, odhady popísané vo vete 1.7 [13, Theorem 2] nie sú konzistentnými odhadmi kovariancií z vety 1.6 [13, Theorem 1]. Dôkaz, že tieto odhady nie sú konzistentné, uvedieme neskôr v kapitole 3.8.

Najprv ale pripomeňme, čo znamená, že odhad je konzistentný.

**Definícia 3.3.** *Nech $\hat{T}_n$ je odhadom parametra $\tau \in \mathbb{R}^k$. Potom $\hat{T}_n$ nazývame konzistentným odhadom parametra $\tau$ ak $\hat{T}_n \xrightarrow{P} \tau$; teda ak pre každé $\varepsilon > 0$ platí, že*

$$\lim_{n\to\infty} P(||\hat{T}_n - \tau|| > \varepsilon) = 0,$$

*kde $||\cdot||$ je norma na $\mathbb{R}^k$*

Teraz prejdime k samotnému kontrapríkladu. Nech $m = J = m_1 = 1$, $F = Alt\left(\frac{1}{2}\right)$ a $G = Alt\left(\frac{1}{2}\right)$. Teda $X = X_{11} \sim F_{11} = F$ a $Y = Y_{11} \sim G_{11} = G$. Majme dva nezávislé náhodné výbery $X_i$ $(i = 1, \ldots, n_x)$ a $Y_\ell$ $(\ell = 1, \ldots, n_y)$ z rozdelení $F$ a $G$. Nech ďalej platí, že $n_x = n_y$, teda $n_x/n_y = 1 = \lambda$.

V takomto prípade je

$$\mathrm{cov}[G_{11}(X_{11}), G_{11}(X_{11})] = \frac{1}{16},$$

ale v kapitole 3.8 ukážeme, že $\hat{c}_{11,11}$ konverguje podľa pravdepodobnosti k 1/8. Analogicky ukážeme, že

$$\mathrm{cov}[F_{11}(Y_{11}), F_{11}(Y_{11})] = \frac{1}{16}$$

ale $\hat{d}_{11,11}$ konverguje podľa pravdepodobnosti k 1/8. Nakoniec ukážeme, že $\Sigma = 1/4$, ale odhad $S \xrightarrow{P} 1/2$.

## 3.6 Dôkaz vety 3.2

V tejto kapitole dokážeme vetu 3.2. Chceme ukázať, že $S^{(4)}$ je konzistentným odhadom $\Sigma^{(4)}$. Dôkaz rozdelíme do troch častí. Začneme dôkazom toho, že

$$\frac{1}{n_x}\sum_{i=1}^{n_x}(\hat{G}_{ak_1}(X_{iak_1}) - \hat{I}^{<}_{ak_1})(\hat{G}_{bk_2}(X_{ibk_2}) - \hat{I}^{<}_{bk_2})$$

je asymptoticky nevychýleným odhadom

$$\operatorname{cov}\left[G_{ak_1}(X_{ak_1}), G_{bk_2}(X_{bk_2})\right].$$

Následne ukážeme, že variancia tohto odhadu je limitne nulová. Z lemy 7.2 následne dostaneme to, že $(1/n_x)\sum_{i=1}^{n_x}(\hat{G}_{ak_1}(X_{iak_1}) - \hat{I}^{<}_{ak_1})(\hat{G}_{bk_2}(X_{ibk_2}) - \hat{I}^{<}_{bk_2})$ je aj konzistentným odhadom $\operatorname{cov}\left[G_{ak_1}(X_{ak_1}), G_{bk_2}(X_{bk_2})\right]$. Nakoniec využitím Slutského vety 7.1 a lemy 7.3 dostaneme, že matica $S^{(4)}$ je konzistentným odhadom matice $\Sigma^{(4)}$ z vety 3.1.

Predtým si ešte pripomeňme označenie z vety 3.2:

$$\hat{I}^{<}_{jk} = \frac{1}{n_x n_y}\sum_{i=1}^{n_x}\sum_{\ell=1}^{n_y} I(Y_{\ell jk} < X_{ijk}) \qquad \hat{I}^{=}_{jk} = \frac{1}{n_x n_y}\sum_{i=1}^{n_x}\sum_{\ell=1}^{n_y} I(Y_{\ell jk} = X_{ijk})$$
$$\hat{I}^{>}_{jk} = \frac{1}{n_x n_y}\sum_{i=1}^{n_x}\sum_{\ell=1}^{n_y} I(Y_{\ell jk} > X_{ijk})$$

a

$$\hat{G}_{jk}(X_{ijk}) = \frac{1}{n_y}\sum_{\ell=1}^{n_y} I(Y_{\ell jk} < X_{ijk}) \qquad \hat{\tilde{G}}_{jk}(X_{ijk}) = \frac{1}{n_y}\sum_{\ell=1}^{n_y} I(Y_{\ell jk} = X_{ijk})$$
$$\hat{F}_{jk}(Y_{\ell jk}) = \frac{1}{n_x}\sum_{i=1}^{n_x} I(Y_{\ell jk} > X_{ijk}) \qquad \hat{\tilde{F}}_{jk}(Y_{\ell jk}) = \frac{1}{n_x}\sum_{i=1}^{n_x} I(Y_{\ell jk} = X_{ijk})$$

### 3.6.1 Asymptotická nevychýlenosť

V tejto kapitole ukážeme, že $(1/n_x)\sum_{i=1}^{n_x}(\hat{G}_{ak_1}(X_{iak_1}) - \hat{I}^{<}_{ak_1})(\hat{G}_{bk_2}(X_{ibk_2}) - \hat{I}^{<}_{bk_2})$ je asymptoticky nevychýleným odhadom $\operatorname{cov}\left[G_{ak_1}(X_{ak_1}), G_{bk_2}(X_{bk_2})\right]$. Počítajme

$$\begin{aligned} &E\left[\frac{1}{n_x}\sum_{i=1}^{n_x}(\hat{G}_{ak_1}(X_{iak_1}) - \hat{I}^{<}_{ak_1})(\hat{G}_{bk_2}(X_{ibk_2}) - \hat{I}^{<}_{bk_2})\right] \\ &= \frac{1}{n_x}\sum_{i=1}^{n_x}\Big[E\Big[\hat{G}_{ak_1}(X_{iak_1})\cdot\hat{G}_{bk_2}(X_{ibk_2})\Big] - E\Big[\hat{I}^{<}_{ak_1}\cdot\hat{G}_{bk_2}(X_{ibk_2})\Big] \\ &\qquad - E\Big[\hat{G}_{ak_1}(X_{iak_1})\cdot\hat{I}^{<}_{bk_2}\Big] + E\Big[\hat{I}^{<}_{ak_1}\cdot\hat{I}^{<}_{bk_2}\Big]\Big]. \end{aligned} \tag{3.11}$$

Počítajme, čomu sa rovnajú jednotlivé členy. Začneme s prvým z priemerov:

$$\begin{aligned} &\frac{1}{n_x}\sum_{i=1}^{n_x} E\Big[\hat{G}_{ak_1}(X_{iak_1})\cdot\hat{G}_{bk_2}(X_{ibk_2})\Big] \\ &= \frac{1}{n_x}\sum_{i=1}^{n_x} E\left[\frac{1}{n_y^2}\sum_{\ell_1=1}^{n_y}\sum_{\ell_2=1}^{n_y} I(Y_{\ell_1 ak_1} < X_{iak_1})I(Y_{\ell_2 bk_2} < X_{ibk_2})\right] \\ &= \frac{1}{n_x n_y^2}\sum_{i=1}^{n_x}\sum_{\ell_1=1}^{n_y}\sum_{\ell_2=1}^{n_y} E\left[I(Y_{\ell_1 ak_1} < X_{iak_1})I(Y_{\ell_2 bk_2} < X_{ibk_2})\right]. \end{aligned}$$

Tento výraz môžeme rozdeliť na dve časti: prvá je keď $\ell_1 \neq \ell_2$ a druhá je keď $\ell_1 = \ell_2 = \ell$. Využitím lemy 2.4 dostávame, že

$$\begin{aligned}
&\frac{1}{n_x}\sum_{i=1}^{n_x} E\Big[\hat{G}_{ak_1}(X_{iak_1})\cdot\hat{G}_{bk_2}(X_{ibk_2})\Big] \\
&= \frac{1}{n_x n_y^2}\sum_{i=1}^{n_x}\sum_{\ell_1=1}^{n_y}\sum_{\ell_2\neq\ell_1} E\Big[I(Y_{\ell_1 ak_1} < X_{iak_1})I(Y_{\ell_2 bk_2} < X_{ibk_2})\Big] \\
&+\frac{1}{n_x n_y^2}\sum_{i=1}^{n_x}\sum_{\ell=1}^{n_y} E\Big[I(Y_{\ell ak_1} < X_{iak_1})I(Y_{\ell bk_2} < X_{ibk_2})\Big] \\
&= \frac{(n_y-1)}{n_y} E\Big[G_{ak_1}(X_{ak_1})G_{bk_2}(X_{bk_2})\Big] \\
&+\frac{1}{n_y} E\Big[I(Y_{ak_1} < X_{ak_1})I(Y_{bk_2} < X_{bk_2})\Big].
\end{aligned}$$

Ďalej sa pozrime na člen $\frac{1}{n_x}\sum_{i=1}^{n_x} E\Big[\hat{G}_{ak_1}(X_{iak_1})\cdot\hat{I}^{<}_{bk_2}\Big]$ z (3.11). Tu dostávame, že

$$\begin{aligned}
&\frac{1}{n_x}\sum_{i=1}^{n_x} E\Big[\hat{G}_{ak_1}(X_{iak_1})\cdot\hat{I}^{<}_{bk_2}\Big] && (3.12)\\
&= \frac{1}{n_x}\sum_{i_1=1}^{n_x} E\Big[\frac{1}{n_x n_y^2}\sum_{i_2=1}^{n_x}\sum_{\ell_1=1}^{n_y}\sum_{\ell_2=1}^{n_y} I(Y_{\ell_1 ak_1} < X_{i_1 ak_1})I(Y_{\ell_2 bk_2} < X_{i_2 bk_2})\Big] \\
&= \frac{1}{n_x^2 n_y^2}\sum_{i_1=1}^{n_x}\sum_{i_2=1}^{n_x}\sum_{\ell_1=1}^{n_y}\sum_{\ell_2=1}^{n_y} E\Big[I(Y_{\ell_1 ak_1} < X_{i_1 ak_1})I(Y_{\ell_2 bk_2} < X_{i_2 bk_2})\Big]. && (3.13)
\end{aligned}$$

Budeme uvažovať štyri prípady

1. $i_1 \neq i_2$ a $\ell_1 \neq \ell_2$,
2. $i_1 = i_2 = i$ a $\ell_1 \neq \ell_2$,
3. $i_1 \neq i_2$ a $\ell_1 = \ell_2 = \ell$,
4. $i_1 = i_2 = i$ a $\ell_1 = \ell_2 = \ell$.

Teda

$$\begin{aligned}
\frac{1}{n_x}\sum_{i=1}^{n_x} E\Big[\hat{G}_{ak_1}(X_{iak_1})\cdot\hat{I}^{<}_{bk_2}\Big] = &\frac{1}{n_x^2 n_y^2}\sum_{i_1=1}^{n_x}\sum_{i_2\neq i_1}\sum_{\ell_1=1}^{n_y}\sum_{\ell_2\neq\ell_1} E\Big[I(Y_{\ell_1 ak_1} < X_{i_1 ak_1})I(Y_{\ell_2 bk_2} < X_{i_2 bk_2})\Big] \\
&+\frac{1}{n_x^2 n_y^2}\sum_{i=1}^{n_x}\sum_{\ell_1=1}^{n_y}\sum_{\ell_2\neq\ell_1} E\Big[I(Y_{\ell_1 ak_1} < X_{iak_1})I(Y_{\ell_2 bk_2} < X_{ibk_2})\Big] \\
&+\frac{1}{n_x^2 n_y^2}\sum_{i_1=1}^{n_x}\sum_{i_2\neq i_1}\sum_{\ell=1}^{n_y} E\Big[I(Y_{\ell ak_1} < X_{i_1 ak_1})I(Y_{\ell bk_2} < X_{i_2 bk_2})\Big] \\
&+\frac{1}{n_x^2 n_y^2}\sum_{i=1}^{n_x}\sum_{\ell=1}^{n_y} E\Big[I(Y_{\ell ak_1} < X_{iak_1})I(Y_{\ell bk_2} < X_{ibk_2})\Big].
\end{aligned}$$

Následne využitím lemy 2.4 dostávame, že

$$\begin{aligned}\frac{1}{n_x}\sum_{i=1}^{n_x} E\Big[\hat{G}_{ak_1}(X_{iak_1})\cdot \hat{I}^{<}_{bk_2}\Big] =& \frac{(n_x-1)(n_y-1)}{n_x n_y}E\Big[G_{ak_1}(X_{ak_1})\Big]E\Big[G_{bk_2}(X_{bk_2})\Big] \\ &+ \frac{n_y-1}{n_x n_y}E\Big[G_{ak_1}(X_{ak_1})G_{bk_2}(X_{bk_2})\Big] \\ &+ \frac{n_x-1}{n_x n_y}E\Big[F'_{ak_1}(Y_{ak_1})F'_{bk_2}(Y_{bk_2})\Big] \\ &+ \frac{1}{n_x n_y}E\Big[I(Y_{ak_1}<X_{ak_1})I(Y_{bk_2}<X_{bk_2})\Big],\end{aligned} \tag{3.14}$$

kde $F'_{jk}(x) = P(x < X_{jk})$. Rovnaký výsledok by sme dostali pre $(1/n_x)\sum_{i=1}^{n_x} E\Big[\hat{I}^{<}_{ak_1}\cdot \hat{G}_{bk_2}(X_{ibk_2})\Big]$ z (3.11). Pozrime sa na posledný člen $\frac{1}{n_x}\sum_{i=1}^{n_x} E\Big[\hat{I}^{<}_{ak_1}\hat{I}^{<}_{bk_2}\Big]$ z (3.11). Vidíme, že táto stredná hodnota nezávisí od $i$ teda $\frac{1}{n_x}\sum_{i=1}^{n_x} E\Big[\hat{I}^{<}_{ak_1}\hat{I}^{<}_{bk_2}\Big] = E\Big[\hat{I}^{<}_{ak_1}\hat{I}^{<}_{bk_2}\Big]$. Rozpísaním $E\Big[\hat{I}^{<}_{ak_1}\hat{I}^{<}_{bk_2}\Big]$ dostávame:

$$E\Big[\hat{I}^{<}_{ak_1}\hat{I}^{<}_{bk_2}\Big] = E\Bigg[\frac{1}{n_x^2 n_y^2}\sum_{i_1=1}^{n_x}\sum_{i_2=1}^{n_x}\sum_{\ell_1=1}^{n_y}\sum_{\ell_2=1}^{n_y} I(Y_{\ell_1 ak_1}<X_{i_1 ak_1})I(Y_{\ell_2 bk_2}<X_{i_2 bk_2})\Bigg]. \tag{3.15}$$

Tu vidíme, že výsledok je rovnaký ako pri (3.13). Spojením týchto výsledkov máme

$$\begin{aligned}&E\Bigg[\frac{1}{n_x}\sum_{i=1}^{n_x}(\hat{G}_{ak_1}(X_{iak_1})-\hat{I}^{<}_{ak_1})(\hat{G}_{bk_2}(X_{ibk_2})-\hat{I}^{<}_{bk_2})\Bigg] \\ &= \frac{(n_y-1)}{n_y}E\Big[G_{ak_1}(X_{ak_1})G_{bk_2}(X_{bk_2})\Big] - \frac{(n_x-1)(n_y-1)}{n_x n_y}E\Big[G_{ak_1}(X_{ak_1})\Big]E\Big[G_{bk_2}(X_{bk_2})\Big] \\ &+\frac{1}{n_y}E\Big[I(Y_{ak_1}<X_{ak_1})I(Y_{bk_2}<X_{bk_2})\Big] - \frac{1}{n_x n_y}E\Big[I(Y_{ak_1}<X_{ak_1})I(Y_{bk_2}<X_{bk_2})\Big] \\ &-\frac{n_y-1}{n_x n_y}E\Big[G_{ak_1}(X_{ak_1})G_{bk_2}(X_{bk_2})\Big] - \frac{n_x-1}{n_x n_y}E\Big[F'_{ak_1}(X_{ak_1})F'_{bk_2}(X_{bk_2})\Big].\end{aligned}$$

Pre $\min\{n_x, n_y\} \to \infty$ dostávame

$$\begin{aligned}&E\Bigg[\frac{1}{n_x}\sum_{i=1}^{n_x}(\hat{G}_{ak_1}(X_{iak_1})-\hat{I}^{<}_{ak_1})(\hat{G}_{bk_2}(X_{ibk_2})-\hat{I}^{<}_{bk_2})\Bigg] \\ &\to E\Big[G_{ak_1}(X_{ak_1})G_{bk_2}(X_{bk_2})\Big] - E\Big[G_{ak_1}(X_{ak_1})\Big]E\Big[G_{bk_2}(X_{bk_2})\Big] \\ &= \mathrm{cov}[G_{ak_1}(X_{ak_1}), G_{bk_2}(X_{bk_2})].\end{aligned}$$

Týmto je asymptotická nevychýlenosť odhadu

$$\frac{1}{n_x}\sum_{i=1}^{n_x}(\hat{G}_{ak_1}(X_{iak_1})-\hat{I}^{<}_{ak_1})(\hat{G}_{bk_2}(X_{ibk_2})-\hat{I}^{<}_{bk_2})$$

dokázaná.

### 3.6.2 Limitne nulová variancia

V tejto časti ukážeme, že pre $\min\{n_x, n_y\} \to \infty$ a $n_x/n_y \to \lambda \in (0, \infty)$ je variancia odhadu

$$\frac{1}{n_x}\sum_{i=1}^{n_x}(\hat{G}_{ak_1}(X_{iak_1}) - \hat{I}^{<}_{ak_1})(\hat{G}_{bk_2}(X_{ibk_2}) - \hat{I}^{<}_{bk_2})$$

asymptoticky nulová. Najprv si ale tento odhad zapíšeme iným spôsobom. Začneme výrazom $\hat{G}_{ak_1}(X_{iak_1}) - \hat{I}^{<}_{ak_1}$:

$$\begin{aligned}\hat{G}_{ak_1}(X_{iak_1}) - \hat{I}^{<}_{ak_1} &= \frac{1}{n_y}\sum_{\ell_1=1}^{n_y} I(Y_{\ell_1 ak_1} < X_{iak_1}) - \frac{1}{n_x n_y}\sum_{i_3=1}^{n_x}\sum_{\ell_3=1}^{n_y} I(Y_{\ell_3 ak_1} < X_{i_3 ak_1}) \\ &= \frac{1}{n_x n_y^2}\sum_{\ell_1=1}^{n_y}\sum_{i_3=1}^{n_x}\sum_{\ell_3=1}^{n_y}[I(Y_{\ell_1 ak_1} < X_{iak_1}) - I(Y_{\ell_3 ak_1} < X_{i_3 ak_1})].\end{aligned}$$

Z toho dostávame, že

$$\begin{aligned}&\frac{1}{n_x}\sum_{i=1}^{n_x}(\hat{G}_{ak_1}(X_{iak_1}) - \hat{I}^{<}_{ak_1})(\hat{G}_{bk_2}(X_{ibk_2}) - \hat{I}^{<}_{bk_2}) \\ &= \frac{1}{n_x^3 n_y^4}\sum_{i=1}^{n_x}\sum_{\ell_1=1}^{n_y}\sum_{\ell_2=1}^{n_y}\sum_{i_3=1}^{n_x}\sum_{\ell_3=1}^{n_y}\sum_{i_4=1}^{n_x}\sum_{\ell_4=1}^{n_y}[I(Y_{\ell_1 ak_1} < X_{iak_1}) - I(Y_{\ell_3 ak_1} < X_{i_3 ak_1})] \\ &\qquad\qquad \cdot [I(Y_{\ell_2 bk_2} < X_{ibk_2}) - I(Y_{\ell_4 bk_2} < X_{i_4 bk_4})]\,.\end{aligned}$$

Zaveďme nasledovné označenie

$$(i, \ell_1, \ell_2, i_3, \ell_3, i_4, \ell_4) =: \gamma$$

$$\Big[I(Y_{\ell_1 ak_1} < X_{iak_1}) - I(Y_{\ell_3 ak_1} < X_{i_3 ak_1})\Big]\Big[I(Y_{\ell_2 bk_2} < X_{ibk_2}) - I(Y_{\ell_4 bk_2} < X_{i_4 bk_4})\Big] =: \Phi_{\gamma, ak_1, bk_2}.$$

Z toho dostávame, že

$$\frac{1}{n_x}\sum_{i=1}^{n_x}(\hat{G}_{ak_1}(X_{iak_1}) - \hat{I}^{<}_{ak_1})(\hat{G}_{bk_2}(X_{ibk_2}) - \hat{I}^{<}_{bk_2}) = \frac{1}{n_x^3 n_y^4}\sum_{\gamma}\Phi_{\gamma, ak_1, bk_2}.$$

Teraz prejdeme k výpočtu variancie. Z vety o variancii súčtu dostávame, že

$$\begin{aligned}&\operatorname{var}\left[\frac{1}{n_x}\sum_{i=1}^{n_x}(\hat{G}_{ak_1}(X_{iak_1}) - \hat{I}^{<}_{ak_1})(\hat{G}_{bk_2}(X_{ibk_2}) - \hat{I}^{<}_{bk_2})\right] \\ &= \operatorname{var}\left[\frac{1}{n_x^3 n_y^4}\sum_{\gamma}\Phi_{\gamma, ak_1, bk_2}\right] \\ &= \frac{1}{n_x^6 n_y^8}\sum_{\gamma}\sum_{\gamma'}\operatorname{cov}\left[\Phi_{\gamma, ak_1, bk_2}, \Phi_{\gamma', ak_1, bk_2}\right].\end{aligned}$$

Chceme ukázať, že len „relatívne málo“ z týchto kovariancií je nenulových. Označme $\gamma \perp \gamma'$, ak $\{i, i_3, i_4\} \cap \{i', i_3', i_4'\} = \emptyset$ a $\{\ell_1, \ell_2, \ell_3, \ell_4\} \cap \{\ell_1', \ell_2', \ell_3', \ell_4'\} = \emptyset$; teda, keď všetky

$i$-čka majú rôzne hodnoty od $i'$ a všetky $\ell$-ká majú rôzne hodnoty od $\ell'$. Vidíme, že $\Phi_{\gamma,ak_1,bk_2}$ a $\Phi_{\gamma',ak_1,bk_2}$ sú určite nezávislé v prípade, keď $\gamma \perp \gamma'$. Počítajme, koľko je takých dvojíc $(\gamma, \gamma')$ kedy $\gamma \not\perp \gamma'$. Možností, kedy $\{i, i_3, i_4\} \cap \{i', i'_3, i'_4\} \neq \emptyset$ teda, keď sa aspoň jedno $i$-čko rovná $i'$, určite nie je viac ako $9 \cdot N^{13}$, kde hodnota 9 je počet kombinácií, kedy sa dva z týchto indexov rovnajú a platí, že ak máme vo vektore $(\gamma, \gamma')$ jednu hodnotu $i$-čka zafixovanú a ostatné sú voľné, tak takýchto dvojíc vektorov je $n_x^5 n_y^8 \leq N^{13}$. Analogickým spôsobom dostaneme, že počet možností kedy $\{\ell_1, \ell_2, \ell_3, \ell_4\} \cap \{\ell'_1, \ell'_2, \ell'_3, \ell'_4\} \neq \emptyset$ je určite menší alebo rovný ako $16 \cdot N^{13}$. Dostávame, že

$$\#\{(\gamma, \gamma') \colon \gamma \not\perp \gamma'\} \leq 25 \cdot N^{13}. \tag{3.16}$$

Ďalej využijeme fakt, že náhodná premenná $\Phi_{\gamma,ak_1,bk_2}$ nadobúda len hodnoty $-1$, 0 a 1. Z toho dostávame, že

$$|\operatorname{cov}[\Phi_{\gamma,ak_1,bk_2}, \Phi_{\gamma',ak_1,bk_2}]| \leq 1. \tag{3.17}$$

Z (3.16) a (3.17) vyplýva, že

$$\frac{1}{n_x^6 n_y^8} \sum_{\gamma} \sum_{\gamma'} |\operatorname{cov}[\Phi_{\gamma,ak_1,bk_2}, \Phi_{\gamma',ak_1,bk_2}]| \leq \frac{25 \cdot N^{13}}{n_x^6 n_y^8} = \frac{25 n_y^{13} \left(1 + \frac{n_x}{n_y}\right)^{13}}{n_x^6 n_y^8} = \frac{25 n_y^5 \left(1 + \frac{n_x}{n_y}\right)^{13}}{n_x^6}.$$

Keďže $\min\{n_x, n_y\} \to \infty$ a $n_x/n_y \to \lambda \in (0, \infty)$, dostávame, že

$$\lim_{\min\{n_x,n_y\}\to\infty} \frac{1}{n_x^6 n_y^8} \sum_{\gamma} \sum_{\gamma'} |\operatorname{cov}[\Phi_{\gamma,ak_1,bk_2}, \Phi_{\gamma',ak_1,bk_2}]| \leq 25 \left(\frac{1}{\lambda}\right)^5 (1+\lambda)^{13} \cdot \lim_{n_x\to\infty} \frac{1}{n_x} = 0.$$

Zhŕňajúc výsledky kapitol 3.6.1 a 3.6.2 sme teda dostali, že

$$\frac{1}{n_x} \sum_{i=1}^{n_x} (\hat{G}_{ak_1}(X_{iak_1}) - \hat{I}^{<}_{ak_1})(\hat{G}_{bk_2}(X_{ibk_2}) - \hat{I}^{<}_{bk_2})$$

je asymptoticky nevychýleným odhadom

$$\operatorname{cov}[G_{ak_1}(X_{ak_1}), G_{bk_2}(X_{bk_2})]$$

s limitne nulovou varianciou. Z lemy 7.2 máme, že

$$\frac{1}{n_x} \sum_{i=1}^{n_x} (\hat{G}_{ak_1}(X_{iak_1}) - \hat{I}^{<}_{ak_1})(\hat{G}_{bk_2}(X_{ibk_2}) - \hat{I}^{<}_{bk_2})$$

je aj konzistentným odhadom tejto kovariancie.

### 3.6.3 Záverečná časť dôkazu vety 3.2

Analogickým spôsobom ako v predchádzajúcich dvoch kapitolách vieme ukázať, že aj ostatné sčítance v odhadoch $\hat{\mathcal{C}}_{ak_1,bk_2}$ a $\hat{\mathcal{D}}_{ak_1,bk_2}$ sú asymptoticky nevychýlenými odhadmi príslušných sčítancov v $\mathcal{C}_{ak_1,bk_2}$ a $\mathcal{D}_{ak_1,bk_2}$ a aj to, že ich variancia je v limite nulová. Z toho podľa lemy 7.2 dostávame, že tieto sčítance v odhadoch $\hat{\mathcal{C}}_{ak_1,bk_2}$ a $\hat{\mathcal{D}}_{ak_1,bk_2}$ sú konzistentnými odhadmi príslušných sčítancov v $\mathcal{C}_{ak_1,bk_2}$ a $\mathcal{D}_{ak_1,bk_2}$.

Ďalej zo Slutského vety 7.1 vyplýva, že aj $\hat{\mathcal{C}}_{ak_1,bk_2}$ je konzistentným odhadom $\mathcal{C}_{ak_1,bk_2}$ a $\hat{\mathcal{D}}_{ak_1,bk_2}$ je konzistentným odhadom $\mathcal{D}_{ak_1,bk_2}$. Znova zo Slutského vety 7.1 dostávame aj to, že

$$s_{ab}^{(4)} = \frac{1}{m_a m_b} \sum_{k_1=1}^{m_a} \sum_{k_2=1}^{m_b} \left\{ \left(1 + \frac{1}{\lambda}\right) \hat{\mathcal{C}}_{ak_1,bk_2} + (1+\lambda) \hat{\mathcal{D}}_{ak_1,bk_2} \right\}$$

je konzistentným odhadom $\sigma_{ab}^{(4)}$. Teda jednotlivé zložky matice $S^{(4)}$ sú konzistentnými odhadmi zložiek matice $\Sigma^{(4)}$. Nakoniec z lemy 7.3 vyplýva, že $S^{(4)}$ je konzistentným odhadom matice $\Sigma^{(4)}$, čím je veta 3.2 dokázaná.

*Poznámka*: Dôkazom vety 3.2 sme dokázali aj vetu 2.2, lebo za platnosti predpokladu (1.1) je $\hat{\mathcal{C}}_{ak_1,bk_2}$ skoro isto $\hat{c}_{ak_1,bk_2}$ a $\hat{\mathcal{D}}_{ak_1,bk_2}$ je skoro isto $\hat{d}_{ak_1,bk_2}$.

## 3.7 Vyšetrenie nekonzistentnosti odhadov z vety 1.7 [13, Theorem 2] v prípade nesplnenia predpokladu (1.1)

V tejto sekcii ukážeme, že odhady $\hat{c}_{ak_1,bk_2}$, $\hat{d}_{ak_1,bk_2}$ a $s_{ak_1,bk_2}$ z vety 1.7 [13, Theorem 2] nie sú konzistentnými odhadmi $\mathrm{cov}[G_{ak_1}(X_{ak_1}), G_{bk_2}(X_{bk_2})]$, $\mathrm{cov}[F_{ak_1}(Y_{ak_1}), F_{bk_2}(Y_{bk_2})]$ a $\sigma_{ak_1,bk_2}$ z vety 1.6 [13, Theorem 1] v prípade, že predpoklad (1.1)nie je splnený. Pre jednoduchosť, nech $m = J = 1$. Pre všeobecné $m$, $J$ by bol dôkaz veľmi podobný.

Začneme dôkazom pre $\hat{c}_{11,11}$. Z vety 1.7 [13, Theorem 2] vieme, že

$$\hat{c}_{11,11} = \frac{1}{n_x} \sum_{i=1}^{n_x} \left( \hat{G}_{11}(X_{i11}) - \frac{1-\hat{\theta}_{11}}{2} \right) \left( \hat{G}_{11}(X_{i11}) - \frac{1-\hat{\theta}_{11}}{2} \right)$$

Z kapitoly 3.4 vieme, že

$$\frac{1-\hat{\theta}_{11}}{2} = \hat{I}_{11}^{<} + \frac{1}{2}\hat{I}_{11}^{=}.$$

Dostávame teda, že

$$\hat{c}_{11,11} = \frac{1}{n_x} \sum_{i=1}^{n_x} \left( \hat{G}_{11}(X_{i11}) - \hat{I}_{11}^{<} - \frac{1}{2}\hat{I}_{11}^{=} \right) \left( \hat{G}_{11}(X_{i11}) - \hat{I}_{11}^{<} - \frac{1}{2}\hat{I}_{11}^{=} \right).$$

Tento výraz vieme rozpísať nasledovne

$$\hat{c}_{11,11} = \frac{1}{n_x}\sum_{i=1}^{n_x}\left[\left(\hat{G}_{11}(X_{i11}) - \hat{I}_{11}^{<}\right)\left(\hat{G}_{11}(X_{i11}) - \hat{I}_{11}^{<}\right) - \hat{G}_{11}(X_{i11})\hat{I}_{11}^{=} + \hat{I}_{11}^{<}\hat{I}_{11}^{=} + \frac{1}{4}\hat{I}_{11}^{=}\hat{I}_{11}^{=}\right].$$

Najprv chceme ukázať, že tento odhad je asymptoticky vychýlený. Počítajme strednú hodnotu $\hat{c}_{11,11}$. Tá je rovná

$$\begin{aligned}
E[\hat{c}_{11,11}] &= E\left[\frac{1}{n_x}\sum_{i=1}^{n_x}\left(\hat{G}_{11}(X_{i11}) - \hat{I}_{11}^{<}\right)\left(\hat{G}_{11}(X_{i11}) - \hat{I}_{11}^{<}\right)\right] \\
&\quad - E\left[\frac{1}{n_x}\sum_{i=1}^{n_x}\hat{G}_{11}(X_{i11})\hat{I}_{11}^{=}\right] \\
&\quad + E\left[\frac{1}{n_x}\sum_{i=1}^{n_x}\hat{I}_{11}^{<}\hat{I}_{11}^{=}\right] + E\left[\frac{1}{n_x}\sum_{i=1}^{n_x}\frac{1}{4}\hat{I}_{11}^{=}\hat{I}_{11}^{=}\right] \\
&= E\left[\frac{1}{n_x}\sum_{i=1}^{n_x}\left(\hat{G}_{11}(X_{i11}) - \hat{I}_{11}^{<}\right)\left(\hat{G}_{11}(X_{i11}) - \hat{I}_{11}^{<}\right)\right] \\
&\quad - E\left[\frac{1}{n_x}\sum_{i=1}^{n_x}\hat{G}_{11}(X_{i11})\hat{I}_{11}^{=}\right] + E\left[\hat{I}_{11}^{<}\hat{I}_{11}^{=}\right] + \frac{1}{4}E\left[\hat{I}_{11}^{=}\hat{I}_{11}^{=}\right]
\end{aligned}$$

Z kapitoly 3.6.1 vieme, že

$$\lim_{\min\{n_x,n_y\}\to\infty} E\left[\frac{1}{n_x}\sum_{i=1}^{n_x}\left(\hat{G}_{11}(X_{i11}) - \hat{I}_{11}^{<}\right)\left(\hat{G}_{11}(X_{i11}) - \hat{I}_{11}^{<}\right)\right] = \mathrm{cov}[G_{11}(X_{11}), G_{11}(X_{11})]. \tag{3.18}$$

Využitím analogického postupu ako v (3.14) vieme ukázať, že platí

$$\lim_{\min\{n_x,n_y\}\to\infty} E\left[\frac{1}{n_x}\sum_{i=1}^{n_x}\hat{G}_{11}(X_{i11})\hat{I}_{11}^{=}\right] = E[G_{11}(X_{11})]E[\tilde{G}_{11}(X_{11})]. \tag{3.19}$$

Ďalej, podobným spôsobom ako v (3.15) by sme vedeli ukázať, že

$$\begin{aligned}
E\left[\hat{I}_{11}^{<}\hat{I}_{11}^{=}\right] &= E\left[\frac{1}{n_x}\sum_{i=1}^{n_x}\hat{G}_{11}(X_{i11})\hat{I}_{11}^{=}\right] \\
E\left[\hat{I}_{11}^{=}\hat{I}_{11}^{=}\right] &= E\left[\frac{1}{n_x}\sum_{i=1}^{n_x}\hat{\tilde{G}}_{11}(X_{i11})\hat{I}_{11}^{=}\right].
\end{aligned}$$

Z toho dostávame, že

$$\begin{aligned}
\lim_{\min\{n_x,n_y\}\to\infty} E\left[\hat{I}_{11}^{<}\hat{I}_{11}^{=}\right] &= E[G_{11}(X_{11})]E[\tilde{G}_{11}(X_{11})] \\
\lim_{\min\{n_x,n_y\}\to\infty} E\left[\hat{I}_{11}^{=}\hat{I}_{11}^{=}\right] &= E[\tilde{G}_{11}(X_{11})]E[\tilde{G}_{11}(X_{11})] = E^2[\tilde{G}_{11}(X_{11})].
\end{aligned} \tag{3.20}$$

Využitím výsledkov z (3.18), (3.19) a (3.20) máme, že

$$\begin{aligned}\lim_{\min\{n_x,n_y\}\to\infty} E[\hat{c}_{11,11}] &= \mathrm{cov}[G_{11}(X_{11}), G_{11}(X_{11})] - E[G_{11}(X_{11})]E[\tilde{G}_{11}(X_{11})] \\ &\quad + E[G_{11}(X_{11})]E[\tilde{G}_{11}(X_{11})] + \frac{1}{4}E^2[\tilde{G}_{11}(X_{11})] \\ &= \mathrm{cov}[G_{11}(X_{11}), G_{11}(X_{11})] + \frac{1}{4}E^2[\tilde{G}_{11}(X_{11})].\end{aligned}$$

Ak neplatí predpoklad (1.1), tak

$$E[\tilde{G}_{11}(X_{11})] = P(X_{11} = Y_{11}) > 0$$

(využívame, že $m = J = 1$). Vidíme teda, že $\hat{c}_{11,11}$ je asymptoticky vychýleným odhadom pre $\mathrm{cov}[G_{11}(X_{11}), G_{11}(X_{11})]$.

To, že variancia $\hat{c}_{11,11}$ je v limite nulová, by sa dokazovalo analogicky, ako pre odhad v kapitole 3.6.2. Jedinými zmenami by boli

$$(i, \ell_1, \ell_2, i_3, \ell_3, i_4, \ell_4, i_5, \ell_5, i_6, \ell_6) =: \gamma$$

$$\begin{aligned}&\left[I(Y_{\ell_1 11} < X_{i11}) - I(Y_{\ell_3 11} < X_{i_3 11}) - \frac{1}{2}I(Y_{\ell_5 11} = X_{i_5 11})\right] \\ &\cdot \left[I(Y_{\ell_2 11} < X_{i11}) - I(Y_{\ell_4 11} < X_{i_4 11}) - \frac{1}{2}I(Y_{\ell_6 11} = X_{i_6 11})\right] =: \Phi_{\gamma,11,11}.\end{aligned}$$

Z vety 7.2 teda dostávame, že $\hat{c}_{11,11}$ je konzistentným odhadom pre

$$\mathrm{cov}[G_{11}(X_{11}), G_{11}(X_{11})] + \frac{1}{4}E^2[\tilde{G}_{11}(X_{11})].$$

Analogickým spôsobom by sme vedeli ukázať, že $\hat{d}_{11,11}$ je konzistentným odhadom pre

$$\mathrm{cov}[F_{11}(Y_{11}), F_{11}(Y_{11})] + \frac{1}{4}E^2[\tilde{F}_{11}(Y_{11})].$$

Z vety 7.1 dostávame, že

$$s_{11} = \left(1 + \frac{1}{\lambda}\right)\hat{c}_{ak_1,bk_2} + (1+\lambda)\hat{d}_{ak_1,bk_2} - \left(\sqrt{\lambda} + \frac{1}{\sqrt{\lambda}}\right)^2 \hat{\theta}_{ak_1}\hat{\theta}_{bk_2}$$

je konzistentným odhadom pre

$$\begin{aligned}&\left(1 + \frac{1}{\lambda}\right)\left[\mathrm{cov}[G_{11}(X_{11}), G_{11}(X_{11})] + \frac{1}{4}E^2[\tilde{G}_{11}(X_{11})]\right] \\ &+ (1+\lambda)\left[\mathrm{cov}[F_{11}(Y_{11}), F_{11}(Y_{11})] + \frac{1}{4}E^2[\tilde{F}_{11}(Y_{11})]\right] - \left(\sqrt{\lambda} + \frac{1}{\sqrt{\lambda}}\right)^2 \theta_{11}\theta_{11}.\end{aligned}$$

Tu vidíme, že sme sa nedostali k výsledkom, ktoré sú uvedené v znení vety 1.7 [13, Theorem 1]. Podľa vety 1.7 [13, Theorem 2] by totiž malo platiť, že

$$\begin{aligned}
\hat{c}_{11,11} &\xrightarrow{P} \operatorname{cov}[G_{11}(X_{11}), G_{11}(X_{11})] \\
\hat{d}_{11,11} &\xrightarrow{P} \operatorname{cov}[F_{11}(Y_{11}), F_{11}(Y_{11})] \\
s_{11} &\xrightarrow{P} \left(1+\frac{1}{\lambda}\right) \operatorname{cov}[G_{11}(X_{11}), G_{11}(X_{11})] + (1+\lambda) \operatorname{cov}[F_{11}(Y_{11}), F_{11}(Y_{11})] \\
&\quad - \left(\sqrt{\lambda}+\frac{1}{\sqrt{\lambda}}\right)^2 \theta_{11}\theta_{11}.
\end{aligned}$$

*Poznámka*: K podobným výsledkom by sme sa dopracovali pri vete 2.2, ak by sme vynechali predpoklad (1.1). V takomto prípade by sme dostali, že $s_{11}$ je konzistentným odhadom pre

$$\begin{aligned}
&\left(1+\frac{1}{\lambda}\right)\left[\operatorname{cov}[G_{11}(X_{11}), G_{11}(X_{11})] + \frac{1}{4}E^2[\tilde{G}_{11}(X_{11})]\right] \\
&+ (1+\lambda)\left[\operatorname{cov}[F_{11}(Y_{11}), F_{11}(Y_{11})] + \frac{1}{4}E^2[\tilde{F}_{11}(Y_{11})]\right],
\end{aligned}$$

pričom podľa vety 2.2 by malo platiť, že

$$s_{11} \xrightarrow{P} \left(1+\frac{1}{\lambda}\right) \operatorname{cov}[G_{11}(X_{11}), G_{11}(X_{11})] + (1+\lambda) \operatorname{cov}[F_{11}(Y_{11}), F_{11}(Y_{11})].$$

## 3.8 Kontrapríklad k vete 1.7 [13, Theorem 2] - dokončenie kapitoly 3.5

Pripomeňme ako vyzerá kontrapríklad, ktorým sa zaoberáme v kapitole 3.5. Nech $m = J = m_1 = 1$, $F = G = Alt\left(\frac{1}{2}\right)$. Teda $X = X_{11} \sim F_{11} = F$ a $Y = Y_{11} \sim G_{11} = G$. Majme dva nezávislé náhodné výbery $X_i$ $(i = 1, \ldots, n_x)$ a $Y_\ell$ $(\ell = 1, \ldots, n_y)$ z rozdelení $F$ a $G$. Nech ďalej platí, že $n_x = n_y$, teda $n_x/n_y = 1 = \lambda$.

Najprv si odvodíme, čomu sa rovnajú $E[\tilde{G}_{11}(X_{11})]$ a $E[\tilde{F}_{11}(Y_{11})]$. Keďže $X_{11}, Y_{11} \sim Alt\left(\frac{1}{2}\right)$, tak

$$X_{11} = \begin{cases} 0 & \text{pr. } \frac{1}{2}, \\ 1 & \text{pr. } \frac{1}{2}, \end{cases} \qquad \tilde{G}_{11}(y) = \begin{cases} \frac{1}{2} & y \in \{0,1\}, \\ 0 & \text{inak.} \end{cases}$$

Z toho dostávame, že

$$\tilde{G}_{11}(X_{11}) = \frac{1}{2}.$$

Tu už ľahko vidíme, že $E[\tilde{G}_{11}(X_{11})] = 1/2$. Keďže $X_{11}$ a $Y_{11}$ majú rovnaké rozdelenia, tak aj $E[\tilde{F}_{11}(Y_{11})] = 1/2$.

V nasledujúcich výpočtoch využijeme výsledky, ktoré sme ukázali v kapitole 3.2, a to konkrétne rovnosti (3.5), (3.6), (3.7) a (3.9). Tie hovoria, že ak $X_{11} \sim Alt(p)$ a $Y_{11} \sim Alt(q)$, kde $p, q \in (0,1)$, tak

$$\begin{aligned}
\theta_{11} &= q - p, \\
\operatorname{cov}[G_{11}(X_{11}), G_{11}(X_{11})] &= p(1-p)(1-q)^2, \\
\operatorname{cov}[F_{11}(Y_{11}), F_{11}(Y_{11})] &= q(1-q)(1-p)^2, \\
\Sigma &= \left(1 + \frac{1}{\lambda}\right) \operatorname{cov}\left[G_{11}(X_{11}), G_{11}(X_{11})\right] \\
&\quad + (1+\lambda) \operatorname{cov}\left[F_{11}(Y_{11}), F_{11}(Y_{11})\right] \\
&\quad - \left(\sqrt{\lambda} + \frac{1}{\sqrt{\lambda}}\right)^2 \theta_{11}\theta_{11}, \\
&= 2(1-p)(1-q)[p(1-q) + q(1-p)] - 4(q-p)^2.
\end{aligned}$$

Prejdime k výpočtu $\hat{c}_{11,11}$ a $\operatorname{cov}[G_{11}(X_{11}), G_{11}(X_{11})]$. Z (3.6) vieme, že

$$\operatorname{cov}[G_{11}(X_{11}), G_{11}(X_{11})] = \frac{1}{2}\left(1 - \frac{1}{2}\right)\left(1 - \frac{1}{2}\right)^2 = \frac{1}{16}. \tag{3.21}$$

V minulej kapitole sme ukázali, že $\hat{c}_{11,11}$ konverguje podľa pravdepodobnosti k

$$\operatorname{cov}[G_{11}(X_{11}), G_{11}(X_{11})] + \frac{1}{4}E^2[\tilde{G}_{11}(X_{11})] = \frac{1}{16} + \frac{1}{4} \cdot \left(\frac{1}{2}\right)^2 = \frac{1}{8}.$$

Tu už jasne vidíme, že $\hat{c}_{11,11}$ nie je konzistentným odhadom pre $\operatorname{cov}[G_{11}(X_{11}), G_{11}(X_{11})]$.

Keďže $X_{11}$ a $Y_{11}$ majú rovnaké rozdelenia, tak aj

$$\operatorname{cov}[F_{11}(Y_{11}), F_{11}(Y_{11})] = \frac{1}{16} \tag{3.22}$$

a

$$\operatorname{cov}[F_{11}(Y_{11}), F_{11}(Y_{11})] + \frac{1}{4}E^2[\tilde{F}_{11}(Y_{11})] = \frac{1}{8}.$$

Teda ani $\hat{d}_{11,11}$ nie je konzistentným odhadom pre $\operatorname{cov}[F_{11}(Y_{11}), F_{11}(Y_{11})]$.

Nakoniec využitím (3.21) a (3.22) máme, že

$$\begin{aligned}
\sigma_{11} &= \left(1 + \frac{1}{\lambda}\right) \operatorname{cov}\left[G_{11}(X_{11}), G_{11}(X_{11})\right] + (1+\lambda) \operatorname{cov}\left[F_{11}(Y_{11}), F_{11}(Y_{11})\right] \\
&\quad - \left(\sqrt{\lambda} + \frac{1}{\sqrt{\lambda}}\right)^2 \theta_{11}\theta_{11} \\
&= 2 \cdot \frac{1}{16} + 2 \cdot \frac{1}{16} - 4 \cdot \left(\frac{1}{2} - \frac{1}{2}\right)^2 = \frac{1}{4}.
\end{aligned}$$

Ak sa pozrieme na odhad $s_{11}$, tak z vety 1.7 [13, Theorem 2] máme, že

$$s_{11} = \left(1 + \frac{1}{\lambda}\right) \hat{c}_{11,11} + (1+\lambda)\hat{d}_{11,11} - \left(\sqrt{\lambda} + \frac{1}{\sqrt{\lambda}}\right)^2 \hat{\theta}_{11}\hat{\theta}_{11}.$$

Využitím výsledkov z minulej kapitoly vieme, že $s_{11}$ konverguje podľa pravdepodobnosti ku

$$\begin{aligned}
&\left(1 + \frac{1}{\lambda}\right)\left[\mathrm{cov}[G_{11}(X_{11}), G_{11}(X_{11})] + \frac{1}{4}E^2[\tilde{G}_{11}(X_{11})]\right] \\
&+(1+\lambda)\left[\mathrm{cov}[F_{11}(Y_{11}), F_{11}(Y_{11})] + \frac{1}{4}E^2[\tilde{F}_{11}(Y_{11})]\right] - \left(\sqrt{\lambda} + \frac{1}{\sqrt{\lambda}}\right)^2 \theta_{11}\theta_{11} \\
=&2 \cdot \frac{1}{8} + 2 \cdot \frac{1}{8} - 4^2 \cdot 0 \cdot 0 = \frac{1}{2}.
\end{aligned}$$

Teda vidíme, že ani $S = s_{11}$ nie je konzistentným odhadom pre $\Sigma = \sigma_{11}$.

## 3.9 Simulačné štúdie - diskrétne dáta

V tejto kapitole sa pozrieme na porovnanie pravdepodobnosti chyby prvého druhu a sily testov pri testovaní zovšeobecnenej Behrensovej-Fisherovej hypotézy na diskrétnych dátach s prekrývajúcimi sa supportami. Naše dáta teda nebudú spĺňať predpoklad (1.1). Porovnávať budeme tri testy:

1. $T_{cMAX}$ - originálna verzia testu z článku [13] využívajúca platnosť predpokladu (1.1) s odhadom kovariančnej matice popísaným vo vete 1.7 [13, Theorem 2.],

2. $T_{cMAX2}$ - opravená verzia testu, ktorá tiež využíva platnosť predpokladu (1.1) s odhadom kovariančnej matice popísaným vo vete 2.2,

3. $T_{cMAX4}$ - všeobecná verzia testu s odhadom kovariančnej matice popísaným vo vete 3.2.

Behrensovu-Fisherovu hypotézu budeme testovať na hladine významnosti 0.05. Pravdepodobnosti chýb prvého druhu a sily testov odhadneme simulačne. Rovnako ako v článku [13, str. 825] sa pozrieme na prípad, kedy marginálne rozdelenia čŕt náhodných vektorov $X$ a $Y$ majú diskrétne rozdelenie s konečným supportom, kde prvé dve črty týchto náhodných vektorov patria do prvého clustera, druhé dve črty patria do druhého clustera a zvyšné dve črty patria do tretieho clustera. Jednotlivé črty vektora $X$, ako aj jednotlivé črty

vektora $Y$, sú navzájom nezávislé. Diskrétne rozdelenie čŕt bude z triedy parametrických rozdelení $D(c_1, c_2, c_3, c_4)$ so supportom $\{-2, -1, 0, 1, 2\}$ definovanej nasledovne: Povieme, že náhodná premenná $Z$ má $D(c_1, c_2, c_3, c_4)$ rozdelenie (kde $-1 < c_1 < c_2 < c_3 < c_4 < 1$)[9], ak má rovnaké rozdelenie ako náhodná premenná

$$-2 \cdot I\left[\tilde{Z} < c_1\right] + -1 \cdot I\left[c_1 \leq \tilde{Z} < c_2\right] + 0 \cdot I\left[c_2 \leq \tilde{Z} < c_3\right]$$
$$+1 \cdot I\left[c_3 \leq \tilde{Z} < c_4\right] + 2 \cdot I\left[c_4 \leq \tilde{Z}\right],$$

kde $\tilde{Z} \sim U[-1; 1]$.

Pracovať budeme so štyrmi rôznymi nastaveniami (viď [13, str. 825-827]):

1. $X_{11}, \ldots, X_{32} \sim D(-0.2, -0.1, 0.1, 0.2)$,
   $Y_{11}, \ldots, Y_{32} \sim D(-0.9, -0.8, 0.8, 0.9)$,

2. $X_{11}, X_{12} \sim D(0.1, 0.3, 0.5, 0.7)$, $X_{21}, X_{22} \sim D(-0.3, -0.2, 0, 0.2)$,
   $X_{31}, X_{32} \sim D(0.2, 0.4, 0.6, 0.8)$,
   $Y_{11}, Y_{12} \sim D(-0.1, 0.0, 0.1, 0.3)$, $Y_{21}, Y_{22} \sim D(-0.2, 0.0, 0.2, 0.4)$,
   $Y_{31}, Y_{32} \sim D(0.1, 0.4, 0.6, 0.8)$,

3. $X_{11}, X_{12} \sim D(-0.1, 0.3, 0.5, 0.7)$, $X_{21}, X_{22} \sim D(0, 0.1, 0.3, 0.5)$,
   $X_{31} \sim D(0.2, 0.4, 0.6, 0.7)$, $X_{32} \sim D(0.1, 0.4, 0.6, 0.8)$,
   $Y_{11}, Y_{12} \sim D(-0.1, 0.1, 0.3, 0.5)$, $Y_{21}, Y_{22} \sim D(-0.1, 0.1, 0.3, 0.4)$,
   $Y_{31} \sim D(0.1, 0.4, 0.6, 0.8)$, $Y_{32} \sim D(0.2, 0.4, 0.6, 0.7)$,

4. $X_{11}, X_{12} \sim D(-0.1, 0.3, 0.5, 0.8)$, $X_{21}, X_{22} \sim D(0, 0.1, 0.3, 0.5)$,
   $X_{31}, X_{32} \sim D(0.2, 0.4, 0.6, 0.7)$,
   $Y_{11}, Y_{12} \sim D(-0.2, 0.3, 0.5, 0.8)$,$Y_{21}, Y_{22} \sim D(-0.1, 0.0, 0.3, 0.5)$,
   $Y_{31}, Y_{32} \sim D(-0.1, 0.1, 0.3, 0.4)$.

Pre jednotlivé nastavenia sú príslušné vektory parametrov $\theta$ nasledovné:

1. $\theta = (0, 0;\ 0, 0;\ 0, 0)$,

2. $\theta = (0.185, 0.185;\ -0.105, -0.105;\ 0.035, 0.035)$,

[9]Definíciu tohto rozdelenia uvádzame v takomto tvare, aby bolo jednoduché porovnávanie s článkom [13].

3. $\theta = (0.075, 0.075;\ 0.045, 0.045;\ 0.025, -0.025)$, [10]

4. $\theta = (0.0325, 0.0325;\ 0.035, 0.035;\ 0.1875, 0.1875)$.

Vo všetkých prípadoch budú veľkosti náhodných výberov $n_x$ a $n_y$ v simuláciách rovnaké. Konkrétne veľkosti $n_x$ a $n_y$, s ktorými sme pracovali, sú uvedené v kapitole 8.

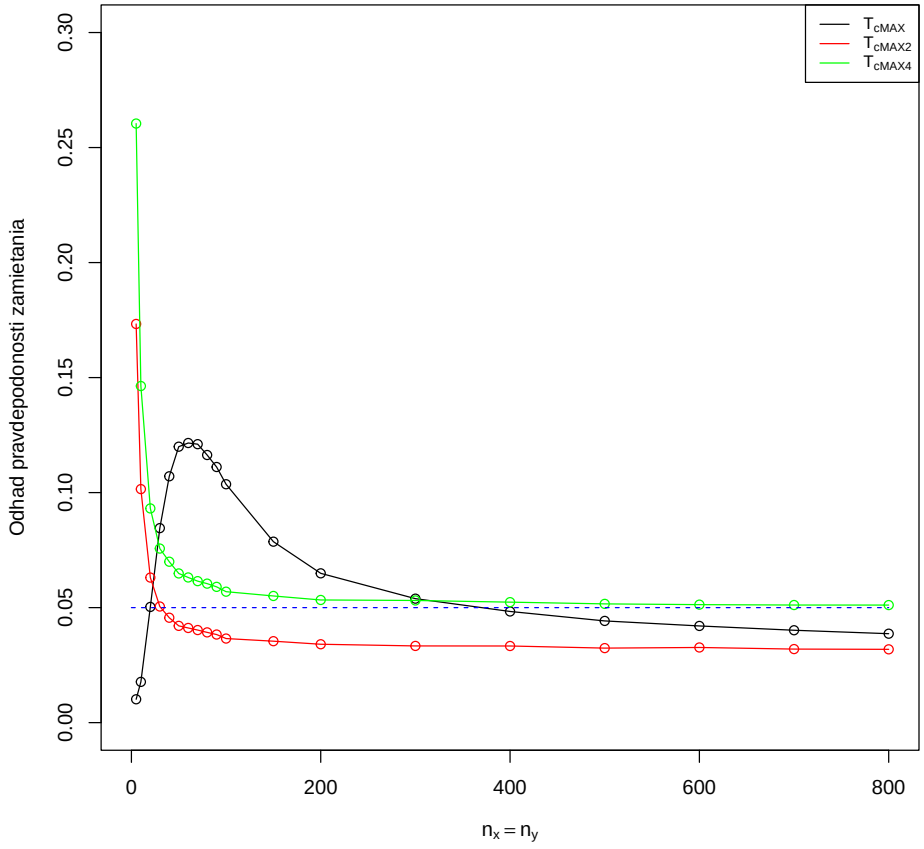
Pri prvom nastavení je naším cieľom odhadnúť pravdepodobnosť chyby prvého druhu jednotlivých testov, nakoľko pri tomto nastavení zovšeobecnená Behrensova-Fisherova hypotéza platí. Na obrázku 6 vidíme, že pravdepodobnosť chyby prvého druhu pre všeobecný $T_{cMAX4}$ test konverguje k želanej hladine 0.05 pre rastúce veľkosti náhodných výberov $n_x$ a $n_y$. Ak sa pozrieme na testy $T_{cMAX}$ a $T_{cMAX2}$, tak vidíme, že pre narastajúce $n_x$ a $n_y$ sa pravdepodobnosť chyby prvého druhu blíži k hladine 0.03. Toto je pravdepodobne následkom toho, že tieto dva testy na odhad kovariančnej matice využívajú predpoklad (1.1), ktorý v simulovaných dátach nie je splnený.

Pri zvyšných nastaveniach sa pozrieme na odhad sily spomínaných testov, nakoľko ani v jednom prípade neplatí zovšeobecnená Behrensova-Fisherova hypotéza. Na obrázkoch 7, 8 a 9 vidíme, že $T_{cMAX4}$ je vo všetkých nami testovaných nastaveniach a veľkostiach náhodných výberov silnejší ako testy $T_{cMAX}$ a $T_{cMAX2}$. Rozdiel síl týchto testov je vo väčšine prípadov dosť výrazný. Napríklad v druhom nastavení (keď vektor parametrov $\theta$ je $(0.185, 0.185;\ -0.105, -0.105;\ 0.035, 0.035)$) a $n_x = n_y = 100$, odhadovaná sila $T_{cMAX4}$ testu je v priemere 91.26%, kým odhady síl $T_{cMAX}$ a $T_{cMAX2}$ testov sú len 34.22% a 50.63%.

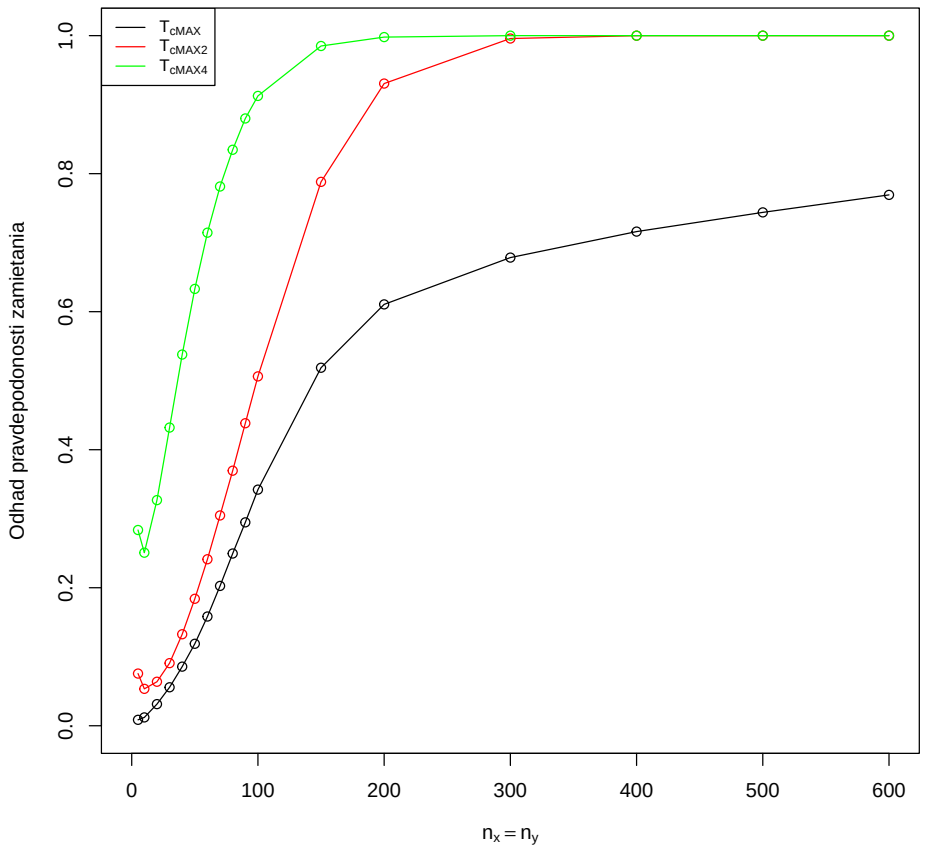
*Poznámka*: Konkrétne hodnoty odhadov pravdepodobnosti zamietania pre jednotlivé testy sú vypísané v tabuľkách v kapitole 8.

*Poznámka*: Podobne ako v kapitole 2.10 si môžeme všimnúť, že sila testov $T_{cMAX2}$ a $T_{cMAX4}$ je pri veľkosti náhodných výberov $n_x = n_y = 10$ nižšia ako pri $n_x = n_y = 5$. Toto je spôsobené tým, že pri menších veľkostiach náhodných výberov sa častejšie stáva, že v niektorom zo zlomkov z (1.5) je čitateľ nenulový a menovateľ nulový. V takom prípade test automaticky zovšeobecnenú Behrensovu-Fisherovu hypotézu zamieta.

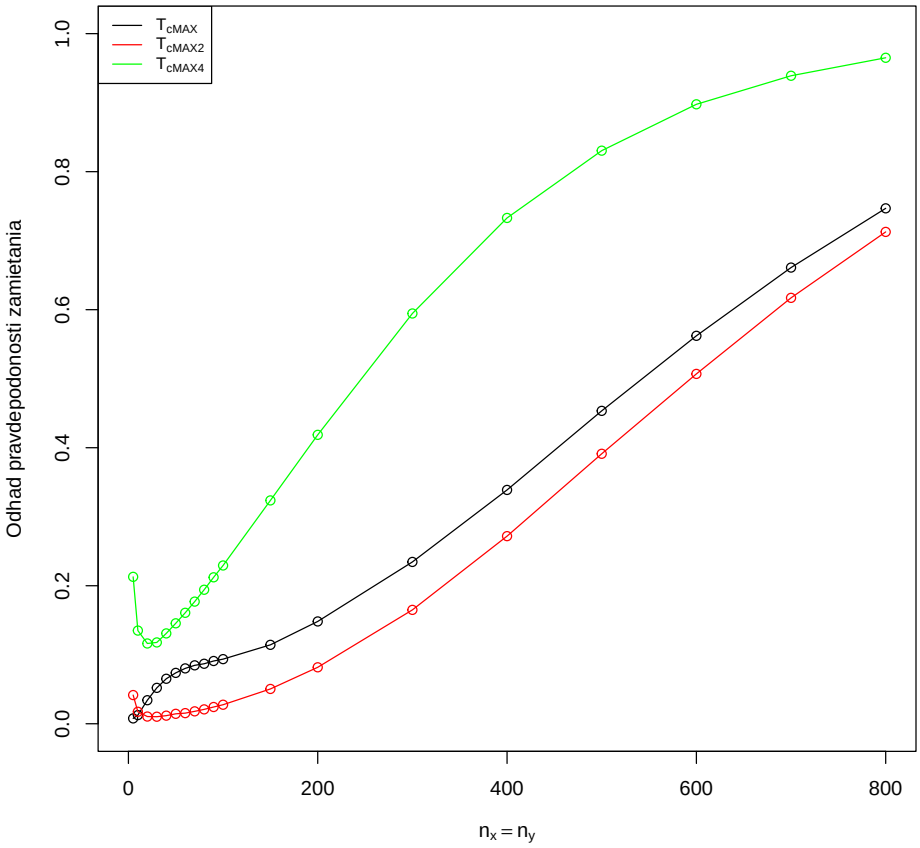
[10] V článku [13, str. 828] sú uvedené iné hodnoty parametrov $\theta$ ako vyšli nám pre toto nastavenie. Simulované sily testov sa v tomto prípade veľmi výrazne líšili od tých, ktoré autori uvádzali v [13, str. 828]



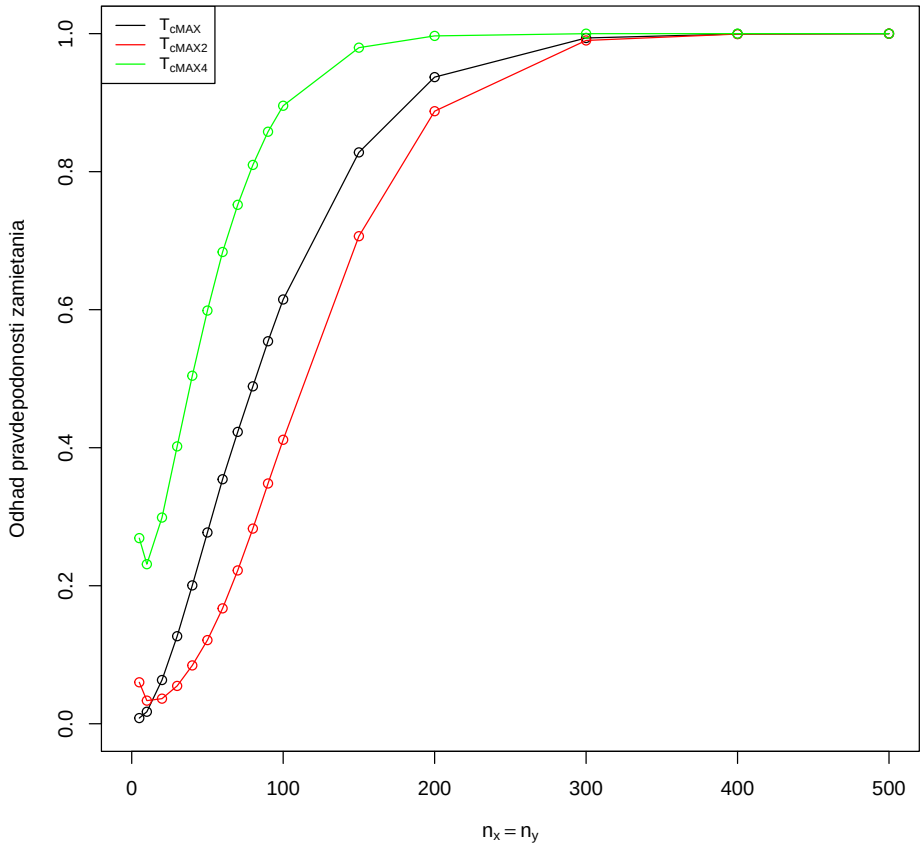

**Obr. 6:** Odhad pravdepodobnosti chyby prvého druhu pre jednotlivé testy pri rastúcej veľkosti náhodných výberov $n_x$ a $n_y$, keď $X$ a $Y$ majú šesť-rozmerné diskrétne rozdelenie a $\theta = (0,0;\ 0,0;\ 0,0)$.




**Obr. 7:** Odhad sily pre jednotlivé testy pri rastúcej veľkosti náhodných výberov $n_x$ a $n_y$, keď $X$ a $Y$ majú šesť-rozmerné diskrétne rozdelenie a $\theta = (0.185, 0.185;\ -0.105, -0.105;\ 0.035, 0.035)$.




**Obr. 8:** Odhad sily pre jednotlivé testy pri rastúcej veľkosti náhodných výberov $n_x$ a $n_y$, keď $X$ a $Y$ majú šesť-rozmerné diskrétne rozdelenie a $\theta = (0.075, 0.075;\ 0.045, 0.045;\ 0.025, -0.025)$.




**Obr. 9:** Odhad sily pre jednotlivé testy pri rastúcej veľkosti náhodných výberov $n_x$ a $n_y$, keď $X$ a $Y$ majú šesť-rozmerné diskrétne rozdelenie a $\theta = (0.0325, 0.0325;\ 0.035, 0.035;\ 0.1875, 0.1875)$.

# 4 Príloha A: Súvis rozdielu priemerných midrankov a priemeru rozdielov indikátorov

V tejto prílohe ukážeme platnosť algebraickej rovnosti, ktorá dáva do súvisu rozdiel priemerných mid-rankov s priemerom rozdielu indikátorov. Veľmi stručný dôkaz je uvedený v [14, str. 2]. My uvedieme jeho podrobnejšiu verziu. Pre reálne čísla $x_1, \ldots, x_n$ a $y_1, \ldots, y_m$ definujeme $R_{xi}$ $(i = 1, \ldots, n)$ a $R_{y\ell}$ $(\ell = 1, \ldots, m)$ ako midranky bodov $x_1, \ldots, x_n$ a $y_1, \ldots, y_m$ vrámci vektora $(x_1, \ldots, x_n, y_1, \ldots, y_m)$.

**Veta 4.1** ([14, str. 2])**.** *Majme reálne čísla $x_1, x_2, \ldots, x_n$ a $y_1, y_2, \ldots, y_m$. Potom platí*

$$\frac{1}{nm}\sum_{i=1}^{n}\sum_{\ell=1}^{m} I(x_i < y_\ell) - I(x_i > y_\ell) = \frac{2}{N}\left(\bar{R}_{y\cdot} - \bar{R}_{x\cdot}\right),$$

*kde $\bar{R}_{x\cdot} = (1/n)\sum_{i=1}^{n} R_{xi}$, $\bar{R}_{y\cdot} = (1/m)\sum_{\ell=1}^{m} R_{y\ell}$ a $N = n + m$.*

*Dôkaz.* Označme

$$\hat{\theta} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{\ell=1}^{m} I(x_i < y_\ell) - I(x_i > y_\ell).$$

Zavedieme aj nasledovné označenie

| | | | |
|---|---|---|---|
| $\alpha_i^- = \#\{i' : x_{i'} < x_i\}$ | $\beta_i^- = \#\{\ell : y_\ell < x_i\}$ | $\gamma_\ell^- = \#\{\ell' : y_{\ell'} < y_\ell\}$ | $\delta_\ell^- = \#\{i : x_i < y_\ell\}$ |
| $\alpha_i^0 = \#\{i' : x_{i'} = x_i\}$ | $\beta_i^0 = \#\{\ell : y_\ell = x_i\}$ | $\gamma_\ell^0 = \#\{\ell' : y_{\ell'} = y_\ell\}$ | $\delta_\ell^0 = \#\{i : x_i = y_\ell\}$ |
| $\alpha_i^+ = \#\{i' : x_{i'} > x_i\}$ | $\beta_i^+ = \#\{\ell : y_\ell > x_i\}$ | $\gamma_\ell^+ = \#\{\ell' : y_{\ell'} > y_\ell\}$ | $\delta_\ell^+ = \#\{i : x_i > y_\ell\}$ |
| $\sum = n$ | $\sum = m$ | $\sum = m$ | $\sum = n$ |

Potom pre midranky (v rámci vektora $x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n$) bodov $x_i$ a $y_\ell$ platí

$$R_{xi} = \alpha_i^- + \beta_i^- + \frac{\alpha_i^0 + \beta_i^0 + 1}{2}$$
$$R_{y\ell} = \gamma_\ell^- + \delta_\ell^- + \frac{\gamma_\ell^0 + \delta_\ell^0 + 1}{2}.$$

Pre $\sum_{\ell=1}^{m} I(x_i < y_\ell)$ platí nasledovná séria rovností

$$\begin{aligned}
\sum_{\ell=1}^{m} I(x_i < y_\ell) &= \beta_i^+ = m - \beta_i^- - \beta_i^0 + (n - \alpha_i^- - \alpha_i^0 - \alpha_i^+) \\
&= (n+m) - \beta_i^- - \alpha_i^- - \left(\frac{\beta_i^0 + \alpha_i^0 + 1}{2}\right) - \left(\frac{\beta_i^0 + \alpha_i^0 - 1}{2}\right) - \alpha_i^+ \\
&= N - R_{xi} - \left(\frac{\beta_i^0 + \alpha_i^0 - 1}{2}\right) - \alpha_i^+. \qquad (4.1)
\end{aligned}$$

Analogicky pre $\sum_{i=1}^{n} I(x_i > y_\ell)$ platí

$$\sum_{i=1}^{n} I(x_i > y_\ell) = N - R_{y\ell} - \left(\frac{\delta_\ell^0 + \gamma_\ell^0 - 1}{2}\right) - \gamma_\ell^+. \tag{4.2}$$

Z (4.1) a (4.2) dostávame, že

$$\begin{aligned} nm\hat{\theta} =& \sum_{i=1}^{n}\sum_{\ell=1}^{m} I(x_i < y_\ell) - I(x_i > y_\ell) \\ =& (n-m)N + m\bar{R}_{y\cdot} - n\bar{R}_{x\cdot} - \sum_{i=1}^{n}\alpha_i^+ + \sum_{\ell=1}^{m}\gamma_\ell^+ \\ & -\frac{1}{2}\sum_{i=1}^{n}(\beta_i^0 + \alpha_i^0 - 1) + \frac{1}{2}\sum_{\ell=1}^{m}(\delta_\ell^0 + \gamma_\ell^0 - 1). \end{aligned} \tag{4.3}$$

Pozrime sa teraz na to, čomu sa rovnajú členy $\sum_{i=1}^{n}\alpha_i^+$ a $\sum_{\ell=1}^{m}\gamma_i^+$. Bez ujmy na všeobecnosti môžeme predpokladať, že

$$x_1 \le x_2 \le \cdots \le x_n.$$

Môžeme teda písať, že

$$x_1 = \cdots = x_{i_1} < x_{i_1+1} = \cdots = x_{i_2} < \cdots < x_{i_{p-1}+1} = \cdots = x_{i_p}, \tag{4.4}$$

kde $i_p = n$. Položme ešte $i_0 = 0$. Platí teda, že

$$0 = i_0 < i_1 < i_2 < \cdots < i_p = n.$$

Ďalej ešte označme

$$\Delta i_h = i_h - i_{h-1}.$$

Pre všetky $i \in (i_{h-1}; i_h]$ platí, že

$$\alpha_i^- = i_{h-1}, \qquad \alpha_i^0 = i_h - i_{h-1} = \Delta i_h, \qquad \alpha_i^+ = n - i_h. \tag{4.5}$$

Z tohto dostávame, že

$$\begin{aligned} \sum_{i=1}^{n}\alpha_i^+ =& \sum_{h=1}^{p}\sum_{i\in(i_{h-1};i_h]}\alpha_i^+ = \sum_{h=1}^{p}\sum_{i\in(i_{h-1};i_h]}(n-i_h) = \sum_{h=1}^{p-1}(i_h - i_{h-1})(n-i_h) \\ =& \sum_{h=1}^{p-1}\Delta i_h\,(\Delta i_{h+1} + \cdots + \Delta i_p) = \frac{1}{2}\left[\left(\sum_{h=1}^{p}\Delta i_h\right)^2 - \sum_{h=1}^{p}(\Delta i_h)^2\right] \\ =& \frac{1}{2}\left[n^2 - \sum_{h=1}^{p}(\Delta i_h)^2\right]. \end{aligned} \tag{4.6}$$

Analogicky označíme $\Delta j_g = j_g - j_{g-1}$ a platí

$$\sum_{\ell=1}^{m} \gamma_\ell^+ = \frac{1}{2}\left[m^2 - \sum_{g=1}^{q} (\Delta j_g)^2\right]. \tag{4.7}$$

Ďalej platí, že

$$\sum_{i=1}^{n} R_{xi} + \sum_{\ell=1}^{m} R_{y\ell} = 1 + 2 + \cdots + N = \frac{N(N+1)}{2},$$

z čoho dostávame, že

$$\bar{R}_{y\cdot} = \frac{N(N+1)}{2m} - \frac{n}{m}\bar{R}_{x\cdot}. \tag{4.8}$$

Do (4.3) dosaďme (4.8), (4.6) a (4.7). Dostávame, že

$$\begin{aligned} nm\hat{\theta} =& (n^2 - m^2) + \frac{N(N+1)}{2} - n\bar{R}_{x\cdot} - n\bar{R}_{x\cdot} \\ & -\frac{1}{2}\left[n^2 - \sum_{h=1}^{p} (\Delta i_h)^2\right] + \frac{1}{2}\left[m^2 - \sum_{g=1}^{q} (\Delta j_g)^2\right] \\ & -\frac{1}{2}\sum_{i=1}^{n}(\beta_i^0 + \alpha_i^0 - 1) + \frac{1}{2}\sum_{\ell=1}^{m}(\delta_\ell^0 + \gamma_\ell^0 - 1) \\ =& \frac{1}{2}(n^2 - m^2) + \frac{N(N+1)}{2} - 2n\bar{R}_{x\cdot} + \frac{1}{2}\sum_{h=1}^{p} (\Delta i_h)^2 - \frac{1}{2}\sum_{g=1}^{q} (\Delta j_g)^2 \\ & -\frac{1}{2}\sum_{i=1}^{n}(\beta_i^0 + \alpha_i^0 - 1) + \frac{1}{2}\sum_{\ell=1}^{m}(\delta_\ell^0 + \gamma_\ell^0 - 1). \end{aligned}$$

Ďalej počítajme $nm\frac{2}{N}\left(\bar{R}_{y\cdot} - \bar{R}_{x\cdot}\right)$. Dosadením (4.8) dostávame, že

$$\begin{aligned} nm\frac{2}{N}\left(\bar{R}_{y\cdot} - \bar{R}_{x\cdot}\right) =& \frac{2mn}{N}\left(\frac{N(N+1)}{2m} - \frac{n}{m}\bar{R}_{x\cdot} - \bar{R}_{x\cdot}\right) \\ =& n(N+1) - \frac{2n}{N}(n\bar{R}_{x\cdot} + m\bar{R}_{x\cdot}) \\ =& n(N+1) - 2n\bar{R}_{x\cdot}. \end{aligned}$$

Označme $\Phi = 2nm\left[\hat{\theta} - \frac{2}{N}\left(\bar{R}_{y\cdot} - \bar{R}_{x\cdot}\right)\right]$. Potom platí, že

$$\begin{aligned} \Phi =& (n^2 - m^2) + N(N+1) - 4n\bar{R}_{x\cdot} + \sum_{h=1}^{p} (\Delta i_h)^2 - \sum_{g=1}^{q} (\Delta j_g)^2 \\ & -\sum_{i=1}^{n}(\beta_i^0 + \alpha_i^0 - 1) + \sum_{\ell=1}^{m}(\delta_\ell^0 + \gamma_\ell^0 - 1) - 2n(N+1) + 4n\bar{R}_{x\cdot} \\ =& (m - n) + \sum_{h=1}^{p} (\Delta i_h)^2 - \sum_{g=1}^{q} (\Delta j_g)^2 - \sum_{i=1}^{n}(\beta_i^0 + \alpha_i^0 - 1) + \sum_{\ell=1}^{m}(\delta_\ell^0 + \gamma_\ell^0 - 1). \end{aligned}$$

Chceme ukázať, že $\Phi = 0$. Ak popreskupujeme členy vo výraze $\Phi$ dostávame, že

$$\begin{aligned}\Phi = &\left(m - \sum_{\ell=1}^{m} 1\right) - \left(n - \sum_{i=1}^{n} 1\right) - \left(\sum_{i=1}^{n} \beta_i^0 - \sum_{\ell=1}^{m} \delta_i^0\right) \\ &- \left(\sum_{i=1}^{n} \alpha_i^0 - \sum_{h=1}^{p} (\Delta i_h)^2\right) + \left(\sum_{\ell=1}^{m} \gamma_\ell^0 - \sum_{g=1}^{q} (\Delta j_g)^2\right).\end{aligned}$$

Prvé dve zátvorky sú nulové triviálne. Pre tretiu zátvorku z definícií $\beta_i^0$ a $\delta_i^0$ vidíme, že počet všetkých $y$-ov, ktoré sa rovnajú $x$-om a všetkých $x$-ov, ktoré sa rovnajú $y$-om je rovnaký. Pre predposlednú zátvorku využijeme (4.4):

$$\sum_{i=1}^{n} \alpha_i^0 = (\Delta i_1 + \cdots + \Delta i_1) + (\Delta i_2 + \ldots \Delta i_2) + \Delta i_3 + \cdots + (\Delta i_p + \cdots + \Delta i_p),$$

pričom v prvej zátvorke je $\Delta i_1$ sčítancov, v druhej $\Delta i_2$, atď. Teda

$$\begin{aligned}\sum_{i=1}^{n} \alpha_i^0 &= (\Delta i_1)^2 + \cdots + (\Delta i_p)^2 \\ &= \sum_{h=1}^{p} (\Delta i_h)^2.\end{aligned}$$

Analogicky $\sum_{\ell=1}^{m} \gamma_\ell^0 = \sum_{g=1}^{q} (\Delta j_g)^2$. Týmto je veta 4.1 dokázaná. □

# 5 Príloha B: Výsledky simulácií z kapitoly 2.10

V tejto prílohe sú uvedené hodnoty výsledkov simulácií uvedených na obrázkoch 2, 3, 4 a 5 v kapitole 2.10. Pri každom z nastavení a dvojici veľkostí náhodných výberov sme simulovali dáta 10 000 krát. Keďže testujeme zovšeobecnenú Behrensovu-Fisherovu hypotézu na hladine 0.05, tak odhadom pravdepodobnosti zamietania pri daných nastaveniach je proporcia testových štatistík, ktorých $p$-hodnota je menšia ako 5%. Každú takúto sériu 10 000 simulácií spustíme 10 krát a ako odhad pravdepodobnosti zamietania zvolíme priemer z odhadov. V tabuľkách 1, 2, 3 a 4 uvádzame aj štandardné odchýlky pre tieto odhady.

| $n_x$ | $n_y$ | $T_{cMAX}$ | $T_{cMAX2}$ | $T_{cMAX2e}$ | $T_{cMAX3}$ | $T_{cMAX3e}$ |
|---|---|---|---|---|---|---|
| 5 | 5 | 2.71% (±0.35%) | 18.05% (±0.34%) | 16.97% (±0.35%) | 0.00% (±0.00%) | 0.00% (±0.00%) |
| 10 | 10 | 4.89% (±0.32%) | 11.02% (±0.33%) | 10.57% (±0.30%) | 0.00% (±0.00%) | 0.00% (±0.00%) |
| 20 | 20 | 9.82% (±0.36%) | 7.95% (±0.20%) | 7.73% (±0.20%) | 1.26% (±0.07%) | 1.29% (±0.07%) |
| 30 | 30 | 12.77% (±0.25%) | 6.81% (±0.22%) | 6.68% (±0.21%) | 2.33% (±0.15%) | 2.33% (±0.15%) |
| 40 | 40 | 14.19% (±0.17%) | 6.36% (±0.31%) | 6.28% (±0.31%) | 2.89% (±0.23%) | 2.87% (±0.23%) |
| 50 | 50 | 14.58% (±0.50%) | 5.98% (±0.20%) | 5.92% (±0.21%) | 3.23% (±0.12%) | 3.21% (±0.13%) |
| 60 | 60 | 14.01% (±0.44%) | 5.93% (±0.27%) | 5.87% (±0.25%) | 3.67% (±0.13%) | 3.64% (±0.13%) |
| 70 | 70 | 13.45% (±0.39%) | 5.86% (±0.24%) | 5.82% (±0.23%) | 3.83% (±0.15%) | 3.80% (±0.16%) |
| 80 | 80 | 13.02% (±0.29%) | 5.66% (±0.24%) | 5.61% (±0.22%) | 3.94% (±0.17%) | 3.91% (±0.18%) |
| 90 | 90 | 12.44% (±0.41%) | 5.53% (±0.26%) | 5.49% (±0.26%) | 4.02% (±0.19%) | 4.00% (±0.18%) |
| 100 | 100 | 11.62% (±0.44%) | 5.54% (±0.17%) | 5.50% (±0.17%) | 4.19% (±0.15%) | 4.16% (±0.15%) |
| 150 | 150 | 9.25% (±0.29%) | 5.31% (±0.20%) | 5.28% (±0.19%) | 4.35% (±0.16%) | 4.34% (±0.16%) |
| 200 | 200 | 8.27% (±0.45%) | 5.25% (±0.29%) | 5.23% (±0.28%) | 4.59% (±0.27%) | 4.58% (±0.28%) |

**Tabuľka 1:** Odhad pravdepodobnosti chyby prvého druhu pre jednotlivé testy pri rastúcej veľkosti náhodných výberov $n_x$ a $n_y$, keď $X$ a $Y$ majú lognormálne rozdelenie a $\theta = (0, 0, 0;\ 0, 0, 0)$.

| $n_x$ | $n_y$ | $T_{cMAX}$ | $T_{cMAX2}$ | $T_{cMAX2e}$ | $T_{cMAX3}$ | $T_{cMAX3e}$ |
|---|---|---|---|---|---|---|
| 5 | 5 | 2.84% (±0.41%) | 25.45% (±0.49%) | 24.19% (±0.48%) | 0.00% (±0.00%) | 0.00% (±0.00%) |
| 10 | 10 | 5.38% (±0.39%) | 23.65% (±0.51%) | 22.89% (±0.48%) | 0.00% (±0.00%) | 0.02% (±0.02%) |
| 20 | 20 | 10.66% (±0.64%) | 30.66% (±0.34%) | 30.13% (±0.39%) | 8.78% (±0.14%) | 9.04% (±0.15%) |
| 30 | 30 | 15.56% (±0.62%) | 39.67% (±0.37%) | 39.31% (±0.39%) | 21.46% (±0.39%) | 21.47% (±0.39%) |
| 40 | 40 | 19.61% (±0.54%) | 49.26% (±0.49%) | 49.00% (±0.51%) | 34.13% (±0.21%) | 34.05% (±0.22%) |
| 50 | 50 | 23.77% (±0.52%) | 57.87% (±0.70%) | 57.62% (±0.68%) | 45.33% (±0.58%) | 45.21% (±0.57%) |
| 60 | 60 | 27.16% (±0.61%) | 65.53% (±0.49%) | 65.33% (±0.50%) | 55.50% (±0.46%) | 55.39% (±0.48%) |
| 70 | 70 | 29.32% (±0.47%) | 72.52% (±0.41%) | 72.38% (±0.42%) | 64.68% (±0.37%) | 64.59% (±0.37%) |
| 80 | 80 | 31.55% (±0.65%) | 78.35% (±0.43%) | 78.21% (±0.44%) | 71.99% (±0.52%) | 71.89% (±0.52%) |
| 90 | 90 | 32.87% (±0.44%) | 83.34% (±0.39%) | 83.23% (±0.40%) | 78.53% (±0.54%) | 78.46% (±0.53%) |
| 100 | 100 | 33.66% (±0.41%) | 87.18% (±0.32%) | 87.11% (±0.32%) | 83.53% (±0.30%) | 83.46% (±0.30%) |
| 150 | 150 | 33.16% (±0.35%) | 97.21% (±0.13%) | 97.20% (±0.13%) | 96.38% (±0.16%) | 96.36% (±0.17%) |
| 200 | 200 | 29.64% (±0.76%) | 99.49% (±0.08%) | 99.49% (±0.08%) | 99.36% (±0.09%) | 99.35% (±0.09%) |

**Tabuľka 2:** Odhad sily pre jednotlivé testy pri rastúcej veľkosti náhodných výberov $n_x$ a $n_y$, keď $X$ a $Y$ majú lognormálne rozdelenie a $\theta = (-0.167, -0.167, -0.167;\ 0.096, 0.096, 0.096)$.

| $n_x$ | $n_y$ | $T_{cMAX}$ | $T_{cMAX2}$ | $T_{cMAX2e}$ | $T_{cMAX3}$ | $T_{cMAX3e}$ |
|---|---|---|---|---|---|---|
| 5 | 5 | 2.75% (±0.34%) | 23.63% (±0.56%) | 22.42% (±0.54%) | 0.00% (±0.00%) | 0.00% (±0.00%) |
| 10 | 10 | 5.53% (±0.33%) | 20.84% (±0.54%) | 20.12% (±0.48%) | 0.00% (±0.01%) | 0.02% (±0.02%) |
| 20 | 20 | 11.81% (±0.39%) | 25.98% (±0.25%) | 25.52% (±0.27%) | 7.55% (±0.15%) | 7.75% (±0.15%) |
| 30 | 30 | 19.08% (±0.56%) | 33.52% (±0.52%) | 33.20% (±0.52%) | 18.19% (±0.29%) | 18.19% (±0.29%) |
| 40 | 40 | 25.52% (±0.46%) | 41.84% (±0.55%) | 41.59% (±0.58%) | 28.94% (±0.38%) | 28.87% (±0.38%) |
| 50 | 50 | 32.39% (±0.68%) | 49.50% (±0.58%) | 49.29% (±0.59%) | 38.78% (±0.46%) | 38.67% (±0.45%) |
| 60 | 60 | 38.95% (±0.60%) | 56.68% (±0.37%) | 56.51% (±0.37%) | 47.76% (±0.34%) | 47.67% (±0.34%) |
| 70 | 70 | 44.35% (±0.54%) | 63.44% (±0.53%) | 63.29% (±0.54%) | 56.18% (±0.40%) | 56.07% (±0.41%) |
| 80 | 80 | 49.43% (±0.89%) | 69.56% (±0.44%) | 69.42% (±0.45%) | 63.48% (±0.57%) | 63.40% (±0.55%) |
| 90 | 90 | 53.89% (±0.59%) | 74.68% (±0.47%) | 74.57% (±0.49%) | 69.84% (±0.57%) | 69.76% (±0.55%) |
| 100 | 100 | 57.81% (±0.81%) | 79.38% (±0.43%) | 79.30% (±0.43%) | 75.39% (±0.44%) | 75.33% (±0.44%) |
| 150 | 150 | 68.13% (±0.44%) | 93.19% (±0.16%) | 93.17% (±0.17%) | 91.87% (±0.18%) | 91.84% (±0.17%) |
| 200 | 200 | 71.94% (±0.32%) | 98.00% (±0.19%) | 98.00% (±0.19%) | 97.63% (±0.21%) | 97.62% (±0.21%) |

**Tabuľka 3:** Odhad sily pre jednotlivé testy pri rastúcej veľkosti náhodných výberov $n_x$ a $n_y$, keď $X$ a $Y$ majú lognormálne rozdelenie a $\theta = (-0.167, -0.167, -0.167;\ 0.096, 0.096, -0.096)$.

| $n_x$ | $n_y$ | $T_{cMAX}$ | $T_{cMAX2}$ | $T_{cMAX2e}$ | $T_{cMAX3}$ | $T_{cMAX3e}$ |
|---|---|---|---|---|---|---|
| 5 | 5 | 2.55% (±0.36%) | 22.32% (±0.28%) | 21.11% (±0.33%) | 0.00% (±0.00%) | 0.00% (±0.00%) |
| 10 | 10 | 5.23% (±0.21%) | 18.04% (±0.35%) | 17.39% (±0.34%) | 0.01% (±0.00%) | 0.03% (±0.02%) |
| 20 | 20 | 12.36% (±0.39%) | 20.32% (±0.48%) | 19.96% (±0.46%) | 5.48% (±0.21%) | 5.63% (±0.20%) |
| 30 | 30 | 19.83% (±0.30%) | 25.09% (±0.48%) | 24.85% (±0.46%) | 13.01% (±0.33%) | 13.00% (±0.33%) |
| 40 | 40 | 27.18% (±0.79%) | 30.69% (±0.33%) | 30.45% (±0.32%) | 20.18% (±0.35%) | 20.12% (±0.35%) |
| 50 | 50 | 34.74% (±0.80%) | 36.60% (±0.48%) | 36.41% (±0.48%) | 27.32% (±0.38%) | 27.24% (±0.39%) |
| 60 | 60 | 41.53% (±0.61%) | 42.52% (±0.40%) | 42.34% (±0.41%) | 34.61% (±0.34%) | 34.52% (±0.35%) |
| 70 | 70 | 48.22% (±0.82%) | 48.42% (±0.37%) | 48.29% (±0.40%) | 41.42% (±0.36%) | 41.32% (±0.35%) |
| 80 | 80 | 54.28% (±0.66%) | 53.78% (±0.51%) | 53.66% (±0.52%) | 47.70% (±0.50%) | 47.58% (±0.50%) |
| 90 | 90 | 59.47% (±0.51%) | 58.98% (±0.38%) | 58.85% (±0.38%) | 53.61% (±0.44%) | 53.54% (±0.42%) |
| 100 | 100 | 64.34% (±0.53%) | 63.90% (±0.46%) | 63.80% (±0.46%) | 59.27% (±0.55%) | 59.18% (±0.55%) |
| 150 | 150 | 81.76% (±0.53%) | 82.04% (±0.52%) | 81.99% (±0.53%) | 79.88% (±0.50%) | 79.83% (±0.50%) |
| 200 | 200 | 90.69% (±0.47%) | 91.82% (±0.30%) | 91.80% (±0.29%) | 90.85% (±0.25%) | 90.84% (±0.25%) |

**Tabuľka 4:** Odhad sily pre jednotlivé testy pri rastúcej veľkosti náhodných výberov $n_x$ a $n_y$, keď $X$ a $Y$ majú lognormálne rozdelenie a $\theta = (0.024, 0.024, 0.024;\ 0.143, 0.143, 0.143)$.

# 6 Príloha C: Pomocné tvrdenia o podmienenej strednej hodnote ku kapitole 2.7

V tejto sekcii uvedieme niekoľko liem, ktoré sa pri počítaní s podmienenou strednou hodnotou často intuitívne používajú, avšak málokedy sú explicitne spomenuté.

**Lema 6.1** (Doob)**.** *Nech $X$ je $m$-rozmerný náhodný vektor a $Y$ je $n$-rozmerný náhodný vektor. Potom $Y$ je $X$-merateľný vtedy a len vtedy, keď existuje borelovské zobrazenie $h : \mathbb{R}^m \to \mathbb{R}^n$ také, že*

$$Y = h(X).$$

*Dôkaz.* Viď [4, Lemma 1.14]. □

**Lema 6.2.** *Nech $g : \mathbb{R}^m \to \mathbb{R}^n$ je borelovská funckia. Nech $X$,$X'$ sú $m$-rozmerné náhodné vektory. Ak $X \stackrel{d}{=} X'$ tak $g(X) \stackrel{d}{=} h(X')$.*

*Dôkaz.* Označme $Y = h(X)$ a $Y' = h(X')$. Vezmime ľubovoľnú borelovskú $B \subseteq \mathbb{R}^n$. Potom

$$P^Y(B) = P(\{\omega \in \Omega \colon h(X(\omega)) \in B\}) = P(\{\omega \in \Omega \colon X(\omega) \in h^{-1}(B)\}) = P^X(h^{-1}(B)).$$

Analogicky $P^{Y'}(B) = P^{X'}(h^{-1}(B))$. Keďže $P^X = P^{X'}$, tak aj $P^Y = P^{Y'}$. □

**Lema 6.3.** *Nech $X$ je $m$-rozmerný náhodný vektor a $Y$ je $n$-rozmerný náhodný vektor. Nech $X$, $Y$ sú nezávislé. Nech $g : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^k$ je borelovská funkcia taká, že $E[g(X, Y)]$ je konečná. Potom pre každé nezávislé náhodné vektory $X'$, $Y'$ také, že $X \stackrel{d}{=} X'$ a $Y \stackrel{d}{=} Y'$ platí, že*

$$E[g(X, Y)|X] \stackrel{d}{=} E[g(X', Y')|X'].$$

*Dôkaz.* Na základe lemy 6.1 existuje borelovské zobrazenie $h : \mathbb{R}^m \to \mathbb{R}^k$ také, že

$$E[g(X, Y)|X] = h(X).$$

Podľa [4, str. 167] $h$ závisí iba na distribúcii $(X, g(X, Y))$, teda závisí iba na distribúcii $(X, Y)$. Keďže $(X, Y)$ má rovnaké rozdelenie ako $(X', Y')$, tak $E[g(X', Y')|X'] = h(X')$. Na dokončenie dôkazu stačí využiť lemu 6.2. □

**Lema 6.4.** *Nech $X$ je $m$-rozmerný náhodný vektor a $Y$ je $n$-rozmerný náhodný vektor. Nech $X$, $Y$ sú nezávislé. Ďalej nech $\bigcirc$ je borelovské zobrazenie z $\mathbb{R}^m \times \mathbb{R}^n$ do $\{Pravda, Nepravda\}$. Nech $h\colon \mathbb{R}^m \to \mathbb{R}$ je funkcia definovaná vzťahom*

$$h(x) = P(x \bigcirc Y).$$

*Potom*

$$P(X \bigcirc Y \mid X) = h(X).$$

*Dôkaz.* Chceme ukázať, že $P(X \bigcirc Y \mid X) = E[I(X \bigcirc Y) \mid X]$ je skoro isto rovné $h(X)$. To dosiahneme tak, že dokážeme, že $h(X)$ spĺňa dve podmienky z definície podmienenej strednej hodnoty. Tie sú

1. $h(X)$ je $X$-merateľná;
2. pre každé $A \in \sigma(X)$ platí, že $\int_A h(X)\,dP = \int_A I(X \bigcirc Y)\,dP$.

Z lemy 6.1 vieme, že $h(X)$ je $X$-merateľná. Vezmime ľubovoľné $A \in \sigma(X)$; chceme ukázať, že

$$E[h(X) \cdot 1_A] = E[I(X \bigcirc Y) \cdot 1_A],$$

kde $1_A(\omega) = I(\omega \in A)$. Keďže $\sigma(X) = X^{-1}(\mathcal{B})$, existuje borelovská $B \subseteq \mathbb{R}^m$ taká, že $A = X^{-1}(B)$. Jednoducho máme, že

$$\begin{aligned} E[h(X) \cdot 1_A] &= \int_\Omega h(X(\omega)) \cdot I(X(\omega) \in B)\,dP(\omega) \\ &= \int_{\mathbb{R}^m} h(x) I(x \in B)\,dF_X(x). \end{aligned}$$

Ďalej,

$$\begin{aligned} E[I(X \bigcirc Y) \cdot 1_A] &= \int_\Omega I(X(\omega) \bigcirc Y(\omega)) \cdot I(X(\omega) \in B)\,dP(\omega) \\ &= \int_{\mathbb{R}^m \times \mathbb{R}^n} I(x \bigcirc y) I(x \in B)\,dF_{(X,Y)}(x,y). \end{aligned}$$

Keďže $X$ a $Y$ sú nezávislé, tak $F_{(X,Y)}(x,y) = F_X(x)F_Y(y)$. Z Fubiniho vety máme

$$\begin{aligned} E[I(X \bigcirc Y) \cdot 1_A] &= \int_{\mathbb{R}^m} \left( I(x \in B) \cdot \int_{\mathbb{R}^n} I(x \bigcirc y)\,dF_Y(y) \right) dF_X(x) \\ &= \int_{\mathbb{R}^m} I(x \in B) \cdot E[I(x \bigcirc Y)]\,dF_X(x) \\ &= \int_{\mathbb{R}^m} I(x \in B) \cdot h(x)\,dF_X(x) \\ &= E[h(X) \cdot 1_A]. \end{aligned}$$

Tým je lema dokázaná. □

**Lema 6.5.** *Nech $X$, $Y$, $Z$ sú postupne $m, n, o$-rozmerné náhodné vektory také, že $Y$ nezávisí od $(X, Z)$. Nech $g : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ je borelovská funkcia taká, že $E[g(X, Y)]$ je konečná. Potom platí, že*

$$E[g(X,Y)|(X,Z)] = E[g(X,Y)|X] \qquad \textit{skoro isto.}$$

K tejto leme uvedieme tri dôkazy, pričom prvé dva z nich budú vyžadovať dodatočné predpoklady o $X$, $Y$, $Z$. Najprv ukážeme, že lema 6.5 platí v prípade, ak rozdelenie $(m+n+o)$-rozmerného náhodného vektora $(X, Y, Z)$ je spojité. Následne ju dokážeme pre diskrétny prípad. Nakoniec uvedieme všeobecný dôkaz bez dodatočných predpokladov na rozdelenia náhodných vektorov $X$,$Y$ a $Z$. Pripomeňme špeciálne vyjadrenia podmienenej strednej hodnoty pre spojitý a diskrétny prípad.

Nech $S$, $T$ sú náhodné vektory (s dimenziami $m$, $n$) také, že $(S, T)$ má hustotu $f_{(S,T)}$. Pre každé $s \in \mathbb{R}^m$ a $t \in \mathbb{R}^n$ položme

$$f_{S|T}(s|t) = \begin{cases} f_{(S,T)}(s,t)/f_T(t) & \text{ak } f_T(t) > 0; \\ 0 & \text{ak } f_T(t) = 0, \end{cases}$$

kde $f_T$ je marginálna hustota $T$. Nech $\phi : \mathbb{R}^m \to \mathbb{R}$ je borelovská funkcia taká, že $E[\phi(S)]$ existuje a je konečná. Potom, podľa [11, str. 87],

$$E[\phi(S)|T] = \psi(T) \qquad \text{skoro isto,} \tag{6.1}$$

kde $\psi(t) = \int_{\mathbb{R}^m} \phi(s) f_{S|T}(s|t)\, ds$. (Autor v [11, str. 87] uvažuje len prípad $n = m = 1$. Dôkaz pre náhodné vektory je však analogický.)

Nech $S$, $T$ sú náhodné vektory také, že $T$ je diskrétny. Nech $\text{supp}(T) = \{t_1, t_2, \dots\}$. Označme $A_i = \{\omega \in \Omega : T(\omega) = t_i\}$. Potom, podľa [4, str. 181, cv. 10], pre každé $i$ platí

$$E[S|T](\omega) = E[S \cdot 1_{A_i}]/P(A_i) \qquad (\omega \in A_i). \tag{6.2}$$

Teda špeciálne ak aj $S$ je diskrétny, tak

$$E[S|T](\omega) = \sum_{s \in \text{supp}(S)} s \cdot P(S = s \mid T = t) \qquad (t = T(\omega)). \tag{6.3}$$

*Dôkaz.* (spojitý prípad) Predpokladajme najskôr, že $(X, Y, Z)$ je spojitý náhodný vektor. Označme výrazom $f_X$ hustotu vektora $X$, $f_Y$ hustotu vektora $Y$ a $f_{(X,Y)}$ združenú hustotu $(X, Y)$. Ďalej počítajme $E[g(X, Y)|X]$. Podľa (6.1) skoro isto platí, že (kde $x = X(\omega)$)

$$E[g(X,Y)|X](\omega) = \int_{\mathbb{R}^n} g(x,y) f(y|x)\ dy = \int_{\mathbb{R}^n} g(x,y) \frac{f_{(X,Y)}(x,y)}{f_X(x)}\ dy.$$

Keďže $X$ a $Y$ sú nezávislé, tak platí

$$E[g(X,Y)|X](\omega) = \int_{\mathbb{R}^n} g(x,y)\frac{f_Y(y)f_X(x)}{f_X(x)}dy = \int_{\mathbb{R}^n} g(x,y)f_Y(y)\ dy.$$

Počítajme $E[g(X,Y)|(X,Z)]$, pričom $f_{(X,Y,Z)}$ je združená hustota vektora $(X,Y,Z)$ a $f_{(X,Z)}$ je marginálna hustota $(X,Z)$. Ak označíme $x = X(\omega)$, $z = Z(\omega)$, tak podľa (6.1) skoro isto platí

$$E[g(X,Y)|(X,Z)](\omega) = \int_{\mathbb{R}^n} g(x,y)f(y|(x,z))dy = \int_{\mathbb{R}^n} g(x,y)\frac{f_{(X,Y,Z)}(x,y,z)}{f_{(X,Z)}(x,z)}\ dy.$$

Keďže $(X,Z)$ a $Y$ sú nezávislé, tak skoro isto platí

$$\begin{aligned} E[g(X,Y)|(X,Z)](\omega) &= \int_{\mathbb{R}^n} g(x,y)\frac{f_Y(y)f_{(X,Z)}(x,z)}{f_{(X,Z)}(x,z)}\ dy = \int_{\mathbb{R}^n} g(x,y)f_Y(y)\ dy \\ &= E[g(X,Y)|X](\omega). \end{aligned}$$

□

*Dôkaz.* (diskrétny prípad) Diskrétny prípad je analogický. Zmení sa len to, že nahradíme integrál sumou a hustoty budú nahradené pravdepodobnostnými funkciami. Pre úplnosť uvádzame aj detaily. Nech $(X,Y)$ je diskrétny náhodný vektor. Z (6.3) dostávame, že

$$E[g(X,Y)|X](\omega) = \sum_{y\in\text{supp}(Y)} g(x,y)P(Y=y|X=x) = \sum_{y\in\text{supp}(Y)} g(x,y)\frac{P(Y=y,X=x)}{P(X=x)}.$$

Keďže $X$ a $Y$ sú nezávislé, tak platí

$$E[g(X,Y)|X](\omega) = \sum_{y\in\text{supp}(Y)} g(x,y)\frac{P(Y=y)P(X=x)}{P(X=x)} = \sum_{y\in\text{supp}(Y)} g(x,y)P(Y=y).$$

Nech $(X,Y,Z)$ je diskrétny náhodný vektor. Počítajme $E[g(X,Y)|(X,Z)]$. Z (6.3) dostávame, že

$$\begin{aligned} E[g(X,Y)|(X,Z)](\omega) &= \sum_{y\in\text{supp}(Y)} g(x,y)P(Y=y|X=x,Z=z) \\ &= \sum_{y\in\text{supp}(Y)} g(x,y)\frac{P(Y=y,X=x,Z=z)}{P(X=x,Z=z)}. \end{aligned}$$

Keďže $(X,Z)$ a $Y$ sú nezávislé, tak skoro isto platí

$$\begin{aligned} E[g(X,Y)|(X,Z)](\omega) &= \sum_{y\in\text{supp}(Y)} g(x,y)\frac{P(Y=y)P(X=x,Z=z)}{P(X=x,Z=z)} \\ &= \sum_{y\in\text{supp}(Y)} g(x,y)P(Y=y) \\ &= E[g(X,Y)|X](\omega). \end{aligned}$$

□

*Dôkaz.* (všeobecný prípad) Teraz sa pozrieme na všeobecný prípad rozdelení náhodných vektorov $X$, $Y$ a $Z$. Označme $E[g(x, Y)] = h(x)$.

V prvej časti dôkazu tejto lemy ukážeme, že $h(X) = E[g(X, Y)|X]$ skoro isto. Ukážeme to dôkazom toho, že $h(X)$ spĺňa dve podmienky z definície podmienenej strednej hodnoty. Z lemy 6.1 vieme, že $h(X)$ je $X$-merateľná. Vezmime ľubovoľné $A \in \sigma(X)$; chceme ukázať, že

$$E[h(X) \cdot 1_A] = E[g(X, Y) \cdot 1_A],$$

kde $1_A(\omega) = I(\omega \in A)$. Keďže $\sigma(X) = X^{-1}(\mathcal{B})$, existuje borelovská $B \subseteq \mathbb{R}^m$ taká, že $A = X^{-1}(B)$. Jednoducho máme, že

$$\begin{aligned} E[h(X) \cdot 1_A] &= \int_\Omega h(X(\omega)) \cdot I(X(\omega) \in B)\, dP(\omega) \\ &= \int_{\mathbb{R}^m} h(x) I(x \in B)\, dF_X(x). \end{aligned}$$

Ďalej,

$$\begin{aligned} E[g(X, Y) \cdot 1_A] &= \int_\Omega g(X(\omega), Y(\omega)) \cdot I(X(\omega) \in B)\, dP(\omega) \\ &= \int_{\mathbb{R}^m \times \mathbb{R}^n} g(x, y) I(x \in B)\, dF_{(X,Y)}(x, y). \end{aligned}$$

Keďže $X$ a $Y$ sú nezávislé, tak $F_{(X,Y)}(x, y) = F_X(x) F_Y(y)$. Z Fubiniho vety máme

$$\begin{aligned} E[g(X, Y) \cdot 1_A] &= \int_{\mathbb{R}^m} \left( I(x \in B) \cdot \int_{\mathbb{R}^n} g(x, y)\, dF_Y(y) \right) dF_X(x) \\ &= \int_{\mathbb{R}^m} I(x \in B) \cdot E[g(x, Y)]\, dF_X(x) \\ &= \int_{\mathbb{R}^m} I(x \in B) \cdot h(x)\, dF_X(x) \\ &= E[h(X) \cdot 1_A]. \end{aligned}$$

Tým sme dokázali, že $h(X) = E[g(X, Y)|X]$ skoro isto.

Dôkaz, že $h(X) = E[g(X, Y)|(X, Z)]$ skoro isto, je analogický. Z toho, že $h(X)$ je $X$-merateľná vyplýva, že $h(X)$ je aj $(X, Z)$-merateľná. Vezmime ľubovoľné $A \in \sigma(X, Z)$; chceme ukázať, že

$$E[h(X) \cdot 1_A] = E[g(X, Y) \cdot 1_A],$$

kde $1_A(\omega) = I(\omega \in A)$. Keďže $\sigma(X, Z) = (X, Z)^{-1}(\mathcal{B})$, existuje borelovská $B \subseteq \mathbb{R}^m \times \mathbb{R}^o$ taká, že $A = (X, Z)^{-1}(B)$. Jednoducho máme, že

$$\begin{aligned} E[h(X) \cdot 1_A] &= \int_\Omega h(X(\omega)) \cdot I(X(\omega) \in B)\, dP(\omega) \\ &= \int_{\mathbb{R}^m} h(x) I(x \in B)\, dF_{(X,Z)}(x, z). \end{aligned}$$

Ďalej,

$$\begin{aligned} E[g(X,Y)\cdot 1_A] &= \int_\Omega g(X(\omega),Y(\omega))\cdot I(X(\omega)\in B)\,dP(\omega) \\ &= \int_{\mathbb{R}^m\times\mathbb{R}^n\times\mathbb{R}^o} g(x,y)I(x\in B)\,dF_{(X,Y,Z)}(x,y,z). \end{aligned}$$

Keďže $(X,Z)$ a $Y$ sú nezávislé, tak $F_{(X,Y,Z)}(x,y,Z) = F_{(X,Z)}(x,z)F_Y(y)$. Z Fubiniho vety máme

$$\begin{aligned} E[g(X,Y)\cdot 1_A] &= \int_{\mathbb{R}^m\times\mathbb{R}^o} \left( I(x\in B)\cdot \int_{\mathbb{R}^n} g(x,y)\,dF_Y(y)\right) dF_{(X,Z)}(x,z) \\ &= \int_{\mathbb{R}^m\times\mathbb{R}^o} I(x\in B)\cdot E[g(x,Y)]\,dF_{(X,Z)}(x,z) \\ &= \int_{\mathbb{R}^m\times\mathbb{R}^o} I(x\in B)\cdot h(x)\,dF_{(X,Z)}(x,z) \\ &= E[h(X)\cdot 1_A]. \end{aligned}$$

Týmto je lema dokázaná. □

# 7 Príloha D: Pomocné tvrdenia pre dôkaz konzistentnosti odhadu kovariančnej matice z kapitoly 3.6

V tejto sekcii uvedieme niekoľko pomocných tvrdení, ktoré sa využili pri dôkaze konzistencie odhadu asymptotickej kovariančnej matice v kapitole 3.6.

**Veta 7.1** (Špeciálny prípad Slutského vety)**.** *Nech* $\{T_n\}_{n=1}^{\infty}$, $\{U_n\}_{n=1}^{\infty}$ *sú postupnosti náhodných premenných. Nech* $T_n$ *konverguje podľa pravdepodobnosti ku konštante* $\tau$ *a nech* $U_n$ *konverguje ku konštante* $\upsilon$*. Potom*

$$T_n + U_n \xrightarrow{P} \tau + \upsilon$$
$$T_n \cdot U_n \xrightarrow{P} \tau \cdot \upsilon.$$

*Dôkaz.* Toto tvrdenie je dôsledkom Slutského vety [7, str. 19, Theorem (Slutsky)] a vety o tom, že konvergencia podľa pravdepodobnosti ku konštante je ekvivalentná s konvergenciou v distribúcii ku konštante [7, str. 19, Corollary B]. □

**Lema 7.2.** *Nech* $\hat{T}_n$ *je odhad parametra* $\tau$*. Potom ak je odhad* $\hat{T}_n$ *asymptoticky nevychýlený (teda* $\lim_{n\to\infty} E\left[\hat{T}_n\right] = \tau$*) a jeho variancia konverguje k nule, tak* $\hat{T}_n$ *je konzistentným odhadom* $\tau$*.*

*Dôkaz.* Dôkaz, ktorý uvádzame je podobný dôkazu, ktorý je uvedený v [9, str. 1].

Nech $\hat{T}_n$ je odhad parametra $\tau$. Ďalej využijeme Markovovu nerovnosť [1, str. 120], ktorá hovorí, že

$$P(|X| > \varepsilon) \leq \frac{E\left(|X|^k\right)}{\varepsilon^k}.$$

Chceme ukázať, že pre každé $\varepsilon > 0$ platí

$$\lim_{n\to\infty} P(|\hat{T}_n - \tau| > \varepsilon) = 0.$$

Z Markovovej nerovnosti vieme, že

$$P(|\hat{T}_n - \tau| > \varepsilon) \leq \frac{E[|\hat{T}_n - \tau|^2]}{\varepsilon^2}.$$

Pre strednú hodnotu na pravej strane platí

$$\begin{aligned} E\left[\left(\hat{T}_n - \tau\right)^2\right] &= E\left[\left(\hat{T}_n - E\left[\hat{T}_n\right] + E\left[\hat{T}_n\right] - \tau\right)^2\right] \\ &= E\left[\left(\hat{T}_n - E\left[\hat{T}_n\right]\right)^2\right] + E\left[\left(E\left[\hat{T}_n\right] - \tau\right)^2\right] \\ &= \operatorname{var}\left[\hat{T}_n\right] + \left(E\left[\hat{T}_n\right] - \tau\right)^2. \end{aligned}$$

Teda dostávame, že

$$P(|\hat{T}_n - \tau| > \varepsilon) \leq \frac{\text{var}\left[\hat{T}_n\right]}{\varepsilon^2} + \frac{\left(E\left[\hat{T}_n\right] - \tau\right)^2}{\varepsilon^2}.$$

V prípade, že je odhad asymptoticky nevychýlený a jeho variancie konverguje k nule, platí nasledovné

$$0 \leq \lim_{n\to\infty} P(|\hat{T}_n - \tau| > \varepsilon) \leq \lim_{n\to\infty} \frac{\text{var}\left[\hat{T}_n\right]}{\varepsilon^2} + \frac{\left(E\left[\hat{T}_n\right] - \tau\right)^2}{\varepsilon^2} = 0.$$

Teda sme ukázali, že takýto odhad je aj konzistentný. □

**Lema 7.3.** *Majme maticu $\Sigma = (\sigma_{ab})_{J\times J}$ a maticu $S_n = (s_{ab,n})_{J\times J}$, ktorá je jej odhadom. Potom ak pre všetky $a, b \in \{1, \ldots, J\}$ platí, že $s_{ab,n}$ je konzistentným odhadom $\sigma_{ab}$, tak $S_n$ je konzistentným odhadom $\Sigma$.*

*Dôkaz.* Keďže pre každé $a, b \in \{1, \ldots, J\}$ je $s_{ab,n}$ konzistentným odhadom $\sigma_{ab}$, tak platí nasledovné: pre každé $\varepsilon > 0$, $\delta > 0$ a $a, b \in \{1, \ldots, J\}$ existuje $n_{ab}$ také, že

$$P(|s_{ab,n} - \sigma_{ab}| > \varepsilon) < \delta$$

pre každé $n \geq n_{ab}$. Potom pre každé $N \geq \max_{a,b}\{n_{ab}\}$ platí, že

$$P(||S_N - \Sigma||_\infty > \varepsilon) < \delta,$$

kde pre maticu $H = (h_{ab})_{J\times J}$ je

$$||H||_\infty = \max_{a,b\in\{1,\cdots,J\}} |h_{ab}|.$$

Dostávame teda, že

$$\lim_{N\to\infty} P(||S_N - \Sigma||_\infty > \varepsilon) = 0,$$

Keďže všetky maticové normy sú ekvivalentné, tak dostávame, že

$$\lim_{N\to\infty} P(||S_N - \Sigma|| > \varepsilon) = 0,$$

a teda vidíme, že $S_N$ je konzistentným odhadom matice $\Sigma$. □

# 8 Príloha E: Výsledky simulácií z kapitoly 3.9

V tejto prílohe sú uvedené hodnoty výsledkov simulácií uvedených na obrázkoch 6, 7, 8 a 9 v kapitole 3.9. Pri každom z nastavení a dvojici veľkostí náhodných výberov sme simulovali dáta 10 000 krát. Keďže testujeme zovšeobecnenú Behrensovu-Fisherovu hypotézu na hladine 0.05, tak odhadom pravdepodobnosti zamietania pri daných nastaveniach je proporcia testových štatistík, ktorých $p$-hodnota je menšia ako 5%. Každú takúto sériu 10 000 simulácií spustíme 10 krát a ako odhad pravdepodobnosti zamietania zvolíme priemer z odhadov. V tabuľkách 5, 6, 7 a 8 uvádzame aj štandardné odchýlky pre tieto odhady.

| $n_x$ | $n_y$ | $T_{cMAX}$ | $T_{cMAX2}$ | $T_{cMAX4}$ |
|---|---|---|---|---|
| 5 | 5 | 1.02% (±0.27%) | 17.33% (±0.35%) | 26.04% (±0.46%) |
| 10 | 10 | 1.77% (±0.16%) | 10.15% (±0.34%) | 14.63% (±0.41%) |
| 20 | 20 | 5.03% (±0.29%) | 6.31% (±0.16%) | 9.31% (±0.29%) |
| 30 | 30 | 8.46% (±0.23%) | 5.05% (±0.15%) | 7.57% (±0.20%) |
| 40 | 40 | 10.71% (±0.31%) | 4.56% (±0.13%) | 7.00% (±0.19%) |
| 50 | 50 | 12.00% (±0.30%) | 4.21% (±0.17%) | 6.49% (±0.19%) |
| 60 | 60 | 12.16% (±0.28%) | 4.12% (±0.16%) | 6.31% (±0.20%) |
| 70 | 70 | 12.11% (±0.23%) | 4.03% (±0.16%) | 6.15% (±0.18%) |
| 80 | 80 | 11.63% (±0.33%) | 3.92% (±0.24%) | 6.04% (±0.24%) |
| 90 | 90 | 11.11% (±0.30%) | 3.83% (±0.20%) | 5.90% (±0.19%) |
| 100 | 100 | 10.36% (±0.39%) | 3.66% (±0.15%) | 5.69% (±0.17%) |
| 150 | 150 | 7.87% (±0.20%) | 3.54% (±0.11%) | 5.51% (±0.14%) |
| 200 | 200 | 6.49% (±0.21%) | 3.41% (±0.21%) | 5.33% (±0.20%) |
| 300 | 300 | 5.39% (±0.19%) | 3.34% (±0.18%) | 5.31% (±0.18%) |
| 400 | 400 | 4.83% (±0.19%) | 3.33% (±0.17%) | 5.24% (±0.18%) |
| 500 | 500 | 4.42% (±0.17%) | 3.24% (±0.14%) | 5.16% (±0.18%) |
| 600 | 600 | 4.20% (±0.17%) | 3.27% (±0.18%) | 5.13% (±0.17%) |
| 700 | 700 | 4.02% (±0.28%) | 3.20% (±0.23%) | 5.11% (±0.33%) |
| 800 | 800 | 3.87% (±0.21%) | 3.19% (±0.20%) | 5.11% (±0.26%) |

**Tabuľka 5:** Odhad pravdepodobnosti chyby prvého druhu pre jednotlivé testy pri rastúcej veľkosti náhodných výberov $n_x$ a $n_y$, keď $X$ a $Y$ majú šesť-rozmerné diskrétne rozdelenie a $\theta = (0,0;\ 0,0;\ 0,0)$.

| $n_x$ | $n_y$ | $T_{cMAX}$ | $T_{cMAX2}$ | $T_{cMAX4}$ |
|---|---|---|---|---|
| 5 | 5 | 0.86% (±0.20%) | 7.56% (±0.25%) | 28.36% (±0.43%) |
| 10 | 10 | 1.21% (±0.15%) | 5.35% (±0.20%) | 25.08% (±0.32%) |
| 20 | 20 | 3.14% (±0.23%) | 6.38% (±0.23%) | 32.69% (±0.45%) |
| 30 | 30 | 5.58% (±0.36%) | 9.07% (±0.30%) | 43.21% (±0.57%) |
| 40 | 40 | 8.57% (±0.34%) | 13.26% (±0.37%) | 53.79% (±0.58%) |
| 50 | 50 | 11.88% (±0.38%) | 18.40% (±0.40%) | 63.30% (±0.45%) |
| 60 | 60 | 15.83% (±0.36%) | 24.13% (±0.41%) | 71.45% (±0.53%) |
| 70 | 70 | 20.26% (±0.34%) | 30.47% (±0.31%) | 78.14% (±0.41%) |
| 80 | 80 | 24.96% (±0.52%) | 36.96% (±0.56%) | 83.46% (±0.45%) |
| 90 | 90 | 29.47% (±0.44%) | 43.84% (±0.51%) | 88.00% (±0.29%) |
| 100 | 100 | 34.22% (±0.44%) | 50.63% (±0.43%) | 91.26% (±0.24%) |
| 150 | 150 | 51.88% (±0.45%) | 78.81% (±0.47%) | 98.50% (±0.15%) |
| 200 | 200 | 61.06% (±0.70%) | 93.06% (±0.30%) | 99.79% (±0.03%) |
| 300 | 300 | 67.84% (±0.40%) | 99.58% (±0.08%) | 100.00% (±0.00%) |
| 400 | 400 | 71.61% (±0.41%) | 99.99% (±0.01%) | 100.00% (±0.00%) |
| 500 | 500 | 74.39% (±0.41%) | 100.00% (±0.00%) | 100.00% (±0.00%) |
| 600 | 600 | 76.92% (±0.37%) | 100.00% (±0.00%) | 100.00% (±0.00%) |

**Tabuľka 6:** Odhad sily pre jednotlivé testy pri rastúcej veľkosti náhodných výberov $n_x$ a $n_y$, keď $X$ a $Y$ majú šesť-rozmerné diskrétne rozdelenie a $\theta = (0.185, 0.185;\ -0.105, -0.105;\ 0.035, 0.035)$.

| $n_x$ | $n_y$ | $T_{cMAX}$ | $T_{cMAX2}$ | $T_{cMAX4}$ |
|---|---|---|---|---|
| 5 | 5 | 0.77% (±0.11%) | 4.16% (±0.21%) | 21.29% (±0.45%) |
| 10 | 10 | 1.25% (±0.11%) | 1.74% (±0.09%) | 13.50% (±0.50%) |
| 20 | 20 | 3.41% (±0.15%) | 1.04% (±0.07%) | 11.63% (±0.23%) |
| 30 | 30 | 5.20% (±0.18%) | 1.02% (±0.12%) | 11.78% (±0.34%) |
| 40 | 40 | 6.52% (±0.16%) | 1.18% (±0.10%) | 13.09% (±0.28%) |
| 50 | 50 | 7.37% (±0.15%) | 1.44% (±0.12%) | 14.56% (±0.36%) |
| 60 | 60 | 8.02% (±0.26%) | 1.54% (±0.09%) | 16.08% (±0.43%) |
| 70 | 70 | 8.46% (±0.27%) | 1.80% (±0.15%) | 17.70% (±0.35%) |
| 80 | 80 | 8.69% (±0.20%) | 2.08% (±0.11%) | 19.42% (±0.37%) |
| 90 | 90 | 9.09% (±0.29%) | 2.43% (±0.14%) | 21.22% (±0.28%) |
| 100 | 100 | 9.36% (±0.30%) | 2.76% (±0.16%) | 22.93% (±0.59%) |
| 150 | 150 | 11.44% (±0.30%) | 5.04% (±0.18%) | 32.36% (±0.32%) |
| 200 | 200 | 14.82% (±0.24%) | 8.18% (±0.24%) | 41.87% (±0.38%) |
| 300 | 300 | 23.45% (±0.26%) | 16.50% (±0.31%) | 59.45% (±0.43%) |
| 400 | 400 | 33.88% (±0.42%) | 27.19% (±0.49%) | 73.29% (±0.38%) |
| 500 | 500 | 45.34% (±0.31%) | 39.12% (±0.37%) | 83.04% (±0.24%) |
| 600 | 600 | 56.22% (±0.44%) | 50.70% (±0.40%) | 89.76% (±0.23%) |
| 700 | 700 | 66.10% (±0.42%) | 61.71% (±0.37%) | 93.90% (±0.31%) |
| 800 | 800 | 74.69% (±0.34%) | 71.27% (±0.21%) | 96.50% (±0.25%) |

**Tabuľka 7:** Odhad sily pre jednotlivé testy pri rastúcej veľkosti náhodných výberov $n_x$ a $n_y$, keď $X$ a $Y$ majú šesť-rozmerné diskrétne rozdelenie a $\theta = (0.075, 0.075;\ 0.045, 0.045;\ 0.025, -0.025)$.

| $n_x$ | $n_y$ | $T_{cMAX}$ | $T_{cMAX2}$ | $T_{cMAX4}$ |
|---|---|---|---|---|
| 5 | 5 | 0.82% (±0.19%) | 6.01% (±0.15%) | 26.90% (±0.41%) |
| 10 | 10 | 1.74% (±0.16%) | 3.35% (±0.15%) | 23.12% (±0.37%) |
| 20 | 20 | 6.34% (±0.19%) | 3.64% (±0.20%) | 29.88% (±0.55%) |
| 30 | 30 | 12.68% (±0.24%) | 5.48% (±0.21%) | 40.19% (±0.57%) |
| 40 | 40 | 20.05% (±0.38%) | 8.45% (±0.32%) | 50.43% (±0.53%) |
| 50 | 50 | 27.72% (±0.68%) | 12.12% (±0.11%) | 59.87% (±0.69%) |
| 60 | 60 | 35.43% (±0.72%) | 16.72% (±0.41%) | 68.36% (±0.66%) |
| 70 | 70 | 42.29% (±0.48%) | 22.22% (±0.34%) | 75.20% (±0.59%) |
| 80 | 80 | 48.89% (±0.52%) | 28.28% (±0.44%) | 80.98% (±0.37%) |
| 90 | 90 | 55.42% (±0.38%) | 34.82% (±0.51%) | 85.79% (±0.22%) |
| 100 | 100 | 61.48% (±0.43%) | 41.15% (±0.36%) | 89.55% (±0.19%) |
| 150 | 150 | 82.79% (±0.34%) | 70.65% (±0.31%) | 97.97% (±0.10%) |
| 200 | 200 | 93.70% (±0.17%) | 88.76% (±0.32%) | 99.67% (±0.06%) |
| 300 | 300 | 99.37% (±0.08%) | 99.01% (±0.08%) | 99.99% (±0.01%) |
| 400 | 400 | 99.95% (±0.02%) | 99.94% (±0.01%) | 100.00% (±0.00%) |
| 500 | 500 | 99.99% (±0.01%) | 100.00% (±0.00%) | 100.00% (±0.00%) |

**Tabuľka 8:** Odhad sily pre jednotlivé testy pri rastúcej veľkosti náhodných výberov $n_x$ a $n_y$, keď $X$ a $Y$ majú šesť-rozmerné diskrétne rozdelenie a $\theta = (0.0325, 0.0325;\ 0.035, 0.035;\ 0.1875, 0.1875)$.

# 9 Príloha F: Porovnanie výsledkov simulácií s výsledkami uvedenými v článku [13]

## 9.1 Porovnanie výsledkov simulácií s výsledkami uvedenými v článku [13] pre spojité rozdelenia

V tejto kapitole uvedieme niekoľko tabuliek s odhadmi pravdepodobnosti zamietania testov pri šesť-rozmernom lognormálnom rozdelení, pri nastaveniach popísaných v kapitole 2.10. Uvedené budú výsledky simulácií

1. $T_{cMAX}$ testu s odhadom kovariančnej matice popísaným vo vete 1.7 [13, Theorem 2],
2. $T_{cMAX2}$ testu s odhadom kovariančnej matice popísaným vo vete 2.2 (a aj v kapitole 2.9),
3. $T_{cMAX3}$ testu s odhadom kovariančnej matice popísaným v kapitole 2.9,
4. ktoré autori uvádzajú v článku [13, str. 826].

*Poznámka*: Hodnoty odhadov pravdepodobnosti zamietania, ktoré uvádzajú autori v článku [13, str. 826], sú pravdepodobne počítané pomocou upravenej verzie $T_{cMAX}$ testu popísanej v kapitole 2.9, nie pomocou všeobecnej verzie uvedenej vo vete 1.7 [13, Theorem 2]. Toto dedukujeme na základe toho, že výsledky, ktoré uvádzajú v článku pre sily testov (v tabuľkách 10,11,12), sú podobné s výsledkami, ktoré sme dostali my pri simuláciách $T_{cMAX2}$ testu a značne sa líšia od našich simulácií všeobecného znenia $T_{cMAX}$.

| $n_x$ | $n_y$ | $T_{cMAX}$ | $T_{cMAX2}$ | $T_{cMAX3}$ | Článok |
|---|---|---|---|---|---|
| 50 | 50 | 14.58% (±0.50%) | 5.98% (±0.20%) | 3.23% (±0.12%) | 5.6% |
| 50 | 100 | 13.31% (±0.41%) | 5.44% (±0.18%) | 3.52% (±0.13%) | 5.7% |
| 50 | 200 | 12.42% (±0.41%) | 5.38% (±0.24%) | 3.73% (±0.19%) | 5.4% |
| 100 | 50 | 14.11% (±0.49%) | 6.08% (±0.26%) | 3.91% (±0.21%) | 5.3% |
| 100 | 100 | 11.62% (±0.44%) | 5.54% (±0.17%) | 4.19% (±0.15%) | 4.5% |
| 100 | 200 | 9.83% (±0.38%) | 5.23% (±0.25%) | 4.21% (±0.21%) | 5.2% |
| 200 | 50 | 13.32% (±0.50%) | 6.14% (±0.24%) | 4.40% (±0.17%) | 5.6% |
| 200 | 100 | 10.25% (±0.25%) | 5.52% (±0.19%) | 4.47% (±0.19%) | 5.1% |
| 200 | 200 | 8.27% (±0.45%) | 5.25% (±0.29%) | 4.59% (±0.27%) | 4.7% |

**Tabuľka 9:** Odhad pravdepodobnosti chyby prvého druhu pre jednotlivé testy pri rastúcej veľkosti náhodných výberov $n_x$ a $n_y$, keď $X$ a $Y$ majú lognormálne rozdelenie a $\theta = (0, 0, 0;\ 0, 0, 0)$. Porovnávame s výsledkami z článku [13, str. 826].

| $n_x$ | $n_y$ | $T_{cMAX}$ | $T_{cMAX2}$ | $T_{cMAX3}$ | Článok |
|---|---|---|---|---|---|
| 50 | 50 | 23.77% (±0.52%) | 57.87% (±0.70%) | 45.33% (±0.58%) | 58.7% |
| 50 | 100 | 17.14% (±0.69%) | 83.31% (±0.26%) | 76.23% (±0.44%) | 84.7% |
| 50 | 200 | 3.18% (±0.26%) | 97.61% (±0.17%) | 95.87% (±0.14%) | 97.6% |
| 100 | 50 | 33.71% (±0.40%) | 60.54% (±0.27%) | 51.71% (±0.47%) | 60.7% |
| 100 | 100 | 33.66% (±0.41%) | 87.18% (±0.32%) | 83.53% (±0.30%) | 87.6% |
| 100 | 200 | 11.89% (±0.41%) | 99.11% (±0.11%) | 98.70% (±0.13%) | 99.1% |
| 200 | 50 | 41.15% (±0.72%) | 61.83% (±0.56%) | 54.56% (±0.61%) | 63.0% |
| 200 | 100 | 49.38% (±0.51%) | 89.02% (±0.40%) | 86.64% (±0.46%) | 89.8% |
| 200 | 200 | 29.64% (±0.76%) | 99.49% (±0.08%) | 99.36% (±0.09%) | 99.5% |

**Tabuľka 10:** Odhad sily pre jednotlivé testy pri rastúcej veľkosti náhodných výberov $n_x$ a $n_y$, keď $X$ a $Y$ majú lognormálne rozdelenie a $\theta = (-0.167, -0.167, -0.167;\ 0.096, 0.096, 0.096)$. Porovnávame s výsledkami z článku [13, str. 826].

| $n_x$ | $n_y$ | $T_{cMAX}$ | $T_{cMAX2}$ | $T_{cMAX3}$ | Článok |
|---|---|---|---|---|---|
| 50 | 50 | 32.39% (±0.68%) | 49.50% (±0.58%) | 38.78% (±0.46%) | 50.4% |
| 50 | 100 | 38.10% (±0.55%) | 74.30% (±0.25%) | 67.28% (±0.45%) | 75.3% |
| 50 | 200 | 22.80% (±0.66%) | 92.75% (±0.21%) | 89.80% (±0.26%) | 93.3% |
| 100 | 50 | 41.84% (±0.43%) | 52.13% (±0.39%) | 44.43% (±0.48%) | 54.1% |
| 100 | 100 | 57.81% (±0.81%) | 79.38% (±0.43%) | 75.39% (±0.44%) | 80.4% |
| 100 | 200 | 48.40% (±0.68%) | 96.65% (±0.12%) | 95.76% (±0.16%) | 96.7% |
| 200 | 50 | 48.40% (±0.50%) | 53.50% (±0.45%) | 47.11% (±0.59%) | 54.3% |
| 200 | 100 | 70.23% (±0.34%) | 81.81% (±0.33%) | 79.16% (±0.47%) | 82.3% |
| 200 | 200 | 71.94% (±0.32%) | 98.00% (±0.19%) | 97.63% (±0.21%) | 97.8% |

**Tabuľka 11:** Odhad sily pre jednotlivé testy pri rastúcej veľkosti náhodných výberov $n_x$ a $n_y$, keď $X$ a $Y$ majú lognormálne rozdelenie a $\theta = (-0.167, -0.167, -0.167;\ 0.096, 0.096, -0.096)$. Porovnávame s výsledkami z článku [13, str. 826].

| $n_x$ | $n_y$ | $T_{cMAX}$ | $T_{cMAX2}$ | $T_{cMAX3}$ | Článok |
|---|---|---|---|---|---|
| 50 | 50 | 34.74% (±0.80%) | 36.60% (±0.48%) | 27.32% (±0.38%) | 37.6% |
| 50 | 100 | 51.14% (±0.54%) | 57.92% (±0.43%) | 50.48% (±0.35%) | 57.9% |
| 50 | 200 | 60.02% (±0.57%) | 80.13% (±0.51%) | 75.44% (±0.46%) | 80.5% |
| 100 | 50 | 41.33% (±0.60%) | 39.22% (±0.54%) | 32.32% (±0.42%) | 39.6% |
| 100 | 100 | 64.34% (±0.53%) | 63.90% (±0.46%) | 59.27% (±0.55%) | 64.0% |
| 100 | 200 | 81.02% (±0.31%) | 87.91% (±0.20%) | 86.04% (±0.18%) | 88.2% |
| 200 | 50 | 45.48% (±0.58%) | 40.62% (±0.54%) | 34.85% (±0.57%) | 40.8% |
| 200 | 100 | 70.88% (±0.50%) | 67.08% (±0.45%) | 63.80% (±0.41%) | 68.1% |
| 200 | 200 | 90.69% (±0.47%) | 91.82% (±0.30%) | 90.85% (±0.25%) | 91.2% |

**Tabuľka 12:** Odhad sily pre jednotlivé testy pri rastúcej veľkosti náhodných výberov $n_x$ a $n_y$, keď $X$ a $Y$ majú lognormálne rozdelenie a $\theta = (0.024, 0.024, 0.024;\ 0.143, 0.143, 0.143)$. Porovnávame s výsledkami z článku [13, str. 826].

## 9.2 Porovnanie výsledkov simulácií s výsledkami uvedenými v článku [13] pre diskrétne rozdelenia

V tejto kapitole uvedieme niekoľko tabuliek s odhadmi pravdepodobnosti zamietania testov pri šesť-rozmernom diskrétnom rozdelení, pri nastaveniach popísaných v kapitole 2.10[11]. Uvedené budú výsledky simulácií

1. $T_{cMAX}$ testu s odhadom kovariančnej matice popísaným vo vete 1.7 [13, Theorem 2],
2. $T_{cMAX2}$ testu s odhadom kovariančnej matice popísaným vo vete 2.2 (a aj v kapitole 2.9),
3. $T_{cMAX4}$ testu s odhadom kovariančnej matice popísaným vo vete 3.1,
4. ktoré autori uvádzajú v článku [13, str. 827-828].

*Poznámka*: Je pozoruhodné, že odhady sily testov uvádzané v článku [13, str. 827-828] pre druhé a štvrté nastavenie (kedy $\theta = (0.185, 0.185;\ -0.105, -0.105;\ 0.035, 0.035)$ resp. $\theta = (0.0325, 0.0325;\ 0.035, 0.035;\ 0.1875, 0.1875)$) (v tabuľkách 14 a 16) sú veľmi podobné výsledkom, ktoré sme dostali pri $T_{cMAX4}$ teste a výrazne sa líšia od výsledkov testov $T_{cMAX}$ a $T_{cMAX2}$.

[11] Ako sme spomínali v kapitole 3.9, tak v článku [13, str. 828] sú uvedené iné hodnoty parametrov $\theta$ ako vyšli nám pre druhé nastavenie nastavenie. Simulované sily testov sa v tomto prípade veľmi výrazne líšili od tých, ktoré autori uvádzali v [13, str. 828] a teda porovnanie s článkom v tabuľke 15 je uvedené len pre kompletnosť výsledkov.

| $n_x$ | $n_y$ | $T_{cMAX}$ | $T_{cMAX2}$ | $T_{cMAX4}$ | Článok |
|---|---|---|---|---|---|
| 50 | 50 | 12.00% (±0.30%) | 4.21% (±0.17%) | 6.49% (±0.19%) | 5.8% |
| 50 | 100 | 12.65% (±0.39%) | 4.52% (±0.16%) | 6.53% (±0.18%) | 5.5% |
| 50 | 200 | 12.58% (±0.27%) | 4.65% (±0.14%) | 6.51% (±0.22%) | 5.4% |
| 100 | 50 | 11.35% (±0.35%) | 3.33% (±0.09%) | 5.79% (±0.13%) | 4.9% |
| 100 | 100 | 10.36% (±0.39%) | 3.66% (±0.15%) | 5.69% (±0.17%) | 4.8% |
| 100 | 200 | 9.09% (±0.27%) | 3.93% (±0.16%) | 5.67% (±0.19%) | 4.9% |
| 200 | 50 | 9.14% (±0.28%) | 2.48% (±0.21%) | 5.33% (±0.31%) | 4.5% |
| 200 | 100 | 7.71% (±0.21%) | 3.04% (±0.19%) | 5.40% (±0.24%) | 4.1% |
| 200 | 200 | 6.49% (±0.21%) | 3.41% (±0.21%) | 5.33% (±0.20%) | 4.0% |

**Tabuľka 13:** Odhad pravdepodobnosti chyby prvého druhu pre jednotlivé testy pri rastúcej veľkosti náhodných výberov $n_x$ a $n_y$, keď $X$ a $Y$ majú šesť-rozmerné diskrétne rozdelenie a $\theta = (0,0;\ 0,0;\ 0,0)$. Porovnávame s výsledkami z článku [13, str. 827].

| $n_x$ | $n_y$ | $T_{cMAX}$ | $T_{cMAX2}$ | $T_{cMAX4}$ | Článok |
|---|---|---|---|---|---|
| 50 | 50 | 11.88% (±0.38%) | 18.40% (±0.40%) | 63.30% (±0.45%) | 64% |
| 50 | 100 | 19.22% (±0.53%) | 31.14% (±0.37%) | 79.55% (±0.50%) | 79.7% |
| 50 | 200 | 25.32% (±0.61%) | 44.74% (±0.61%) | 88.84% (±0.38%) | 89.5% |
| 100 | 50 | 17.96% (±0.23%) | 26.53% (±0.31%) | 72.95% (±0.27%) | 73.4% |
| 100 | 100 | 34.22% (±0.44%) | 50.63% (±0.43%) | 91.26% (±0.24%) | 91.9% |
| 100 | 200 | 45.74% (±0.61%) | 75.65% (±0.31%) | 98.09% (±0.18%) | 98.2% |
| 200 | 50 | 23.32% (±0.46%) | 33.18% (±0.56%) | 79.31% (±0.44%) | 79.5% |
| 200 | 100 | 46.79% (±0.37%) | 67.02% (±0.51%) | 96.40% (±0.18%) | 96.0% |
| 200 | 200 | 61.06% (±0.70%) | 93.06% (±0.30%) | 99.79% (±0.03%) | 99.8% |

**Tabuľka 14:** Odhad sily pre jednotlivé testy pri rastúcej veľkosti náhodných výberov $n_x$ a $n_y$, keď $X$ a $Y$ majú šesť-rozmerné diskrétne rozdelenie a $\theta = (0.185, 0.185;\ -0.105, -0.105;\ 0.035, 0.035)$. Porovnávame s výsledkami z článku [13, str. 828].

| $n_x$ | $n_y$ | $T_{cMAX}$ | $T_{cMAX2}$ | $T_{cMAX4}$ | Článok |
|---|---|---|---|---|---|
| 50 | 50 | 7.37% (±0.15%) | 1.44% (±0.12%) | 14.56% (±0.36%) | 63.9% |
| 50 | 100 | 8.26% (±0.24%) | 1.68% (±0.13%) | 18.59% (±0.33%) | 79.4% |
| 50 | 200 | 8.92% (±0.17%) | 2.01% (±0.22%) | 22.55% (±0.33%) | 88.6% |
| 100 | 50 | 8.33% (±0.24%) | 1.98% (±0.12%) | 16.53% (±0.43%) | 73.5% |
| 100 | 100 | 9.36% (±0.30%) | 2.76% (±0.16%) | 22.93% (±0.59%) | 91.2% |
| 100 | 200 | 10.91% (±0.28%) | 4.25% (±0.18%) | 31.44% (±0.56%) | 97.9% |
| 200 | 50 | 8.61% (±0.32%) | 2.41% (±0.09%) | 18.16% (±0.33%) | 79.0% |
| 200 | 100 | 10.57% (±0.27%) | 4.29% (±0.23%) | 27.88% (±0.44%) | 96.4% |
| 200 | 200 | 14.82% (±0.24%) | 8.18% (±0.24%) | 41.87% (±0.38%) | 99.7% |

**Tabuľka 15:** Odhad sily pre jednotlivé testy pri rastúcej veľkosti náhodných výberov $n_x$ a $n_y$, keď $X$ a $Y$ majú šesť-rozmerné diskrétne rozdelenie a $\theta = (0.075, 0.075;\ 0.045, 0.045;\ 0.025, -0.025)$. Porovnávame s výsledkami z článku [13, str. 828].

| $n_x$ | $n_y$ | $T_{cMAX}$ | $T_{cMAX2}$ | $T_{cMAX4}$ | Článok |
|---|---|---|---|---|---|
| 50 | 50 | 27.72% (±0.68%) | 12.12% (±0.11%) | 59.87% (±0.69%) | 60.3% |
| 50 | 100 | 42.13% (±0.88%) | 22.86% (±0.33%) | 75.22% (±0.52%) | 75.3% |
| 50 | 200 | 53.33% (±0.70%) | 35.27% (±0.41%) | 84.30% (±0.41%) | 84.8% |
| 100 | 50 | 37.67% (±0.31%) | 18.78% (±0.32%) | 71.20% (±0.34%) | 72.7% |
| 100 | 100 | 61.48% (±0.43%) | 41.15% (±0.36%) | 89.55% (±0.19%) | 89.6% |
| 100 | 200 | 79.44% (±0.29%) | 66.99% (±0.39%) | 96.90% (±0.08%) | 96.8% |
| 200 | 50 | 43.91% (±0.51%) | 24.42% (±0.44%) | 78.16% (±0.55%) | 78.9% |
| 200 | 100 | 73.98% (±0.45%) | 57.79% (±0.48%) | 95.76% (±0.25%) | 95.8% |
| 200 | 200 | 93.70% (±0.17%) | 88.76% (±0.32%) | 99.67% (±0.06%) | 99.7% |

**Tabuľka 16:** Odhad sily pre jednotlivé testy pri rastúcej veľkosti náhodných výberov $n_x$ a $n_y$, keď $X$ a $Y$ majú šesť-rozmerné diskrétne rozdelenie a $\theta = (0.0325, 0.0325;\ 0.035, 0.035;\ 0.1875, 0.1875)$. Porovnávame s výsledkami z článku [13, str. 828].

# Záver

V diplomovej práci sme sa venovali istému už známemu neparametrickému testu polohy zohľadňujúcemu clusterovanie, ktorý je označovaný ako $T_{cMAX}$ test. V prvej kapitole sme začali úvodom do problematiky viacrozmerných neparametrických testov polohy. Uviedli sme pojem relatívneho efektu $\theta$, zovšeobecnenú viacrozmernú Behrensovu-Fisherovu hypotézu a priblížili sme históriu jej testovania pomocou pozložkových rankov. Venovali sme sa teórii viacrozmerných zovšeobecnených $U$-štatistík, ktoré sú kľúčové pri dôkaze vety 1.6 [13, Theorem 1]. V ďalšej časti sme sa zaoberali už samotnou vetou 1.6 [13, Theorem 1], kde sme uviedli jej znenie spolu s podrobnejším dôkazom. Na konci prvej kapitoly sme uviedli vetu 1.7 [13, Theorem 2] spolu s ideou jej dôkazu.

V druhej kapitole sme poukázali na chyby v znení aj v dôkaze vety 1.6 [13, Theorem 1], čo má negatívny dopad na praktické fungovanie testu $T_{cMAX}$. Uviedli sme aj kontrapríklad, pri ktorom podľa tvrdenia vety vychádza kovariančná matica negatívne definitná. V ďalšej časti kapitoly sme predstavili opravy chýb v dôkaze vety [13, Theorem 1] a následne sme podľa našich viet 2.1 a 2.2 navrhli dve možnosti, ako by opravený $T_{cMAX}$ test mohol vyzerať. Nakoniec sme sa v simulačných štúdiách pozreli na pravdepodobnosti chyby prvého druhu a sily nami navrhnutých testov. Porovnávali sme test $T_{cMAX2}$, kde sme pri výpočte kovariančnej matice parameter $\theta$ odhadovali, a test $T_{cMAX3}$, pri ktorom sme parameter $\theta$ dosadzovali podľa testovanej nulovej hypotézy. Pravdepodobnosť chyby prvého druhu s narastajúcou veľkosťou náhodných výberov konvergovala k želanej hladine 0.05 pre oba nami navrhnuté testy. Pri skúmaní sily testov sa $T_{cMAX2}$ ukázal ako silnejší pri všetkých nami testovaných nastaveniach a veľkostiach náhodných výberov. Pri finalizácii diplomovej práce sme našli preprint [12], kde sú uvedené tvrdenia[12, Theorem 1 a Theorem 2], ktorých znenie je veľmi podobné zneniu našich viet 2.1 a 2.2.

V tretej kapitole sme sa zaoberali všeobecným prípadom, zahŕňajúci prípad diskrétnych rozdelení s prekrývajúcimi sa supportami. Na kontrapríklade sme ukázali, že v takomto prípade veta 1.6 [13, Theorem 1] neplatí. Naša veta 3.1, ktorú sme dokázali v kapitole 3.3, platí vo všeobecnosti, bez dodatočných predpokladov na rozdelenia náhodných vektorov. Ďalej sme ukázali, že odhady z vety 1.7 [13, Theorem 2] nemusia byť konzistentné. Problémy vo vetách 1.6 a 1.7 [13, Theorem 1 a Theorem 2] znamenajú opäť problémy s fungovaním $T_{cMAX}$ testu v prípade diskrétnych dát s výskytom zhôd. Vo

vete 3.2 sme predstavili nové odhady a dokázali ich konzistentnosť v kapitole 3.6. Výsledkom toho bola nami navrhnutá nová verzia $T_{cMAX4}$ pôvodného testu $T_{cMAX}$. Nakoniec sme sa venovali simulačným štúdiám a skúmali sme pravdepodobnosť chyby prvého druhu a silu testov na diskrétnych dátach. Porovnávali sme originálne znenie $T_{cMAX}$ testu z článku [13], testu navrhnutého v druhej kapitole $T_{cMAX2}$ a nami navrhnutého zovšeobecneného testu z tretej kapitoly $T_{cMAX4}$. V našich simuláciách sme ukázali, že pravdepodobnosť chyby prvého druhu pre $T_{cMAX4}$ test konverguje k želanej hladine 0.05. Toto sa už nedalo povedať o testoch $T_{cMAX}$ a $T_{cMAX2}$, kde sa pravdepodobnosť chyby prvého druhu blížila k hladine 0.03. Simuláciami sily testu sme ukázali, že $T_{cMAX4}$ test je výrazne silnejší oproti testom $T_{cMAX}$ a $T_{cMAX2}$ vo všetkých nami testovaných nastaveniach a veľkostiach náhodných výberov. Opravy, ktoré sme uviedli v tretej kapitole a teda aj z nich plynúci test $T_{cMAX4}$, sme v čase finalizácie diplomovej práce nenašli publikované. Problémy, ktoré sme v tejto kapitole popísali, sú prítomné nielen v $T_{cMAX}$ teste z článku [13] ale aj v teste popísanom v preprinte [12], ktorý chyby z druhej kapitoly neobsahuje.

# Zoznam použitej literatúry


[1] Gut, A. *Probability: a graduate course*, vol. 200. Springer, 2006.

[2] Hoeffding, W. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics 19* (1948), 193–325.

[3] Huang, P., Tilley, B. C., Woolson, R. F., and Lipsitz, S. Adjusting O'Brien's test to control type I error for the generalized nonparametric Behrens–Fisher problem. *Biometrics 61*, 2 (2005), 532–539.

[4] Kallenberg, O. *Foundations of modern probability, Third edition*, vol. 1. Springer, 2021.

[5] Liu, A., Li, Q., Liu, C., Yu, K., and Yu, K. F. A rank-based test for comparison of multidimensional outcomes. *Journal of the American Statistical Association 105*, 490 (2010), 578–587.

[6] O'Brien, P. C. Procedures for comparing samples with multiple endpoints. *Biometrics 40* (1984), 1079–1087.

[7] Serfling, R. J. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 2009.

[8] Sugiura, N. Multisample and multivariate nonparametric tests based on U statistics and their asymptotic efficiencies. *Osaka Journal of Mathematics 2*, 2 (1965), 385–426.

[9] Thompson, E. A. Asymptotic unbiasedness and consistency, poznámky k predmetu Stat341. dostupné na internete (08.05.2024): `https://sites.stat.washington.edu/thompson/S341_10/Notes/week3.pdf`, University of Washington, Seattle, 2010.

[10] Van der Vaart, A. W. *Asymptotic statistics*, vol. 3. Cambridge university press, 2000.

[11] Williams, D. *Probability with martingales*. Cambridge university press, 1991.

[12] Xu, X., and Luo, S. dostupné na internete (08.05.2024): `https://www.medrxiv.org/content/early/2023/06/29/2023.06.24.23291858`, Duke University, Durham, 2023.

[13] Zhang, W., Liu, A., Tang, L. L., and Li, Q. A cluster-adjusted rank-based test for a clinical trial concerning multiple endpoints with application to dietary intervention assessment. *Biometrics 75*, 3 (2019), 821–830.

[14] Zhang, W., Liu, A., Tang, L. L., and Li, Q. Supporting information for "A cluster-adjusted rank-based test for a clinical trial concerning multiple endpoints with application to dietary intervention assessment". *Biometrics 75*, 3 (2019), 1–23.