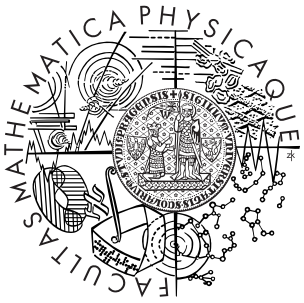
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Zdeněk Veselý

Jednovýběrový Wilcoxonův test

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: Mgr. Tomáš Jurczyk
Studijní program: Matematika, Obecná matematika

2010

Rád bych poděkoval vedoucímu práce za jeho trpělivou pomoc a kolegům Martinu Formánkovi, Michaele Tiché a Julii Klačanské za technické rady týkající se LaTexu a programu R.

Prohlašuji, že jsem svou bakalářskou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

V Praze dne 4. srpna 2010 Zdeněk Veselý

Obsah

Úvod **5**

1 Konstrukce Wilcoxonova testu **7**
1.1 Motivace k Wilcoxonovu testu 7
1.2 Základní pojmy testování hypotéz a jiné používané pojmy . 8
1.3 Pořadové testy hypotézy o středu symetrie 10
1.4 Vlastnosti statistiky W^+ 13
1.5 Konstrukce kritických hodnot 19
1.6 Jiné alternativy Wilcoxonova testu 29
1.7 Shody a nuly 30

2 Vlastnosti Wilcoxonova testu **34**
2.1 T-test a jeho srovnání s Wilcoxonovým testem 34
2.2 Síla Wilcoxonova testu 37
2.3 Použitá rozdělení 40

Závěr **44**

Literatura **45**

Název práce: Jednovýběrový Wilcoxonův test
Autor: Zdeněk Veselý
Katedra (ústav): Katedra pravděpodobnosti a matematické statistiky
Vedoucí bakalářské práce: Mgr. Tomáš Jurczyk
e-mail vedoucího: jurczyk@karlin.mff.cuni.cz

Abstrakt: Práce se zabývá jednovýběrovým Wilcoxonovým testem a jeho variantou pro párová data. Vychází ze základních pojmů testování hypotéz a obsahuje zavedení Wilcoxonova testu jako konkrétního příkladu pořadového testu. Je zde popsán postup provedení testu a vlastnosti testové statistiky jak přesné, tak asymptotické, aby bylo možno použít test i bez znalosti přesného rozdělení testové statistiky. Krátce se věnuje i shodám a nulám v reálných pozorováních. Práce ukazuje vlastnosti testu na simulovaných datech, snaží se porovnat účinnost testování pro několik vybraných rozdělení a také nabízí srovnání Wilcoxonova testu s jeho parametrickou obdobou, t-testem.

Klíčová slova: Wilcoxonův test, pořadové testy, t-test, Frank Wilcoxon

Title: One-sample Wilcoxon test
Author: Zdeněk Veselý
Department: Department of Probability and Mathematical Statistics
Supervisor: Mgr. Tomáš Jurczyk
Supervisor's e-mail address: jurczyk@karlin.mff.cuni.cz

Abstract: In the present thesis we study one-sample and paired Wilcoxon signed rank test. This thesis is based on the fundamental concept of statistical hypotheses testing. The Wilcoxon signed rank test is presented as an example of rank test. There can be found the description of the testing procedure as well as characteristics of the test statistic. The asymptotic ones are essential for realizing test without knowledge of the exact distribution of the test statistic. The thesis briefly deals with ties and nulls which are usually present in real data. There are demonstrations of properties of the test on simulations, comparison between the efficacies of the test under a few selected distributions and confrontation between signed rank test and the t-test which is a parametric version of signed rank test.

Keywords: Wilcoxon signed-rank test, rank tests, t-test, Frank Wilcoxon

Úvod

Práce bude pojednávat o jednovýběrovém Wilcoxonově testu a jeho obměně pro párová pozorování. Ten patří mezi pořadové testy, proto zde budou nastíněny také základy testování hypotéz a pořadových testů. Celá práce se bude skládat ze dvou částí. V první bude ukázána potřeba Wilcoxonova testu, celá konstrukce tohoto testu jakožto pořadového testu a nezbytnost jeho předpokladů, jimiž jsou spojitost a symetrie rozdělení. Pozornost bude věnována také rozdělení statistiky, na níž je Wilcoxonův test založen, a to jak přesnému, tak i asymptotickému či simulovanému. Práce se snaží představit tento test tak, aby jej čtenář mohl ihned používat a to i bez tabulek přesných kvantilů rozdělení testové statistiky. Zmíněny budou i problémy, které mohou nastat v praxi s reálnými daty.

Simulace pak budou tvořit hlavní náplň druhé části. Ta se bude věnovat vlastnostem Wilcoxonova testu, pokusí se jej porovnat ho s t-testem a nabídnout jako náhradu t-testu v případě nedodržení předpokladů. Jedna z podkapitol se bude věnovat také srovnání efektivity Wilcoxonova testu pro různá rozdělení. Zde čeká otázka, jak vůbec mezi sebou porovnat testování dvou různých rozdělení.

Ideálním čtenářem je člověk, který zná základy pravděpodobnosti a statistiky a používá t-test. Tato práce mu nabízí rozšíření obzorů. Znalost t-testu není podmínkou, v práci bude popsán. Pro ve statistice zběhlého čtenáře může být přínosem obzvláště druhá část, ve které si může ověřit, jak dobře funguje Wilcoxonův test v různých případech. U čtenáře se předpokládá jen základní znalosti oboru, pojmy jako pravděpodobnostní prostor zde definovány nebudou, v práci používaná pravděpodobnostní rozdělení budou pro přehled uvedeny v podkapitole 2.3.

Za pilíř teoretické části práce poslouží kniha profesorky Jurečkové [2], ze které se zde objeví jen stručný výtah vedoucí přímo k Wilcoxonovu testu. Bude se zavádět jen nutné množství nových pojmů, aby konstrukce tohoto testu byla co nejpřehlednější, ale zároveň úplná. Simulace budou prováděny

v programu R, zde se objeví jen výstupy programu, kód bude k nahlédnutí na přiloženém CD.

Kapitola 1

Konstrukce Wilcoxonova testu

1.1 Motivace k Wilcoxonovu testu

Roku 1945 Frank Wilcoxon publikoval článek [1], v němž navrhl mimo jiné i test, jenž dnes nese jeho jméno - Jednovýběrový Wilcoxonův test, který testuje hypotézu o poloze mediánu symetrického spojitého rozdělení. Pokud $Z_1, \ldots, Z_N$ je výběr z rozdělení s hustotou symetrickou kolem $\tilde{x}$ a my bychom rádi věděli, zda lze zamítnout hypotézu, že $\tilde{x} = 0$, pak můžeme test založit na statistice

$$W^+ = \sum_{Z_i \geq 0} R_i^+,$$

kde R_i^+ je pořadí $|Z_i|$ mezi $|Z_1|, \ldots, |Z_N|$ a N délka výběru. Vlastnosti této statistiky a její použití pro testování si odvodíme v dalších kapitolách.

Přínos Wilcoxonova testu spočívá v absenci podmínky normality testovaného rozdělení. Uvědomíme-li si, že pokud v rozdělení se symetrickou hustotou existuje střední hodnota, pak už se nutně musí střed symetrie rovnat této střední hodnotě. Připomeňme, že $\tilde{x} \in \mathbb{R}$ je medián náhodné veličiny X právě tehdy, když $P(X \leq \tilde{x}) \geq \frac{1}{2}$ a zároveň $P(X \geq \tilde{x}) \geq \frac{1}{2}$. Střed symetrie je tedy zároveň mediánem.

Lemma 1. *Nechť X je náhodná veličina s konečnou střední hodnotou a hustotou f. Nechť existuje $\tilde{x} \in \mathbb{R}$ takové, že $f(\tilde{x} - x) = f(\tilde{x} + x)$ platí skoro všude na $\mathbb{R}$. Pak $EX = \tilde{x}$.*

Důkaz. S použitím substitucí $y = x - \tilde{x}$ a následně $z = y + \tilde{x}$

$$EX = \int_{\mathbb{R}} x f(x) dx = \int_{\mathbb{R}} (y + \tilde{x}) f(y + \tilde{x}) dy =$$

$$= \int_{\mathbb{R}} yf(y+\tilde{x})dy + \int_{\mathbb{R}} \tilde{x}f(y+\tilde{x})dy =$$

$$= 0 + \tilde{x}\int_{\mathbb{R}} f(y+\tilde{x})dy = \tilde{x}\int_{\mathbb{R}} f(z)dz = \tilde{x},$$

neboť $yf(y+\tilde{x})$ je lichá funkce v proměnné y a integrál $\int_{\mathbb{R}} yf(y+\tilde{x})dy$ vždy existuje, protože předpokládáme konečnost EX, musí tedy platit $\int_{\mathbb{R}} yf(y+\tilde{x})dy = 0$. □

Podle Lemmatu 1 tedy lze Wilcoxonův test použít i jako test o poloze střední hodnoty symetrického rozdělení. Jedním z nejpoužívanějších spojitých symetrických rozdělení je normální rozdělení. U něj testujeme hypotézu o poloze střední hodnoty (zároveň tedy i mediánu) pomocí jednovýběrového t-testu.

Kdyby Z_i pocházel z normálního rozdělení, jistě bychom sáhli po t-testu, neboť testová statistika t-testu využívá více informací z náhodného výběru než statistika W^+. Avšak pro jiná rozdělení neznáme přesné rozdělení statistiky t-testu, zatímco rozdělení statistiky W^+ se nemění. Více se t-testu budeme věnovat v kapitole 2.1. Wilcoxonův test nás tedy bude zajímat zejména ve spojení s rozděleními, které normální nejsou, které nemusejí mít střední hodnotu, popřípadě je ani nemusíme blíže znát.

1.2 Základní pojmy testování hypotéz a jiné používané pojmy

V této podkapitole si ukážeme všeobecné základy statistických testů.

Mějme náhodný vektor $\mathbf{X} = (X_1, \ldots, X_n)^T$ s rozdělením, které závisí na *parametru* $\theta = (\theta_1, \ldots, \theta_k)^T$, který leží v nějaké množině Θ, která má nejméně dva prvky. Pod parametrem si můžeme představit různé informace o rozdělení, například střední hodnotu, rozptyl, ale také třeba hustotu. Často značíme $P_\theta(B)$ pravděpodobnost, že $\mathbf{X} \in B$ za podmínky, že θ je skutečná hodnota parametru rozdělení vektoru $\mathbf{X}$.

Na počátku máme domněnku, že parametr θ leží v neprázdné množině $\Theta_0 \subset \Theta$, kterou nazveme *nulovou hypotézou* a označíme H_0. Zjednodušeně napíšeme $H_0 : \theta \in \Theta_0$. Opačná možnost, tedy že $\theta \notin \Theta_0$, se nazývá *alternativní hypotéza*, značí se H_1, zkráceně opět můžeme psát $H_1 : \theta \in \Theta \backslash \Theta_0$. O tom, zda nulovou hypotézu zamítneme nebo ne, rozhoduje konkrétní pozorování vektoru $\mathbf{X}$. Obvykle zvolíme množinu $K \subseteq \mathbb{R}^n$, kterou nazveme

kritický obor. Nulovou hypotézu zamítneme, právě když $\mathbf{X} \in K$. Pokud hypotézu H_0 zamítneme, ač je správná, nastane *chyba prvního druhu*. Také se může stát, že hypotézu H_0 nezamítáme, ač není správná, v tom případě se dopouštíme *chyby druhého druhu*.

Nejčastěji jsou testy konstruovány tak, že si určíme (respektive dostaneme zadánu) pravděpodobnost chyby prvního druhu. Nazýváme ji *hladina testu*, značí se α a v praxi je její nejčastější hodnota 0,05. Pak se snažíme minimalizovat pravděpodobnost chyby druhého druhu vhodnou volbou kritického oboru tak, aby zároveň test nepřesáhl zvolenou hladinu α. Doplněk pravděpodobnosti chyby druhého druhu, tedy pravděpodobnost, že hypotézu H_0 zamítneme za podmínky, že platí H_1 (tj. za podmínky, že ji zamítnout máme), nazýváme *silou testu* a značíme $\beta = P(\mathbf{X} \in K | H_1)$. Tím, že minimalizujeme pravděpodobnost chyby druhého druhu, maximalizujeme sílu testu, proto čím větší síla testu, tím lépe. Můžeme si situaci představit jako pokus někoho odsoudit. Platí presumpce neviny (nulová hypotéza), vina se musí dokázat tak, aby nebylo pochyb o tom, že odsouzený je opravdu vinný (tedy zamítnutí nulové hypotézy s velmi malou možností chyby). Důležitější a závažnější pro nás je chyba prvního druhu, v naší analogii je to odsouzení nevinného. Uvažujeme tak, že chyba prvního druhu je jen velmi málo pravděpodobná, proto když zamítáme H_0, tak jsme si poměrně jistí tím, že $\theta \notin \Theta_0$, tedy tím, že obžalovaný je vinný. Na druhou stranu, pokud nedojde k zamítnutí hypotézy H_0, neřekneme, že hypotéza H_0 platí, jen to, že nejde zamítnout. Stejně tak když obžalovanému není dokázána vina, ještě ho nemůžeme považovat za úplně nevinného.

Ve výše uvedeném textu jsme použili pojem statistika, proto si nyní připomeňme, že statistikou se myslí měřitelná funkce vektoru $\mathbf{X}$.

Definice 2. *Řekneme, že statistika S je* sufficientní *pro parametr θ, pokud podmíněné rozdělení vektoru $\mathbf{X}$ při pevném S nezávisí na parametru θ.*

Sufficientní statistiky se někdy také nazývají postačující. Sufficientní statistiku používáme proto, abychom si zjednodušili vektor $\mathbf{X}$ a přitom neztratili žádnou informaci o parametru θ.

1.3 Pořadové testy hypotézy o středu symetrie

V této podkapitole si odvodíme, jak se obecně konstruují testy založené na pořadí absolutních hodnot pozorování z jednorozměrného symetrického rozdělení. Na začátku si předvedeme, jak lze test (párových) dat z dvourozměrného rozdělení převést na jednorozměrný případ.

Mějme $\{(X_i, Y_i)^T, i = 1, 2, \ldots, N\}$ náhodný výběr z dvourozměrného rozdělení se spojitou distribuční funkcí $F(x, y)$, kde $x \in \mathbb{R}$, $y \in \mathbb{R}$. Na F neklademe, kromě spojitosti, žádné nároky. Chceme zjišťovat, zda $F(x, y)$ je symetrická podle osy $x = y$ (tedy zda platí $F(x, y) = F(y, x)$), nebo zda rozdělení náhodného vektoru $(X, Y)^T$ je posunuto směrem k polorovině $y > x$, tedy zda existuje $\delta > 0$, že rozdělení je symetrické podle přímky $y = x + \delta$ ($F(x, y + \delta) = F(y, x + \delta)$). Pro ilustraci si můžeme představit za vektorem $(X_i, Y_i)^T$ hodnoty krevního tlaku i-tého pacienta před podáním a po podání experimentálního léku, o kterém se chceme dozvědět, zda má vliv na zvýšení krevního tlaku. V případě, že $F(x, y)$ je distribuční funkce dvourozměrného normálního rozdělení, se využije párový t-test. My ale normalitu F nepředpokládáme, tedy budeme muset použít obecnější testy.

Lemma 3. *Nechť náhodný vektor $(X, Y)^T$ má spojité rozdělení s distribuční funkcí F, a platí $F(x, y + \delta) = F(y, x + \delta)$. Definujme náhodnou veličinu $Z = Y - X$. Pak Z má symetrické rozdělení kolem bodu δ.*

Důkaz.

$$P(Z < z + \delta) = P(Y - X < z + \delta) =$$

$$= \int_{y-x<z+\delta} dF(x, y) = \int_{y-x<z+\delta} dF(y - \delta, x + \delta) =$$

poslední rovnost plyne ze $F(x, y+\delta) = F(y, x+\delta)$, dále použijeme substituce $u = y - \delta$ a $v = x + \delta$

$$= \int_{u-v<z-\delta} dF(u, v) = P(X - Y < z - \delta) =$$

$$= P(Y - X > -z + \delta) = P(Z > -z + \delta)$$

□

Na začátek provedeme transformaci $Z_i = Y_i - X_i$. Potom je $Z_1, \ldots, Z_N$ výběr z jednorozměrného rozdělení se spojitou distribuční funkcí $G(z)$. Označme hustotu tohoto rozdělení $g(z)$. Problém testování symetrie funkce $F(x, y)$ proti jejímu jednostrannému posunutí je podle Lemmatu 3 ekvivalentní testu nulové hypotézy $H_0 : g(z) = g(-z)$, že rozdělení G je symetrické kolem nuly, proti alternativě $H_1 : g(z + \delta) = g(-z + \delta)$ pro skoro všechny $z \in \mathbb{R}$, kde $\delta > 0$ je střed symetrie rozdělení G, v našem případě očekávaný nárůst krevního tlaku.

V případě, že vycházíme přímo z jednorozměrného rozdělení, musíme kromě spojitosti rozdělení předpokládat ještě jeho symetrii kolem bodu δ. Tím rovnou dostaneme výše popsaný případ.

Nechť $z_1, \ldots, z_N$ jsou hodnoty z výběru $Z_1, \ldots, Z_N$. Najdeme celá čísla $m, n \geq 0, m+n = N$ a $(i_1, \ldots, i_m, j_1, \ldots, j_n)^T$ permutaci $(1, \ldots, N)^T$ takové, že $z_{i_1}, \ldots, z_{i_m} < 0 < z_{j_1}, \ldots, z_{j_n}$, dále označíme $(S_1, \ldots, S_m)^T$ *vektor pořadí* $(|z_{i_1}|, \ldots, |z_{i_m}|)^T$ mezi $(|z_1|, \ldots, |z_N|)^T$, kde

$$S_j = \sum_{k=1}^{N} u(z_k), \text{ kde } u(z_k) = \begin{cases} 1, & |z_{i_j}| \geq |z_k| \\ 0, & |z_{i_j}| < |z_k| \end{cases} .$$

Obdobně označíme $(R_1, \ldots, R_n)^T$ vektor pořadí $(|z_{j_1}|, \ldots, |z_{j_n}|)^T$ mezi $(|z_1|, \ldots, |z_N|)^T$. U *pořadových testů* založených na vektoru $(|z_1|, \ldots, |z_N|)^T$ dále pracujeme výhradně s vektory pořadí $(S_1, \ldots, S_m)^T, (R_1, \ldots, R_n)^T$. U vektoru $(S_1, \ldots, S_m)^T$ navíc nezáleží na pořadí, neboli vektor $(S_1, \ldots, S_m)^T$ nabývá všech permutací jednoho vektoru se stejnou pravděpodobností (tedy

$$P((S_1, \ldots, S_m)^T = (k_1, \ldots, k_m)^T) = P((S_1, \ldots, S_m)^T = (k_{c_1}, \ldots, k_{c_m})^T)$$

pro všechny $(c_1, \ldots, c_m)^T$ permutace vektoru $(1, 2, \ldots, m)^T$). Tento fakt plyne přímo z nezávislosti náhodných veličin $Z_1, \ldots, Z_N$, platí i při nesymetrické hustotě rozdělení.

Vektor $(R_1, \ldots, R_n)^T$ má stejné vlastnosti jako $(S_1, \ldots, S_m)^T$, navíc jeden vektor jednoznačně určuje druhý (až na permutaci), neboť $(S_1, \ldots, S_m, R_1, \ldots, R_n)^T$ nabývá hodnot permutací vektoru $(1, 2, \ldots, N)^T$. Označme vektor uspořádaných pořadí absolutních hodnot záporných pozorování

$$(S'_1, \ldots, S'_m)^T : S'_1 < \ldots < S'_m,$$

kde $(S'_1, \ldots, S'_m)^T$ je permutací $(S_1, \ldots, S_m)^T$. Důsledkem výše uvedených skutečností je patrné, že $(S'_1, \ldots, S'_m)^T$ je suficientní statistikou pro vektor $(S_1, \ldots, S_m, R_1, \ldots, R_n)^T$, neboť $(S'_1, \ldots, S'_m)^T$ určuje tento vektor přesně až

na permutaci prvních m prvků a permutaci posledních n prvků, které nemění pravděpodobnosti. Symetricky to platí i pro vektor uspořádaných pořadí absolutních hodnot kladných pozorování $(R'_1, \ldots, R'_n)^T$. Nakonec jsme pořadové testy založené na vektoru $(|z_1|, \ldots, |z_N|)^T$ zredukovali pouze na testy závislé na vektoru $(R'_1, \ldots, R'_n)^T$.

Lemma 4. *Za platnosti hypotézy H_0 platí:*

$$P(R'_1 = r_1, \ldots, R'_n = r_n) = \frac{1}{\binom{N}{n}}$$

pro všechny n-tice $r_1, \ldots, r_n$ takové, že $1 \leq r_1 < \ldots < r_n \leq N$.

Důkaz. Nejprve si uvědomme, že počet všech takových n-tic $r_1, \ldots, r_n$ odpovídá počtu kombinací n-té třídy z N prvků, tedy $\binom{N}{n}$. Nyní musíme dokázat, že všechny kombinace jsou stejně pravděpodobné.
Definujme $Z_i^+ = max(0, Z_i)$ a $Z_i^- = max(0, -Z_i)$. Veličiny $Z_1, \ldots, Z_N$ jsou nezávislé stejně rozdělené, stejně tak musejí být i jejich funkce, tedy $Z_1^+, \ldots, Z_N^+$ jsou nezávislé stejně rozdělené a $Z_1^-, \ldots, Z_N^-$ jsou nezávislé stejně rozdělené. Pro $i \neq j$ jsou Z_i^+ a Z_j^- nezávislé (z nezávislosti Z_i a Z_j $\forall i \neq j$) a kvůli symetrii rozdělení a tomu, že jsou Z_i a Z_j nezávislé a stejně rozdělené, jsou i Z_i^+ a Z_j^- stejně rozdělené. Můžeme tedy říci, že veličiny $Z_1^{G_1}, \ldots, Z_N^{G_N}$, kde $G_i \in \{+, -\}$, jsou nezávislé stejně rozdělené pro všechny možné hodnoty $\{G_i\}_{i=1}^N$.

Platí tedy:

$$P(Z_i^+ < Z_j^+) = P(Z_i^+ > Z_j^+)$$
$$P(Z_i^- < Z_j^-) = P(Z_i^- > Z_j^-)$$
$$P(Z_i^+ < Z_j^-) = P(Z_i^+ > Z_j^-).$$

Označme $(V_1, \ldots, V_N)^T$ vektor pořadí vektoru $(Z_1^{G_1}, \ldots, Z_N^{G_N})^T$. Vidíme, že

$$P(V_1 = 1, \ldots, V_N = N) = P(Z_1 < Z_2 < \ldots < Z_N) =$$
$$= P(Z_{i_1} < Z_{i_2} < \ldots < Z_{i_N}) = P(V_{i_1} = 1, \ldots, V_{iN} = N),$$

kde $(i_1, \ldots, i_N)^T$ je permutace vektoru $(1, \ldots, N)^T$.

Pak tedy $(V_1, \ldots, V_N)^T$ nabývá všech permutací vektoru $(1, \ldots, N)^T$ se stejnou pravděpodobností. Jelikož pro všechny i, j : $P(G_i = \{+\}) = P(G_j = \{+\})$, je tedy každá n-tice vybraná z $(V_1, \ldots, V_N)^T$ tvořící vektor $(R_1, \ldots, R_n)^T$ opět stejně pravděpodobná. Vidíme, že vektor $(R_1, \ldots, R_n)^T$ nabývá každé neuspořádané kombinace se stejnou pravděpodobností. Po uspořádání do vektoru $(R'_1, \ldots, R'_n)^T$ získáváme požadované tvrzení. □

Vidíme, že v Lemmatu 4 je již nutná podmínka symetrie rozdělení, bez níž by tvrzení neobstálo.

Při konstrukci testů si uvědomujeme, že výše použitá konstanta n se liší při jednotlivých pozorováních podle toho, kolik pozorování bylo kladných. Označme tedy ν počet kladných pozorování mezi $Z_1, \ldots, Z_N$, pak vidíme, že ν je náhodná veličina s binomickým rozdělením, neboť $P(Z_i > 0) = p$ pro $i = 1, \ldots, N$ a $Z_1, \ldots, Z_N$ jsou nezávislé. Za platnosti hypotézy H_0 je $p = \frac{1}{2}$. Tedy pro pevné $n \in \mathbb{N}$ takové, že $0 \leq n \leq N$, za platnosti hypotézy H_0 platí:

$$P(R'_1 = r_1, \ldots, R'_\nu = r_\nu, \nu = n) =$$
$$= P(R'_1 = r_1, \ldots, R'_\nu = r_\nu | \nu = n) \cdot P(\nu = n) =$$
$$= \frac{1}{\binom{N}{n}} \cdot \binom{N}{n} (\frac{1}{2})^N = (\frac{1}{2})^N,$$

kde $(r_1, \ldots, r_n)$ je n-tice přirozených čísel takových, že

$$1 \leq r_1 < \ldots < r_n \leq N.$$

Rovnosti plynou z Lemmatu 4 a z toho, že ν pochází z binomického rozdělení $Bi(N, \frac{1}{2})$.

Vidíme, že $P(R'_1 = r_1, \ldots, R'_\nu = r_\nu, \nu = n)$ nezávisí na n. Tedy všechny testy založené na vektoru $(R'_1, \ldots, R'_\nu)^T$, které mají kritický obor o právě k prvcích tvaru $(r_1, \ldots, r_j)^T, j \in \{0, 1, \ldots, N\}$ hladině testu $\alpha = \frac{k}{2^N}$. Prvky kritického oboru jsou vektory různých délek (dokonce i nulové délky), definujeme rovnost tak, že $(R'_1, \ldots, R'_\nu)^T = (r_1, \ldots, r_j)^T$ právě tehdy, když $\nu = j$ a $R'_i = r_i, \forall i \in \{1, \ldots, j\}$. Jednotlivé pořadové testy hypotézy o symetrii se liší svými kritickými obory. Ty se obvykle uvádějí ve tvaru:

$$h(R'_1) + \ldots + h(R'_\nu) > C,$$

kde ν je náhodná veličina závisející na pozorováních, konstanta C je dána hladinou testu a h je vhodná neklesající funkce, jíž se liší jednotlivé testy.

1.4 Vlastnosti statistiky W^+

Wilcoxonův test patří mezi testy založené na pořadích absolutních hodnot pozorování. Jeho statistiku získáme, pokud si za funkci h z konce kapitoly 1.3

zvolíme identitu, tedy $h(k) = k$. Tuto statistiku značíme

$$W^+ = \sum_{i=1}^{\nu} R'_i.$$

Označme R_i^+ pořadí $|Z_i|$ mezi $|Z_1|, \ldots, |Z_N|$. Pak statistiku W^+ můžeme zapsat ve tvaru:

$$W^+ = \sum_{Z_i \geq 0} R_i^+.$$

V kapitole 1.1 jsme tuto statistiku již zmínili jako statistiku, na které zakládáme jednovýběrový Wilcoxonův test. Nyní tedy vidíme, že Wilcoxonův test se dá použít pro test polohy střední hodnoty (nebo mediánu, pokud střední hodnota neexistuje) jednorozměrného symetrického rozdělení. Je tedy alternativou jednovýběrového t-testu. Zároveň jím lze testovat posun osy symetrie ve dvourozměrném rozdělení, což jej činí alternativou i pro párový t-test. Více si uvedeme v podkapitole 2.1.

Lemma 5. *Za platnosti H_0 jsou vektory $(sign\ Z_1, \ldots, sign\ Z_N)^T$ a $(|Z_1|, \ldots, |Z_N|)^T$ nezávislé.*

Důkaz. Veličiny Z_i pochází z náhodného výběru, jsou tedy mezi sebou nezávislé, stejně tak jejich funkce, speciálně vektory $(sign\ Z_i, |Z_i|)^T$ jsou nezávislé. Ze spojitosti rozdělení vyplývá:

$$P(sign\ Z_i = 0) = (Z_i = 0) = 0$$

a ze symetrie pak:

$$P(sign\ Z_i = 1) = P(sign\ Z_i = -1) = \frac{1}{2}.$$

Pro libovolné $z > 0$ platí:

$$P(sign\ Z_i = 1, |Z_i| < z) = P(0 < Z_i < z) = \frac{1}{2}P(-z < Z_i < z) =$$

$$= \frac{1}{2}P(|Z_i| < z) = P(sign\ Z_i = 1)P(|Z_i| < z).$$

Stejným způsobem dostaneme:

$$P(sign\ Z_i = -1, |Z_i| < z) = P(sign\ Z_i = -1)P(|Z_i| < z),$$

veličiny $sign\ Z_i$ a $|Z_i|$ jsou tedy pro každé i nezávislé. Pak tedy i náhodné vektory $(sign\ Z_1, \ldots, sign\ Z_N)^T$ a $(|Z_1|, \ldots, |Z_N|)^T$ jsou nezávislé. □

Důsledek 6. *Náhodné veličiny sign* Z_i *a* R_j^+ *jsou nezávislé pro všechny* $i, j \in \{1, \ldots, N\}$.

Důkaz. Tvrzení plyne přímo z Lemmatu 5, neboť pro všechny $j \in \{1, \ldots, N\}$ je R_j^+ funkcí $(|Z_1|, \ldots, |Z_N|)^T$. □

Nadále budeme potřebovat následující statistiky, které se statistikou W^+ úzce souvisí:

Definice 7.

$$W^- = \sum_{Z_i<0} R_i^+$$

$$W = \sum_{i=1}^{N} R_i^+ \ sign \ Z_i$$

Tyto statistiky jsou navzájem provázané. Jak uvidíme v následující větě, hodnota každé z nich určuje hodnoty zbylých dvou.

Věta 8. *Platí*

$$W^+ + W^- = \frac{N(N+1)}{2}$$

$$W^+ = \frac{1}{2}W + \frac{N(N+1)}{4}.$$

Důkaz. Zřejmě platí $W^+ + W^- = \sum_{i=1}^{N} R_i^+ = \frac{N(N+1)}{2}$ a také $W^+ - W^- = W$. Sečtením obou rovnic vypočteme W^+. □

Odvodíme si některé vlastnosti statistiky W^+ za platnosti H_0. Tyto vlastnosti jsou důležité pro konstrukci testu.

Věta 9. *Platí-li* H_0*, pak*

$$EW^+ = \frac{N(N+1)}{4}$$

$$var \ W^+ = \frac{1}{24}N(N+1)(2N+1).$$

Důkaz. Ze symetrie rozdělení plyne $Esign\ Z_i = 0$ a z důsledku 6 pak

$$E(R_i^+\ sign\ Z_i) = ER_i^+\ Esign\ Z_i = 0$$

pro všechny $i = 1, \ldots, N$, tedy $EW = \sum_{i=1}^{N} E(R_i^+\ sign\ Z_i) = 0$. Použitím Věty 8 dostáváme

$$EW^+ = \frac{1}{2}EW + \frac{N(N+1)}{4} = \frac{N(N+1)}{4}.$$

Ze stejných poznatků vychází

$$var(R_i^+\ sign\ Z_i) = E(R_i^+\ sign\ Z_i)^2 - (E(R_i^+\ sign\ Z_i))^2 = E(R_i^+)^2(sign\ Z_i)^2 =$$

$$= E(R_i^+)^2 = 1^2\frac{1}{N} + 2^2\frac{1}{N} + \ldots + N^2\frac{1}{N} = \frac{1}{6}(N+1)(2N+1),$$

$$cov(R_i^+\ sign\ Z_i, R_j^+\ sign\ Z_j) = E(R_i^+\ sign\ Z_i \cdot R_j^+\ sign\ Z_j) =$$

$$= E(R_i^+ R_j^+)E(sign\ Z_i)E(sign\ Z_j) = 0, \forall i \neq j$$

pak tedy

$$var\ W = \sum_{i=1}^{N} var(R_i^+\ sign\ Z_i) + \sum_{i=1,\ j=1,\ i\neq j}^{N}\sum^{N} cov(R_i^+\ sign\ Z_i, R_j^+\ sign\ Z_j) =$$

$$= \sum_{i=1}^{N} var(R_i^+\ sign\ Z_i) = \frac{1}{6}N(N+1)(2N+1)$$

$$var\ W^+ = var\frac{1}{2}W = \frac{1}{4}var\ W = \frac{1}{24}N(N+1)(2N+1).$$

□

Výše uvedené vlastnosti statistiky W^+ jsou převzaty z [4], strany 233–235.

Věta 10. *Za platnosti hypotézy H_0 má W^- stejné rozdělení jako W^+, které je symetrické kolem $\frac{N(N+1)}{4}$.*

Důkaz. Ze symetrie rozdělení Z_i a jeho spojitosti plyne

$$P\left(W^+ = c\right) = P\left(\sum_{Z_i>0} R_i^+ = c\right) = P\left(\sum_{i=1}^{N} I(Z_i > 0)\cdot R_i^+ = c\right) =$$

$$= P\left(\sum_{i=1}^{N} I(Z_i < 0)\cdot R_i^+ = c\right) = P\left(\sum_{Z_i<0} R_i^+ = c\right) = P\left(W^- = c\right).$$

S využitím rovnosti $W^+ + W^- = \frac{N(N+1)}{2}$ z Věty 8 a první části tvrzení máme

$$P\left(W^+ = \frac{N(N+1)}{4} + k\right) = P\left(\frac{N(N+1)}{2} - W^- = \frac{N(N+1)}{4} + k\right) =$$

$$= P\left(-W^- = -\frac{N(N+1)}{4} + k\right) = P\left(W^- = \frac{N(N+1)}{4} - k\right) =$$

$$= P\left(W^+ = \frac{N(N+1)}{4} - k\right).$$

□

Tvrzení Vět 9 a 10 ilustruje obrázek 1.1. Ten vznikl z 10 000 simulací náhodného výběru o 100 prvcích (tedy $N = 100$) z logistického rozdělení s parametry $a = 0, b = 1$ a reprezentuje výskyt hodnot statistiky W^+. Přehled použitých rozdělení najdeme v kapitole 2.3.

Dle Věty 9 platí $EW^+ = \frac{N(N+1)}{4} = \frac{100\cdot 101}{4} = 2525$ a dle Věty 10 je rozdělení W^+ symetrické okolo této hodnoty, což odpovídá tomu, co na obrázku 1.1 můžeme vidět. Pokud není dodržen předpoklad symetrie rozdělení, pak statistika W^+ nemá vlastnosti popsané ve Větách 9 a 10. Tuto skutečnost máme k nahlédnutí na obrázku 1.2. Vidíme tedy, že předpoklad symetrie je velmi důležitý.

Přestože můžeme pro každé $N \in \mathbb{N}$ spočítat přesné rozdělení statistiky W^+, jak si ukážeme v podkapitole 1.5, mnohdy se nám mnohdy hodí aproximovat rozdělení náhodné veličiny W^+ normálním rozdělením. Věta 11, která o této aproximaci pojednává, má význam hlavně pro veliká N, neboť v tom případě může být výpočet přesného rozdělení časově náročný a aproximace normálním rozdělením je už dostatečně přesná, což si ukážeme v podkapitole 1.5.

Věta 11. *Definujme náhodnou veličinu*

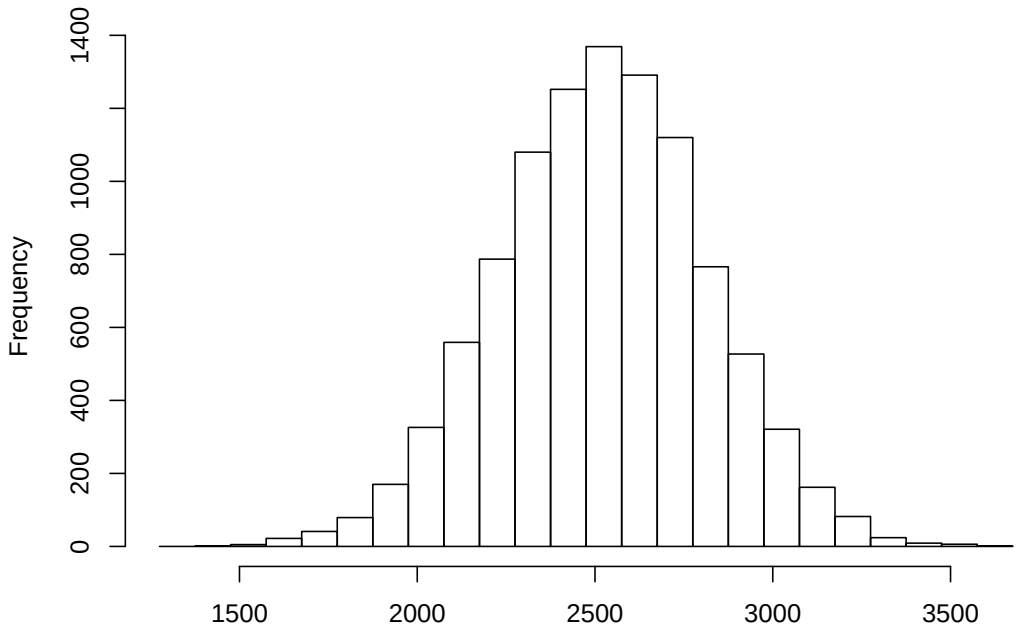
$$U = \frac{W^+ - EW^+}{\sqrt{varW^+}}.$$

Pak U má asymptoticky normální rozdělení $N(0,1)$.

Důkaz. Důkaz silnějšího tvrzení lze nalézt ve [3], strany 166–167. □

W^+ má tedy přibližně rozdělení $N(EW^+, varW^+)$.

Obrázky 1.3 a 1.4 ukazují nasimulované hodnoty náhodné veličiny U z Věty 11, v obou případech jde o 10 000 simulací náhodného výběru z logistického rozdělení s parametry $a = 0, b = 1$, v prvním případě je délka výběru 15, ve druhém 100. K porovnání je do obou obrázků vložena hustota normovaného normálního rozdělení. Vidíme, že v obrázku 1.4 rozložení



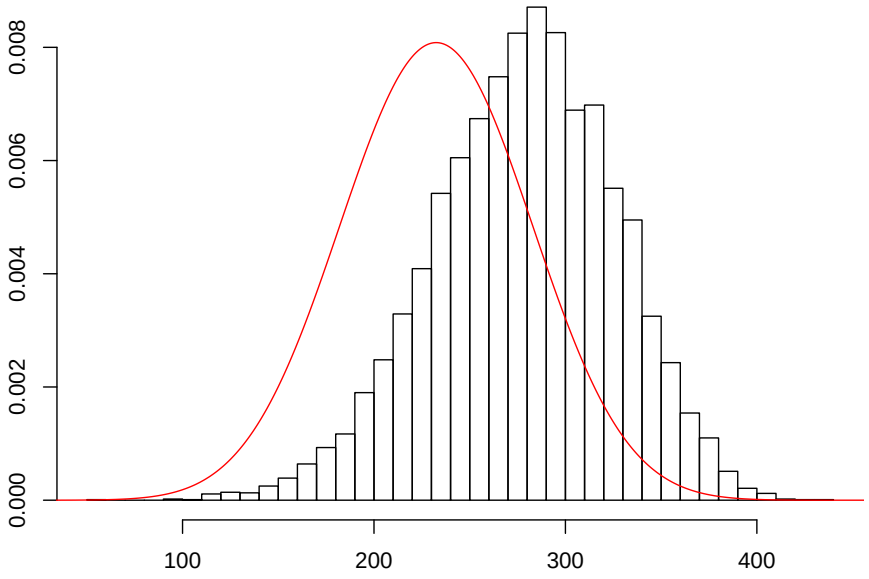
Obrázek 1.1: Histogram hodnot statistiky W^+ založený na 10 000 simulací výběru z logistického rozdělení o délce 100.

nasimulovaných hodnot věrně kopíruje hustotu, ale také pro poměrně malou hodnotu $N = 15$ je již podobnost z obrázku 1.3 čitelná.

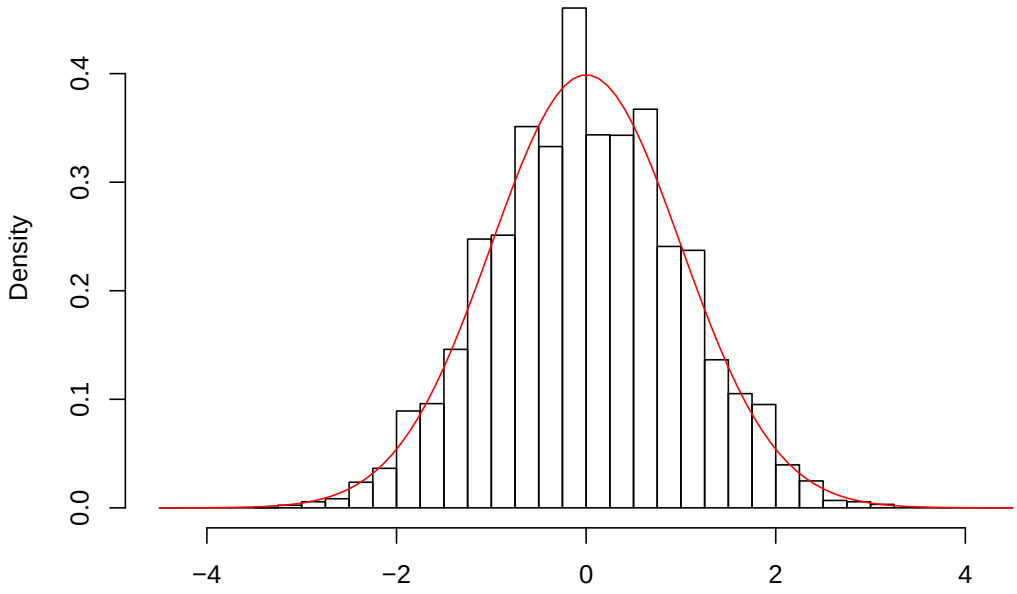
1.5 Konstrukce kritických hodnot

Jak jsme již uvedli v podkapitole 1.1, Wilcoxonův test zakládáme na statistice W^+. Vlastnosti této statistiky jsme si odvodili v podkapitole 1.4, nyní si ukážeme, kdy Wilcoxonův test zamítá nulovou hypotézu.

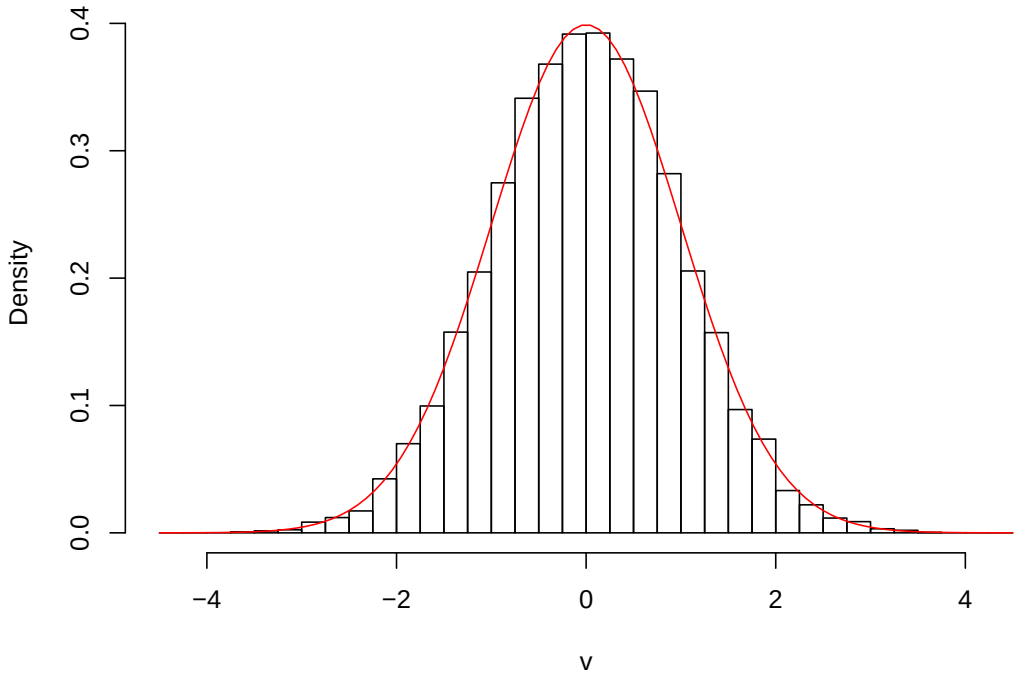
Prakticky provádíme test tak, že z naměřených hodnot spočteme W^+. V případě, že platí H_1, tedy že střed symetrie je posunut do kladných hodnot, ν bude mít binomické rozdělení s koeficientem $p > \frac{1}{2}$, pak očekáváme, že počet kladných pozorování bude větší než počet záporných. Tím nám vzroste



Obrázek 1.2: Histogram hodnot statistiky W^+ založený na 10 000 simulacích náhodného výběru délky $N = 20$ z rozdělení $LN(0,1) - 1$, který je proložený křivkou znázorňující rozdělení pravděpodobnosti statistiky W^+ platnosti H_0. Rozdělení $LN(0,1) - 1$ má medián $\tilde{x} = 0$, ale není symetrické.



Obrázek 1.3: Histogram relativních četností hodnot veličiny $U = \frac{W^+ - EW^+}{\sqrt{varW^+}}$ založený na 10 000 simulací výběru z logistického rozdělení o délce 15 proložený hustotou normovaného normálního rozdělení.



Obrázek 1.4: Histogram relativních četností hodnot veličiny $U = \frac{W^+ - EW^+}{\sqrt{varW^+}}$ založený na 10 000 simulací výběru z logistického rozdělení o délce 100 proložený hustotou normovaného normálního rozdělení.

počet sčítanců ve $W^+ = \sum_{i=1}^{\nu} R'_i$. Navíc platí pro všechny $x \in \mathbb{R}$:

$$P(Z \leq -x) = F(-x) \leq F(-x+\delta) = 1 - F(x+\delta) \leq 1 - F(x) = P(Z \geq x)$$

tedy budeme očekávat, že kladná měření budou v absolutní hodnotě větší než záporná. Obě tyto skutečnosti tedy za hypotézy H_1 zvětšují očekávanou hodnotu W^+, právě to je důvodem zvolení této statistiky a test tedy založíme právě na tom, že když hodnota W^+ překročí určitou konstantu, upustíme od hypotézy H_0 ve prospěch alternativy. Tuto konstantu nazveme *kritickou hodnotou testu*. O konstrukci kritických hodnot pojednáme v této podkapitole.

Hypotézu H_0 zamítáme tedy v případě, že $W^+ \geq C_\alpha$, kde C_α je kritická hodnota závisející na hladině testu α a velikosti výběru N.

Dovolte nám nejdříve vše předvést na příkladě pro konkrétní hodnotu $N = 4$. Označíme si ν počet kladných pozorování, $R' = (R'_1, \ldots, R'_\nu)^T$ a $S' = (S'_1, \ldots, S'_{N-\nu})^T$ vektory uspořádaných pořadí absolutních hodnot kladných, respektive záporných pozorování mezi všemi pozorováními, které jsme definovali v podkapitole 1.3. Tam jsme zjistili, že počet všech přípustných vektorů je 2^N a že test má hladinu $\alpha = \frac{k}{2^N}$, pokud jeho kritický obor obsahuje k prvků. My si vypíšeme všechny možné vektory R' a S', jim odpovídající hodnoty statistiky $W^+ = \sum_{i=1}^{\nu} R'_i$ pro $N = 4$ (viz tabulka 1.1).

Z tabulky 1.1 můžeme pro každou kritickou hodnotu spočíst hladinu testu. Označme si K_α kritický obor testu s hladinou testu α. V tabulce 1.2 vidíme jednotlivé kritické hodnoty, k nim příslušející hladiny testu a kritické obory pro $N = 4$.

Platí tedy $P(W^+ \geq C_\alpha) = \alpha$ pro α výše uvedeného tvaru (z definice hladiny testu). Ale obvykle nám někdo zadá nejvyšší povolenou pravděpodobnost chyby prvního druhu (označme ji α_0), která nemusí odpovídat právě našemu tvaru. Například, jak jsme si řekli v kapitole 1.3, to často bývá hodnota $0,05$. V tom případě hledáme nejnižší C_{α_0}, pro které platí $P(W^+ \geq C_{\alpha_0}) \leq \alpha_0$. Jelikož už máme pro pevné N vypočtené všechny dvojice hodnot α a C_α (kde α má výše uvedený tvar a C_α můžeme považovat za klesající funkci α) a víme, že W^+ nabývá celočíselných hodnot, pak $C_{\alpha_0} = C_{\alpha'}$, kde $\alpha' = \max_{\alpha \leq \alpha_0} \alpha$.

Kritické hodnoty nemusíme pokaždé ručně vypočítávat, bylo by to pracné, pro menší N je najdeme v tabulkách v různých tvarech pro několik používaných hladin. Zde uvádíme tabulka 1.3 převzatá z [2] (strana 128), ve které jsou vybrané kvantily rozdělení statistiky W^+ za platnosti H_0. V programu R k dosažení kritických hodnot můžeme využít funkci $qsignrank(p, N)$, která

R'	S'	W^+
(1,2,3,4)	()	10
(2,3,4)	(1)	9
(1,3,4)	(2)	8
(1,2,4)	(3)	7
(3,4)	(1,2)	7
(1,2,3)	(4)	6
(2,4)	(1,3)	6
(1,4)	(2,3)	5
(2,3)	(1,4)	5
(1,3)	(2,4)	4
(4)	(1,2,3)	4
(1,2)	(3,4)	3
(3)	(1,2,4)	3
(2)	(1,3,4)	2
(1)	(2,3,4)	1
()	(1,2,3,4)	0

Tabulka 1.1: Všechny hodnoty vektorů R', S' a statistiky W^+, které mohou být nabývány při $N = 4$. Symbol „()“ značí vektor délky 0.

C_α	α	K_α
10	1/16	$\{(1,2,3,4)^T\}$
9	2/16	$\{(1,2,3,4)^T, (2,3,4)^T\}$
8	3/16	$\{(1,2,3,4)^T, (2,3,4)^T, (1,3,4)^T\}$
7	5/16	$\{(1,2,3,4)^T, (2,3,4)^T, (1,3,4)^T, (1,2,4)^T, (3,4)^T\}$
6	7/16	$\{(1,2,3,4)^T, \ldots, (1,2,4)^T, (3,4)^T, (1,2,3)^T, (2,4)^T\}$
5	9/16	$\{(1,2,3,4)^T, \ldots, (2,4)^T, (1,4)^T, (2,3)^T\}$
4	11/16	$\{(1,2,3,4)^T, \ldots, (1,4)^T, (2,3)^T, (1,3)^T, (4)^T\}$
3	13/16	$\{(1,2,3,4)^T, \ldots, (1,4)^T, (2,3)^T, (1,3)^T, (4), (1,2)^T, (3)\}$
2	14/16	$\{(1,2,3,4)^T, \ldots, (1,4)^T, (2,3)^T, (1,3)^T, (4), (1,2)^T, (3), (2)\}$
1	15/16	$\{(1,2,3,4)^T, \ldots, (2,3)^T, (1,3)^T, (4), (1,2)^T, (3), (2), (1)\}$
0	16/16	$\{(1,2,3,4)^T, \ldots, (1,3)^T, (4), (1,2)^T, (3), (2), (1), ()\}$

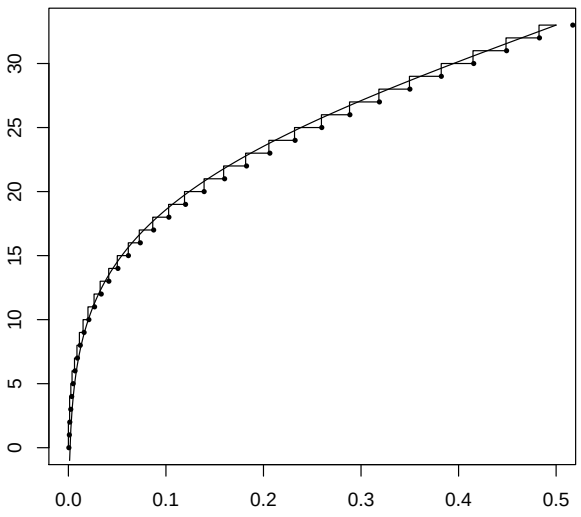
Tabulka 1.2: Kritické hodnoty, hladiny a kritické obory pro $N = 4$. Platí $P(W^+ \geq C_\alpha) = \alpha = \frac{|K_\alpha|}{2^N}$.

N	$w_{0,005}$	$w_{0,01}$	$w_{0,025}$	$w_{0,05}$	$w_{0,1}$	$w_{0,2}$	$w_{0,3}$	$w_{0,4}$
4	0	0	0	0	0	2	2	3
5	0	0	0	0	2	3	4	5
6	0	0	0	2	3	5	7	8
7	0	0	2	3	5	8	10	11
8	0	1	3	5	8	11	13	15
9	1	3	5	8	10	14	17	19
10	3	5	8	10	14	18	21	24
11	5	7	10	13	17	22	26	29
12	7	9	13	17	21	27	31	35
13	9	12	17	21	26	32	37	41
14	12	15	21	25	31	38	43	47
15	15	19	25	30	36	44	50	54
16	19	23	29	35	42	50	57	62
17	23	27	34	41	48	57	64	70
18	27	32	40	47	55	65	72	79
19	32	37	46	53	62	73	81	88
20	37	43	52	60	69	81	90	97

Tabulka 1.3: Přesné hodnoty kvantilů. Tabulka udává kvantily rozdělení náhodné veličiny W^+ za platnosti hypotézy H_0. Platí $P(W^+ \leq w_\alpha) \leq \alpha$, $P(W^+ \leq w_\alpha + 1) > \alpha$. Další kvantily spočteme ze vztahu $w_\alpha = \frac{N(N+1)}{2} - w_{1-\alpha}$ plynoucího z Věty 10: $P(W^+ \leq w_\alpha) = P(W^+ \geq \frac{N(N+1)}{2} - w_\alpha)$

N	$w'_{0,005}$	$w'_{0,01}$	$w'_{0,025}$	$w'_{0,05}$	$w'_{0,1}$	$w'_{0,2}$	$w'_{0,3}$	$w'_{0,4}$
4	0	0	0	0	1	2	3	4
5	0	0	0	1	2	4	5	6
6	0	0	1	2	4	6	7	9
7	0	0	2	4	6	9	10	12
8	0	1	4	6	8	11	14	16
9	0	2	5	8	11	15	18	20
10	2	4	8	11	14	19	22	25
11	4	6	10	14	18	23	27	30
12	6	9	14	18	22	28	32	35
13	8	12	17	21	27	33	37	41
14	11	15	21	26	32	39	44	48
15	14	19	25	31	37	45	50	55
16	18	23	30	36	43	51	57	63
17	22	27	35	41	49	58	65	71
18	26	32	40	47	56	66	73	79
19	30	37	46	54	63	74	81	88
20	36	42	52	60	70	82	90	98

Tabulka 1.4: Asymptotické hodnoty kvantilů. Kvantily $w'_\alpha = \lfloor EW^+ + \Phi^{-1}(\alpha)\sqrt{varW^+} \rfloor$ jsou spočtené na základě Věty 11. (Symbol $\lfloor\ \rfloor$ znamená dolní celá část.)



Obrázek 1.5: Porovnání přesných a asymptotických hodnot kvantilu pro délku výběru $N = 11$. Na ose x jsou hodnoty p, na ose y pak hodnoty p-kvantilu. Skoková křivka je kvantilová funkce rozdělení W^+ za platnosti hypotézy H_0, hladká křivka je kvantilová funkce rozdělení $N\left(\frac{N(N+1)}{4}, \frac{N(N+1)(2N+1)}{24}\right)$.

dává hodnotu kvantilové funkce rozdělení statistiky W^+ pro délku výběru N za platnosti H_0 v bodě p.

Pro větší hodnoty N využíváme normální aproximaci $U = \frac{W^+ - EW^+}{\sqrt{var\ W^+}}$, jak jsme si již uvedli na konci podkapitoly 1.4.

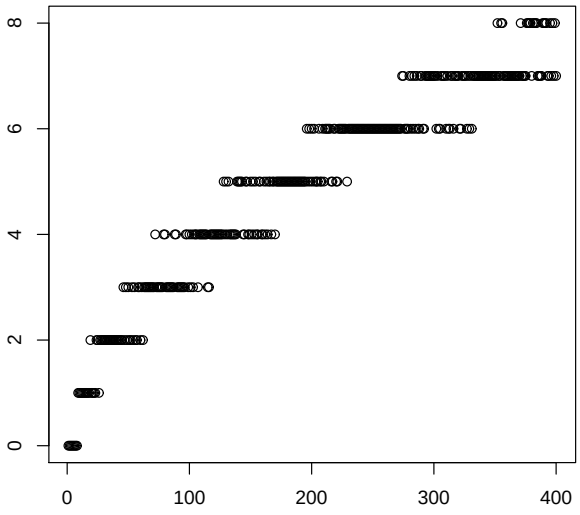
Tedy pro větší hodnoty N platí:

$$1 - \alpha \doteq P\left(U < \Phi^{-1}(1-\alpha)\right) = P\left(\frac{W^+ - EW^+}{\sqrt{var\ W^+}} < \Phi^{-1}(1-\alpha)\right) =$$

$$= P\left(W^+ < EW^+ + \Phi^{-1}(1-\alpha)\sqrt{var\ W^+}\right),$$

kde $\Phi^{-1}(\alpha)$ je kvantil normovaného normálního rozdělení. EW^+ a $var\ W^+$ jsou funkcí pouze N, tedy pro určitý test jsou to konstanty. Kritické hodnoty tedy spočteme ze vztahu:

$$\alpha = P\left(W^+ \geq EW^+ + \Phi^{-1}(1-\alpha)\sqrt{var\ W^+}\right).$$



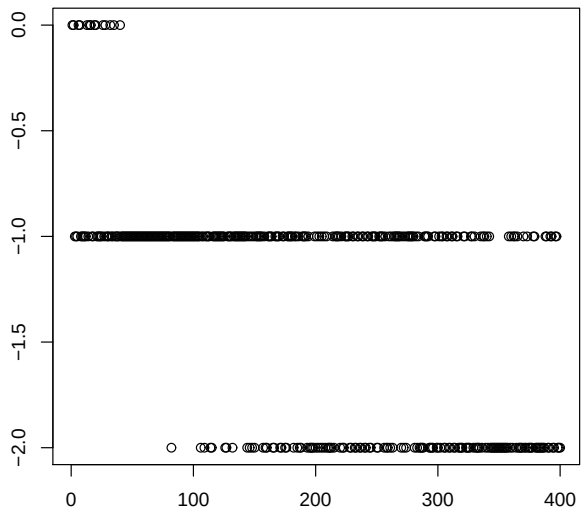
Obrázek 1.6: Rozdíl přesných a asymptotických hodnot 0.005-kvantilu. Na ose x jsou hodnoty délky výběru N, na ose y potom rozdíl přesné hodnoty a asymptotické hodnoty 0.005-kvantilu rozdělení statistiky W^+ za platnosti hypotézy H_0.

Mohu označit $C'_\alpha = EW^+ + \Phi^{-1}(1-\alpha)\sqrt{var\ W^+}$ kritické hodnoty normalizované metody.

Tyto hodnoty (přesněji řečeno kvantily rozdělení $N(EW^+, varW^+)$ zaokrouhlené dolů tak, aby odpovídali tabulce 1.3) jsou pro $N \in \{4, \cdots, 20\}$ k nahlédnutí v tabulce 1.4.

Nyní si provedeme porovnání přesných a asymptotických hodnot kvantilů, abychom zjistili, jak těsná aproximace pomocí normálního rozdělení. Pro $N = 11$ vidíme porovnání těchto kvantilů na obrázku 1.5. Ten naznačuje, že pro menší p jsou hodnoty přesných p-kvantilů větší než asymptotických, zatímco pro p blíže hodnotě 0.5 je tomu opačně. Pro $p > 0.5$ to platí symetricky, neboť rozdělení statistiky W^+ za platnosti hypotézy H_0 i rozdělení $N\left(\frac{N(N+1)}{4}, \frac{N(N+1)(2N+1)}{24}\right)$ jsou symetrická.

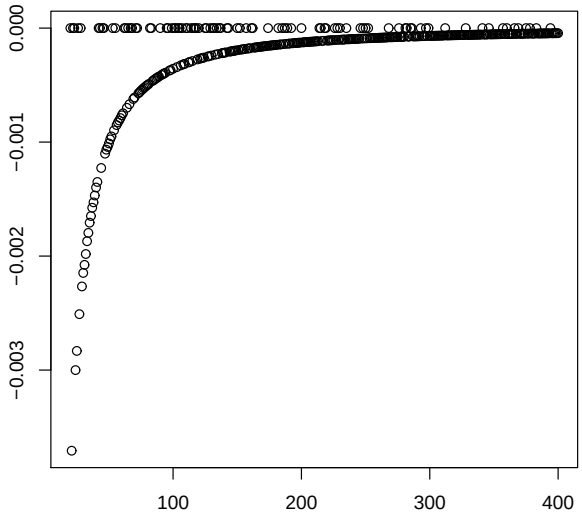
Pro vyšší hodnoty N pak obrázky 1.6 a 1.7 ukazují, jak mnoho se liší asymptotické hodnoty od přesných. Obrázky naznačují, že se chyba s rostoucím N zvětšuje, ale musíme brát také v potaz velikost hodnot, které kvantily nabývají. Například 0.005-kvantil pro $N = 200$ je roven 7944, pak tedy



Obrázek 1.7: Rozdíl přesných a asymptotických hodnot 0.3-kvantilu. Na ose x jsou hodnoty délky výběru N, na ose y potom rozdíl přesné hodnoty a asymptotické hodnoty 0.3-kvantilu rozdělení statistiky W^+ za platnosti hypotézy H_0.

chyba velikosti 5 je zanedbatelná. Z těchto obrázků vidíme, že menší rozdíly jsou nabývány pro p-kvantily, kde p je blízko 0.5, zatímco pro menší (a symetricky taky větší) hodnoty p se rozdíly zvětšují. Z obrázku 1.8 můžeme vyčíst, že vliv rozdílu přesných a asymptotických hodnot 0.05-kvantilu s rostoucím N rychle klesá k nule, hodnoty na ose y můžeme považovat za odchylku od přesných hodnot pravděpodobností v případě užití asymptotických hodnot na místo přesných. Vidíme, že takováto odchylka pro $N = 50$ je přibližně 0.001, tedy nijak zvláště významná. Asymptotické hodnoty kvantilů lze pro větší N celkem bezpečně používat.

Třetí možností, jak získat kritické hodnoty, je pomocí simulací. Nepotřebujeme žádné tabulky hodnot. Pro konkrétní N nasimulujeme k realizací statistiky W^+ z náhodného vektoru délky N z rozdělení, které splňuje předpoklady Wilcoxonova testu. Zvolme si $C \in \{1, \cdots, \frac{N(N+1)}{2}\}$ a počet realizací, které jsou větší nebo rovny C, označme a, dále pak $\alpha' = \frac{a}{k}$. Zřejmě pro dostatečně vysoký počet pozorování k platí $P(W^+ \geq C) \doteq \alpha'$. C je tedy kritická hodnota testu s hladinou α'. Čím větší k zvolíme, tím přesnější



Obrázek 1.8: Porovnání přesných a asymptotický hodnot 0.05-kvantilu. Na ose x jsou délky výběru $N = 20, \ldots, 400$, na ose y je pak hodnota $P(W^+ \leq C_{0.05}) - P(W^+ \leq C'_{0.05})$, kde $C_{0.05}$ je přesná a $C'_{0.05}$ asymptotická hodnota 0.05-kvantilu pro délky výběru N.

budou simulované kritické hodnoty.

1.6 Jiné alternativy Wilcoxonova testu

Doposud jsme zkoumali jen Wilcoxonův test s alternativní hypotézou $H_1 : \ g(z+\delta) = g(-z+\delta), \ \delta > 0$, nyní si popíšeme i jiné alternativy. Kdybychom v kapitole 1.3 omezili posunutí δ opačnou podmínkou $\delta < 0$ (tedy testovali hypotézu o posunutí středu symetrie rozdělení náhodné veličiny Z doleva namísto doprava), pak bychom dostali podobnou alternativní hypotézu $H_2 : \ g(z+\delta) = g(-z+\delta), \ \delta < 0$, celé odvození by pak bylo symetrické, místo statistiky W^+ užijeme statistiku W^-, kvantily (tedy i kritické hodnoty) budou stejné jako kvantily pro statistiku W^+ v případě testu pro alternativní hypotézu H_1. Hypotézu H_0 zamítneme, bude-li mít W^- dostatečně vysokou hodnotu a tedy W^+ dostatečně nízkou (to plyne ze vztahu $W^+ + W^- = \frac{N(N+1)}{2}$ z Věty 8).

Stejné řešení jiným způsobem dostaneme, kdybychom si vzali místo ná-

hodné veličiny Z veličinu $-Z$ a pak již pokračovali stejně jako při testování s alternativou H_1. Testování s alternativními hypotézami H_1 a H_2 se nazývá testování *jednostranné alternativy*.

O *oboustranné alternativě* mluvíme v případě, že alternativou je posunutí kterýmkoliv z obou směrů, tedy $H_3: \ g(z+\delta) = g(-z+\delta), \ \delta \in \mathbb{R}, \ \delta \neq 0$. Složením obou dvou postupů dojdeme k tomu, že H_0 zamítáme, bude-li hodnota W^+ dostatečně velká nebo malá. Nechť α je tvaru $\alpha = \frac{k}{2^N}$, kde k je sudý počet prvků kritického oboru. Pak zamítneme nulovou hypotézu na hladině α, pokud $W^+ \geq C_{\frac{\alpha}{2}}$ nebo $W^- \geq C_{\frac{\alpha}{2}}$. Tedy ekvivalentně s využitím Věty 8 zamítáme H_0, pokud $W^+ \notin \left[\frac{N(N+1)}{2} - C_{\frac{\alpha}{2}}, \ C_{\frac{\alpha}{2}}\right]$. Kritický obor má právě k prvků, neboť množiny $\{R' : W^+ \geq C_{\frac{\alpha}{2}}\}$ a $\{R' : W^- \geq C_{\frac{\alpha}{2}}\}$ mají obě $\frac{k}{2}$ prvků a jsou disjunktní pro všechny α výše uvedeného tvaru. Pokud je zadána maximální hodnota hladiny testu obecně, pak se kritický obor (kritická hodnota) spočte obdobně jako v kapitole 1.5.

Ještě doplňme, že testování, zda rozdělení $\mathcal{D}$ je symetrické podle mediánu $\tilde{x_0} \in \mathbb{R}$ odpovídá testování, zda rozdělení $\mathcal{D} - \tilde{x_0}$ je symetrické podle nuly.

V programu R můžeme popsané verze jednovýběrových testů provést příkazem

$$wilcox.test(\mathbf{Z}, alternative, conf.level, mu, exact),$$

kde $\mathbf{Z}$ je vektor pozorování, zvolíme $alternative = greater$ pro alternativu H_1, $alternative = less$ pro alternativu H_2 nebo $alternative = two.sided$ pro alternativu H_3, dále pak $conf.level = 1 - \alpha$, $mu = \tilde{x_0}$ je střed symetrie, pro který chceme rozdělení testovat, nastavíme $exact = TRUE$, pokud chceme, aby se testová statistika porovnávala s přesnými hodnotami kvantilu, při $exact = FALSE$ bude test pracovat s asymptotickými kvantily. Kromě $\mathbf{Z}$ jsou parametry testu nepovinné, přednastaveny jsou takto:

$$alternative = two.sided, conf.level = 0.95, mu = 0, exact = FALSE.$$

1.7 Shody a nuly

Doposud jsme Wilcoxonův test používali jen pro spojitá rozdělení. Statistika W^+ je v tomto případě s pravděpodobností 1 dobře definována. Každému pozorování totiž přiřazujeme dvě vlastnosti - zaprvé mu dáváme vlastní pořadí podle velikosti absolutní hodnoty, za druhé rozlišujeme, zda je kladné

či záporné. Ale pokud se rovnají absolutní hodnoty dvou pozorování, nemůžeme jednoznačně určit pořadí. V případě, že je některé pozorování rovno 0, pak není ani kladné ani záporné. Takovýmto pozorováním budeme říkat *shody* a *nuly*. Sice tyto situace nastanou s pravděpodobností 0 v případě testu výběru ze spojitého rozdělení, ale v praxi se často vyskytují. Reálná data jsou totiž obvykle zaokrouhlená, bereme je sice jako data ze spojitého rozdělení, pochází však kvůli zaokrouhlení z diskrétního rozdělení. Wilcoxonův test není vhodný pro diskrétní rozdělení, právě kvůli nenulové pravděpodobnosti výskytu shod a nul. Přesto si budeme muset ukázat, jak modifikovat Wilcoxonův test na diskrétní rozdělení, aby byl reálně použitelný i v případě, kdy dojde k nějakým shodám či nulám.

Existuje několik metod, které problémy shod řeší. V případě, že máme shody jen u pozorování se stejným znaménkem, pak se nemusíme uchylovat k žádným metodám, protože ať už ke shodným pozorováním přiřadíme jakákoliv odpovídající pořadí (tedy ať už shodná pozorování seřadíme jakkoliv), statistika W^+ bude mít stejnou hodnotu. Ovšem hodnota statistiky se bude lišit pro různá pořadí shod, které mají různé znaménko.

My si nyní nastíníme metodu *průměrných pořadí*. Ta se zakládá na myšlence, že pozorování se stejnými absolutními hodnotami mají mít stejné pořadí, které je průměrem pořadí připadajících na skupinu pozorování se stejnou absolutní hodnotou. Mějme tedy počet pozorování N, mezi nimi je $l \leq N$ různých absolutních hodnot pozorování. d_1 z nich má nejmenší hodnotu, d_2 druhou nejmenší hodnotu, ..., d_l největší hodnotu. Platí zřejmě $\sum_{i=1}^{l} d_i = N$. Pořadí pozorování s nejmenšími hodnotami je rovno

$$\frac{1}{d_1}\sum_{i=1}^{d_1} i = \frac{1}{d_1}\frac{d_1(d_1+1)}{2} = \frac{d_1+1}{2},$$

s k-tými nejmenšími hodnotami

$$\frac{1}{d_k}\sum_{i=d_1+d_2+\ldots+d_{k-1}+1}^{d_k} i = d_1 + d_2 + \ldots + d_{k-1} + \frac{d_k+1}{2}.$$

Označme R_i^+ takto získané pořadí náhodné veličiny $|Z_i|$ mezi $(|Z_1|, \ldots, |Z_N|)^T$ a $(R_1^+, \ldots, R_N^+)^T$ vektor modifikovaných pořadí. Statistika má pak tvar $W^+ = \sum_{Z_i>0} R_i^+$. Tato definice je zobecněním definice v podkapitole 1.1, pro pozorování beze shod má W^+ stejnou hodnotu jako v původní definici, EW^+ a $varW^+$ také zůstávají nezměněny. Ale nově definovaný vektor

$(R_1^+, \ldots, R_N^+)^T$ nenabývá jen permutací vektoru $(1, \ldots, N)^T$, ale jeho složky mohou mít i neceločíselné hodnoty. Statistika W^+ tedy také nabývá neceločíselných hodnot, nemůžeme pro ni užívat tabulku 1.3. Obvykle se užívá normální aproximace, pokud není rozsah některé skupiny pozorování se shodnými absolutními hodnotami blízký N.

Máme několik způsobů jak se vyrovnat s nulami. Je-li nul mezi pozorováními jen málo vzhledem k N, můžeme je vynechat a provést Wilcoxonův test pro data bez nulových pozorování. Jiná možnost je, že si vezmeme statistiku

$$W^+ = \sum_{Z_i>0} R_i^+ + \frac{1}{2} \sum_{Z_i=0} R_i^+.$$

Idea je taková, že polovina nul připadne ke kladným pozorováním, druhá polovina k záporným.

Ošetření shod a nul můžeme samozřejmě kombinovat.

Pro ilustraci postupu celého testu si ukážeme příklad, který se neřešený vyskytuje v [2], strany 80–81.

Příklad. *Bylo vybráno 24 dětí, vždy po dvou homogenních párech. Jedno z dětí v každém páru dostávalo pravidelné dávky vitamínu B_1, druhé placebo. Každé dítě prošlo před provedením i po provedení pokusu IQ testem a rozdíl měření je zanesen v tabulce 1.5. První tři řádky tabulky 1.5 jsou zadány, další řádky už jsou součástí našeho výpočtu. Otázka je, zdali má vitamín B_1 pozitivní vliv na učení dětí.*

Zadání si přeformulujeme do řeči statistiky. Označíme si X_i nárůst IQ dítěte z i-té dvojice, které dostávalo vitamín, a Y_i nárůst IQ dítěte z i-té dvojice, které dostávalo placebo. Předpokládáme, že $(X_1, Y_1)^T, \ldots, (X_{12}, Y_{12})^T$ je náhodný výběr z dvourozměrného rozdělení s distribuční funkcí $F(x, y)$. V případě, že vitamín nemá na učení vliv, je toto rozdělení symetrické podle osy $y = x$. Pokud ale vitamín napomáhá učení, pak je osa posunuta směrem k polorovině $x > y$. Vytvoříme veličinu $Z_i = X_i - Y_i$, v tabulce 1.5

Číslo dvojce	1	2	3	4	5	6	7	8	9	10	11	12
S vitamínem	14	18	2	4	-5	14	-3	-1	1	6	3	3
S placebem	8	26	-7	-1	2	9	0	-4	13	3	3	4
Rozdíl	6	-8	9	5	-7	5	-3	3	-12	3	0	-1
R_i^+	**8**	10	**11**	**6.5**	9	**6.5**	4	**4**	12	**4**	**1**	2

Tabulka 1.5: Zadání a výpočty příkladu.

ji říkáme Rozdíl. $Z_1, \ldots, Z_{12}$ *je pak náhodný výběr z rozdělení symetrického kolem neznámého bodu* $\tilde{x}$.

Nulová hypotéza má tvar $H_0 : \tilde{x} = 0$, *alternativní potom* $H_1 : \tilde{x} > 0$.

Čtvrtý řádek tabulky 1.5 udává vektor vektor modifikovaných pořadí absolutních hodnot $\left(R_1^+, \ldots, R_{12}^+\right)^T$, *přičemž tučně označené hodnoty přísluší k nezáporným pozorováním. Užijeme modifikovaný vzorec pro testovou statistiku* $W^+ = \sum_{Z_i>0} R_i^+ + \frac{1}{2}\sum_{Z_i=0} R_i^+$, *neboť se v těchto datech vyskytují jak shody, tak i nuly. Výsledná hodnota testové statistiky pro tento příklad tedy vychází* $W^+ = 40.5$. *Porovnáme ji s asymptotickou kritickou hodnotou pro hladinu testu* $\alpha = 0.05$ *a délku výběru* $N = 12$, *tedy* $C'_{0.05} \doteq 59.97$. *Vidíme, že hodnota testové statistiky kritickou hodnotu nedosahuje, nulovou hypotézu tedy zamítnout nemůžeme.*

Výsledek příkladu je tedy takový, že vitamín B_1 *nemá prokazatelný pozitivní vliv na učení dětí.*

Kapitola 2

Vlastnosti Wilcoxonova testu

2.1 T-test a jeho srovnání s Wilcoxonovým testem

T-test existuje v několika verzích, my si zde uvedeme jen *jednovýběrový* t-test a *párový* t-test. Jednovýběrový t-test je test hypotézy o poloze střední hodnoty μ náhodné veličiny, která má normální rozdělení s neznámým rozptylem. Mějme tedy $\{Z_i, i = 1, ..., N\}$ náhodný výběr z tohoto rozdělení $N(\mu, \sigma^2)$. Testová statistika má tvar:

$$T = \frac{\bar{Z} - \mu_0}{S}\sqrt{N},$$

kde $\bar{Z} = \frac{1}{N}\sum_{i=1}^{N} Z_i$ je průměr pozorování a $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Z_i - \bar{Z})^2$ výběrový rozptyl.

Nulovou hypotézou tedy myslíme $H_0' : \mu = \mu_0$. Ukážeme si jen test s jednostrannou alternativou $H_1' : \mu > \mu_1$. Normální rozdělení je symetrické a dle Lemmatu 1 je $EZ = \tilde{x}$, tedy H_0' je ekvivalentní H_0, jen v H_0' předpokládáme, že výběr pochází z normálního rozdělení. Podobně H_1' je ekvivalentní H_1. Statistika T má za platnosti hypotézy H_0' *Studentovo t rozdělení o* $N - 1$ *stupních volnosti*, jak můžeme vidět v [4], strana 74.

Hypotézu H_0' zamítáme na hladině α tehdy, je-li $T > t_{N-1}(1 - \alpha)$, kde t_{N-1} je kvantilová funkce Studentova rozdělení o $N - 1$ stupních volnosti. Vidíme, že $P(T > t_{N-1}(1 - \alpha)|H_0') = \alpha$.

Párový t-test je verze t-testu o rozdílu středních hodnot pro párová data. Mějme tedy náhodný výběr $\left\{(X_i, Y_i)^T, i = 1, ..., N\right\}$ z dvourozměrného

	$\beta_{0,005}$	$\beta_{0.01}$	$\beta_{0,025}$	$\beta_{0,05}$	$\beta_{0,1}$	$\beta_{0,2}$
t-test	0.2981	0.3955	0.5630	0.6911	0.8144	0.9139
Wilcox	0.2884	0.3953	0.5599	0.6873	0.7989	0.9076

Tabulka 2.1: Porovnání sil t-testu a Wilcoxonova testu. Tabulka vznikla z 10 000 simulací náhodného výběru délky $N = 20$ z normálního rozdělení $N(0.5, 1)$. Hodnota β_α pro daný test udává relativní četnost výběrů, v nichž test zamítl hypotézu H_0 na hladině α.

	$a_{0,005}$	$a_{0.01}$	$a_{0,025}$	$a_{0,05}$	$a_{0,1}$	$a_{0,2}$
$N(0,1)$	0.0047	0.0100	0.0256	0.0499	0.1017	0.2023
$Laplace(0,1)$	0.0033	0.0080	0.0239	0.0517	0.1024	0.2075
$t(2)$	0.0020	0.0052	0.0154	0.0407	0.0985	0.2167
$U(-1,1)$	0.0062	0.0119	0.0273	0.0496	0.0963	0.1923
$Logist(0,1)$	0.0043	0.0095	0.0229	0.0482	0.1039	0.1986
$Cauchy(0,1)$	0.0013	0.0033	0.0111	0.0308	0.0936	0.2666

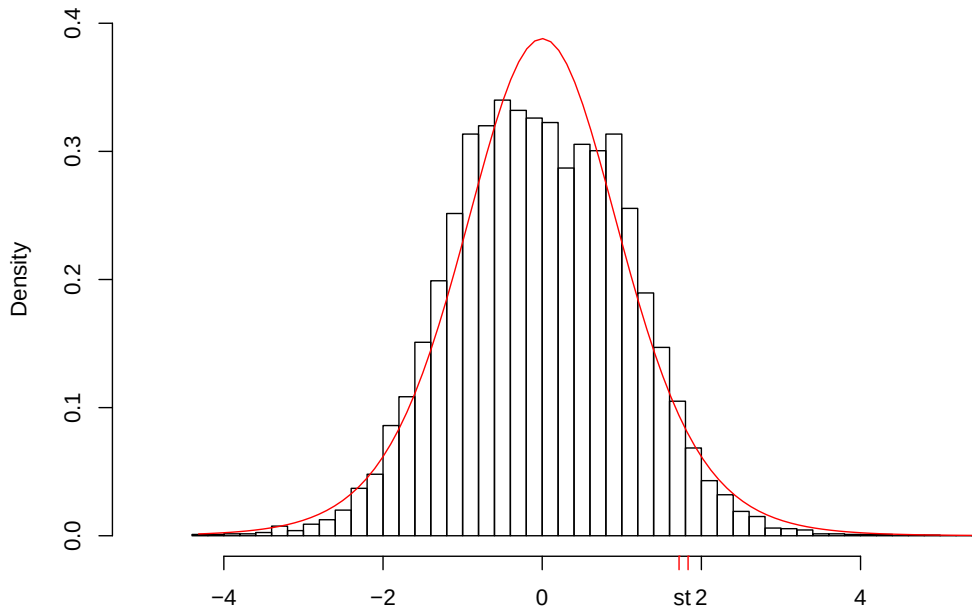
Tabulka 2.2: Tabulka simulovaných pravděpodobností výskytu hodnot statistiky T za několika vybranými kvantily t-rozdělení pro různá rozdělení symetrická kolem nuly. Tabulka vznikla z 10 000 simulací náhodného výběru délky $N = 20$ z rozdělení uvedených v řádku nalevo. Pro každý výběr byla vypočtena hodnota statistiky T. Relativní počet hodnot T větších než $t_{N-1}(\alpha)$ je označen a_α.

	$N(\frac{1}{2},1)$	$Laplace(\frac{1}{2},1)$	$t(2)+\frac{1}{2}$	$U(-0.8,1.2)$	$Logist(-\frac{1}{2},1)$	$Cauchy(\frac{1}{2},1)$
t-test	0.8434	0.6187	0.4046	0.8516	0.4391	0.1329
Wilcox	0.8317	0.7116	0.5928	0.8341	0.4596	0.4158

Tabulka 2.3: Porovnání sil t-testu a Wilcoxonova testu pro vybraná rozdělení. Tabulka vznikla z 10 000 simulací náhodného výběru délky $N = 30$ z rozdělení uvedených v prvním řádku. Hodnoty ve sloupcích udávají pro daný test relativní četnost výběrů, v nichž test zamítl hypotézu H_0 na hladině 0.05.

normálního rozdělení $N\left(\binom{\mu_1}{\mu_2}, \mathbf{V}\right)$. Hypotéza $H_0'' : \mu_2 - \mu_1 = \Delta$ je ekvivalentní hypotéze $H_0' : \mu = \mu_0$, pokud si vezmeme $Z = Y - X$ a $\mu_0 = \Delta$, alternativní hypotéza $H_1'' : \mu_2 - \mu_1 > \Delta$ je pak ekvivalentní H_1'. Párový test se tedy převede na jednovýběrový (Z má normální rozdělení) podobně jako v případě Wilcoxonova testu v kapitole 1.3.

T-test se v praxi často používá i v případech, kdy data nepochází z normálního rozdělení. Pak ale statistika T nemá t-rozdělení s $N-1$ stupni volnosti, jak můžeme vidět například na obrázku 2.1. Tím pádem se nemůžeme spolehnout na vlastnosti statistiky T, změnit se mohou i hladiny testu. Posuny hladin t-testu v případě nedodržení podmínky normality můžeme



Obrázek 2.1: Histogram relativních četností hodnot statistiky T založený na 10 000 simulací výběru ze Studentova rozdělení o 2 stupních volnosti a délce 10, proložený hustotou Studentova rozdělení o 9 stupních volnosti, jež by měla statistika T mít, pokud by simulace byly založeny na normálním rozdělení. Hodnoty na ose x jsou: $t = t_9(0.95)$ a s je hodnota takové realizace statistiky T, že 5% všech pozorování má hodnotu větší než s ($t = 1.833113, s = 1.719420$). Realizací statistiky T s hodnotou větší než t bylo 3.7%.

pro několik rozdělení porovnat v tabulce 2.2. V této tabulce je i test hodnot z Cauchyho rozdělení, které nemá konečnou střední hodnotu, v tom případě testujeme polohu středu symetrie.

Je tedy otázkou, kdy použít t-test a kdy Wilcoxonův test. Pokud máme data z rozdělení, o němž bezpečně víme, že je normální, jistě je vhodnější sáhnout po t-testu. V tabulce 2.1 vidíme, že t-test je silnější než Wilcoxonův. Pokud ale máme výběr z jiného rozdělení nebo dokonce nevíme, jaké to rozdělení je, nemůžeme garantovat, že test bude mít takové vlastnosti, jaké požadujeme. Dosti velké posuny hladiny testu pozorujeme v tabulce 2.2. Pak je bezpečnější využít Wilcoxonova testu, neboť v tom případě rozdělení z tabulky 2.2 splňují jeho předpoklady. Wilcoxonův test také funguje lépe pro některá rozdělení, z tabulky 2.3 je patrné, že Wilcoxonův test má větší sílu než t-test obzvláště v případě Cauchyho, Studentova či dvojitého exponenciálního rozdělení, slabší je pro rovnoměrné a samozřejmě i normální rozdělení, ale ne o mnoho. Tedy na datech z normálního rozdělení, domácím hřišti t-testu, Wilcoxonův test jen slabě zaostává za t-testem, jinde je ale znatelně lepším. Z těchto důvodů je vhodné sáhnout po Wilcoxonově testu pokaždé, pokud nemáme jistotu, že námi zkoumaná data nepocházejí z normálního rozdělení.

2.2 Síla Wilcoxonova testu

Tato podkapitola se věnuje srovnání efektivity Wilcoxonova testu pro různá rozdělení. Budeme zde zkoumat jen rozdělení symetrická podle mediánu $\tilde{x} > 0$, aby jsme vyhověli podmínkám Wilcoxonova testu. Pro dvě konkrétní rozdělení, první s mediánem $\tilde{x_1} > 0$ a druhé s mediánem $\tilde{x_2} > 0$, a zvolenou maximální hladinu testu α nám mírou účinnosti testu bude síla testu. Ta udává, s jakou pravděpodobností test udělá, co od něj chceme, tedy zamítne nulovou hypotézu. Za H_1 budeme brát jednostrannou alternativu. Otázkou ale je, jaká zvolit posunutí (mediány) pro jednotlivá rozdělení. Konstanta není šťastná volba, neboť některá rozdělení jsou více soustředěna do malých hodnot než jiná, co je pak pro jedno rozdělení posunem zásadním, pro jiné může být nepatrným. Například posuneme-li o 1 náhodnou veličinu $U(-1, 1)$, pak bude statistika $W^+ = \frac{N(N+1)}{2}$ a zamítat H_0 budeme na téměř všech hladinách (mimo $\alpha \leq P\left(W^+ = \frac{N(N+1)}{2} | H_0\right)$). Stejný posun pro rozdělení $U(-100, 100)$ způsobí, že se síla testu nebude mnoho lišit od hladiny.

Z toho důvodu je vhodnější si zvolit za posun do kladných hodnot

	posun	$\beta_{0,005}$	$\beta_{0.01}$	$\beta_{0,025}$	$\beta_{0,05}$	$\beta_{0,1}$	$\beta_{0,2}$
$N(0,1)$	0.2533	0.0944	0.1455	0.2567	0.3721	0.5218	0.6891
$Laplace(0,1)$	0.2231	0.0577	0.0903	0.1728	0.2572	0.3977	0.5738
$t(2)$	0.2886	0.0733	0.1123	0.2025	0.2959	0.4407	0.6086
$U(-1,1)$	0.2000	0.1895	0.2638	0.4113	0.5365	0.6848	0.8191
$Logist(0,1)$	0.4054	0.0911	0.1379	0.2412	0.3491	0.4957	0.6631
$Cauchy(0,1)$	0.3249	0.0525	0.0853	0.1600	0.2448	0.3752	0.5453

Tabulka 2.4: Simulované síly Wilcoxonova testu při posunutí o 0.6-kvantil. V každém řádku tabulky vidíme pro vybrané hladiny testu α simulované síly testu β_α pro uvedené rozdělení posunuté o 0.6-kvantil daného rozdělení vpravo. Je provedeno 10 000 realizací statistiky W^+ z náhodného výběru délky $N = 30$ z uvedeného posunutého rozdělení. β_α je rovno relativní četnosti pozorování W^+, které jsou větší nebo rovny kritické hodnotě C_α. 0.6-kvantily jednotlivých rozdělení obsahuje sloupec *posun*.

p	0.51	0.55	0.60	0.65	0.70
$N(0,1)$	0.0643	0.1472	0.3324	0.5626	0.8026
$Laplace(0,1)$	0.0607	0.1170	0.2536	0.4359	0.6583
$t(2)$	0.0621	0.1282	0.2662	0.4595	0.6630
$U(-1,1)$	0.0674	0.1955	0.4647	0.7425	0.9246
$Logist(0,1)$	0.0619	0.1427	0.3144	0.5344	0.7638
$Cauchy(0,1)$	0.0627	0.1170	0.2265	0.3757	0.5578

Tabulka 2.5: Simulované síly testu pro vybraná posunutí a hladinu testu $\alpha = 0.05$. V každém řádku tabulky vidíme simulované síly testu $\beta_{0.05}$ pro uvedené rozdělení posunuté o p-kvantil daného rozdělení vpravo. Hodnoty p jsou uvedeny pro každý sloupec v prvním řádku. Je provedeno 10 000 realizací statistiky W^+ z náhodného výběru délky $N = 25$ z uvedeného posunutého rozdělení. $\beta_{0.05}$ je rovno relativní četnosti pozorování W^+, které jsou větší nebo rovny kritické hodnotě $C_{0.05}$.

p-kvantil příslušného rozdělení, kde $p > 0.5$. U symetrického rozdělení je 0.5-kvantil roven nule. Mějme tedy náhodný výběr $Z_1, \ldots, Z_N$ z takto posunutého rozdělení. Pak pro všechny $i \in \{1, 2, \ldots, N\}$ platí $P(Z_i > 0) = p$ a $P(Z_i < 0) = 1 - p$. Náhodná veličina $\nu = \sum_{i=1}^{N} I(Z_i > 0)$ z konce podkapitoly 1.3 má binomické rozdělení s parametrem p, tedy $Bi(N, p)$. Tím se pro všechna zkoumaná rozdělení stejně změní rozdělení počtu sčítanců ve vzorci $W^+ = \sum_{i=1}^{\nu} R'_i$ z podkapitoly 1.4. Můžeme vyjít z poznatků ze začátku kapitoly 1.5, kde jsme si odvodily dvě kritéria, která zvětšují očekávanou hodnotu statistiky W^+ za platnosti alternativní hypotézy. První kritérium se týkalo rozdělení počtu sčítanců ve statistice W^+. Výše popsa-

	posun	p	$\beta_{0,005}$	$\beta_{0.01}$	$\beta_{0,025}$	$\beta_{0,05}$	$\beta_{0,1}$
$N(0,1)$	$1/2$	0.6915	0.4907	0.5955	0.7399	0.8335	0.9126
$Laplace(0,1)$	1	0.8161	0.9092	0.9465	0.9791	0.9903	0.9973
$t(4)$	1	0.8130	0.9294	0.9608	0.9849	0.9929	0.9977
$U(-1,1)$	$1/6$	0.5833	0.1193	0.1780	0.3038	0.4231	0.5776
$Logist(0,1)$	$\pi^2/6$	0.8382	0.9834	0.9932	0.9983	0.9994	0.9998

Tabulka 2.6: Simulované síly Wilcoxonova testu při posunutí o polovinu rozptylu. V každém řádku tabulky vidíme pro vybrané hladiny testu α simulované síly testu β_α pro uvedené rozdělení posunuté o polovinu rozptylu daného rozdělení vpravo. Je provedeno 10 000 realizací statistiky W^+ z náhodného výběru délky $N = 30$ z uvedeného posunutého rozdělení. β_α je rovno relativní četnosti pozorování W^+, které jsou větší nebo rovny kritické hodnotě C_α. Velikosti posunutí jednotlivých rozdělení obsahuje sloupec *posun*, koeficient binomického rozdělení počtu sčítanců ve statistice W^+ sloupec p.

ným posunutím jsme si ale sjednotili jeho účinek na všechna rozdělení, nebude mít tedy vliv pro porovnávání jednotlivých rozdělení. Druhým kritériem je fakt, že pro všechny $x \in \mathbb{R}$ platí $F(-x) \leq 1 - F(x)$ ($F(x)$ je distribuční funkce posunutého rozdělení), což zvyšuje pravděpodobnost vyšších pořadí pro kladná pozorování. Přesto pomocí p-kvantilu dosáhneme sjednocení posunutí tak, aby to bylo alespoň trochu spravedlivé. Můžeme tedy srovnávat síly testů pro jednotlivá posunutá rozdělení.

Z náhledu do tabulek 2.4 a 2.5 můžeme jednoduše usoudit, že z testovaných rozdělení funguje Wilcoxonův test výrazně nejlépe podle výše popsaného nastavení pro rovnoměrné rozdělení, také normální a logistické rozdělení slibuje vyšší hodnoty síly testu.

Ovšem důležitou informací pro nás je i to, že se síly hodně liší. Z toho usuzujeme, že na sílu mají pozorovatelný vliv i jiná kritéria. Tak jako máme hodnotu p pro první kritérium, tak hodnotou k porovnání pro druhé kritérium může být $E\left[1 - F(x) - F(-x)\right]$, tedy střední hodnota rozdílů pravděpodobností $P(Z > x) - P(Z < -x)$.

Jinou možností posunu může být násobek rozptylu. Zde si jednoduše spočteme p, koeficient binomického rozdělení počtu sčítanců ve statistice W^+. Je-li rozdělení posunuto o d, pak $p = P(Z < d)$. Můžeme pak zjistit, jak velký vliv má v tomto nastavení první kritérium. Toto nastavení je použito v tabulce 2.6, z něj nejlépe vychází logistické, Studentovo a dvojitě exponenciální rozdělení. Vidíme, že síly jsou seřazeny až na jedinou výjimku podle velikostí koeficientu p.

2.3 Použitá rozdělení

- Diskrétní rozdělení

  1. Binomické rozdělení
     Označení: $X \sim Bi(n, p)$
     X nabývá pouze hodnot $k \in \{0, 1, \cdots, n\}$ a to s pravděpodobností $p_k = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$
     Střední hodnota:
     $$EX = np$$
     Rozptyl:
     $$varX = np(1-p)$$

- Spojitá rozdělení

  1. Normální rozdělení
     Označení: $X \sim N(\mu, \sigma^2)$
     Parametry:
     $$\mu \in \mathbb{R}, \sigma^2 > 0$$
     Hustota:
     $$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, x \in \mathbb{R}$$
     Střední hodnota (medián):
     $$EX = \tilde{x} = \mu$$
     Rozptyl:
     $$varX = \sigma^2$$
     Rozdělení je symetrické se středem symetrie $\tilde{x}$.
  2. Dvojité exponenciální rozdělení
     Nazývá se také Laplaceovo rozdělení.
     Označení: $X \sim Laplace(a, b)$
     Parametry:
     $$a \in \mathbb{R}, b > 0$$

Hustota:

$$f(x) = \frac{1}{2b} \exp\left\{-\frac{|x-a|}{b}\right\}, x \in \mathbb{R}$$

Střední hodnota (medián):

$$EX = \tilde{x} = a$$

Rozptyl:

$$varX = 2b^2$$

Rozdělení je symetrické se středem symetrie $\tilde{x}$.

3. Rozdělení t

   Nazývá se také Studentovo rozdělení.

   Označení: $X \sim t_n$

   Parametry:

$$n \geq 1$$

   Hustota:

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{\pi n}} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}, x \in \mathbb{R},$$

   kde $\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx, a \geq 1$

   Střední hodnota (medián):

   Medián $\tilde{x} = 0$. Pro $n > 1$ existuje EX a platí $EX = 0$

   Rozptyl: Pro $n > 2$ existuje konečný rozptyl a platí $varX = \frac{n}{n-2}$

   Rozdělení je symetrické se středem symetrie $\tilde{x}$.

4. Spojité rovnoměrné rozdělení

   Označení: $X \sim U(a, b)$

   Parametry:

$$a, b \in \mathbb{R}, a < b$$

   Hustota:

$$f(x) = \frac{1}{b-a}, x \in (a, b)$$

   Střední hodnota (medián):

$$EX = \tilde{x} = \frac{a+b}{2}$$

Rozptyl:

$$varX = \frac{(b-a)^2}{12}$$

Rozdělení je symetrické se středem symetrie $\tilde{x}$.

5. Logistické rozdělení
   Označení: $X \sim Logis(a, b)$
   Parametry:

$$a \in \mathbb{R}, b > 0$$

   Hustota:

$$f(x) = \frac{be^{-(a+bx)}}{\left[1 + e^{-(a+bx)}\right]^2}, x \in \mathbb{R}$$

   Střední hodnota (medián):

$$EX = \tilde{x} = -\frac{a}{b}$$

   Rozptyl:

$$varX = \frac{1}{b^2}\frac{\pi^2}{3}$$

   Rozdělení je symetrické se středem symetrie $\tilde{x}$.

6. Cauchyho rozdělení
   Označení: $X \sim Cauchy(a, b)$
   Parametry:

$$a \in \mathbb{R}, b > 0$$

   Hustota:

$$f(x) = \frac{1}{\pi}\frac{b}{b^2 + (x-a)^2}, x \in \mathbb{R}$$

   Střední hodnota ani rozptyl neexistují.
   Medián:

$$\tilde{x} = a$$

   Rozdělení je symetrické se středem symetrie $\tilde{x}$.

7. Logaritmicko normální rozdělení
   Označení: $X \sim LN(\mu, \sigma^2)$
   Parametry:

$$\mu \in \mathbb{R}, \sigma^2 > 0$$

   Hustota:

$$f(x) = \frac{1}{x\sqrt{2\pi}\sigma} \exp\left\{-\frac{(\ln x - \mu)^2}{2\sigma^2}\right\}, x \in \mathbb{R}$$

   Střední hodnota (medián):

$$EX = e^{\mu + \frac{\sigma^2}{2}},\ \tilde{x} = e^{\mu}$$

   Rozptyl:

$$varX = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2}$$

   Rozdělení není symetrické.

Závěr

V první části je poměrně přímočaře odvozena statistika Wilcoxonova testu a přehledně podány její vlastnosti. Způsobům, jakými lze dojít ke kritickým hodnotám rozdělení této statistiky za nulové hypotézy, je věnováno hodně pozornosti. Uvedeny jsou zde i postupy, jakými přibližné kritické hodnoty získat bez patřičných tabulek, a vlastnosti těchto aproximací.

Co se týče kapitoly o shodách a nulách, ta je pouze informativní. Toto téma, totiž použití pořadových testů pro diskrétní rozdělení, by možná vydalo na samostatnou práci. Šířeji se o této problematice pojednává například i v [3] a [2].

Nad očekávání dopadlo srovnání Wilcoxonova testu s t-testem. Tam se jasně ukázalo, že i když Wilcoxonův test je založen pouze na pořadích a tedy se vzdává zdánlivě spousty informací, když na rozdíl od t-testu nebere v potaz hodnoty pozorování, může svou silou dobře konkurovat t-testu i v datech z normálního rozdělení. Práce tedy naznačuje, že Wilcoxonův test by mohl t-test, který se běžně používá i pro jiná než normální data, plně zastoupit. Není však dořešeno, jak dobře by fungoval Wilcoxonův test, byla-li by slabě porušena symetrie rozdělení. Jistě by šlo zabývat se dále porovnáváním sil Wilcoxonova testu pro různá rozdělení a zabývat se zdůvodňováním rozdílů sil, ovšem ucelenou představu o fungování testu mohl čtenář získat z textu zde uvedeného. Poslouží jistě i vědomí, jaká je řádově síla testu pro testované rozdělení s určitým posunem.

Ovšem zda byl dosažen cíl této práce přiblížit čtenáři Wilcoxonův test natolik, aby ho mohl zasvěceně používat, kdykoliv se to hodí, to už záleží na každém individuálně.

Literatura

[1] Wilcoxon F.: *Individual comparisons by ranking methods*, Biometric bulletin, Vol. 1, No. 6. (Dec., 1945), 80–83.

[2] Jurečková J.: *Pořadové testy*, Státní pedagogické nakladatelství, Praha, 1981.

[3] Hájek J., Šidák Z.: *Theory of rank tests*, Academia, Praha, and Academic Press, New York, 1967.

[4] Anděl J.: *Základy matematické statistiky*, Matfyzpress, Praha, 2007.