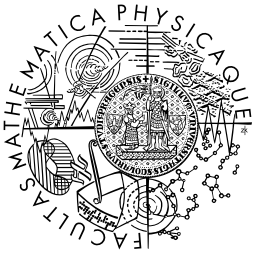


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁRSKA PRÁCA

Filip Kulla

Dvojvýberový Wilcoxonov test v prípade existencie zhôd

Katedra pravděpodobnosti a matematické statistiky, MFF UK

Vedúci bakalárskej práce: doc. Ing. Omelka Marek, Ph.D.

Študijný program: Matematika

Študijný obor: Obecná matematika

Praha 2020

Prehlasujem, že som túto bakalársku prácu vypracoval samostatne a výhradne s použitím citovaných prameňov, literatúry a ďalších odborných zdrojov. Táto práca nebola využitá k získaniu iného alebo rovnakého titulu.

Beriem na vedomie, že sa na moju prácu vzťahujú práva a povinnosti vyplývajúce zo zákona č. 121/2000 Sb., autorského zákona v platnom znení, najmä skutočnosť, že Univerzita Karlova má právo na uzavrenie licenčnej zmluvy o použití tejto práce ako školského diela podľa §60 odst. 1 autorského zákona.

V ..... dňa ....

Podpis autora

Ďakujem môjmu školiteľovi, doc. Ing. Marekovi Omelkovi, Ph.D., za vedenie, užitočné rady a prejavenú ochotu. Rovnako ďakujem aj mojej rodine a priateľom.

Názov práce: Dvojvýberový Wilcoxonov test v prípade existencie zhôd

Autor: Filip Kulla

Katedra: Katedra pravděpodobnosti a matematické statistiky, MFF UK

Vedúci bakalárskej práce: doc. Ing. Omelka Marek, Ph.D., Katedra pravděpodobnosti a matematické statistiky, MFF UK

Abstrakt: Táto práca sa venuje dvojvýberovému Wilcoxonovmu testu v prípade existencie zhôd. V práci je s využitím známych výsledkov o U-štatistikách odvodené asymptotické rozdelenie Wilcoxonovej testovej štatistiky v prípade existencie zhôd. S pomocou tohto výsledku je navrhnutá korekcia testu pre dáta obsahujúce zhody. Ďalej sa v práci skúma súvislosť odvodenej korekcie a podmieneného rozptylu testovej štatistiky (a podmienenej strednej hodnoty) za H_0. Nakoniec je v práci pomocou simulácií preskúmané, aký vplyv má odvodená korekcia na skutočnú hladinu testu pri zvyšujúcom sa počte zhodných pozorovaní.

Kľúčové slová: dvojvýberový Wilcoxonov test, Mannov-Whitneyov U test, zhody.

Title: Two-sample Wilcoxon test in the presence of ties

Author: Filip Kulla

Department: Department of Probability and Mathematical Statistics, MFF UK

Supervisor: doc. Ing. Omelka Marek, Ph.D., Department of Probability and Mathematical Statistics, MFF UK

Abstract: This work is devoted to the Wilcoxon rank-sum test in the presence of ties. In this work, asymptotic distribution of the Wilcoxon test statistic in the presence of ties is derived using well-known results about U-statistics. With the aid of this result, a corrected test statistic for data containing ties is proposed. Furthermore, the thesis examines the relation between the derived correction and conditional variance of the test statistic (and conditional expectation). Finally, by means of simulation, the effect of the derived correction on the actual significance level is examined as the number of tied observations increases.

Keywords: Wilcoxon rank-sum test, Mann–Whitney U test, ties.

Obsah

Úvod

Táto práca sa venuje dvojvýberovému Wilcoxonovmu testu. Ide o neparametrický test založený na poradiach, pomocou ktorého porovnávame 2 nezávislé náhodné výbery (napr. pacienti, ktorým bolo podané liečivo a pacienti, ktorí dostali placebo). Štandardná teória poradových testov predpokladá, že dáta pochádzajú zo spojitých rozdelení, čo zaručuje, že pravdepodobnosť zhodných pozorovaní je nulová. V praxi sa ale častokrát stáva, že predpoklad spojitosti nie je splnený a v dátach sa zhody nachádzajú (napr. nevieme rozhodnúť či je stav prvého pacienta lepší ako stav druhého pacienta alebo telesná teplota prvého pacienta je po zaokrúhlení na 1 desatinné miesto rovnaká ako zaokrúhlená teplota druhého pacienta).

V tejto práci budeme skúmať vplyv prítomnosti zhodných pozorovaní na dvojvýberový Wilcoxonov test. Naším cieľom bude odvodiť asymptotické rozdelenie Wilcoxonovej testovej štatistiky v prípade existencie zhôd, na základe dosiahnutého výsledku navrhnúť korekciu testu pre prípad, že naše dáta obsahujú zhody a nakoniec pomocou simulácií porovnať skutočnú hladinu testu s korekciou a testu bez korekcie v prítomnosti zhodných pozorovaní.

Hlavná časť práce je rozdelená do štyroch kapitol. Prvá kapitola je venovaná predstaveniu dvojvýberového Wilcoxonovho testu, pričom zatiaľ predpokladáme spojitosť rozdelení, z ktorých dáta pochádzajú. V druhej kapitole upúšťame od predpokladu spojitosti. Jej podstatnú časť tvorí odvodenie asymptotického rozdelenia testovej štatistiky v prípade existencie zhôd, ktoré nás vedie k návrhu korekcie testu pre dáta so zhodami. Tretia kapitola ukazuje súvislosť medzi odvodenou korekciou a podmieneným rozptylom testovej štatistiky (a podmienenou strednou hodnotou) za H_0. Štvrtá kapitola obsahuje simulačnú štúdiu, v ktorej zisťujeme skutočnú hladinu testu s korekciou a testu bez korekcie.

1. Dvojvýberový Wilcoxonov test bez zhôd

Dvojvýberový Wilcoxonov test je neparametrický test založený na poradiach. Uvažujme 2 nezávislé náhodné výbery $\boldsymbol{X} = (X_1, \ldots, X_n)$ a $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$. Nech $\boldsymbol{X}$ je náhodný výber z rozdelenia s distribučnou funkciou F_X a $\boldsymbol{Y}$ je náhodný výber z rozdelenia s distribučnou funkciou F_Y. Predpokladajme, že F_X a F_Y sú spojité. Tento predpoklad implikuje, že pravdepodobnosť zhodných pozorovaní je 0. Formálnejšie, ak zavedieme značenie $N = n + m$ a $\boldsymbol{Z} = (Z_1, \ldots, Z_N) = (X_1, \ldots, X_n, Y_1, \ldots, Y_m)$, tak

$$\mathsf{P}(Z_i = Z_j, \text{ pre nejaké } i, j \in \{1, \ldots, N\}) = 0. \tag{1.1}$$

Predpokladajme, že navyše

$$\exists \delta \in \mathbb{R}\ \forall x \in \mathbb{R} : F_X(x) = F_Y(x - \delta).$$

Takýto model sa niekedy nazýva model posunutia v polohe.

Zaujíma nás posunutie δ. Presnejšie, chceme testovať hypotézu

$$H_0 : \delta = 0, \tag{1.2}$$

proti alternatíve

$$H_1 : \delta \neq 0.$$

V modeli posunutia v polohe teda dvojvýberový Wilcoxonov test testuje, či náhodné výbery $\boldsymbol{X}$ a $\boldsymbol{Y}$ pochádzajú z rovnakého rozdelenia.

1.1 Zavedenie testovej štatistiky

Zoraďme všetky náhodné veličiny v združenom výbere $\boldsymbol{Z}$ od najmenšej po najväčšiu: $Z_{(1)} < Z_{(2)} < \ldots < Z_{(N)}$. Symbol $Z_{(k)}$ označuje k-tu najmenšiu hodnotu medzi pozorovaniami zo $\boldsymbol{Z}$. Vďaka predpokladu (1.1), vyplývajúcemu zo spojitosti F_X a F_Y, je usporiadanie pozorovaní zo $\boldsymbol{Z}$ dobre definované a nemusíme uvažovať neostré nerovnosti.

Definícia 1. *Vzostupne usporiadaný výber $\boldsymbol{Z}$, $Z_{(1)} < Z_{(2)} < \ldots < Z_{(N)}$, budeme nazývať usporiadaný výber a označovať $\boldsymbol{Z}_{(.)}$*

Definícia 2. *Poradím náhodnej veličiny Z_i, $i \in \{1, \ldots, N\}$, v združenom výbere $\boldsymbol{Z}$ rozumieme prirodzené číslo $R_i \in \{1, \ldots, N\}$ také, že $Z_i = Z_{(R_i)}$.*

Platí, že

$$R_i = \sum_{j=1}^{N} 1\{Z_j \leq Z_i\}. \tag{1.3}$$

Špeciálne, pre $i \in \{1, \ldots, n\}$ je $R_i = \sum_{j=1}^{n} 1\{X_j \leq X_i\} + \sum_{j=1}^{m} 1\{Y_j \leq X_i\}$.

Wilcoxonova testová štatistika je definovaná ako suma poradí náhodných veličín z $\boldsymbol{X}$ v združenom výbere $\boldsymbol{Z}$, teda

$$W_X = \sum_{i=1}^{n} R_i.$$

Testová štatistika W_X môže nadobúdať hodnoty $n(n+1)/2, \ldots, nm+n(n+1)/2$. Proti H_0 zrejme svedčia príliš malé a zároveň príliš veľké hodnoty W_X.

Pokiaľ nie sú n a m príliš veľké, je možné určiť presné rozdelenie W_X za H_0. Toto presné rozdelenie sa dá odvodiť z toho, že za H_0 je ľubovoľné usporiadanie $\boldsymbol{Z}$ rovnako pravdepodobné, a teda

$$\mathsf{P}(R_1 = r_1, \ldots, R_n = r_n) = \frac{m!}{N!},$$

pre všetky $r_1, \ldots, r_n \in \{1, \ldots, N\}$ rôzne.

V prípade, že n a m sú veľké, využijeme na zvolenie kritického oboru tvrdenie 1 a vetu 3 zo sekcie 1.2. Aby sme dostali test s asymptotickou hladinou α, zvolíme kritický obor tak, že

$$H_0 \text{ zamietame} \Leftrightarrow \frac{\left|W_X - \frac{n(N+1)}{2}\right|}{\sqrt{\frac{nm(N+1)}{12}}} \geq u_{1-\alpha/2},$$

kde $u_{1-\alpha/2}$ značí $(1-\alpha/2)$-kvantil rozdelenia $N(0,1)$.

Tvrdenie 1. *Za platnosti H_0 je*

$$\mathsf{E}\, W_X = \frac{n(n+m+1)}{2} \quad a \quad var W_X = \frac{nm(n+m+1)}{12}.$$

Dôkaz. Postup ako určiť strednú hodnotu a rozptyl sa dá nájsť v knihe Anděl (2007b, strana 102).

□

Na odvodenie asymptotického rozdelenia W_X použijeme teóriu U-štatistík, viď sekcia 5.1 v kapitole Dodatky.

1.2 Odvodenie asymptotického rozdelenia testovej štatistiky v prípade bez zhôd

V tejto sekcii použijeme vetu 12 zo sekcie 5.1 na odvodenie asymptotického rozdelenia W_X. Definujme

$$W_{XY} = \sum_{i=1}^{n} \sum_{j=1}^{m} 1\{X_i < Y_j\}.$$

Štatistika W_{XY} sa často označuje ako Mannova-Whitneyova. S Wilcoxonovou štatistikou ju spája deterministický vzťah.

Tvrdenie 2. *Platí, že $W_{XY} = nm + \frac{n(n+1)}{2} - W_X$.*

Dôkaz. Toto tvrdenie sa dokáže analogicky ako tvrdenie 4 v sekcii 2.2.

□

Veta 3. *Ak platí H_0 a $n/N \to p \in (0,1)$, keď $n+m \to \infty$, tak*

$$\frac{W_X - \mathsf{E}_{H_0} W_X}{\sqrt{var_{H_0} W_X}} \xrightarrow{D} N(0,1).$$

Dôkaz. Dôkaz bude aplikáciou vety 12. Definujme jadro $h : \mathbb{R}^2 \to \mathbb{R}$ predpisom

$$h(x,y) = \begin{cases} 1, & \text{ak} \quad x < y, \\ 0, & \text{inak.} \end{cases}$$

Potom korešpondujúca U-štatistika pre jadro h je tvaru

$$U_{n,m} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} h(X_i,Y_j) = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} 1\{X_i < Y_j\}.$$

Zrejme platí, že

$$U_{n,m} = \frac{W_{XY}}{nm}. \tag{1.4}$$

Teraz spočítame $\theta, \sigma_{10}^2, \sigma_{01}^2$ z vety 12 a overíme konečnosť σ_{11}^2.

Platí, že

$$\theta = \mathsf{E}\,[h(X_i,Y_j)] = \mathsf{E}\; 1\{X_i < Y_j\} = \mathsf{P}(X_i < Y_j) \overset{H_0}{=} \frac{1}{2},$$

pričom v kroku označenom $\overset{H_0}{=}$ používame predpoklad, že platí H_0.

Pre σ_{10}^2 a σ_{01}^2 dostávame nasledujúce vzťahy:

$$\begin{aligned} \sigma_{10}^2 &= cov\,(1\{X_1 < Y_1\}, 1\{X_1 < Y_2\}) = \mathsf{E}\left[1\{X_1 < Y_1\} 1\{X_1 < Y_2\}\right] - \theta^2 \\ &= \mathsf{P}(X_1 < Y_1, X_1 < Y_2) - \theta^2 \overset{H_0}{=} \frac{1}{3} - \frac{1}{4} = \frac{1}{12}, \end{aligned}$$

$$\begin{aligned} \sigma_{01}^2 &= cov\,(1\{X_1 < Y_1\}, 1\{X_2 < Y_1\}) = \mathsf{E}\left[1\{X_1 < Y_1\} 1\{X_2 < Y_1\}\right] - \theta^2 \\ &= \mathsf{P}(X_1 < Y_1, X_2 < Y_1) - \theta^2 \overset{H_0}{=} \frac{1}{3} - \frac{1}{4} = \frac{1}{12}, \end{aligned}$$

pričom predposledná rovnosť v každom vzťahu plynie z toho, že máme 6 možností ako usporiadať X_1, Y_1 a Y_2 (respektíve X_1, X_2 a Y_1) a každá z nich je rovnako pravdepodobná.

Konečnosť σ_{11}^2 je zrejmá, keďže ide o rozptyl $1\{X_1 < Y_1\}$.

Z vety 12 dostávame, že ak $n/N \to p \in (0,1)$, keď $n+m \to \infty$, tak

$$\sqrt{N}\left(U_{n,m} - \frac{1}{2}\right) \xrightarrow{D} N\left(0, \frac{1}{12}\left(\frac{1}{p} + \frac{1}{1-p}\right)\right).$$

Veta o spojitej transformácii dá, že

$$\frac{\sqrt{N}\left(U_{n,m}-\frac{1}{2}\right)}{\sqrt{\frac{1}{12}\left(\frac{1}{p}+\frac{1}{1-p}\right)}} \xrightarrow{D} N(0,1).$$

Vďaka Cramérovej-Sluckého vete môžeme výraz $\frac{1}{p}+\frac{1}{1-p}$ v menovateli nahradiť výrazom $\frac{N^2}{nm}$, čím je konvergencia v distribúcii zachovaná, a teda

$$\frac{\sqrt{N}\left(U_{n,m}-\frac{1}{2}\right)}{\sqrt{\frac{N^2}{12nm}}} \xrightarrow{D} N(0,1). \tag{1.5}$$

Výraz na ľavej strane (1.5) teraz prevedieme na tvar, v ktorom vystupuje W_X. Postupnosťou úprav s využitím tvrdenia 2 a vzťahu (1.4) dostávame, že

$$\begin{aligned}\frac{\sqrt{N}\left(U_{n,m}-\frac{1}{2}\right)}{\sqrt{\frac{N^2}{12nm}}} &= \frac{\sqrt{N}\left(\frac{W_{XY}}{nm}-\frac{1}{2}\right)}{\sqrt{\frac{N^2}{12nm}}} = \frac{\sqrt{N}\left(\frac{nm+n(n+1)-2W_X}{2nm}\right)}{\sqrt{\frac{N^2}{12nm}}} \\ &= -\frac{1}{\sqrt{N}}\frac{\left(\frac{W_X-\frac{n(n+m+1)}{2}}{nm}\right)}{\sqrt{\frac{1}{12nm}}} = -\frac{W_X-\frac{n(N+1)}{2}}{\sqrt{\frac{nmN}{12}}}.\end{aligned}$$

Z tvrdenia 1 a Cramérovej-Sluckého vety konečne dostávame, že ak $n/N \to p \in (0,1)$, keď $n+m \to \infty$, tak

$$\begin{aligned}\frac{W_X-\mathsf{E}_{H_0} W_X}{\sqrt{\operatorname{var}_{H_0} W_X}} &= \frac{W_X-\mathsf{E}_{H_0} W_X}{\sqrt{\frac{nmN}{12}}}\frac{\sqrt{\frac{nmN}{12}}}{\sqrt{\operatorname{var}_{H_0} W_X}} \\ &= \frac{W_X-\mathsf{E}_{H_0} W_X}{\sqrt{\frac{nmN}{12}}}\frac{\sqrt{\frac{nmN}{12}}}{\sqrt{\frac{nm(N+1)}{12}}} \xrightarrow{D} N(0,1).\end{aligned}$$

□

2. Dvojvýberový Wilcoxonov test so zhodami

V tejto kapitole opäť uvažujeme 2 nezávislé náhodné výbery $\boldsymbol{X} = (X_1, \ldots, X_n)$ a $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$ z rozdelení s distribučnými funkciami F_X a F_Y, avšak už nepredpokladáme, že F_X a F_Y sú spojité. Už teda nemôžeme predpokladať, že pravdepodobnosť zhodných pozorovaní je 0, ako sme predpokladali v (1.1).

2.1 Zavedenie testovej štatistiky v prípade zhôd

V prípade existencie zhôd prestáva dávať definícia poradí z prvej kapitoly zmysel, pretože už nie je možné zoradiť náhodné veličiny v združenom výbere $\boldsymbol{Z}$ od najmenšej po najväčšiu tak, aby boli splnené ostré nerovnosti $Z_{(1)} < Z_{(2)} < \ldots < Z_{(N)}$. Z tohto dôvodu prejdeme k definícii takzvaných priemerných poradí. Napriek nemožnosti usporiadať združený výber $\boldsymbol{Z}$ tak, aby platili ostré nerovnosti, ho stále môžeme usporiadať tak, aby platili neostré nerovnosti $Z_{(1)} \leq Z_{(2)} \leq \ldots \leq Z_{(N)}$. Uvažujme blok k zhodných pozorovaní: $\ldots < Z_{(i)} = Z_{(i+1)} = \ldots = Z_{(i+k-1)} < \ldots$. Z množiny $1, \ldots, N$ by mali tomuto bloku zrejme prislúchať čísla $i, i+1, \ldots, i+k-1$. Definície 2 však každej náhodnej veličine z tohto bloku priraďuje všetky tieto čísla, zatiaľ čo vzťah (1.3) každej náhodnej veličine z tohto bloku priraďuje číslo $i+k-1$. Priemerné poradia definujeme tak, aby priemerné poradie každej náhodnej veličiny z tohto bloku bolo aritmetickým priemerom čísel $i, i+1, \ldots, i+k-1$, teda $(2i+k-1)/2$.

Definícia 3. *Priemerné poradie náhodnej veličiny Z_i, $i \in \{1, \ldots, N\}$, v združenom výbere $\boldsymbol{Z}$ definujeme ako*

$$R_i^* = 1 + \sum_{j=1}^{N} 1\{Z_j < Z_i\} + \frac{1}{2} \sum_{\substack{j=1 \\ j \neq i}}^{N} 1\{Z_j = Z_i\}.$$

Poznámka. Usporiadaným výberom $\boldsymbol{Z}_{(.)}$ budeme stále rozumieť vzostupne usporiadaný výber $\boldsymbol{Z}$, len s neostrými nerovnosťami: $Z_{(1)} \leq Z_{(2)} \leq \ldots \leq Z_{(N)}$.

Wilcoxonova testová štatistika je v prípade existencie zhôd definovaná ako suma priemerných poradí náhodných veličín z $\boldsymbol{X}$ v združenom výbere $\boldsymbol{Z}$, teda

$$W_X^* = \sum_{i=1}^{n} R_i^*.$$

Rovnako ako v prípade bez zhôd svedčia proti H_0 (1.2) príliš malé a zároveň príliš veľké hodnoty testovej štatistiky W_X^*.

Pokiaľ nie sú n a m príliš veľké, je možné určiť presné rozdelenie W_X^* za H_0 pri danom $\boldsymbol{Z}_{(.)}$. Toto presné rozdelenie sa dá odvodiť z toho, že za H_0 je ľubovoľné usporiadanie $\boldsymbol{Z}$ rovnako pravdepodobné. Nasledujúci príklad ilustruje, ako by sme postupovali, pre n a m dostatočne malé.

Príklad. Predpokladajme, že platí H_0, že $n = m = 2$, a že $\boldsymbol{Z}_{(.)} = (2, 7, 7, 10)$. Vektor priemerných poradí je potom $(1, \frac{5}{2}, \frac{5}{2}, 4)$. Za tejto informácie je 12 možností

ako usporiadať (X_1, X_2, Y_1, Y_2), pretože máme $\binom{4}{2}$ možností ako vybrať 2 náhodné veličiny, ktoré sa budú rovnať a v každej z týchto možností máme 2 možnosti, ktorá zo zvyšných náhodných veličín bude najmenšia. Potom

$$\begin{aligned}
&\mathsf{P}(R_1^* = 1, R_2^* = \frac{5}{2}) = \frac{2}{12}, \quad \mathsf{P}(R_1^* = \frac{5}{2}, R_2^* = 1) = \frac{2}{12}, \quad \mathsf{P}(R_1^* = 4, R_2^* = 1) = \frac{1}{12},\\
&\mathsf{P}(R_1^* = 1, R_2^* = 4) = \frac{1}{12}, \quad \mathsf{P}(R_1^* = \frac{5}{2}, R_2^* = \frac{5}{2}) = \frac{2}{12}, \quad \mathsf{P}(R_1^* = 4, R_2^* = \frac{5}{2}) = \frac{2}{12},\\
&\qquad\qquad\qquad\qquad \mathsf{P}(R_1^* = \frac{5}{2}, R_2^* = 4) = \frac{2}{12}.
\end{aligned}$$

Vidíme, že W_X^* môže nadobúdať hodnoty $\frac{7}{2}, 5$ a $\frac{13}{2}$. Každú z nich s pravdepodobnosťou $\frac{1}{3}$.

V prípade, že n a m sú veľké, využijeme na zvolenie kritického oboru vetu 8 zo sekcie 2.2. Aby sme dostali test s asymptotickou hladinou α, zvolíme kritický obor tak, že

$$H_0 \text{ zamietame} \Leftrightarrow \frac{\left|W_X^* - \frac{n(N+1)}{2}\right|}{\sqrt{\frac{nm(N+1)}{12} - \frac{nm\kappa}{(N-1)N}}} \geq u_{1-\alpha/2},$$

kde

$$\kappa = \sum_z \frac{k_z^3 - k_z}{12},$$

pričom k_z značí počet, koľkokrát sa medzi hodnotami $Z_1, \ldots, Z_N$ vyskytla hodnota z. Suma $\sum_z$ značí sčítanie cez všetky rôzne hodnoty $Z_1, \ldots, Z_N$.

Pre odvodenie asymptotického rozdelenia testovej štatistiky W_X^* budeme postupovať analogicky ako v prvej kapitole.

2.2 Odvodenie asymptotického rozdelenia testovej štatistiky v prípade zhôd

V tejto sekcii opäť použijeme vetu 12 zo sekcie 5.1 na odvodenie asymptotického rozdelenia W_X^*. Definujme

$$W_{XY}^* = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathit{1}\{X_i < Y_j\} + \frac{1}{2}\mathit{1}\{X_i = Y_j\}.$$

Podobne ako v prípade bez zhôd spája Mannovu-Whitneyovu štatistiku W_{XY}^* s Wilcoxonovou štatistikou deterministický vzťah.

Tvrdenie 4. *Platí, že* $W_{XY}^* = nm + \frac{n(n+1)}{2} - W_X^*$.

Dôkaz. Podľa definície 3, pre $i \in \{1, \ldots, n\}$ máme, že

$$\begin{aligned}
R_i^* = 1 &+ \sum_{j=1}^{n} \mathit{1}\{X_j < X_i\} + \sum_{j=1}^{m} \mathit{1}\{Y_j < X_i\}\\
&+ \frac{1}{2}\sum_{\substack{j=1\\ j\neq i}}^{n} \mathit{1}\{X_j = X_i\} + \frac{1}{2}\sum_{j=1}^{m} \mathit{1}\{Y_j = X_i\}.
\end{aligned} \tag{2.1}$$

Označme S_i^* priemerné poradie náhodnej veličiny X_i v $\boldsymbol{X}$. S využitím (2.1) je možné upraviť súčet W_{XY}^* a W_X^* nasledovne:

$$\begin{aligned}
W_{XY}^* + W_X^* &= \sum_{i=1}^{n}\sum_{j=1}^{m} 1\{X_i < Y_j\} + \frac{1}{2}1\{X_i = Y_j\} + \sum_{i=1}^{n} R_i^* \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} 1\{X_i < Y_j\} + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m} 1\{X_i = Y_j\} \\
&\quad + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m} 1\{Y_j = X_i\} + \sum_{i=1}^{n}\sum_{j=1}^{m} 1\{Y_j < X_i\} \\
&\quad + \sum_{i=1}^{n} 1 + \sum_{i=1}^{n}\sum_{j=1}^{n} 1\{X_j < X_i\} + \frac{1}{2}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} 1\{X_j = X_i\} \\
&= nm + \sum_{i=1}^{n}\left(1 + \sum_{j=1}^{n} 1\{X_j < X_i\} + \frac{1}{2}\sum_{\substack{j=1\\ j\neq i}}^{n} 1\{X_j = X_i\}\right) \\
&= nm + \sum_{i=1}^{n} S_i^* = nm + \frac{n(n+1)}{2}.
\end{aligned}$$

□

Veta 5. *Ak platí H_0 a $n/N \to p \in (0,1)$, keď $n+m \to \infty$, tak*

$$\sqrt{N}\left(\frac{W_{XY}^*}{nm} - \frac{1}{2}\right) \xrightarrow{D} N\left(0, \xi\frac{1}{p(1-p)}\right),$$

kde $\xi = var\left(F(Z_i) + F(Z_i-)\right)/4$, pričom F značí spoločnú distribučnú funkciu pre náhodné veličiny z $\boldsymbol{X}$ aj z $\boldsymbol{Y}$.

Dôkaz. K dôkazu použijeme vetu 12, rovnako ako sme ju použili v dôkaze vety 3. Definujme jadro $h : \mathbb{R}^2 \to \mathbb{R}$ predpisom

$$h(x,y) = \begin{cases} 1, & \text{ak} \quad x < y, \\ 1/2, & \text{ak} \quad x = y, \\ 0, & \text{inak.} \end{cases}$$

Potom korešpondujúca U-štatistika pre jadro h je tvaru

$$U_{n,m} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} h(X_i,Y_j) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} 1\{X_i < Y_j\} + \frac{1}{2}1\{X_i = Y_j\}.$$

Zrejme platí, že

$$U_{n,m} = \frac{W_{XY}^*}{nm}.$$

Teraz spočítame θ z vety 12 a overíme konečnosť σ_{11}^2.

Platí, že

$$\begin{aligned}
\theta &= \mathsf{E}\left[h(X_i,Y_j)\right] = \mathsf{E}\left[1\{X_i < Y_j\} + \frac{1}{2}1\{X_i = Y_j\}\right] \\
&= \mathsf{P}(X_i < Y_j) + \frac{1}{2}\mathsf{P}(X_i = Y_j) \overset{H_0}{=} \frac{1}{2},
\end{aligned}$$

pričom v kroku označenom $\overset{H_0}{=}$ používame predpoklad, že platí H_0, a teda náhodné veličiny X_i a Y_j pochádzajú z toho istého rozdelenia.

Ďalej,

$$\begin{aligned}\sigma_{11}^2 &= cov\left(1\{X_1 < Y_1\} + \frac{1}{2}1\{X_1 = Y_1\}, 1\{X_1 < Y_1\} + \frac{1}{2}1\{X_1 = Y_1\}\right)\\ &= \mathsf{E}\left[1\{X_1 < Y_1\} + \frac{1}{4}1\{X_1 = Y_1\}\right] - \theta^2 < \infty.\end{aligned}$$

Pre σ_{10}^2 a σ_{01}^2 dostávame nasledujúce vzťahy:

$$\begin{aligned}\sigma_{10}^2 &= cov\left(1\{X_1 < Y_1\} + \frac{1}{2}1\{X_1 = Y_1\}, 1\{X_1 < Y_2\} + \frac{1}{2}1\{X_1 = Y_2\}\right)\\ &= \mathsf{E}\left[1\{X_1 < Y_1\}1\{X_1 < Y_2\} + \frac{1}{2}1\{X_1 < Y_1\}1\{X_1 = Y_2\}\right.\\ &\quad \left. + \frac{1}{2}1\{X_1 < Y_2\}1\{X_1 = Y_1\} + \frac{1}{4}1\{X_1 = Y_1\}1\{X_1 = Y_2\}\right] - \theta^2,\end{aligned} \tag{2.2}$$

$$\begin{aligned}\sigma_{01}^2 &= cov\left(1\{X_1 < Y_1\} + \frac{1}{2}1\{X_1 = Y_1\}, 1\{X_2 < Y_1\} + \frac{1}{2}1\{X_2 = Y_1\}\right)\\ &= \mathsf{E}\left[1\{X_1 < Y_1\}1\{X_2 < Y_1\} + \frac{1}{2}1\{X_1 < Y_1\}1\{X_2 = Y_1\}\right.\\ &\quad \left. + \frac{1}{2}1\{X_2 < Y_1\}1\{X_1 = Y_1\} + \frac{1}{4}1\{X_1 = Y_1\}1\{X_2 = Y_1\}\right] - \theta^2.\end{aligned} \tag{2.3}$$

Pokračujme ďalej s úpravou σ_{01}^2 s využitím podmienenej strednej hodnoty. Špeciálne, zamerajme sa na člen $\mathsf{E}\left[1\{X_1 < Y_1\}1\{X_2 < Y_1\}\right]$. Postupnosťou úprav dostávame, že

$$\begin{aligned}\mathsf{E}\left[1\{X_1 < Y_1\}1\{X_2 < Y_1\}\right] &= \mathsf{E}\left[\mathsf{E}\left[1\{X_1 < Y_1\}1\{X_2 < Y_1\} \mid Y_1\right]\right]\\ &= \int_{\mathbb{R}} \mathsf{E}\left[1\{X_1 < Y_1\}1\{X_2 < Y_1\} \mid Y_1 = y\right] d\,\mathsf{P}_{Y_1}(y)\\ &= \int_{\mathbb{R}} \mathsf{E}\left[1\{X_1 < y\}1\{X_2 < y\}\right] d\,\mathsf{P}_{Y_1}(y)\\ &= \int_{\mathbb{R}} \mathsf{E}\left[1\{X_1 < y\}\right]\mathsf{E}\left[1\{X_2 < y\}\right] d\,\mathsf{P}_{Y_1}(y)\\ &= \int_{\mathbb{R}} F_X(y-)F_X(y-)\, d\,\mathsf{P}_{Y_1}(y) = \mathsf{E}\left[F_X(Y_1-)F_X(Y_1-)\right].\end{aligned} \tag{2.4}$$

Analogicky, pre zvyšné členy vystupujúce v σ_{01}^2 dostávame, že

$$\begin{aligned}\mathsf{E}\left[1\{X_1 < Y_1\}1\{X_2 = Y_1\}\right] &= \int_{\mathbb{R}} \mathsf{E}\left[1\{X_1 < y\}\right]\mathsf{E}\left[1\{X_2 = y\}\right] d\,\mathsf{P}_{Y_1}(y)\\ &= \int_{\mathbb{R}} F_X(y-)\Big(F_X(y) - F_X(y-)\Big)\, d\,\mathsf{P}_{Y_1}(y)\\ &= \mathsf{E}\left[F_X(Y_1-)\Big(F_X(Y_1) - F_X(Y_1-)\Big)\right],\end{aligned} \tag{2.5}$$

$$\begin{aligned}\mathsf{E}\left[1\{X_1 = Y_1\}1\{X_2 = Y_1\}\right] &= \int_{\mathbb{R}} \mathsf{E}\left[1\{X_1 = y\}\right]\mathsf{E}\left[1\{X_2 = y\}\right] d\,\mathsf{P}_{Y_1}(y) \\ &= \int_{\mathbb{R}} \Big(F_X(y) - F_X(y-)\Big)^2 d\,\mathsf{P}_{Y_1}(y) \\ &= \mathsf{E}\left[\Big(F_X(Y_1) - F_X(Y_1-)\Big)^2\right]. \end{aligned} \tag{2.6}$$

S využitím podmienenej strednej hodnoty navyše rovnako (podmienením náhodnou veličinou Y_1) upravíme θ:

$$\begin{aligned}\theta &= \mathsf{E}\left[1\{X_1 < Y_1\}\right] + \frac{1}{2}\mathsf{E}\left[1\{X_1 = Y_1\}\right] \\ &= \int_{\mathbb{R}} \mathsf{E}\left[1\{X_1 < y\}\right] d\,\mathsf{P}_{Y_1}(y) + \frac{1}{2}\int_{\mathbb{R}} \mathsf{E}\left[1\{X_1 = y\}\right] d\,\mathsf{P}_{Y_1}(y) \\ &= \int_{\mathbb{R}} F_X(y-)\, d\,\mathsf{P}_{Y_1}(y) + \frac{1}{2}\int_{\mathbb{R}} F_X(y) - F_X(y-)\, d\,\mathsf{P}_{Y_1}(y) \\ &= \mathsf{E}\left[F_X(Y_1-)\right] + \frac{1}{2}\mathsf{E}\left[F_X(Y_1) - F_X(Y_1-)\right] \\ &= \frac{1}{2}\mathsf{E}\left[F_X(Y_1) + F_X(Y_1-)\right]. \end{aligned} \tag{2.7}$$

Konečne, dosadením rovností (2.4), (2.5), (2.6) a (2.7) do (2.3) dostávame, že

$$\begin{aligned}\sigma_{01}^2 &= \mathsf{E}\left[F_X(Y_1-)F_X(Y_1-)\right] + \mathsf{E}\left[F_X(Y_1-)\Big(F_X(Y_1) - F_X(Y_1-)\Big)\right] \\ &\quad + \frac{1}{4}\mathsf{E}\left[\Big(F_X(Y_1) - F_X(Y_1-)\Big)^2\right] - \Big(\frac{1}{2}\mathsf{E}\left[F_X(Y_1) + F_X(Y_1-)\right]\Big)^2 \\ &= \frac{1}{4}\mathsf{E}\left[\Big(F_X(Y_1) + F_X(Y_1-)\Big)^2\right] - \frac{1}{4}\Big(\mathsf{E}\left[F_X(Y_1) + F_X(Y_1-)\right]\Big)^2 \\ &= \frac{1}{4}\operatorname{var}\Big(F_X(Y_1) + F_X(Y_1-)\Big). \end{aligned}$$

Analogicky, podmienením náhodnou veličinou X_1 dostávame, že

$$\sigma_{10}^2 = \frac{1}{4}\operatorname{var}\Big(F_Y(X_1) + F_Y(X_1-)\Big).$$

Táto rovnosť je podrobne odvodená v kapitole Dodatky.

Predpokladáme, že platí H_0. Potom platí, že $\boldsymbol{Z}$ je náhodný výber z rozdelenia s distribučnou funkciou F, kde $F = F_X = F_Y$. Ďalej platí, že $\sigma_{10}^2 = \sigma_{01}^2$. Označme túto spoločnú hodnotu ξ.

□

Problémom však zatiaľ ostáva, že nepoznáme spoločnú distribučnú funkciu F z tvrdenia vety 5, a teda ani ξ. Z tohto dôvodu prejdeme k odhadu ξ pomocou výberového rozptylu a empirickej distribučnej funkcie. Zaveďme empirickú distribučnú funkciu.

Definícia 4. *Funkciu*

$$\widehat{F}(x) = \frac{1}{N}\sum_{i=1}^{N} 1\{Z_i \le x\} = \frac{1}{N}\left(\sum_{i=1}^{n} 1\{X_i \le x\} + \sum_{j=1}^{m} 1\{Y_j \le x\}\right)$$

budeme nazývať empirickou distribučnou funkciou $\boldsymbol{Z}$.

Teraz s využitím $\widehat{F}$ zaveďme odhad ξ ako

$$\widehat{\xi_N} = \frac{1}{4}\frac{1}{N-1}\sum_{i=1}^{N}(U_i - \overline{U_N})^2, \text{ kde } U_i = \widehat{F}(Z_i) + \widehat{F}(Z_i-).$$

Vidíme, že $\widehat{\xi_N}$ sme skonštruovali ako zloženie výberového rozptylu a empirickej distribučnej funkcie, teda dvoch konzistentných odhadov.

Tvrdenie 6. *Ak platí H_0 a $N = n + m \to \infty$, tak $\widehat{\xi_N} \xrightarrow{P} \xi$.*

Dôkaz. K dôkazu tohto tvrdenia použijeme Glivenkovu-Cantelliho vetu, ktorá hovorí, že empirická distribučná funkcia konverguje rovnomerne v pravdepodobnosti k skutočnej distribučnej funkcii, z ktorej pochádza náhodný výber, teda, že

$$\sup_{t\in\mathbb{R}} \left|\widehat{F}(t) - F(t)\right| \xrightarrow{P} 0.$$

Dôkaz Glivenkovej-Cantelliho vety je uvedený v knihe van der Vaart (1998, strana 266).

Štandardná úprava výberového rozptylu dá, že

$$4\widehat{\xi_N} = \frac{1}{N-1}\sum_{i=1}^{N}(U_i - \overline{U_N})^2 = \frac{1}{N-1}\sum_{i=1}^{N}U_i^2 - \frac{N}{N-1}\overline{U_N}^2. \tag{2.8}$$

Najprv vyšetrime konvergenciu členu $\frac{N}{N-1}\overline{U_N}^2$. Platí, že

$$\frac{1}{N}\sum_{i=1}^{N}\widehat{F}(Z_i) = \frac{1}{N}\sum_{i=1}^{N}\Big(\widehat{F}(Z_i) - F(Z_i)\Big) + \frac{1}{N}\sum_{i=1}^{N}F(Z_i) \xrightarrow{P} \mathsf{E}\ F(Z_i),$$

pretože

$$\left|\frac{1}{N}\sum_{i=1}^{N}\Big(\widehat{F}(Z_i) - F(Z_i)\Big)\right| \le \sup_{t\in\mathbb{R}}\left|\widehat{F}(t) - F(t)\right| \xrightarrow{P} 0$$

a podľa zákona veľkých čísel

$$\frac{1}{N}\sum_{i=1}^{N}F(Z_i) \xrightarrow{P} \mathsf{E}\ F(Z_i).$$

Potom pre $\frac{N}{N-1}\overline{U_N}^2$ dostávame, že

$$\begin{aligned}\frac{N}{N-1}\overline{U_N}^2 &= \frac{N}{N-1}\Big(\frac{1}{N}\sum_{i=1}^{N}\widehat{F}(Z_i) + \frac{1}{N}\sum_{i=1}^{N}\widehat{F}(Z_i-)\Big)^2 \\ &\xrightarrow{P} \Big(\mathsf{E}\left[F(Z_i) + F(Z_i-)\right]\Big)^2.\end{aligned} \tag{2.9}$$

Teraz vyšetríme konvergenciu členu $\frac{1}{N-1}\sum_{i=1}^{N}U_i^2$.

$$\frac{1}{N-1}\sum_{i=1}^{N}U_i^2 = \frac{1}{N-1}\sum_{i=1}^{N}\widehat{F}(Z_i)^2 + \frac{2}{N-1}\sum_{i=1}^{N}\widehat{F}(Z_i)\widehat{F}(Z_i-) + \frac{1}{N-1}\sum_{i=1}^{N}\widehat{F}(Z_i-)^2$$

Platí, že

$$\frac{1}{N-1}\sum_{i=1}^{N}\widehat{F}(Z_i)^2 = \frac{1}{N-1}\sum_{i=1}^{N}\Big(\widehat{F}(Z_i)-F(Z_i)\Big)^2 + \frac{2}{N-1}\sum_{i=1}^{N}F(Z_i)\Big(\widehat{F}(Z_i)-F(Z_i)\Big) + \frac{1}{N-1}\sum_{i=1}^{N}F(Z_i)^2 \xrightarrow{P} \mathsf{E}\ F(Z_i)^2,$$

pretože

$$\left|\frac{1}{N-1}\sum_{i=1}^{N}\Big(\widehat{F}(Z_i)-F(Z_i)\Big)^2\right| \le \frac{N}{N-1}\Big(\sup_{t\in\mathbb{R}}\left|\widehat{F}(t)-F(t)\right|\Big)^2 \xrightarrow{P} 0,$$

$$\left|\frac{2}{N-1}\sum_{i=1}^{N}F(Z_i)\Big(\widehat{F}(Z_i)-F(Z_i)\Big)\right| \le \frac{2}{N-1}\sum_{i=1}^{N}F(Z_i)\sup_{t\in\mathbb{R}}\left|\widehat{F}(t)-F(t)\right| \le 2\frac{N}{N-1}\sup_{t\in\mathbb{R}}\left|\widehat{F}(t)-F(t)\right| \xrightarrow{P} 0$$

a podľa zákona veľkých čísel

$$\frac{1}{N-1}\sum_{i=1}^{N}F(Z_i)^2 \xrightarrow{P} \mathsf{E}\ F(Z_i)^2.$$

Analogicky,

$$\frac{1}{N-1}\sum_{i=1}^{N}\widehat{F}(Z_i-)^2 \xrightarrow{P} \mathsf{E}\ F(Z_i-)^2.$$

Zostáva sa pozrieť na člen $\frac{2}{N-1}\sum_{i=1}^{N}\widehat{F}(Z_i)\widehat{F}(Z_i-)$, pre ktorý platí, že

$$\frac{2}{N-1}\sum_{i=1}^{N}\widehat{F}(Z_i)\widehat{F}(Z_i-) = \frac{2}{N-1}\sum_{i=1}^{N}\Big(\widehat{F}(Z_i)\widehat{F}(Z_i-)-\widehat{F}(Z_i)F(Z_i-)\Big) + \frac{2}{N-1}\sum_{i=1}^{N}\Big((\widehat{F}(Z_i)F(Z_i-)-F(Z_i)F(Z_i-)\Big) + \frac{2}{N-1}\sum_{i=1}^{N}F(Z_i)F(Z_i-) \xrightarrow{P} 2\,\mathsf{E}\ F(Z_i)F(Z_i-),$$

pretože

$$\left|\frac{2}{N-1}\sum_{i=1}^{N}\Big(\widehat{F}(Z_i)\widehat{F}(Z_i-)-\widehat{F}(Z_i)F(Z_i-)\Big)\right| \le \frac{2}{N-1}\sum_{i=1}^{N}\widehat{F}(Z_i)\sup_{t\in\mathbb{R}}\left|\widehat{F}(t-)-F(t-)\right| \xrightarrow{P} 0,$$

$$\left|\frac{2}{N-1}\sum_{i=1}^{N}\Big((\widehat{F}(Z_i)F(Z_i-)-F(Z_i)F(Z_i-)\Big)\right| \le \frac{2}{N-1}\sum_{i=1}^{N}F(Z_i-)\sup_{t\in\mathbb{R}}\left|\widehat{F}(t)-F(t)\right| \xrightarrow{P} 0$$

a podľa zákona veľkých čísel

$$\frac{2}{N-1}\sum_{i=1}^{N} F(Z_i)F(Z_i-) \xrightarrow{P} 2\,\mathsf{E}\; F(Z_i)F(Z_i-).$$

Pre $\frac{1}{N-1}\sum_{i=1}^{N} U_i^2$ teda dostávame, že

$$\begin{aligned}\frac{1}{N-1}\sum_{i=1}^{N} U_i^2 &\xrightarrow{P} \mathsf{E}\; F(Z_i)^2 + 2\,\mathsf{E}\; F(Z_i)F(Z_i-) + \mathsf{E}\; F(Z_i-)^2 \\ &= \mathsf{E}\,\Big(F(Z_i) + F(Z_i-)\Big)^2 \end{aligned} \tag{2.10}$$

Konečne, z rozpisu (2.8) pre $4\widehat{\xi_N}$ dostávame pomocou (2.9) a (2.10), že

$$4\widehat{\xi_N} \xrightarrow{P} \mathsf{E}\,\Big(F(Z_i)+F(Z_i-)\Big)^2 - \Big(\mathsf{E}\,\Big[F(Z_i)+F(Z_i-)\Big]\Big)^2 = \operatorname{var}\Big(F(Z_i)+F(Z_i-)\Big),$$

a teda $\widehat{\xi_N} \xrightarrow{P} \xi$.

□

Tvrdenie 7. *Platí, že* $U_i = \frac{1}{N}(2R_i^* - 1)$ *a* $\overline{U_N} = 1$.

Dôkaz. Nech l je počet náhodných veličín zo $\boldsymbol{Z}$ menších ako Z_i a $l+k$ je počet náhodných veličín zo $\boldsymbol{Z}$ menších alebo rovných Z_i. Náhodná veličina Z_i je teda rovná $k-1$ ďalším náhodným veličinám zo $\boldsymbol{Z}$. Potom

$$\begin{aligned} U_i &= \frac{1}{N}\sum_{j=1}^{N} 1\{Z_j \le Z_i\} + \frac{1}{N}\sum_{j=1}^{N} 1\{Z_j < Z_i\} = \frac{1}{N}(l+k) + \frac{1}{N}l \\ &= \frac{1}{N}\Big(2(1 + l + \frac{1}{2}(k-1)) - 1\Big) = \frac{1}{N}(2R_i^* - 1). \end{aligned}$$

Pre výberový priemer U_i potom platí, že

$$\overline{U_N} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{N}(2R_i^* - 1) = \frac{1}{N^2}\left(2\frac{N(N+1)}{2} - N\right) = 1.$$

□

Veta 8. *Ak platí* H_0 *a* $n/N \to p \in (0,1)$, *keď* $n+m \to \infty$, *tak*

$$\frac{W_X^* - \frac{n(N+1)}{2}}{\sqrt{\frac{nm(N+1)}{12} - \frac{nm\kappa}{(N-1)N}}} \xrightarrow{D} N(0,1),$$

kde

$$\kappa = \sum_z \frac{k_z^3 - k_z}{12},$$

pričom k_z *značí počet, koľkokrát sa medzi hodnotami* $Z_1, \ldots, Z_N$ *vyskytla hodnota* z. *Suma* $\sum_z$ *značí sčítanie cez všetky rôzne hodnoty* $Z_1, \ldots, Z_N$.

Dôkaz. Vďaka Cramérovej-Sluckého vete a tvrdeniu 6 dostávame z vety 5, že

$$\frac{\sqrt{N}\left(\frac{W_{XY}^*}{nm}-\frac{1}{2}\right)}{\sqrt{\widehat{\xi_N}\frac{N^2}{nm}}} \xrightarrow{D} N(0,1). \tag{2.11}$$

Z lemy 10 z kapitoly 3 vieme, že

$$\sum_{i=1}^{N} R_i^{*2} = \frac{N(N+1)(2N+1)}{6} - \kappa.$$

S využitím tejto rovnosti a tvrdenia 7 sa $\widehat{\xi_N}$ zjednoduší nasledovne:

$$\begin{aligned}
\widehat{\xi_N} &= \frac{1}{4}\frac{1}{N-1}\sum_{i=1}^{N}(U_i-\overline{U_N})^2 \\
&= \frac{1}{4}\frac{1}{N-1}\left(\sum_{i=1}^{N}\frac{1}{N^2}(2R_i^*-1)^2 - 2\sum_{i=1}^{N}\frac{1}{N}(2R_i^*-1)+N\right) \\
&= \frac{1}{4}\frac{1}{N-1}\left(\sum_{i=1}^{N}\frac{1}{N^2}(2R_i^*-1)^2 - \frac{2}{N}N^2+N\right) \\
&= \frac{1}{4}\frac{1}{N-1}\left(\frac{4}{N^2}\sum_{i=1}^{N}R_i^{*2} - \frac{4}{N^2}\sum_{i=1}^{N}R_i^* + \frac{1}{N} - N\right) \\
&= \frac{1}{4}\frac{1}{N-1}\left(\frac{4}{N^2}\left(\frac{N(N+1)(2N+1)}{6}-\kappa\right) - \frac{4}{N^2}\frac{N(N+1)}{2} + \frac{1}{N} - N\right) \\
&= \frac{1}{4}\frac{1}{N-1}\left(\frac{4(N+1)(2N+1)-12(N+1)+6-6N^2}{6N} - \frac{4}{N^2}\kappa\right) \\
&= \frac{1}{4}\frac{1}{N-1}\frac{2N^2-2}{6N} - \frac{1}{N-1}\frac{1}{N^2}\kappa \\
&= \frac{1}{12}\frac{N+1}{N} - \frac{1}{N-1}\frac{1}{N^2}\kappa.
\end{aligned} \tag{2.12}$$

Výraz na ľavej strane (2.11) už len prevedieme na tvar, v ktorom vystupuje W_X^*. Postupnosťou úprav s využitím tvrdenia 4 a vzťahu (2.12) dostávame, že

$$\begin{aligned}
\frac{\sqrt{N}\left(\frac{W_{XY}^*}{nm}-\frac{1}{2}\right)}{\sqrt{\widehat{\xi_N}\frac{N^2}{nm}}} &= \frac{\sqrt{N}\left(\frac{W_{XY}^*}{nm}-\frac{1}{2}\right)}{N\sqrt{\frac{1}{nm}}\sqrt{\frac{N+1}{12N}-\frac{1}{(N-1)N^2}\kappa}} = \frac{\sqrt{N}\left(\frac{nm+\frac{n(n+1)}{2}-W_X^*}{nm}-\frac{1}{2}\right)}{N\sqrt{\frac{1}{nm}}\sqrt{\frac{N+1}{12N}-\frac{1}{(N-1)N^2}\kappa}} \\
&= \frac{\sqrt{N}\frac{1}{nm}\left(\frac{n(n+m+1)}{2}-W_X^*\right)}{N\sqrt{\frac{1}{nm}}\sqrt{\frac{N+1}{12N}-\frac{1}{(N-1)N^2}\kappa}} = -\frac{W_X^*-\frac{n(n+m+1)}{2}}{\sqrt{nm}\sqrt{\frac{N+1}{12}-\frac{1}{(N-1)N}\kappa}} \\
&= -\frac{W_X^*-\frac{n(N+1)}{2}}{\sqrt{\frac{nm(N+1)}{12}-\frac{nm\kappa}{(N-1)N}}}.
\end{aligned}$$

Alternatívny dôkaz je uvedený v knihe Lehmann a D'Abrera (1975, strana 355). □

3. Súvislosť asymptotického a podmieneného rozptylu W_X^*

V tejto kapitole odvodíme

$$\mathsf{E}\left[W_X^* \mid \boldsymbol{Z}_{(.)}\right] \qquad \text{a} \qquad \text{var}\left[W_X^* \mid \boldsymbol{Z}_{(.)}\right],$$

za platnosti H_0 (1.2), čím ukážeme súvislosť s asymptotickým výsledkom z vety 8. K tomu sa nám budú hodiť takzvané lineárne poradové štatistiky a náhodné poradia.

Nech $\boldsymbol{X} = (X_1, \ldots, X_n)$ a $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$ sú nezávislé náhodné výbery z rozdelení s distribučnými funkciami F_X a F_Y. Zatiaľ predpokladajme, že F_X a F_Y sú spojité. Opäť označme $\boldsymbol{Z} = (Z_1, \ldots, Z_N) = (X_1, \ldots, X_n, Y_1, \ldots, Y_m)$.

Nech R_i je poradie náhodnej veličiny Z_i v $\boldsymbol{Z}$. Ďalej, nech pre $i \in \{1, \ldots, N\}$ je $c_i \in \mathbb{R}$ reálny koeficient a $a : \{1, \ldots, N\} \to \mathbb{R}$ je takzvaná skórová funkcia. Položme

$$S = \sum_{i=1}^{N} c_i \, a(R_i)$$

a označme

$$\bar{a} = \frac{1}{N}\sum_{i=1}^{N} a(i), \qquad \sigma_a^2 = \frac{1}{N}\sum_{i=1}^{N} [a(i) - \bar{a}]^2,$$
$$\bar{c} = \frac{1}{N}\sum_{i=1}^{N} c_i, \qquad \sigma_c^2 = \frac{1}{N}\sum_{i=1}^{N} [c_i - \bar{c}]^2.$$

Veta 9. *Nech $\boldsymbol{X}$ a $\boldsymbol{Y}$ sú nezávislé náhodné výbery zo spojitého rozdelenia s distribučnou funkciou $F = F_X = F_Y$. Potom platí, že*

$$\mathsf{E}\, S = N\bar{a}\bar{c} \qquad a \qquad var\, S = \frac{N^2}{N-1}\sigma_a^2\sigma_c^2.$$

Dôkaz. Dôkaz plynie z toho, že náhodné veličiny R_i nadobúdajú každú z hodnôt $1, \ldots, N$ s pravdepodobnosťou $1/N$. Podrobný dôkaz je uvedený v knihe Anděl (2007a, strana 236).

□

V tejto chvíli upustíme od predpokladu spojitosti F_X a F_Y. Už teda nemôžeme predpokladať, že pravdepodobnosť zhodných pozorovaní je 0, ako sme predpokladali v (1.1). Vieme, že v prípade existencie zhôd nedáva definícia poradí z prvej kapitoly zmysel, pretože už nie je možné zoradiť náhodné veličiny v združenom výbere $\boldsymbol{Z}$ od najmenšej po najväčšiu tak, aby boli splnené ostré nerovnosti $Z_{(1)} < Z_{(2)} < \ldots < Z_{(N)}$. Z tohto dôvodu prejdeme k takzvaných náhodným poradiam. Napriek nemožnosti usporiadať združený výber $\boldsymbol{Z}$ tak, aby platili ostré nerovnosti, ho stále môžeme usporiadať tak, aby platili neostré nerovnosti $Z_{(1)} \leq Z_{(2)} \leq \ldots \leq Z_{(N)}$. Uvažujme blok k zhodných pozorovaní: $\ldots < Z_{(i)} = Z_{(i+1)} = \ldots = Z_{(i+k-1)} < \ldots$. Náhodné poradia náhodných veličín z tohto bloku budú dané náhodne vybranou permutáciou množiny

$\{i, i+1, \ldots, i+k-1\}$. Za predpokladu, že $F_X = F_Y$ náhodné poradie $(\widetilde{R}_i)$ stále spĺňa, že je rovnomerne rozdelené na množine $\{1, \ldots, N\}$, a teda veta 9 platí aj s náhodnými poradiami, a to aj v prípade, že spojitosť F_X a F_Y nepredpokladáme. Využijeme ju aby sme spočítali strednú hodnotu a rozptyl štatistiky W_X^* za podmienky, že poznáme $\boldsymbol{Z}_{(.)}$.

Položme

$$c_i = \begin{cases} 1, & \text{ak} \quad i = 1, \ldots, n, \\ 0, & \text{ak} \quad i = n+1, \ldots, N. \end{cases}$$

Označme $z_{(1)}, \ldots, z_{(N)}$ vzostupne usporiadanú realizáciu združeného výberu $\boldsymbol{Z}$ a definujme

$$a(i) = 1 + \sum_{j=1}^{N} 1[z_{(j)} < z_{(i)}] + \frac{1}{2} \sum_{\substack{j=1 \\ j \neq i}}^{N} 1[z_{(j)} = z_{(i)}].$$

Pre takto zvolené konštanty c_i a skórovú funkciu a platí, že $W_X^* = \sum_{i=1}^{N} c_i \, a(\widetilde{R}_i)$.

Zrejme platí, že

$$\bar{a} = \frac{N+1}{2} \qquad \text{a} \qquad \bar{c} = \frac{n}{N}. \tag{3.1}$$

Ďalej,

$$\begin{aligned} \sigma_c^2 &= \frac{1}{N} \left(\sum_{i=1}^{n} (1 - \frac{n}{N})^2 + \sum_{i=n+1}^{N} (-\frac{n}{N})^2 \right) \\ &= \frac{1}{N} \left(n(1 - \frac{2n}{N} + \frac{n^2}{N^2}) + m\frac{n^2}{N^2} \right) \\ &= \frac{1}{N} \left(n - \frac{2n^2}{N} + \frac{n^2}{N} \right) = \frac{nm}{N^2}. \end{aligned} \tag{3.2}$$

Pozrime sa na σ_a^2:

$$\sigma_a^2 = \frac{1}{N} \sum_{i=1}^{N} [a(i) - \bar{a}]^2 = \frac{1}{N} \sum_{i=1}^{N} (a(i))^2 - \bar{a}^2. \tag{3.3}$$

Vidíme, že sa stačí zamerať na $\sum_{i=1}^{N} (a(i))^2$. Uvedomme si, že ide vlastne o sumu druhých mocnín prvkov postupnosti, ktorá vznikne z postupnosti prirodzených čísel $1, \ldots, N$ tak, že túto postupnosť rozdelíme do r blokov, kde r je počet rôznych hodnôt medzi hodnotami $z_1, \ldots, z_N$, a každé číslo v danom bloku nahradíme aritmetickým priemerom daného bloku. S touto sumou nám pomôže nasledujúca lema.

Lema 10. *Nech* $(a_1^{(1)}, \ldots, a_{s_1}^{(1)}, \ldots, a_1^{(r)}, \ldots, a_{s_r}^{(r)}) = (1, \ldots, N)$, *kde* $N \in \mathbb{N}$ *a* $r \in \{1, \ldots, N\}$. *Pre* $i \in \{1, \ldots, r\}$ *a* $j \in \{1, \ldots, s_i\}$ *položme*

$$b_j^{(i)} = \frac{1}{s_i} \sum_{k=1}^{s_i} a_k^{(i)}.$$

Potom

$$\sum_{i=1}^{r} \sum_{j=1}^{s_i} (b_j^{(i)})^2 = \frac{N(N+1)(2N+1)}{6} - \kappa,$$

kde $\kappa = \sum_{i=1}^{r} \frac{s_i^3 - s_i}{12}$.

Dôkaz. Pre $i \in \{1, \ldots, r\}$ platí, že

$$\sum_{k=1}^{s_i} (a_k^{(i)})^2 = \sum_{k=0}^{s_i-1} (a_1^{(i)} + k)^2 = s_i (a_1^{(i)})^2 + a_1^{(i)}(s_i - 1)s_i + \frac{(s_i - 1)s_i(2s_i - 1)}{6} \quad (3.4)$$

a

$$s_i \Big(\frac{2a_1^{(i)} + s_i - 1}{2} \Big)^2 = s_i (a_1^{(i)})^2 + a_1^{(i)}(s_i - 1)s_i + \frac{(s_i - 1)^2 s_i}{4}. \quad (3.5)$$

Z (3.4) a (3.5) dostávame, že

$$s_i \Big(\frac{2a_1^{(i)} + s_i - 1}{2} \Big)^2 = \sum_{k=1}^{s_i} (a_k^{(i)})^2 - \frac{(s_i - 1)s_i(2s_i - 1)}{6} + \frac{(s_i - 1)^2 s_i}{4},$$

vďaka čomu už môžeme dôkaz priamočiaro dokončiť:

$$\begin{aligned}
\sum_{i=1}^{r} \sum_{j=1}^{s_i} (b_j^{(i)})^2 &= \sum_{i=1}^{r} \sum_{j=1}^{s_i} \Big(\frac{1}{s_i} \sum_{k=1}^{s_i} a_k^{(i)} \Big)^2 = \sum_{i=1}^{r} \frac{1}{s_i} \Big(\sum_{k=1}^{s_i} a_k^{(i)} \Big)^2 \\
&= \sum_{i=1}^{r} \frac{1}{s_i} \Big(s_i \frac{2a_1^{(i)} + s_i - 1}{2} \Big)^2 \\
&= \sum_{i=1}^{r} \Big(\sum_{k=1}^{s_i} (a_k^{(i)})^2 - \frac{(s_i - 1)s_i(2s_i - 1)}{6} + \frac{(s_i - 1)^2 s_i}{4} \Big) \\
&= \sum_{i=1}^{N} i^2 - \sum_{i=1}^{r} \frac{s_i^3 - s_i}{12} = \frac{N(N+1)(2N+1)}{6} - \sum_{i=1}^{r} \frac{s_i^3 - s_i}{12}.
\end{aligned}$$

□

Označme

$$\kappa = \sum_z \frac{k_z^3 - k_z}{12},$$

kde k_z značí počet, koľkokrát sa medzi hodnotami $z_1, \ldots, z_N$ vyskytla hodnota z. Suma $\sum_z$ značí sčítanie cez všetky rôzne hodnoty $z_1, \ldots, z_N$. Podľa lemy 10 teda platí, že

$$\sum_{i=1}^{N} (a(i))^2 = \frac{N(N+1)(2N+1)}{6} - \kappa.$$

Dosadením do (3.3) dostaneme, že

$$\begin{aligned}
\sigma_a^2 &= \frac{(N+1)(2N+1)}{6} - \frac{\kappa}{N} - \bar{a}^2 \\
&= \frac{(N+1)(2N+1)}{6} - \frac{\kappa}{N} - \frac{(N+1)^2}{4} \\
&= \frac{(N+1)(N-1)}{12} - \frac{\kappa}{N}.
\end{aligned} \quad (3.6)$$

Veta 11. *Ak platí H_0, tak pre štatistiku W_X^* platí, že*

$$E\left[W_X^* \mid \boldsymbol{Z}_{(.)}\right] = \frac{n(N+1)}{2} \qquad a \qquad var\left[W_X^* \mid \boldsymbol{Z}_{(.)}\right] = \frac{nm(N+1)}{12} - \frac{nm\kappa}{(N-1)N}.$$

Dôkaz. Tvrdenie plynie z dosadenia (3.1), (3.2) a (3.6) do vety 9.
Pre rozptyl dostávame, že

$$\operatorname{var}\left[W_X^* \mid \boldsymbol{Z}_{(.)}\right] = \frac{N^2}{N-1}\sigma_a^2\sigma_c^2 = \frac{nm}{N-1}\sigma_a^2 = \frac{nm}{N-1}\left(\frac{(N+1)(N-1)}{12} - \frac{\kappa}{N}\right)$$
$$= \frac{nm(N+1)}{12} - \frac{nm\kappa}{(N-1)N}.$$

Pre strednú hodnotu dostávame, že

$$\mathsf{E}\left[W_X^* \mid \boldsymbol{Z}_{(.)}\right] = N\bar{a}\bar{c} = \frac{n(N+1)}{2}.$$

□

Vidíme, že menovateľ výrazu z tvrdenia vety 8 je presne $\sqrt{\operatorname{var}\left[W_X^* \mid \boldsymbol{Z}_{(.)}\right]}$.
Ešte si všimnime, že

$$\frac{\operatorname{var}\left[W_X^* \mid \boldsymbol{Z}_{(.)}\right]}{nmN} = \widehat{\xi_N},$$

a teda takto znormovaný podmienený rozptyl konverguje podľa tvrdenia 6 za H_0 s $N = n + m \to \infty$ v pravdepodobnosti ku ξ.

4. Porovnanie testu bez korekcie a testu s korekciou pomocou simulácie

V tejto kapitole sa pokúsime pomocou simulácií nahliadnuť na to, aký je vplyv odvodenej korekcie na dodržovanie predpísanej hladiny α. Zaujíma nás, ako sa správa hladina Wilcoxonovho testu s narastajúcim počtom zhodných pozorovaní v prípade, že sme v priebehu výpočtu zabudli zohľadniť korekciu. V priebehu celej kapitoly budeme predpokladať, že $\alpha = 0{,}05$.

Uvažujme 2 nezávislé náhodné výbery $\boldsymbol{X}$ a $\boldsymbol{Y}$ z toho istého rozdelenia. Keďže obidva náhodné výbery sú z toho istého rozdelenia, je splnená H_0 (1.2). Chceme odhadnúť skutočnú hladinu testu s korekciou, respektíve testu bez korekcie.

Pri simulácii budeme postupovať tak, že 10 000-krát vygenerujeme 2 nezávislé náhodné výbery z $N(100,\sigma^2)$, vygenerované dáta zaokrúhlime na desiatky, aby nám vznikli zhodné pozorovania a zakaždým určíme, či jednotlivé testy (s korekciou alebo bez korekcie) H_0 zamietli. Podiel počtu zamietnutí a 10 000 nám odhaduje skutočnú hladinu testu. Tento postup simuluje častú príčinu, prečo vznikajú zhody v reálnych dátach - zaokrúhľovanie.

So zmenšujúcou sa smerodajnou odchýlkou (σ) klesá variabilita výberu, a teda počet zhodných pozorovaní stúpa. Pre $\sigma = 15$ (relatívne malý počet zhodných pozorovaní) si test s korekciou drží predpísanú 5-percentnú hladinu, zatiaľ čo hladina testu bez korekcie už predpísaných 5 percent nedosahuje (viď Tabuľka 4.1). Pre $\sigma = 10$ nie je na teste s korekciou badateľná žiadna zmena, stále si drží predpísanú hladinu, avšak hladina testu bez korekcie je už výrazne nižšia než požadovaných 5 percent (viď Tabuľka 4.2). Pre $\sigma = 5$ (veľmi veľký počet zhodných pozorovaní) test s korekciou stále funguje správne. Hladina testu bez korekcie už naopak nedosahuje ani 2 percentá (viď Tabuľka 4.3).

Simulácie ukazujú, že hladina testu bez korekcie s klesajúcou smerodajnou odchýlkou systematicky klesá. Simulujme toto správanie ešte jedným grafom. Nech rozsah náhodného výberu $\boldsymbol{X}$ je 50 a rozsah náhodného výberu $\boldsymbol{Y}$ je takisto 50. Budeme simulovať hladinu testu s korekciou a hladinu testu bez korekcie pre smerodajnú odchýlku z intervalu $[3, 17]$. Vezmime delenie tohto intervalu také, že vzdialenosť medzi deliacimi bodmi je 0,1. Pre každý deliaci bod nasimulujeme hladinu testu s korekciou a hladinu testu bez korekcie pre smerodajnú odchýlku rovnú tomuto bodu. Výsledky vynesieme do grafu a vidíme, že test s korekciou si stabilne drží predpísanú hladinu, zatiaľ čo hladina testu bez korekcie s klesajúcou smerodajnou odchýlkou klesá k 0 (viď Obr. 4.1).

Rozsah X	**Rozsah** Y	**Podiel zamietnutí s korekciou** [%]	**Podiel zamietnutí bez korekcie** [%]
10	90	4,90	4,39
90	10	5,01	4,37
30	70	4,75	4,17
70	30	5,06	4,54
50	50	4,86	4,31

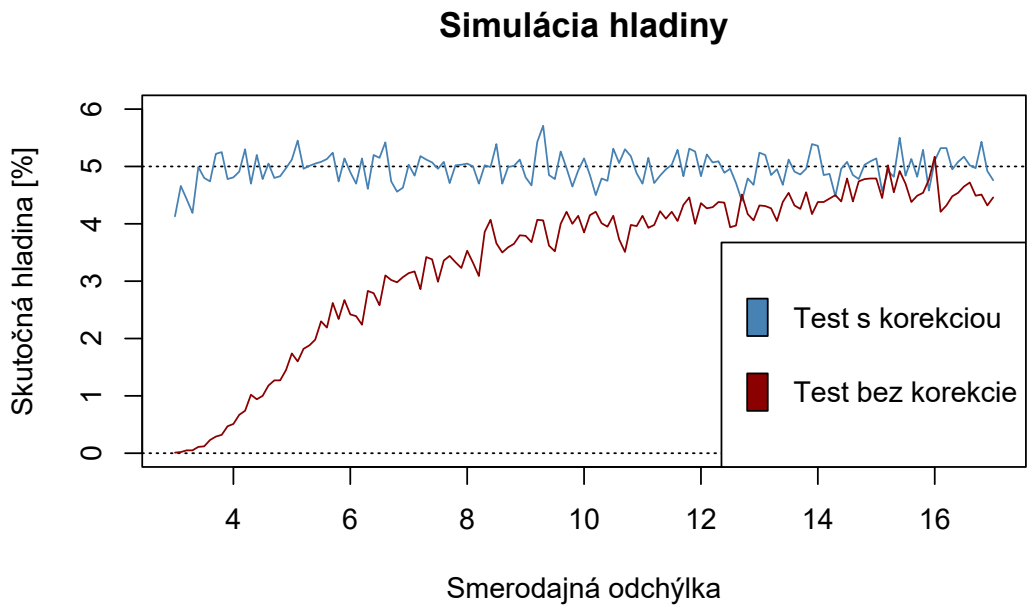
Tabuľka 4.1: Porovnanie hladiny jednotlivých testov pre $\sigma = 15$.

Rozsah X	**Rozsah** Y	**Podiel zamietnutí s korekciou** [%]	**Podiel zamietnutí bez korekcie** [%]
10	90	4,80	3,68
90	10	4,75	3,71
30	70	5,00	3,91
70	30	4,84	3,82
50	50	4,97	3,98

Tabuľka 4.2: Porovnanie hladiny jednotlivých testov pre $\sigma = 10$.

Rozsah X	**Rozsah** Y	**Podiel zamietnutí s korekciou** [%]	**Podiel zamietnutí bez korekcie** [%]
10	90	4,96	1,73
90	10	5,13	1,46
30	70	5,05	1,81
70	30	4,68	1,31
50	50	4,70	1,42

Tabuľka 4.3: Porovnanie hladiny jednotlivých testov pre $\sigma = 5$.



Obr. 4.1: Porovnanie hladiny jednotlivých testov v závislosti na σ.

5. Dodatky

5.1 U-štatistiky

Definícia 5. *Pre 2 nezávislé náhodné výbery* $(X_1,\dots,X_n)$ *a* $(Y_1,\dots,Y_m)$ *a jadro* $h : \mathbb{R}^{r+s} \to \mathbb{R}$ *symetrické v bloku prvých* r *premenných a v bloku posledných* s *premenných je príslušná dvojvýberová U-štatistika definovaná ako*

$$U_{n,m} = \frac{1}{\binom{n}{r}\binom{m}{s}} \sum_{\alpha} \sum_{\beta} h(X_{\alpha_1},\dots,X_{\alpha_r},Y_{\beta_1},\dots,Y_{\beta_s}),$$

kde $\alpha = (\alpha_1,\dots,\alpha_r)$ *značí sčítanie cez všetky* r*-prvkové podmnožiny* $\{1,\dots,n\}$ *a* $\beta = (\beta_1,\dots,\beta_s)$ *značí sčítanie cez všetky* s*-prvkové podmnožiny* $\{1,\dots,m\}$.

Označme

$$\sigma_{ij}^2 = cov\Big(h(X_1,\dots,X_i,X_{i+1},\dots,X_r,Y_1,\dots,Y_j,Y_{j+1},\dots,Y_s),$$
$$h(X_1,\dots,X_i,X_{i+1}^*,\dots,X_r^*,Y_1,\dots,Y_j,Y_{j+1}^*,\dots,Y_s^*)\Big),$$

kde $(X_1,\dots,X_r,X_{i+1}^*,\dots,X_r^*)$ a $(Y_1,\dots,Y_s,Y_{j+1}^*,\dots,Y_s^*)$ sú nezávislé náhodné výbery. Ďalej označme

$$\theta = \mathsf{E}\, h(X_1,\dots,X_r,Y_1,\dots,Y_s).$$

Veta 12. *Ak platí, že* $n/N \to p \in (0,1)$, *keď* $n+m \to \infty$ *a* σ_{11}^2 *je navyše konečné, tak*

$$\sqrt{N}\,(U_{n,m} - \theta) \xrightarrow{D} N(0,\sigma^2),\ kde\ \sigma^2 = \frac{\sigma_{10}^2}{p} + \frac{\sigma_{01}^2}{1-p}.$$

Dôkaz. Dôkaz je uvedený v knihe van der Vaart (1998, strana 166).

□

5.2 Odvodenie σ_{10}^2 z vety 5

Analogicky ako pre σ_{01}^2 pokračujme s úpravou σ_{10}^2 s využitím podmienenej strednej hodnoty. Špeciálne, zamerajme sa na člen $\mathsf{E}\left[1\{X_1 < Y_1\}1\{X_1 < Y_2\}\right]$. Postupnosťou úprav dostávame, že

$$\begin{aligned}
\mathsf{E}\left[1\{X_1 < Y_1\}1\{X_1 < Y_2\}\right] &= \mathsf{E}\left[\mathsf{E}\left[1\{X_1 < Y_1\}1\{X_1 < Y_2\} \mid X_1\right]\right] \\
&= \int_{\mathbb{R}} \mathsf{E}\left[1\{X_1 < Y_1\}1\{X_1 < Y_2\} \mid X_1 = x\right] d\,\mathsf{P}_{X_1}(x) \\
&= \int_{\mathbb{R}} \mathsf{E}\left[1\{x < Y_1\}1\{x < Y_2\}\right] d\,\mathsf{P}_{X_1}(x) \\
&= \int_{\mathbb{R}} \mathsf{E}\left[1\{x < Y_1\}\right] \mathsf{E}\left[1\{x < Y_2\}\right] d\,\mathsf{P}_{X_1}(x) \\
&= \int_{\mathbb{R}} \Big(1 - F_Y(x)\Big)\Big(1 - F_Y(x)\Big)\, d\,\mathsf{P}_{X_1}(x) \\
&= \mathsf{E}\left[\Big(1 - F_Y(X_1)\Big)\Big(1 - F_Y(X_1)\Big)\right]. \qquad (5.1)
\end{aligned}$$

Analogicky, pre zvyšné členy vystupujúce v σ_{10}^2 dostávame, že

$$\begin{aligned}\mathsf{E}\left[1\{X_1 < Y_1\}1\{X_1 = Y_2\}\right] &= \int_{\mathbb{R}} \mathsf{E}\left[1\{x < Y_1\}\right] \mathsf{E}\left[1\{x = Y_2\}\right] d\,\mathsf{P}_{X_1}(x) \\ &= \int_{\mathbb{R}} \Big(1 - F_Y(x)\Big)\Big(F_Y(x) - F_Y(x-)\Big)\, d\,\mathsf{P}_{X_1}(x) \\ &= \mathsf{E}\left[\Big(1 - F_Y(X_1)\Big)\Big(F_Y(X_1) - F_Y(X_1-)\Big)\right], \end{aligned} \tag{5.2}$$

$$\begin{aligned}\mathsf{E}\left[1\{X_1 = Y_1\}1\{X_1 = Y_2\}\right] &= \int_{\mathbb{R}} \mathsf{E}\left[1\{x = Y_1\}\right] \mathsf{E}\left[1\{x = Y_2\}\right] d\,\mathsf{P}_{X_1}(x) \\ &= \int_{\mathbb{R}} \Big(F_Y(x) - F_Y(x-)\Big)\Big(F_Y(x) - F_Y(x-)\Big)\, d\,\mathsf{P}_{X_1}(x) \\ &= \mathsf{E}\left[\Big(F_Y(X_1) - F_Y(X_1-)\Big)\Big(F_Y(X_1) - F_Y(X_1-)\Big)\right]. \end{aligned} \tag{5.3}$$

S využitím podmienenej strednej hodnoty naviac rovnako (podmienením náhodnou veličinou X_1) upravíme θ:

$$\begin{aligned}\theta &= \mathsf{E}\left[1\{X_1 < Y_1\}\right] + \frac{1}{2}\,\mathsf{E}\left[1\{X_1 = Y_1\}\right] \\ &= \int_{\mathbb{R}} \mathsf{E}\left[1\{x < Y_1\}\right] d\,\mathsf{P}_{X_1}(x) + \frac{1}{2}\int_{\mathbb{R}} \mathsf{E}\left[1\{x = Y_1\}\right] d\,\mathsf{P}_{X_1}(x) \\ &= \int_{\mathbb{R}} 1 - F_Y(x)\, d\,\mathsf{P}_{X_1}(x) + \frac{1}{2}\int_{\mathbb{R}} F_Y(x) - F_Y(x-)\, d\,\mathsf{P}_{X_1}(x) \\ &= \mathsf{E}\left[1 - F_Y(X_1)\right] + \frac{1}{2}\,\mathsf{E}\left[F_Y(X_1) - F_Y(X_1-)\right] \\ &= 1 - \frac{1}{2}\,\mathsf{E}\left[F_Y(X_1) + F_Y(X_1-)\right]. \end{aligned} \tag{5.4}$$

Konečne, dosadením (5.1), (5.2), (5.3) a (5.4) do (2.2) dostávame, že

$$\begin{aligned}\sigma_{10}^2 &= \mathsf{E}\left[\Big(1 - F_Y(X_1)\Big)\Big(1 - F_Y(X_1)\Big)\right] \\ &\quad + \mathsf{E}\left[\Big(1 - F_Y(X_1)\Big)\Big(F_Y(X_1) - F_Y(X_1-)\Big)\right] \\ &\quad + \frac{1}{4}\,\mathsf{E}\left[\Big(F_Y(X_1) - F_Y(X_1-)\Big)\Big(F_Y(X_1) - F_Y(X_1-)\Big)\right] \\ &\quad - \left(1 - \frac{1}{2}\,\mathsf{E}\left[F_Y(X_1) + F_Y(X_1-)\right]\right)^2 \\ &= \mathsf{E}\left[F_Y(X_1)F_Y(X_1-)\right] + \frac{1}{4}\,\mathsf{E}\left[F_Y(X_1)^2 - 2F_Y(X_1)F_Y(X_1-) + F_Y(X_1-)^2\right] \\ &\quad - \frac{1}{4}\Big(\,\mathsf{E}\left[F_Y(X_1) + F_Y(X_1-)\right]\Big)^2 \\ &= \frac{1}{4}\,\mathsf{E}\left[\Big(F_Y(X_1) + F_Y(X_1-)\Big)^2\right] - \frac{1}{4}\Big(\,\mathsf{E}\left[F_Y(X_1) + F_Y(X_1-)\right]\Big)^2 \\ &= \frac{1}{4}\,\mathrm{var}\Big(F_Y(X_1) + F_Y(X_1-)\Big). \end{aligned}$$

Záver

Táto práca sa venovala dvojvýberovému Wilcoxonovmu testu. Najprv sme sa zaoberali jednoduchším prípadom, kedy sme predpokladali, že dáta pochádzajú zo spojitých rozdelení. Následne sme sa začali venovať zložitejšiemu prípadu, kedy sme sa vzdali predpokladu spojitosti a začali predpokladať, že v dátach sa môžu vyskytovať zhody.

S využitím známych výsledkov o U-štatistikách sme odvodili asymptotické rozdelenie Wilcoxonovej testovej štatistiky v prípade existencie zhôd, čo nás viedlo k návrhu korekcie testu pre dáta obsahujúce zhody. Ďalej sme poukázali na súvislosť odvodenej korekcie a podmieneného rozptylu testovej štatistiky (a podmienenej strednej hodnoty).

Hlavným prínosom teoretickej časti práce teda bola aplikácia známych výsledkov o U-štatistikách na odvodenie asymptotického rozdelenia Wilcoxonovej testovej štatistiky v prípade zhôd a výpočet podmieneného rozptylu tejto štatistiky v tretej kapitole.

Nakoniec sme pomocou simulácií sledovali aký vplyv má odvodená korekcia na skutočnú hladinu testu pri zvyšujúcom sa počte zhodných pozorovaní. Ukázali sme, že zatiaľ čo test s korekciou si pri rastúcom počte zhodných pozorovaní stabilne držal predpísanú hladinu, hladina testu bez korekcie systematicky klesala až k nule.

Test s korekciou stanovenú hladinu dodržiava len asymptoticky. V práci sme uviedli príklad ako by sme postupovali, keby sme chceli test založený na presnom rozdelení. Predmetom ďalšej štúdie by teda mohla byť otázka, ako súvisia rozsahy výberov (kedy sú už dostatočne veľké) a skutočná hladina testu s korekciou založeného na asymptotickom rozdelení.

Zoznam použitej literatúry

Anděl, J. (2007a). *Základy matematické statistiky*. Druhé opravené vydání. Matfyzpress, Praha. ISBN 80-7378-001-1.

Anděl, J. (2007b). *Statistické metody*. Čtvrté upravené vydání. Matfyzpress, Praha. ISBN 978-80-7378-003-6.

Lehmann, E. L. a D'Abrera, H. J. M. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Holden-day series in probability and statistics. Holden-Day, San Francisco. ISBN 0-8162-4994-6.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge. ISBN 0-521-78450-6.