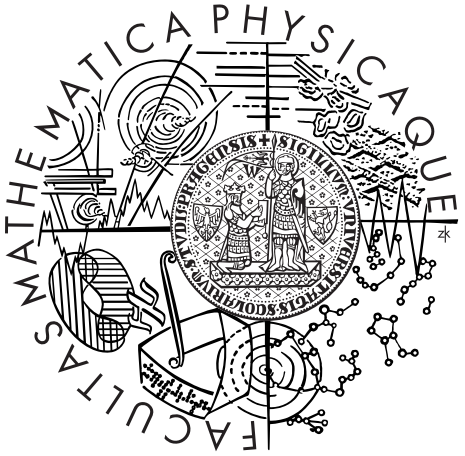
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Natália Prelecová

Úvod do neparametrických metod

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. Mgr. Michal Kulich,Ph.D.

Studijní program: Matematika

Studijní obor: Finanční matematika

Praha 2012

Chcela by som poďakovať všetkým, ktorí mi akýmkoľvek spôsobom pomohli pri spracovaní tejto bakalárskej práce. Moje poďakovanie patrí najmä vedúcemu práce, doc. Mgr. Michalovi Kulichovi,Ph.D., za vedenie a cenné pripomienky pri spracovaní práce.

Osobitné poďakovanie patrí mojej rodine a priateľom, bez ktorých podpory by som to určite nezvládla.

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V Praze dne 18.5.2012 Podpis autora

Název práce: Úvod do neparametrických metod
Autor: Natália Prelecová
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedoucí bakalářské práce: doc. Mgr. Michal Kulich, Ph.D., Katedra pravděpodobnosti a matematické statistiky
Abstrakt: Cílem této bakalářské práce je představit základní neparametrické metody. Neparametrické metody jsou velkou skupinou statistických postupů, které nepředpokládají konkrétní rozdělení dat (například normální). Často jsou jedinou dostupnou metodou pro specifické typy údajů, např. pro zkoumání pořadí nebo četností dat. Slabší předpoklady těchto metod způsobují, že tyto neparametrické testy nejsou tak silné jako testy parametrické.
Práce se zaobírá čtyřmi neparametrickými testy. Jsou to znaménkový test, jednovýběrový Wilcoxonův test, Mann-Whitneyův test a dvouvýběrový Wilcoxonův test. Každý test bude popsán v následující struktuře. Formulace předpokladů, nulové hypotézy a alternativy. Konstrukce testové statistiky a přezkoumání kritických oborů. Také dojde k prozkoumání problémů vyskytujících se při testování jako například problému shodných pozorování. Při dvouvýběrovém Wilcoxonovém testu budou představeny základní charakteristiky testů založených na pořadové statistice.
Klíčová slova: neparametrický, hypotéza, pořadí, konzistence, statistika

Title: Introduction to Nonparametric Methods
Author: Natália Prelecová
Department: Department of Probability and Mathematical Statistics
Supervisor: doc. Mgr. Michal Kulich,Ph.D., Department of Probability and Mathematical Statistics
Abstract: The aim of this thesis is to introduce basic nonparametric methods. Nonparametric methods embrace a large class of statistical procedures which do not assume specific data distribution such as normal distribution. They often represent the only available means of examining specific types of data, for example ranks or counts. Weaker assumptions of these methods make them less powerful than their parametric counterparts.
This thesis describes in detail four nonparametric tests- the Ordinary Sign Test, the Wilcoxon Signed-rank Test, the Mann-Whitney Test and finally the Two-sample Wilcoxon Test. The structure of their description will entail the following: the formulation of assumptions, null hypothesis and alternatives, the construction of the test statistic and the definition of rejection regions. The most essential problems, such as the problem of ties, will be also covered. The basic characteristics of the Linear Rank Statistics will be also explained, followed by the Two-sample Wilcoxon test.
Keywords: nonparametrical, hypothesis, ranks, consistency, statistic

Obsah

Úvod

Štatistika je charakterizovaná ako vedecký obor zaoberajúci sa zberom, analýzou a spracovaním informácií, ktoré kvantitatívne charakterizujú zákonitosti javov v spojitosti s ich kvalitatívnym obsahom. Štatistické metódy sa využívajú v mnohých oboroch a odvetviach ako napríklad v medicíne fyzike či demografii. V súčasnosti sa na analýzu dát vo veľkej miere využívajú počítače a špecifický software. Pre lepšie použitie je nutné porozumieť základným súvislostiam pri používaní testov a ich základným charakteristikám.

Pri štatistických testoch rozlišujeme dva typy modelov: parametrický a neparametrický. Parametrické modely predpokladajú známe rozdelenie náhodnej veličiny, pričom tento model môžeme popísať konečným počtom parametrov.

Témou tejto bakalárskej práce bude veľká skupina neparametrických testov, pri ktorých nepoznáme rozdelenie dát. Často sú jedinou dostupnou metódou pre špecifické typy údajov, napr. pre skúmanie poradí alebo početností dát. Slabšie predpoklady týchto metód spôsobujú, že neparametrické testy nie sú také silné ako ich parametrické náprotivky .

Cieľom práce je popísať základy teórie neparametrických testov pre jednovýberový a dvojvýberový problém.

V prvej kapitole si predstavíme práve dva základné jednovýberové neparametrické testy. Uvedieme ich predpoklady, skonštruujeme testové štatistiky a špecifikujeme kritické obory. Ako prvý si ukážeme znamienkový test, pri ktorom si bližšie vysvetlíme pojem konzistencie testu a demonštrujeme závislosť sily testu na testovanom parametri. Bude nasledovať jednovýberový Wilcoxonov test, kde popíšeme základné postupy na ošetrenie problému zhodných pozorovaní.

Druhá kapitola je zameraná na dvojvýberový problém, kde si opäť predvedieme dva neparametrické testy a ich vlastnosti. Prvým je Mann-Whitneyho test, kde pomocou rekurzívneho vzťahu nájdeme kritické hodnoty. Druhým testom je dvojvýberový Wilcoxonov test, ktorý sa zaraďuje do skupiny testov založených na poradovej štatistike. Taktiež si objasníme všeobecné charakteristiky tejto skupiny testov.

1. Jednovýberový problém

V tejto kapitole si priblížime jednovýberový problém, v ktorom dáta pozostávajú z jedného súboru pozorovaní. Predstavíme dva testy testujúce hodnotu mediánu náhodného výberu.

1.1 Znamienkový test

Predpoklady testu:
Vezmeme náhodný výber n nezávislých pozorovaní $X_1, ..., X_n$ zo spojitého rozdelenia s distribučnou funkciou F_X a neznámym mediánom m_X.
Hypotéza a alternatíva:
Nulovú hypotézu pre znamienkový test zapíšeme ako

$$H_0 : m_X = m_0,$$

kde m_0 je dopredu daná konštanta. Nulovú hypotézu môžeme vyjadriť aj dvoma ďalšími spôsobmi, pre ktoré je nutné definovať parameter θ ako

$$P(X < m_0) = \theta.$$

Potom môžeme písať

$$H_0 : P(X > m_0) = P(X < m_0),$$

$$H_0 : \theta = 1/2.$$

Pre túto nulovú hypotézu formulujeme tri rôzne alternatívy:

$$H_1 : m_X \neq m_0, \qquad H_1 : \theta \neq 1/2, \qquad H_1 : P(X > m_0) \neq P(X < m_0),$$

$$H_2 : m_X > m_0, \qquad H_2 : \theta < 1/2, \qquad H_2 : P(X > m_0) > P(X < m_0),$$

$$H_3 : m_X < m_0, \qquad H_3 : \theta > 1/2, \qquad H_3 : P(X > m_0) < P(X < m_0).$$

Testová štatistika:
Na zadefinovanie testovej štatistiky rozdelíme pozorovania na dve skupiny. Prvá skupina bude obsahovať pozorovania, ktoré ležia pod hodnotou m_0 (čiže rozdiel $X_i - m_0$ bude menší ako 0). Druhá skupina bude pozostávať z dát nad hodnotou m_0 (rozdiel $X_i - m_0 > 0$). Ako testovú štatistiku použijeme počet pozorovaní ležiacich nad teoretickou hodnotou m_0.

Definícia 1. *Testovú štatistiku znamienkového testu definujeme ako*

$$K_n = \sum_{i=1}^{n} R_i^+,$$

$$\text{kde } R_i^+ = \begin{cases} 1 & \text{ak } X_i - m_0 > 0, \\ 0 & \text{ak } X_i - m_0 < 0. \end{cases}$$

V prípade, že prijmeme nulovú hypotézu nám vznikne skupina n nezávislých náhodných veličín s alternatívnym rozdelením s parametrom $1/2$. Keď tieto pozorovania nasčítame do testovej štatistiky dostaneme $K_n \sim Bi(n, 1/2)$.
Treba si všimnúť, že pri konštrukcii testovej štatistiky neberieme do úvahy dáta, pre ktoré je rozdiel $X_i - m_0 = 0$. Takéto dáta sa vyskytujú a my pri nich budeme používať postup, že vynecháme všetky X_i také, ktorých rozdiel $X_i - m_0 = 0$ a o ich počet zmenšíme rozsah výberu.
(Existuje však niekoľko alternatívnych postupov. Prvý je taký, že polovicu nulových rozdielov budeme brať ako kladné rozdiely a polovicu ako záporné. Druhý prístup priraďuje nulovým rozdielom náhodne kladné alebo záporné znamienka. Tretí najkonzervatívnejší prístup je ten, že nulovým rozdielom budeme priraďovať také znamienka, ktoré najmenej ovplyvnia zamietnutie H_0.)
Kritický obor:
Kritický obor (značíme C) bude závisieť od druhu alternatívy.

1. Ako prvej sa budeme venovať alternatíve H_2.

   $$H_2 : m_X > m_0, \qquad H_2 : \theta < 1/2, \qquad H_2 : P(X > m_0) > P(X < m_0).$$

   $K_n \in C$ v prípade, že $K_n \geq k_\alpha$, kde k_α je najmenšie číslo splňujúce

   $$\sum_{j=k_\alpha}^{n} \binom{n}{j} 0.5^n \leq \alpha.$$

   Hladinu obyčajne nemôžeme dosiahnuť presne, napríklad ak chceme vyšetrovať súbor veľkosti 10 na hladine $\alpha = 0.05$, máme

   $$\sum_{j=8}^{10} \binom{10}{j} 0.5^{10} = 0.0546875 > \alpha,$$

   $$\sum_{j=9}^{10} \binom{10}{j} 0.5^{10} = 0.0107422 < \alpha.$$

   Je vidieť, že požadovaná hladina nebola ani zďaleka dosiahnutá presne, toto je spôsobené diskrétnosťou náhodného výberu. K_n bude v tomto prípade patriť do kritického oboru ak $K_n \geq 9$.

2. Podobne pre opačnú jednostrannú alternatívu

   $$H_3 : m_X > m_0, \qquad H_3 : \theta > 1/2, \qquad H_3 : P(X > m_0) < P(X < m_0).$$

   $K_n \in C$ ak $K_n \leq k'_\alpha$, kde k'_α je najväčšie číslo splňujúce

   $$\sum_{j=0}^{k'_\alpha} \binom{n}{j} 0.5^n \leq \alpha$$

   Opäť pre súbor veľkosti 10 na hladine $\alpha = 0.05$.

   $$\sum_{j=0}^{1} \binom{10}{j} 0.5^{10} = 0.0107422 < \alpha,$$

$$\sum_{j=0}^{2} \binom{10}{j} 0.5^{10} = 0.0546875 > \alpha.$$

Z rovníc vyplýva, že pre súbor veľkosti 10 na hladine 0.05 bude testová štatistika patriť do kritického oboru pre hodnoty $K_n \leq 1$, resp. budeme zamietať nulovú hypotézu na hladine $\alpha = 0.05$ ak bude testová štatistika nadobúdať hodnotu 0, alebo 1.

3. Nakoniec ak je alternatíva obojstranná,

$$H_1 : m_X \neq m_0, \qquad H_1 : \theta \neq 1/2, \qquad H_1 : P(X > m_0) \neq P(X < m_0).$$

Keďže binomické rozdelenie je symetrické pre $\theta = 0.5$, najväčšiu silu testu dosiahneme ak zvolíme $k'_{\alpha/2}$ a $k_{\alpha/2}$ symetricky.
$K_n \in C$, keď $K_n \geq k_{\alpha/2}$ alebo $K_n \leq k'_{\alpha/2}$, kde $k_{\alpha/2}$ a $k'_{\alpha/2}$ sú najmenšie a najväčšie číslo splňujúce

$$\sum_{j=k_{\alpha/2}}^{n} \binom{n}{j} 0.5^n \leq \alpha/2, \quad \sum_{j=0}^{k'_{\alpha/2}} \binom{n}{j} 0.5^n \leq \alpha/2.$$

Dostávame vzťah $k'_{\alpha/2} = n - k_{\alpha/2}$.
Napríklad pre vzorku veľkosti 10, na hladine $\alpha = 0.05$.

$$\sum_{j=9}^{10} \binom{10}{j} 0.5^{10} = 0.0107422 \leq \alpha/2 = 0.025,$$

$$\sum_{j=0}^{1} \binom{10}{j} 0.5^{10} = 0.0107422 \leq \alpha/2 = 0.025.$$

Obecne zamietame H_0 pre príliš malé hodnoty K_n(blízko 0) alebo veľké hodnoty, ktoré sa blížia n.
Teraz vyšetríme konzistenciu znamienkového testu pre vyššie uvedené kritické obory. Konzistencia sa ako vlastnosť testu využíva skôr v neparametrických testoch, ktoré majú veľmi obecné alternatívy. Kritérium konzistencie nám pomáha vyberať testy, ktorých podtrieda alternatívy je zaujímavá pre naše rozhodovanie. Nato, aby sme mohli vyšetriť konzistenciu znamienkového testu si sformulujeme vetu a definíciu podľa [5, str. 267].

Definícia 2. *Nech $X_1, ..., X_n$ je náhodný výber veľkosti n z množiny rozdelení s hustotami v tvare $f_X(.;\theta)$, $\theta \in \delta \subseteq R^k$. Definujeme dve disjunktné podmnožiny množiny δ ako ω_1 a ω_2. Definujeme T ako testovú štatistiku, pre hypotézu $\theta \in \omega_1$ oproti alternatíve $\theta \in \omega_2$. Potom test na hladine α s testovou štatistikou T je konzistentý pre $\Psi \subset \omega_2$, ak*

$$\lim_{n\to\infty} \beta(\theta) = 1,$$

pre $\theta \in \Psi$, kde β je sila testu.

Veta 3. *Nech $X_1, ..., X_n$ je náhodný výber veľkosti n z množiny rozdelení s hustotami v tvare $f_X(.;\theta)$, $\theta \in \delta \subseteq R^k$. Definujeme dve disjunktné podmnožiny množiny δ ako ω_1 a ω_2. Potom definujeme T ako testovú štatistiku, pre hypotézu $\theta \in \omega_1$ oproti alternatíve $\theta \in \omega_2$ a nech $g(\theta)$ je funkcia θ, taká, že*

$$\begin{aligned} g(\theta) &= \theta_0 \qquad ak \qquad \theta \in \omega_1, \\ g(\theta) &\neq \theta_0 \qquad ak \qquad \theta \in \Psi \text{ pre } \Psi \subset \omega_2. \end{aligned}$$

Nech pre všetky θ máme

$$E(T) = g(\theta), \qquad \lim_{n\to\infty} var(T) = 0.$$

$a_)$ *Potom test so štatistikou T na hladine α s kritickým oborom*

$$T \in C \qquad ak\ |T - \theta_0| > c_{\alpha,n},$$

kde $c_{\alpha,n}$ je kritická hodnota, je konzistentný pre podmnožinu Ψ.

$b_)$ *Podobne pre jednostrannú alternatívu. Ak je $g(\theta)$ funkciou θ takou, že spĺňuje*

$$\begin{aligned} g(\theta) &= \theta_0 \qquad ak \qquad \theta \in \omega_1, \\ g(\theta) &> \theta_0 \qquad ak \qquad \theta \in \Psi \text{ pre } \Psi \subset \omega_2. \end{aligned}$$

Nech pre všetky θ máme

$$E(T) = g(\theta), \qquad \lim_{n\to\infty} var(T) = 0.$$

Potom konzistentný test so štatistikou T na hladine α má kritický obor

$$T \in C \qquad ak\ T - \theta_0 > c'_{\alpha,n},$$

kde $c'_{\alpha,n}$ je kritická hodnota.

Dôkaz predchádzajúcej vety je podrobne spracovaný v knihe [5, str.267].
Overíme predpoklady predchádzajúcej vety pre Znamienkový test.
Nech $X_1, ..., X_n$ je náhodný výber veľkosti n z množiny spojitých rozdelení. Nech K_n testová štatistika z Definície 1 je testovou štatistikou pre hypotézu H_0 a alternatívu H_1.

$$\begin{aligned} \theta = \theta \qquad &\text{a} \qquad g(\theta) = \theta, \\ g(\theta) = \theta_0 = 1/2 \qquad &\text{ak} \qquad \theta \in H_0, \\ g(\theta) \neq \theta_0 = 1/2 \qquad &\text{ak} \qquad \theta \in H_1, \\ E(K_n/n) = \theta \qquad \lim_{n\to\infty} &var(K_n/n) = \theta(1-\theta)/n \to 0, \text{ pre } n \to \infty. \end{aligned}$$

Potom test so štatistikou K_n na hladine α s kritickým oborom v tvare

$$\begin{aligned} K_n \in C \qquad \text{ak} \qquad & |K_n/n - 1/2| > c_{\alpha,n} \\ & K_n > (c_{\alpha,n} + 1/2)n = k_{\alpha/2} \\ & K_n < -(c_{\alpha,n} - 1/2)n = k'_{\alpha/2} \end{aligned}$$

je konzistentný voči obojstrannej alternatíve.
Pre jednostranný test máme nulovú hypotézu H_0 a alternatívu H_3. Platí

$$g(\theta) = \theta_0 = 1/2 \qquad \text{ak} \qquad \theta \in H_0,$$

$$g(\theta) > \theta_0 = 1/2 \qquad \text{ak} \qquad \theta \in H_3,$$

$$E(K_n/n) = \theta \qquad \lim_{n\to\infty} var(K_n/n) = \theta(1-\theta)/n \to 0, \text{ pre } n \to \infty.$$

Potom test so štatistikou K_n na hladine α s kritickým oborom v tvare

$$K_n \in C \qquad \text{ak} \qquad K_n/n - 1/2 > c'_{\alpha,n}$$
$$K_n > (c'_{\alpha,n} + 1/2)n = k'_\alpha$$

je konzistentný voči jednostrannej alternatíve H_3.
Kritický obor(asymptotický test):
Pre malé hodnoty n sú v literatúre priložené tabuľky napr. pre znamienkový test [1, str. 368], na určenie kritických oborov. Keď chceme vyšetriť test s veľkým rozsahom výberu, je možné použiť postup z [1, str. 21] na určenie kritických oborov, pomocou aproximácie rozdelenia testovej štatistiky. Ak aproximujeme diskrétne rozdelenie spojitým, aproximáciu môžeme vylepšiť zahrnutím takzvanej "opravy pre spojitosť". Táto korekcia spočíva v tom, že testovú štatistiku rozdelenia uvažujeme ako stredný bod intervalu. Keď testová štatistika K_n nadobúda len hodnoty prirodzených čísiel predpokladáme, že pozorovaná hodnota testovej štatistiky bude $k \pm 0.5$, tzn. ak by hodnota testovej štatistiky bola 6 tak skutočná hodnota sa bude pohybovať v intervale $(5.5, 6.5)$.
Pokiaľ je naším rozhodovacím kritériom zamietnuť nulovú hypotézu pre $K_n \geq k_\alpha$ a jedná sa o aproximáciu $[K_n - E(K_n)]/\sqrt{var(K_n)}$ normovaným normálnym rozdelením, kritický obor so zohľadnením opravy o 0.5 dostaneme výpočtom rovnice

$$\frac{k_\alpha - 0.5 - E(K_n)}{\sqrt{var(K_n)}} = z_\alpha,$$

kde z_α splňuje $\Phi(z_\alpha) = 1-\alpha$ a Φ je distribučná funkcia normovaného normálneho rozdelenia.
Opačne ak použijeme $K_n \leq k'_\alpha$, tak k'_α musí splňovať rovnicu

$$\frac{k'_\alpha + 0.5 - E(K_n)}{\sqrt{var(K_n)}} = -z_\alpha,$$

kde z_α splňuje tie isté podmienky ako v predchádzajúcej rovnici.
V prípade znamienkového testu budeme aproximovať binomické rozdelenie. Z Centrálnej limitnej vety vyplýva, že binomické rozdelenie sa blíži normálnemu pre $n \to \infty$. Kedže sa v tomto prípade jedná o aproximáciu disktrétneho rozdelenia spojitým, zahrnieme "opravu pre spojitosť" o 0.5.
Konkrétne pre alternatívu $H_2 : m_X > m_0$, H_0 zamietame pre $K_n \geq k_\alpha$, kde k_α musí splňovať rovnicu

$$\frac{(k_\alpha - 0.5) - 0.5n}{0.5\sqrt{n}} = z_\alpha.$$

Sila znamienkového testu a funkcia sily:
Sila testu je charakterizovaná ako pravdepodobnosť, s akou padne testová štatistika do kritického oboru. Závisí

1. od stupňa nepravdivosti H_0 čo je veľkosť rozdielu medzi nulovou hypotézou H_0 a platnou alternatívou,

2. od hladiny α,

3. od veľkosti testovaného súboru,

4. od rozptylu pozorovaní $var(X_i)$.

O funkcii sily hovoríme vtedy, keď tri z týchto štyroch faktorov držíme konštantné, a meníme číslo 1.
Funkciu sily znamienkového testu je ľahké vypočítať, pretože testová štatistika K_n má binomické rozdelenie s parametrom θ. Sila testu je funkciou θ, definujeme si pomocný parameter $\kappa = 1-\theta$. Môžeme ju vypočítať napríklad pre $H_2 : \theta < 1/2$ ako

$$\beta(\kappa) = \sum_{j=k_\alpha}^{n} \binom{n}{j} \kappa^j (1-\kappa)^{n-j}$$

Ak máme situáciu s vopred daným rozdelením náhodnej veličiny, môžeme parameter θ presne vypočítať. Tento typ výpočtu funkcie sily testu je požadovaný v prípadoch, keď potrebujeme porovnať znamienkový test s nejakým parametrickým testom.
Nasledujúci príklad demonštruje závislosť sily znamienkového testu na veľkosti parametru κ.
Nech $H_0 : m_X = m_0$ a $H_2 : m_X > m_0$, potom na hladine $\alpha = 0.05$ a veľkosti $n = 10$ máme

$$\beta(1-0.4) = \sum_{j=9}^{10} \binom{10}{j} (0.6)^j (1-0.6)^{10-j} = 0.0463574,$$

$$\beta(1-0.1) = \sum_{j=9}^{10} \binom{10}{j} (0.9)^j (1-0.9)^{10-j} = 0.736099.$$

Je vidieť, že pre alternatívu H_2 so zvyšujúcim sa κ sila testu rýchlo rastie. Pre alternatívu $H_3 : m_X < m_0$ počítame pomocou rovnice

$$\beta(\kappa) = \sum_{j=0}^{k'_\alpha} \binom{n}{j} \kappa^j (1-\kappa)^{n-j}.$$

V tomto prípade sila testu rastie pre zmenšujúce sa κ.

1.2 Wilcoxonov test

Predpoklady testu:
Vezmeme náhodný výber n nezávislých pozorovaní $X_1, ..., X_n$ zo spojitého, symetrického rozdelenia s distribučnou funkciou F_X a neznámym mediánom m_X.
Hypotéza a alternatíva:
Nulovou hypotézou bude

$$H_0 : m_X = m_0,$$

kde m_0 je dopredu daná konštanta. H_0 môžeme zapísať aj inými spôsobmi, pre ktoré využijeme parameter θ definovaný v kapitole 1.1, ako

$$H_0 : P(X > m_0) = P(X < m_0),$$

$$H_0 : \theta = 1/2.$$

Obojstranná alternatíva bude

$$H_1 : m_x \neq m_0, \qquad H_1 : \theta \neq 1/2, \qquad H_1 : P(X > m_0) \neq P(X < m_0).$$

Zároveň formulujeme aj jednostranné alternatívy ako

$$H_2 : m_X > m_0, \qquad H_2 : \theta < 1/2, \qquad H_2 : P(X > m_0) > P(X < m_0),$$

$$H_3 : m_X < m_0, \qquad H_3 : \theta > 1/2, \qquad H_3 : P(X > m_0) < P(X < m_0).$$

Testová štatistika:
Testovú štatistiku skonštruujeme podľa [1, str. 107]:

1. Definujeme rozdiely $Z_i = X_i - m_0$. Za platnosti nulovej hypotézy, sú tieto rozdiely symetricky rozmiestnené okolo nuly, čiže

$$P(Z_i \leq -c) = P(Z_i \geq c).$$

2. $|Z_i|$ zoradíme od najmenšieho po najväčšie.

3. Dostaneme $0 < |Z_{(1)}| < |Z_{(2)}| < ... < |Z_{(n)}|$.

4. Každému $Z_{(i)}$ priradíme poradie R_i, pre ktoré platí $|Z_i| = |Z_{(R_i)}|$. Zároveň nezabúdame na pôvodné znamienka rozdielov Z_i.

5. Definujeme súčet poradí kladných rozdielov ako

$$T^+ = \sum_{i=1}^{n} R_i I_{(0,\infty)}(Z_i),$$

kde $I_{(0,\infty)}(Z_i) = \begin{cases} 1 & \text{pre } Z_i > 0, \\ 0 & \text{inak.} \end{cases}$

6. Podobne súčet poradí záporných rozdielov

$$T^- = \sum_{i=1}^{n} R_i I_{(-\infty,0)}(Z_i),$$

kde $I_{(-\infty,0)}(Z_i) = \begin{cases} 1 & \text{pre } Z_i < 0, \\ 0 & \text{inak.} \end{cases}$

7. Za platnosti nulovej hypotézy (tj. $m_X = m_0$) sa $E(T^+) = E(T^-)$. Suma všetkých poradí je konštantná, čiže

$$T^+ + T^- = \sum_{i=1}^{n} i = \frac{n(n+1)}{2}.$$

Keďže T^+, T^- a dokonca aj

$$T^+ - T^- = 2\sum_{i=1}^{n} R_i I_{(0,\infty)}(Z_i) - \frac{n(n+1)}{2}$$

sú lineárne závislé ako testovú štatistiku môžeme použiť ľubovoľné z nich.

Definícia 4. *Testovú štatistiku Wilcoxonovho testu definujeme ako*

$$T^+ = \sum_{i=1}^{n} R_i I_{(0,\infty)}(Z_i),$$

$$\textit{kde } I_{(0,\infty)}(Z_i) = \begin{cases} 1 & \textit{pre } Z_i > 0, \\ 0 & \textit{inak.} \end{cases}$$

Za platnosti nulovej hypotézy sú $I_{(0,\infty)}(Z_i)$ nezávislé, rovnako rozdelené náhodné veličiny s $Alt(1/2)$. Túto skutočnosť využijeme pri odvodení strednej hodnoty a rozptylu testovej štatistiky.

Tvrdenie 5. *Nech $X_1, ..., X_n$ sú náhodné veličiny spĺňujúce predpoklady modelu a nulovú hypotézu, potom*

$$\begin{aligned} E(T^+|H_0) &= \sum_{i=1}^{n} \frac{R_i}{2} = \frac{n(n+1)}{4}, \\ var(T^+|H_0) &= \sum_{i=1}^{n} \frac{R_i^2}{4} = \frac{n(n+1)(2n+1)}{24}. \end{aligned} \tag{1.1}$$

Dôkaz tohoto tvrdenia uvedieme neskôr. Pri konštrukcii testovej štatistiky sme opäť neuvažovali rozdiely $Z_i = 0$. Zachovávame sa ako v prvej kapitole, kde sme tieto rozdiely ignorovali a o ich počet znížili rozsah výberu. Avšak pri testovej štatistike Wilcoxonovho testu sa nám môže ľahko objaviť nový problém tzv. zhodných pozorovaní, kde sa $|Z_i| = |Z_j|$, pre $i \neq j$. Priblížime si tri vybrané metódy ošetrenia zhodných pozorovaní.

1. **Metóda stredných poradí**
   Metóda stredných poradí je asi najpoužívanejšou metódou na ošetrenie zhodných pozorovaní. Spočíva v tom, že každému Z_i z danej skupiny zhodných pozorovaní priradíme namiesto intervalu poradí priemerné poradie z tohoto intervalu (v súbore (1,2,2,2,2,6,8,9) nebude 2-kám priradený interval poradí (2,5), ale iba priemerné poradie 3,5). Použitím tejto metódy dostávajú pozorovania rovnaké poradia, čo ovplyvní rozdelenie pravdepodobnosti poradí. Je zrejmé, že hoci sa stredná hodnota nezmení, rozptyl rozdelenia poradí sa zmenší. Keď používame túto metódu musíme to zaznamenať v testovej štatistike.

2. **Metóda náhodných poradí**
   Keď máme n pozorovaní, pričom neplatí, že $Z_i \neq Z_j$, pre všetky $i \neq j$, vznikne nám m skupín čísiel s rovnakou hodnotou, kde sa i-ta rozdielna hodnota vyskytuje s početnosťou τ_i a platí $\sum_{i=1}^{m} \tau_i = n$. Každá skupina čísiel,

pre ktorú platí $\tau_i \geq 2$ je skupinou zhodných pozorovaní.
V metóde náhodných poradí je jedno z $\prod_{i=1}^{m} \tau_i!$ možných priradení poradí vybraté nejakým náhodným procesom.
Majme súbor pozorovaní $2, 3, 3, 4, 5, 5, 5, 6$, vidíme, že je $2!3! = 12$ možných priradení poradí tomuto súboru. Jedno z týchto poradí bude určené doplnkovým náhodným pokusom a slúži ako jedinečné priradenie poradia. Výhodou tejto metódy oproti metóde stredných poradí je, že zachováva vlastnosti testovej štatistiky tým, že každé poradie sa vyskytuje s rovnakou pravdepodobnosťou. Avšak, pridaním dodatočného prvku náhody ovplyvníme rozdelenie testovej štatistiky v prípade platnosti alternatív.

3. **Metóda priemernej testovej štatistiky**
Ak nechceme zvoliť špecifický súbor poradí ako v predchádzajúcich metódach, je možné vypočítať hodnotu testovej štatistiky pre všetky možné poradia $\prod_{i=1}^{m} \tau_i!$ a použiť ich priemer ako hodnotu finálnej testovej štatistiky.
V tejto metóde platí to, čo v metóde stredných poradí, tj. zmenší sa nám rozptyl testovej štatistiky a stredná hodnota zostane zachovaná.

Pre jednovýberový Wilcoxonov test použijeme metódu stredných poradí na ošetrenie zhodných pozorovaní. Je tiež zrejmé, že rozdelenie testovej štatistiky nemôže byť rovnaké ako pred použitím tejto metódy, avšak pokiaľ zhodných pozorovaní nie je priveľa, nemusíme používať žiadne úpravy testovej štatistiky.
Kritický obor:
Nasleduje dôkaz Tvrdenia 5.

$$T^+ = \sum_{1 \leq i \leq j \leq n} \sum T_{ij},$$
$$\text{kde } T_{ij} = \begin{cases} 1 & \text{ak } Z_i + Z_j > 0, \\ 0 & \text{inak.} \end{cases} \tag{1.2}$$

Tento zápis je ekvivaletný zápisu testovej štatistiky z Definície 4. Dôkaz ekvivalencie zápisov si predvedieme matematickou indukciou.

Dôkaz. Definíciu testovej štatistiky

$$T^+ = \sum_{i=1}^{n} R_i I_{(0,\infty)}(Z_i),$$

$$\text{kde } I_{(0,\infty)}(Z_i) = \begin{cases} 1 & \text{pre } Z_i > 0, \\ 0 & \text{inak,} \end{cases}$$

označíme ako S_L^n.
Alternatívny zápis

$$T^+ = \sum_{1 \leq i \leq j \leq n} \sum T_{ij},$$
$$\text{kde } T_{ij} = \begin{cases} 1 & \text{ak } Z_i + Z_j > 0, \\ 0 & \text{inak,} \end{cases}$$

označíme ako S_R^n.
Nech $n = 2$:
Máme súbor veľkosti 2, s prvkami X_1, X_2 a im príslušnými rozdielmi Z_1, Z_2.

$$\begin{aligned}
X_1 - m_0 > 0 \quad &\& \quad X_2 - m_0 > 0 \quad \Rightarrow \quad S_L = 3 = S_R, \\
X_1 - m_0 < 0 \quad &\& \quad X_2 - m_0 < 0 \quad \Rightarrow \quad S_L = 0 = S_R, \\
X_1 - m_0 > 0 \quad &\& \quad X_2 - m_0 < 0 \quad \& \quad Z_1 < Z_2 \quad \Rightarrow \quad S_L = S_R = 1, \\
X_1 - m_0 > 0 \quad &\& \quad X_2 - m_0 < 0 \quad \& \quad Z_1 > Z_2 \quad \Rightarrow \quad S_L = S_R = 2.
\end{aligned}$$

Nech $n \to k+1$:
Máme súbor veľkosti $k+1$, s prvkami $X_1, ..., X_k, X_{k+1}$ a im príslušnými rozdielmi $Z_1, ...Z_{k+1}$.
BÚNO: Budeme predpokladať, že $|Z_{k+1}| > |Z_1| > ... > |Z_k|$.
Ak je $Z_{k+1} < 0$ hodnoty S_L a S_R sa nezmenia. Vyplýva to z toho, že ak je táto hodnota záporná tak $I_{(0,\infty)}(Z_{k+1}) = 0$ a $T_{i,k+1} = 0$ pre $i = 1, ..., k+1$.
V opačnom prípade, keď $Z_{k+1} > 0$ sa suma poradí kladných rozdielov zvýši o $k+1$ čo je poradie Z_{k+1} zoradenom súbore absolútnych hodnôt. Máme

$$S_L^{k+1} = S_L^k + (k+1).$$

Keďže Z_{k+1} je v absolútnej hodnote väčšie ako $|Z_1|, ..., |Z_k|$ tak

$$\sum_{1 \le i \le k+1} T_{i,k+1} = k+1.$$

To znamená, že celková suma S_R^{k+1} sa zvýši o $(k+1)$.
Máme dokázané, že $S_L^k = S_R^k$, čiže $S_L^{k+1} = S_R^{k+1}$. □

Ekvivalenciu týchto zápisov si tiež ukážeme na nasledujúcom jednoduchom príklade.
Majme súbor pozorovaní $(1, 10, 14, 27)$. Teoretickú hodnotu m_0 predpokladáme $m_0 = 13$. Definujeme si pomocné značenie $a = 2m_0$

Hodnota	1	10	14	27
$X_i - m_0$	-12	-3	1	14
Poradie	3	2	1	4
Rozdiely	$1+1-a<0$	$10+10-a<0$	$14+14-a>0$	$27+27-a>0$
	$1+10-a<0$	$10+14-a<0$	$14+27-a>0$	
	$1+14-a<0$	$10+27-a>0$		
	$1+27-a>0$			

Hodnota testovej štatistiky z Definície 4 vychádza $T^+ = 1+4 = 5$. Počet rozdielov > 0 nám dáva hodnotu testovej štatistiky podľa (1.2.) ako $T^+ = 5$.

Pre všetky navzájom rôzne i, j, k definujeme pravdepodobnosti

$$\begin{aligned}
p_1 &= P(Z_i > 0), \\
p_2 &= P(Z_i + Z_j > 0), \\
p_3 &= P(Z_i > 0 \text{ a } Z_i + Z_j > 0), \\
p_4 &= P(Z_i + Z_j > 0 \text{ a } Z_i + Z_k > 0).
\end{aligned} \tag{1.3}$$

Momenty premenných pre všetky rôzne i, j, k, h sú potom

$$E(T_{ii}) = p_1, \qquad E(T_{ij}) = p_2,$$

$$var(T_{ii}) = p_1 - p_1^2, \qquad var(T_{ij}) = p_2 - p_2^2,$$

$$cov(T_{ii}, T_{ik}) = p_3 - p_1 p_2, \qquad cov(T_{ij}, T_{ik}) = p_4 - p_2^2, \qquad cov(T_{ij}, T_{hk}) = 0.$$

Stredná hodnota a rozptyl lineárnej kombinácie z (1.2) s danými momentmi sú

$$E(T^+) = nE(T_{ii}) + \frac{n(n-1)E(T_{ii})}{2} = np_1 + \frac{n(n-1)p_2}{2}, \tag{1.4}$$

$$\begin{aligned} var(T^+) &= nvar(T_{ii}) + \binom{n}{2} var(T_{ij}) + 2n(n-1)cov(T_{ii}, T_{ik}) \\ &\quad + 2n\binom{n-1}{2} cov(T_{ij}, T_{ik}) + \binom{n}{4} cov(T_{ij}, T_{hk}) \\ &= np_1(1-p_1) + n(n-1)(n-2)(p_4 - p_2^2) \\ &\quad + \frac{n(n-1)}{2}[p_2(1-p_2) + 4(p_3 - p_1 p_2)]. \end{aligned} \tag{1.5}$$

Pravdepodobnostiam z (1.3) pridáme predpoklad symetrie rozdelenia a pravdivosti nulovej hypotézy. Potom môžeme odvodiť

$$\begin{aligned} p_1 =& P(Z_i > 0) = 1/2, \\ p_2 =& P(Z_i + Z_j > 0) = \int_{-\infty}^{\infty} \int_{-v}^{\infty} f_Z(u) f_Z(v) du\, dv = 1/2, \\ p_3 =& P(Z_i > 0 \text{ a } Z_i + Z_j > 0) = \int_{0}^{\infty} \int_{-v}^{\infty} f_Z(u) f_Z(v) du\, dv \\ =& \int_{0}^{\infty} F_Z(v) f_Z(v) dv = \int_{1/2}^{1} x dx = 3/8, \\ p_4 =& P(Z_i + Z_j > 0 \text{ a } Z_i + Z_k > 0) \\ =& 2 \int_{-\infty}^{\infty} \int_{-w}^{\infty} \int_{v}^{\infty} f_Z(u) f_Z(v) f_Z(w) du\, dv\, dw \\ =& 2 \int_{-\infty}^{\infty} \int_{-w}^{\infty} f_Z(v) f_Z(w) dv\, dw - 2 \int_{\infty}^{\infty} \int_{-w}^{\infty} F_Z(v) f_Z(v) f_Z(w) dv\, dw \\ =& 2 \int_{-\infty}^{\infty} F_Z(w) f_Z(w) dw + \int_{\infty}^{\infty} [1 - F_Z(w)]^2 f_Z(w) dw \\ =& 2(1/2) - 1 + 1/3 = 1/3. \end{aligned}$$

Keď tieto výsledky dosadíme do rovníc (1.4.) dostaneme výsledky z Tvrdenia 5.

$$E(T^+) = np_1 + \frac{n(n-1)p_2}{2} = n\frac{1}{2} + \frac{n(n-1)1/2}{2} = \frac{n(n+1)}{4},$$

$$
\begin{aligned}
var(T^+) &= np_1(1-p_1) + n(n-1)(n-2)(p_4 - p_2^2) \\
&+ \frac{n(n-1)}{2}[p_2(1-p_2) + 4(p_3 - p_1p_2)] \\
&= n1/2(1-1/2) + n(n-1)(n-2)(1/3 - (1/2)^2) \\
&+ \frac{n(n-1)}{2}[1/2(1-1/2) + 4(3/8 - 1/2 * 1/2)] \\
&= \frac{n(n+1)(2n+1)}{24}.
\end{aligned}
$$

Tieto novo definované premenné využijeme pri vyšetrovaní konzistencie testu. Na vyšetrenie konzistencie testu použijeme Vetu 3 z kapitoly 1.1.

$$
E\left[\frac{2T^+}{n(n+1)}\right] = \frac{2p_1}{n+1} + \frac{(n-1)p_2}{n+1} = 1/2,
$$

ak prijmeme nulovú hypotézu a

$$
var\left[\frac{2T^+}{n(n+1)}\right] \to 0, \text{ pre } n \to \infty.
$$

Potom test s kritickým oborom

$$
T^+ \in C \qquad \text{pre } \frac{2T^+}{n(n+1)} - 1/2 \geq k,
$$

kde k je kritická hodnota, je konzistentný proti alternatíve H_2, ktorú môžeme zapísať v tvare $p_2 = P(Z_i + Z_j > 0) > 1/2$.
Podobne môžeme vyvodiť záver, že ak je T^+ centrovaná okolo $n(n+1)/4$, test je konzistentný voči alternatíve H_1 v tvare $p_2 \neq 1/2$ s obojstranným kritickým oborom. Presný postup na určenie tohoto kritického oboru preberieme z [1, str. 110].
Aby sme tento kritický obor mohli určiť, musíme stanoviť rozdelenie T^+ za platnosti H_0. Strednú hodnotu a rozptyl T^+ máme zo vzťahu (1.1). T^+ je pevne určené premennými $I_{(0,\infty)}(Z_i)$ v Definícii 4. Za priestor vzoriek môžeme považovať súbor všetkých možných n-tíc $(b_1, b_2, ..., b_n)$ so zložkami jedna alebo nula, ktorých je 2^n. Preto

$$
P(T^+ = t) = \frac{u(t)}{2^n}, \tag{1.6}
$$

kde $u(t)$ je počet možností, ktorými môžeme priradiť plusové alebo mínusové znamienko pre prvých n čísel tak, aby sa súčet kladných čísiel rovnal t. Ku každému priradeniu existuje priradenie so zamenenými plusovými a mínusovými znamienkami a T^+ pre toto priradenie je

$$
\sum_{i=1}^{n} i(1 - I_{(0,\infty)}(Z_i)) = \frac{n(n+1)}{2} - \sum_{i=1}^{n} iI_{(0,\infty)}(Z_i).
$$

Každé priradenie sa vyskytuje s rovnakou pravdepodobnosťou, čo implikuje, že nulové rozdelenie T^+ je symetrické okolo strednej hodnoty $n(n+1)/4$. Z toho vyplýva, že len jedna polovica rozdelenia musí byť určená.
Keďže T^+ a T^- sú lineárne závislé a rozdelenie je symetrické, vieme ukázať,

že T^+ a T^- sú rovnako rozdelené a stačí jeden súbor kritických hodnôt aj pre obojstrannú alternatívu. To, že sú T^+ a T^- rovnako rozdelené ukážeme ako

$$\begin{aligned} P(T^+ \geq c) &= P[T^+ - \frac{n(n+1)}{4} \geq c - \frac{n(n+1)}{4}] \\ &= P[T^+ - \frac{n(n+1)}{4} \leq \frac{n(n+1)}{4} - c] \\ &= P[T^+ - \frac{n(n+1)}{2} \leq -c] = P(-T^- \leq -c) = P(T^- \geq c). \end{aligned} \tag{1.7}$$

Tabuľky ľavých kritických hodnôt sú obecne stanovené pre náhodné premenné T, ktoré znamenajú buď T^+, alebo T^-. Ak číslo t_α spĺňa $P(T \leq t_\alpha) = \alpha$, kritické obory pre test $H_0 : m_X = m_0$ na hladine α sú

$$\begin{aligned} T^- &\leq t_\alpha \qquad \text{pre } H_2 : m_X > m_0, \\ T^+ &\leq t_\alpha \qquad \text{pre } H_3 : m_X < m_0, \\ T^+ \leq t_{\alpha/2} &\qquad \text{alebo } T^- \leq t_{\alpha/2} \qquad \text{pre } H_1 : m_X \neq m_0. \end{aligned}$$

Kritický obor (asymptotický test):
Tabuľky kritických hodnôt sú uvedené v literatúre spolu s príslušnými hladinami, na ktorých test skúmame. Väčšinou sú tieto tabuľky zostavené až po $n \leq 50$. Ak potrebujeme hodnoty pre $n > 50$, z Centrálnej limitnej vety vyplýva, že asymptotické rozdelenie $T^+ \sim N(0,1)$. Odvodenie tohoto tvrdenia nájdeme v [2, str. 5]. Použitím momentov zo vzťahu (1.1) dotaneme novú premennú

$$Y = \frac{4T^+ - n(n+1)}{\sqrt{2n(n+1)(2n+1)/3}} \to N(0,1), \text{ keď } n \to \infty.$$

Pre $H_2 : m_X > m_0$, zamietame H_0, keď $Y \geq z_\alpha$ ($\Phi(z_\alpha) = 1-\alpha$). Túto aproximáciu môžeme použiť pre hodnoty už od $n \geq 15$. Súčasne, keďže sa opäť jedná o aproximáciu disktrétneho rozdelenia spojitým, môžeme použiť metódu vysvetlenú v prvej kapitole tzv. "opravu pre spojitosť" o 0.5.

2. Dvojvýberový problém

Doteraz sme sa zaoberali výhradne jednovýberovým problémom, v nasledujúcej kapitole si vysvetlíme ako postupovať v prípade, že potrebujeme skúmať vzťah dvoch nezávislých výberov a vyvodiť záver o ich vzájomnom vzťahu resp. vzťahu ich distribučných funkcií či mediánov.

2.1 Mann-Whitneyho test

Predpoklady testu:
Nech $X_1, ...X_m$ a $Y_1, ..., Y_n$ sú nezávislé náhodné výbery z ľubovoľných spojitých rozdelení s distribučnými funkciami F_X a F_Y.
Hypotéza a alternatíva:
Nulová hypotéza bude

$$H_0 : P(X_i < Y_j) = P(X_i > Y_j) = 1/2.$$

Zadefinujeme parameter

$$p = P(Y < X) = \int_{-\infty}^{\infty} \int_{-\infty}^{x} f_Y(y) f_X(x) dy dx = \int_{-\infty}^{\infty} F_Y(x) f_X(x) dx.$$

Ak platí nulová hypotéza, potom

$$p = \int_{-\infty}^{\infty} F_X(x) f_X(x) dx = 1/2.$$

Po zadefinovaní parametru p môžeme nulovú hypotézu zapísať ako

$$H_0 : p = 1/2.$$

Alternatívy budú mať tvary

$$H_1 : p \neq 1/2,$$
$$H_2 : p < 1/2,$$
$$H_3 : p > 1/2.$$

Testová štatistika:
Na konštrukciu testovej štatistiky najprv musíme vytvoriť spojený výber z náhodných výberov $X_1, ..., X_m$ a $Y_1, ..., Y_n$. Veľkosť tohoto nového výberu bude $n+m = q$.

Definícia 6. *Testovú štatistiku Mann-Whitneyho testu definujeme ako*

$$W = \sum_{i=1}^{m} \sum_{j=1}^{n} I_{ij},$$

kde

$$I_{ij} = \begin{cases} 1 \ ak \ Y_j < X_i, \\ 0 \ ak \ Y_j > X_i, \end{cases}$$

pre všetky $i = 1, ...m$ *a* $j = 1, ..., n$.

Za platnosti nulovej hypotézy má I_{ij} Alternatívne rozdelenie s parametrom p. Z toho môžeme odvodiť strednú hodnotu a rozptyl ako

$$E(I_{ij}) = E(I_{ij}^2) = p, \qquad var(I_{ij}) = p(1-p).$$

Definujeme p_1 a p_2 ako

$$\begin{aligned} p_1 &= P(Y_j < X_i \cap Y_k < X_i) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{x_i}\int_{-\infty}^{y_k} f_Y(y_j)f_Y(y_k)f_X(x_i)dy_j dy_k dx_i \\ &+ \int_{-\infty}^{\infty}\int_{-\infty}^{x_i}\int_{-\infty}^{y_j} f_Y(y_k)f_Y(y_j)f_X(x_i)dy_k dy_j dx_i \\ &= \int_{-\infty}^{\infty}[F_Y(x)]^2 f_X(x)dx, \end{aligned}$$

$$p_2 = P(X_i > Y_j \cap X_h > Y_j) = \int_{-\infty}^{\infty}[1-F_X(y)]^2 f_Y(y)dy.$$

Teraz odvodíme kovariancie pomocou p_1 a p_2:

a)

pre $i \neq h$ a $j \neq k$ $\qquad cov(I_{ij}, I_{hk}) = E(I_{ij}I_{hk}) - E(I_{ij})E(I_{hk}) = p^2 - p^2 = 0,$

$E(I_{ij}I_{hk}) = E(I_{ij})E(I_{hk})$ vďaka nezávislosti I_{ij} a I_{hk}.

b) pre $j \neq k$

$$cov_{j\neq k}(I_{ij}, I_{ik}) = E(I_{ij}I_{hk}) - E(I_{ij})E(I_{hk}) = E(I_{ij}I_{hk}) - p^2 = p_1 - p^2,$$

c) pre $i \neq h$

$$cov_{i\neq h}(I_{ij}, I_{hj}) = E(I_{ij}I_{hj}) - E(I_{ij})E(I_{hj}) = E(I_{ij}I_{hj}) - p^2 = p_2 - p^2.$$

Testová štatistika W je definovaná ako lineárna kombinácia I_{ij}, takže môžeme odvodiť aj jej strednú hodnotu a rozptyl.

$$E(W) = \sum_{i=1}^{m}\sum_{j=1}^{n} E(I_{ij}) = mnp,$$

$$\begin{aligned} var(W) &= \sum_{i=1}^{m}\sum_{j=1}^{n} var(I_{ij}) + \sum_{i=1}^{m}\sum_{1\leq j\neq k\leq n}\sum cov(I_{ij}, I_{ik}) \\ &+ \sum_{j=1}^{n}\sum_{1\leq i\neq h\leq m}\sum cov(I_{ij}, I_{hj}) + \sum_{1\leq i\neq h\leq m}\sum\sum_{1\leq j\neq k\leq n}\sum cov(I_{ij}, I_{hk}) \\ &= mnp(1-p) + mn(n-1)(p_1 - p^2) + nm(m-1)(p_2 - p^2) \\ &= mn[p - p^2(q-1) + (n-1)p_1 + (m-1)p_2]. \end{aligned}$$

Za platnosti nulovej hypotézy bude

$$E(W|H_0) = \frac{mn}{2},$$

$$p_1 = 1/3 \qquad \text{a} \qquad p_2 = 1/3,$$

$$\begin{aligned} var(W|H_0) =& mnp(1-p) + mn(n-1)(p_1 - p^2) + nm(m-1)(p_2 - p^2) \\ =& \frac{mn(q+1)}{12}. \end{aligned}$$

Pri Mann-Whitneyho teste sa môže vyskytnúť rovnaký problém ako pri jednovýberovom Wilcoxonovom teste, a to problém zhodných pozorovaní. V definícii testovej štatistiky nezohľadňujeme zhodné pozorovania tj. I_{ij} nie je definované pre $X_i = Y_j$. Keď sa vo vzorke vyskytne priveľa zhodných pozorovaní, musíme túto skutočnosť zahrnúť do výpočtu. Najčastejší prístup spočíva v definícii novej testovej štatistiky ako

$$W_T = \sum_{i=1}^{m} \sum_{j=1}^{n} I_{ij},$$

kde

$$I_{ij} = \begin{cases} -1 \text{ ak } Y_j < X_i, \\ 0 \text{ ak } Y_j = X_i, \\ 1 \text{ ak } Y_j > X_i. \end{cases}$$

Na výpočet strednej hodnoty a rozptylu testovej štatistiky definujeme p^+ a p^- ako

$$p^+ = P(X > Y), \qquad p^- = P(X < Y).$$

Teraz môžeme vypočítať

$$E(W_T) = mn(p^+ - p^-).$$

W_T má asymptoticky normované normálne rozdelenie. Ak platí nulová hypotéza, tak $p^+ = p^-$ a platí $E(W_T|H_0) = 0$, aj v prípade, že sa zhodné pozorovania nevyskytnú. Ak sa však objavia, ovplyvní to rozptyl testovej štatistiky.

Kritický obor:

Kritický obor opäť súvisí s vyšetrením konzistencie testu.

$$E(W/mn) = p$$

$$var(W/mn) \to 0 \text{ pre } m, n \to \infty$$

Z uvedených rovníc vyplýva, že W/mn je konzistentným odhadom parametru p. Konzistenciu testu vyšetríme podľa Vety 3, tj. test je konzistetný pre nasledujúce alternatívy s kritickými obormi

Druh alternatívy	*Kritický obor*
$H_1 : p \neq 1/2$	$\left\lvert W - \frac{mn}{2} \right\rvert > k_1,$
$H_2 : p < 1/2$	$W - \frac{mn}{2} < k_2,$
$H_3 : p > 1/2$	$W - \frac{mn}{2} > k_3,$

kde k_1, k_2 a k_3 sú kritické hodnoty.
Na určenie presných kritických oborov testu použijeme rozdelenie pravdepodobnosti testovej štatistiky W. Pri spojenom súbore pozostávajúcom z m X-ov a n Y-ov máme $\binom{m+n}{m}$ možností usporiadania náhodných premenných. Keď platí nulová hypotéza každá z týchto možností sa vyskytuje s rovnakou pravdepodobnosťou, čiže

$$P(W = w) = \frac{s_{m,n}(w)}{\binom{m+n}{m}},$$

kde $s_{m,n}(w)$ je počet usporiadaní m X a n Y, takých, že v každom usporiadaní Y predbehne X presne w-krát. Za platnosti nulovej hypotézy je rozdelenie pravdepodobnosti symetrické okolo strednej hodnoty $mn/2$. Z toho vyplýva, že platí

$$P\left(W - \frac{mn}{2} = w\right) = P\left[W = mn - \left(\frac{mn}{2} + w\right)\right] = P\left(W - \frac{mn}{2} = -w\right).$$

Vďaka tejto symetrickej vlastnosti stačí nájsť len spodné chvosty kritických hodnôt, či už pre jedno- alebo obojstranný test. Definujeme náhodnú premennú W' ako

$$W' = \sum_{i=1}^{m}\sum_{j=1}^{n}(1 - I_{ij}),$$

ktorá symbolizuje udáva koľkokrát X predbehne Y v spojenom súbore.
Kritické obory pre hladinu α priamo súvisiace s konzistenciou testu sú

Alternatíva	*Kritický obor*
$H_1 : p \neq 1/2$	$W \leq c_{\alpha/2}$ alebo $W' \leq c_{\alpha/2}$,
$H_2 : p < 1/2$	$W \leq c_\alpha$,
$H_3 : p > 1/2$	$W' \leq c_\alpha$.

Hodnoty c_α môžeme určiť jednoduchým spôsobom pre hocijaké m a n. Postupujeme vymenovaním možností usporiadania spojeného súboru od $w = 0$ po $\alpha\binom{m+n}{m}$. Ukážeme jednoduchý príklad na výpočet kritických hodnôt pre $m = 3$ a $n = 4$.

$$\binom{m+n}{m} = 35$$

Časť tabuľky hodnôt so zoradenými X a Y, príslušnými hodnotami w a pravdepodobnosťami:

Poradie	w	Pravdepodobnosť
XXXYYYY	0	$P(W \leq 0) = 1/35 = 0.028$
XXYXYYY	1	$P(W \leq 1) = 2/35 = 0.057$
XYXXYYY	2	...

Z tabuľky vyplýva, že kritická hodnota na hladine 0.05 bude $c_\alpha = 0$.
Pre malé hodnoty w je jednoduché vypočítať kritickú hodnotu, avšak so zväčšujúcou sa veľkosťou súboru sa zvyšuje riziko prehliadnutia niektorých poradí a tým sa znižuje kvalita výpočtu. Preto pre veľké súbory môžeme použiť rekurzívny

vzťah, ktorý výpočet zjednoduší. Tento rekurzívny vzťah odvodíme jednoduchou úvahou z [1, str. 144].
Majme postupnosť X-ov a Y-ov veľkosti $m+n-1$, ktorá vznikla pridávaním zložiek na pravú stranu postupnosti. Keď chceme vytvoriť postupnosť veľkosti $m+n$ je nutné doplniť stávajúcu postupnosť o jeden prvok buď X alebo Y. Ak postupnosť $m+n-1$ pozostáva z m X-ov a $n-1$ Y-ov, potom nové písmeno musí byť nutne Y. Keď však pridáme Y na pravú stranu postupnosti počet koľkokrát Y predbehne X zostane nezmenený.
Opačne ak je pridávané písmeno X, v pôvodnej postupnosti bolo $m-1$ X-ov a n Y-ov, všetky Y predbehnú toto nové X. Keďže je ich tam n tak w narastie o n. Tieto dve možnosti sú vzájomne zámenné. Využijeme vzťah $P(W=w)=\frac{s_{m,n}(w)}{\binom{m+n}{m}}$ a rekurzívny vzťah môžeme zapísať ako

$$s_{m,n}(w) = s_{m,n-1}(w) + s_{m-1,n}(w-n),$$

$$\begin{aligned} P(W=w) = p_{m,n}(w) &= \frac{s_{m,n-1}(w) + s_{m-1,n}(w-n)}{\binom{m+n}{m}} \\ &= \frac{n}{m+n}\frac{s_{m,n-1}(w)}{\binom{m+n-1}{n-1}} + \frac{m}{m+n}\frac{s_{m-1,n}(w-n)}{\binom{m+n-1}{m-1}}, \end{aligned}$$

alebo

$$(m+n)p_{m,n}(w) = np_{m,n-1}(w) + mp_{m-1,n}(w-n).$$

Uvedená rekurzívna formula platí pre všetky $w = 0, ..., mn$ a pre všetky prirodzené čísla m, n, ak sú definované počiatočné a medzné podmienky pre všetky $i = 1, ..., m$ a $j = 1, ..., n$:

$$\begin{aligned} s_{i,j}(w) &= 0, && \text{pre všetky } w < 0, \\ s_{i,0}(0) &= 1, && s_{0,i}(0) = 1, \\ s_{i,0}(w) &= 0, && \text{pre všetky } w \neq 0, \\ s_{0,i}(w) &= 0, && \text{pre všetky } w \neq 0. \end{aligned}$$

Kritický obor(asymptotický test):
Pre malé hodnoty m, n sú v literatúre priložené hodnoty kritických hodnôt. Pri veľkých hodnotách môžeme použiť asymptotické rozdelenie pravdepodobnosti.

$$V = \frac{W - mn/2}{\sqrt{mn(q+1)/12}} \sim N(0,1), \text{keď } m, n \to \infty.$$

Aby sme mohli prísť k tomuto záveru, najprv sme museli odvodiť strednú hodnotu a rozptyl W, s podmienkou platnosti nulovej hypotézy.

$$E(W|H_0) = \frac{mn}{2}, \qquad var(E|H_0) = \frac{mn(q+1)}{12}.$$

Opäť aproximujeme diskrétne rozdelenie spojitým a preto je vhodné rátať s "opravou pre spojitosť" veľkosti 0.5.

2.2 Poradová štatistika a dvojvýberový problém

2.2.1 Poradová štatistika

V predchádzajúcej kapitole sme si predstavili Mann-Whitneyho test, v ktorom berieme dva náhodné výbery a zoradíme ich do spojeného súboru, ktorý potom využijeme pri konštrukcii testovej štatistiky. Dvojvýberové testy, ktorých výbery spojíme do jedného zoradeného súboru, a ktorých poradia v tomto súbore budeme využívať pri konštrukcii testovej štatistiky, patria do skupiny testov založených na poradovej štatistike. Priblížime si všeobecné charakteristiky platiace pre túto skupinu testov.

Všeobecné predpoklady:

Nech $X_1, ..., X_m$ a $Y_1, ..., Y_n$ sú dva nezávislé náhodné výbery zo spojitých rozdelení s distribučnými funkciami F_X, F_Y.

Nulová hypotéza má tvar

$$H_0 : F_Y(x) = F_X(x), \qquad \text{pre všetky x.}$$

Vytvoríme si spojený súbor $n + m = q$ pozorovaní s neznámou distribučnou funkciou, v ktorom môžeme jednotlivým pozorovaniam priradiť poradia $1, ..., q$. Poradie náhodnej premennej si môžeme definovať aj ako funkciu, v ktorej definícii nepredpokladáme zhodné pozorovania.

$$R_{XY}(x_i) = \sum_{j=1}^{m} I(x_i - x_j) + \sum_{j=1}^{n} I(x_i - y_j),$$

$$R_{XY}(y_i) = \sum_{j=1}^{m} I(y_i - x_j) + \sum_{j=1}^{n} I(y_i - y_j),$$

kde

$$I(v) = \begin{cases} 1 \text{ ak } v \geq 0, \\ 0 \text{ ak } v < 0. \end{cases}$$

Pre prácu s dátami je jednoduchšie tento spojený súbor zoradiť a priradiť mu vektor ukazateľov. Nech

$$Z = (Z_1, Z_2, ..., Z_q),$$

kde Z_i je indikátor, taký, že $Z_i = 1$, keď i-ta náhodná premenná v zoradenom kombinovanom súbore je X a $Z_i = 0$ ak je to Y.

Definícia 7. *Testovú štatistiku pre ľubovoľný test patriaci do skupiny testov založených na poradovej štatistike definujeme ako*

$$T_q(Z) = \sum_{i=1}^{q} a_i Z_i,$$

$Z = (Z_1, ...Z_q)$ a a_i sú dané čísla.

Keď máme zadefinovanú testovú štatistiku, môžeme odvodiť niekoľko vzorcov. Za platnosti nulovej hypotézy $H_0 : F_X(x) = F_Y(x)$ pre všetky x a všetky $i = 1...q$ platí

1.

$$E(Z_i) = \frac{m}{q}, \qquad var(Z_i) = \frac{mn}{q^2}, \qquad cov(Z_i, Z_j) = \frac{-mn}{q^2(q-1)}, \tag{2.1}$$

Dôkaz.

$$f_{Z_i}(z_i) = \begin{cases} m/q \text{ ak } z_i = 1, \\ n/q \text{ ak } z_i = 0 \qquad \text{pre } i = 1, ..., q, \\ 0 \text{ inak} \end{cases}$$

Z_i má alternatívne rozdelenie so strednou hodnotou a rozptylom

$$E(Z_i) = m/q \qquad var(Z_i) = mn/q^2.$$

Združené momenty budú mať tvar, pre $i \neq j$

$$E(Z_i Z_j) = P(Z_i = 1 \cap Z_j = 1) = \frac{\binom{m}{2}}{\binom{q}{2}} = \frac{m(m-1)}{q(q-1)}$$

Z toho vyplýva, že

$$cov(Z_i, Z_j) = \frac{m(m-1)}{q(q-1)} - (\frac{m}{q})^2 = \frac{-mn}{q^2(q-1)}.$$

□

2.

$$\begin{aligned} E(T_q) &= m \sum_{i=1}^{q} \frac{a_i}{q}, \\ var(T_q) &= \frac{mn}{q^2(q-1)} \left[q \sum_{i=1}^{q} a_i^2 - (\sum_{i=1}^{q} a_i)^2 \right]. \end{aligned} \tag{2.2}$$

Dôkaz.

$$E(T_q) = m \sum_{i=1}^{q} \frac{a_i}{q}$$

$$\begin{aligned} var(T_q) &= \sum_{i=1}^{q} a_i^2 var(Z_i) + \sum \sum_{i \neq j} a_i a_j cov(Z_i, Z_j) \\ &= \frac{mn \sum\limits_{i=1}^{q} a_i^2}{q^2} - \frac{mn \sum \sum_{i \neq j} a_i a_j}{q^2(q-1)} \\ &= \frac{mn}{q^2(q-1)} \left(q \sum_{i=1}^{q} a_i^2 - \sum_{i=1}^{q} a_i^2 - \sum \sum_{i \neq j} a_i a_j \right) \\ &= \frac{mn}{q^2(q-1)} \left(q \sum_{i=1}^{q} a_i^2 - (\sum_{i=1}^{q} a_i)^2 \right) \end{aligned}$$

□

2.2.2 Dvojvýberový Wilcoxonov test

Dvojvýberový Wilcoxonov test zaraďujeme medzi základných predstaviteľov testov založených na poradovej štatistike.
Predpoklady testu:
Nech $X_1, ..., X_m$ a $Y_1, ..., Y_n$ sú dva nezávislé náhodné výbery zo spojitých rozdelení s distribučnými funkciami F_X, F_Y.
(V dvojvýberovom Wilcoxonovom teste sa môžeme obmedziť na menší model, v ktorom budú F_X a F_Y spĺňovať $F_Y(x) = F_X(x - \theta)$, pre všetky x a $\theta = 0$ za platnosti nulovej hypotézy.)
Hypotéza a alternatíva:
Nulovú hypotézu môžeme formulovať ako

$$H_0 : P(X_i < Y_j) = P(X_i > Y_j) = 1/2.$$

Po vytvorení zoradeného spojeného súboru z $X_1, ..., X_m$ a $Y_1, ..., Y_n$, kde $q = m+n$, môžeme formulovať alternatívy. Keď bude suma poradí X v spojenom súbore veľmi veľká, zodpovedá to alternatíve

$$H_2 : P(X_i > Y_j) > 1/2.$$

V opačnom prípade, ak bude suma poradí X malá, máme

$$H_3 : P(X_i > Y_j) < 1/2.$$

Nakoniec, keď je suma poradí X buď veľká, alebo malá, platí alternatíva

$$H_1 : P(X_i > Y_j) \neq 1/2.$$

Testová štatistika:

Definícia 8. *Testovú štatistiku dvojvýberového Wilcoxonovho testu definujeme ako*

$$V_q = \sum_{i=1}^{q} iZ_i,$$

kde Z_i je indikátor, taký, že $Z_i = 1$, keď i-ta náhodná premenná v kombinovanom súbore je X a $Z_i = 0$ ak je to Y.

Strednú hodnotu a rozptyl testovej štatistiky odvodíme podľa rovnice (2.2)

$$E(V_q) = \frac{m(q+1)}{2}, \qquad var(V_q) = \frac{mn(q+1)}{12}.$$

Hodnota W_q má minimum v $\sum_{i=1}^{m} i = \frac{m(m+1)}{2}$ a maximum v $\sum_{i=q-m+1}^{q} i = \frac{m(2q-m+1)}{2}$.

Tvrdenie 9. *Nech $T_q(Z)$ je definovaná ako v Definícii 7. Potom rozdelenie testovej štatistiky $T_q(Z)$ je symetrické okolo strednej hodnoty $\mu = m\sum_{i=1}^{q} \frac{a_i}{q}$, keď a_i spĺňujú vzťah $a_i + a_{q-i+1} = c$, kde c je konštantné, pre $i = 1, ..., q$.*

V našom prípade sa $a_i = i$, pre $i = 1, ..., q$ a $a_i + a_{q-i+1} = q+1$ pre $i = 1, ..., q$ takže testová štatistika je symetrická okolo strednej hodnoty.
Rozdelenie testovej štatistiky môžeme odvodiť klasickým výpočtom alebo použitím nasledovného rekurzívneho vzťahu, v ktorom $s_{m,n}(k)$ predstavuje počet poradí X a Y, v ktorých sa súčet X-ových poradí rovná k.

$$s_{m,n} = s_{m-1,n}(k-q) + s_{m,n-1}(k)$$

a

$$f_{V_q}(k) = p_{m,n}(k) = \frac{s_{m-1,n}(k-q) + s_{m,n-1}(k)}{\binom{m+n}{m}},$$

alebo

$$(m+n)p_{m,n}(k) = mp_{m-1,n}(k-q) + np_{m,n-1}(k).$$

Ukážeme príklad na klasický výpočet pre $m = 2$ a $n = 3$. V tomto prípade máme $\binom{m+n}{n} = \binom{5}{3} = 10$ možností ako môže byť upriadaný vektor núl a jedničiek Z. Hodnota testovej štatistiky V_5 sa bude pohybovať medzi 3 a 9, symetricky okolo 6.

Hodnota	Poradia X	Frekvencia	Pravdepodobnosť
3	1,2	1	1/10
4	1,3	1	1/10
5	1,4;2,3	2	1/5
6	1,5;2,4	2	1/5
7	2,5;3,4	2	1/5
8	3,5	1	1/10
9	4,5	1	1/10

Takisto pre rekurzívny výpočet

$$s_{2,3}(6) = 2 = s_{1,3}(1) + s_{2,2}(6) = 1 + 1,$$

z toho vyplýva, že

$$f_{V_5}(6) = \frac{2}{\binom{5}{3}} = 1/5.$$

Pri tomto neparametrickom teste sa zase objavuje problém zhodných pozorovaní. Na jeho vyriešenie použijeme metódu stredných poradí vysvetlenú v kapitole 1.2.

Dvojvýberový Wilcoxonov test je v podstate Mann-Whitneyho testom z predchádzajúcej kapitoly, vďaka lineárnej závislosti medzi testovými štatistikami. Z tohoto dôvodu má Wilcoxonov test všetky vlastnosti Mann-Whitneyho testu vrátane konzistencie.

Záver

Táto bakalárska práca bola zameraná na základnú teóriu neparametrických testov. V prvej kapitole som rozviedla teóriu jednovýberového problému na znamienkovom a Wilcoxonovom teste. Zamerala som sa na problém zhodných pozorovaní a vyšetrovanie konzistencie testu, čo som potom aplikovala na dané testy. Na znamienkovom teste som tiež predviedla silu testu, jej závislosť na testovanom parametri a tiež postup nazývaný ako "oprava pre spojitosť" pri aproximácii testovej štatistiky pre veľké súbory dát.

Druhá kapitola predstavila dvojvýberové testy, špeciálne Mann-Whitneyho a dvojvýberový Wilcoxonov test. Pri Wilcoxonovom teste som ukázala základné charakteristiky platiace pre testy založené na poradovej štatistike a koniec kapitoly bol venovaný konštatovaniu vzťahu medzi Wilcoxonovým a Mann-Whitneyho testom. Pri Mann-Whitneyho teste som navyše predviedla postup v prípade výskytu zhodných pozorovaní a rekurzívny vzťah na odvodenie kritických hodnôt pre súbory väčšieho rozsahu.

Téma tejto bakalárskej práce bola veľmi zaujímavá, pretože pomáha pochopiť teóriu neparametrických testov od konštrukcie testových štatistík cez kritické obory až po aproximáciu testových štatistík. Základné tvrdenia dopĺňa dôkazmi a jednoduchými odvodeniami, pričom nechýbajú jednoduché príklady na pochopenie problémov.

Zoznam použitej literatúry

[1] Gibbons, J. D. *Nonparametric statistical inference*. Marcel Dekker, INC., 1985. ISBN 0-8247-7327-6.

[2] Lehmann, E. L. *Nonparametrics-Statistical methods based on ranks*. Holden-Day, 1975. ISBN 0-8162-4996-6.

[3] Hušková, M., Dupač, V. *Pravděpodobnost a matematická statistika*. Karolinum, 2009. ISBN 978-80-246-0009-3.

[4] Anděl, J. *Statistické metody*. matfyzpress, 2007. ISBN 80-7378-003-8.

[5] Fraser, D.A.S. *Nonparametric methods in Statistics*. John Wiley and Sons,Inc. , 1957.