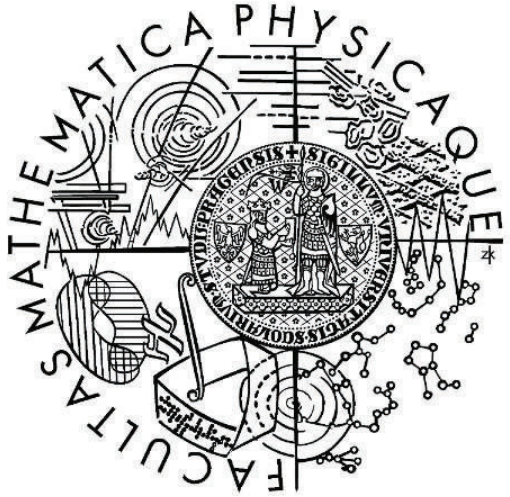
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Barbora Madurkayová

Testy založené na U-statistikách

Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: Prof. RNDr. Marie Hušková, DrSc.
Studijní program: Matematika, Matematická statistika

Poděkování

Děkuji především vedoucí diplomové práce Prof. RNDr. Marii Huškové, DrSc. za mnoho cenných rad, návrhů a usměrnění.

Děkuji také své rodině za podporu při studiu.

Prohlašuji, že jsem svou diplomovou práci napsala samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce.

V Praze dne 28. ledna 2010 Barbora Madurkayová

Obsah

Abstrakt

Název práce: Testy založené na U-statistikách
Autor: Barbora Madurkayová
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: Prof. RNDr. Marie Hušková, DrSc.
E-mail vedoucího: huskova@karlin.mff.cuni.cz
Abstrakt: U-statistiky jsou základem mnoha testových statistik. Táto práce se zabývá testy založenými na U-statistikách pro obecný dvouvýběrový problém. Po popisu dvouvýběrových testů založených na U-statistikách z obecného hlediska, následuje přehled několika konkrétních testů. Všechny uvažované testové statistiky jsou popsány v návaznosti na teorii U-statistik, s důrazem na jejich asymptotické vlastnosti. Zvláštní pozornost je věnována testům založeným na rozdílu dvou jednovýběrových statistik a testům založeným na empirických charakteristických funkcích. Testové statistiky, založené na rozdílu dvou empirických charakteristických funkcí, mají tvar V-statistiky. Od nich je odvozena příbuzná U-statistika a jsou studovány její vlastnosti. Součásti práce je také příklad použití metody bootstrap pro tyto statistiky.
Klíčová slova: U-statistiky, dvouvýběrový problém, neparametrický test, charakteristická funkce

Abstract

Title: Tests based on U-statistics
Author: Barbora Madurkayová
Department: Department of Probability and Mathematical Statistics
Supervisor: Prof. RNDr. Marie Hušková, DrSc.
Supervisor's e-mail address: huskova@karlin.mff.cuni.cz
Abstract: There are many test statistics that are based on U-statistics. This thesis deals with tests based on U-statistics for the general two-sample problem. After describing two-sample tests based on U-statistics from a general viewpoint, a presentation of some particular test statistics follows. All considered test statistics are described in a connection to the theory of U-statistics, with emphasis on their asymptotic properties. Special concern is given to tests based on a difference between two one-sample U-statistics and to tests based on empirical characteristic functions. Test statistics, based on a difference of two empirical characteristic functions, have a form of a V-statistic. A related U-statistic is derived and its properties are studied. An example of applying the bootstrap method to these test statistics is included.
Keywords: U-statistics, two-sample problem, nonparametric test, characteristic function

Úvod

U-statistiky lze považovat za zobecnění pojmu výběrového průměru nezávislých stejně rozdělených náhodných veličin. Teorie U-statistik se začala rozvíjet po tom, co Hoeffding v roce 1948 publikoval svůj základní článek o U-statistikách. Jedním z hlavních důvodů pro zájem statistiků o problematiku U-statistik je skutečnost, že tvoří třídu nestranných odhadů a za určitých podmínek jsou dokonce nejlepšími nestrannými odhady, tj. odhady s nejmenším rozptylem mezi všemi nestrannými odhady určitého parametru. Část U-statistik má přibližně stejné asymptotické chování jako součet nezávislých stejně rozdělených náhodných veličin. Tato podtřída U-statistik má asymptoticky normální rozdělení. V opačném případě mluvíme o tzv. degenerovaných U-statistikách, které mají složitější tvar rozdělení.

Koncept U-statistik byl později zobecněn několika směry. V statistické literatuře se lze setkat například s U-statistikami založenými na několika výběrech, s váženými U-statistikami i s U-statistikami, jejichž jádro je funkcí počtu pozorování. V této práci budeme studovat především zobecnění U-statistik na dva výběry.

Díky svým výhodným vlastnostem jsou U-statistiky základem mnoha testových statistik v oblasti parametrické i neparametrické statistiky. Ne vždy je však souvislost s U-statistikami na první pohled zřejmá. Příklady testových statistik pro různé druhy hypotéz lze najít například v knize [Puri a Sen 1971], nebo také v knize [Lee 1990]. V této práci se zaměříme především na neparametrické testy pro obecný dvouvýběrový problém, tedy na testování hypotézy, že jsou obě zkoumaná rozdělení stejná, proti alternativě, že se nějakým způsobem liší. Mezi nejznámější dvouvýběrové testy založené na U-statistikách patří Mannův-Whitneyův dvouvýběrový test. Funkcí jednoduché U-statistiky – výběrového průměru – je i dvouvýběrový t-test.

Kromě dvouvýběrového Mannova-Whitneyova testu v práci dále popíšeme tzv. Sukhatmův test a test založený na rozdílu výběrových rozptylů. Výběrový rozptyl lze považovat za speciální případ obecnějšího tvaru měr variability, které jsou také U-statistikami. V těchto případech má testová statistika asymptoticky normální rozdělení. Dále popíšeme test založený na vzdálenosti

distribuční funkcí výběrů, který je asymptoticky ekvivalentní s dvouvýběrovým Cramérovým-von Misesovým testem. Ukážeme, že do třídy testů založených na U-statistikách je možné zařadit i testy založené na empirických charakteristických funkcích, které byly studovány v práci [Meintanis 2005]. Přesněji řečeno, tyto testy mají tvar V-statistik, které tvoří příbuznou třídu ke třídě U-statistik. Od tohoto tvaru V-statistik odvodíme příbuzný tvar U-statistik a budeme studovat jejich vlastnosti.

První kapitola shrnuje teorii jednovýběrových a dvouvýběrových U-statistik. V druhé kapitole je popsán obecný tvar testových statistik pro dvouvýběrový problém a speciální tvar testové statistiky založené na rozdílu dvou výběrů. Tématem třetí kapitoly jsou konkrétní testové statistiky pro dvouvýběrový problém. Nejvíce pozornosti je věnováno testům založeným na charakteristických funkcích – jejich popis je proto zařazen do samostatné kapitoly. Pro ilustraci chování testů založených na empirických charakteristických funkcích je připojen krátký příklad aplikace testu na simulovaná data, pomocí metody bootstrap s vracením.

Použité značení

Často se budeme setkávat se situací, kdy máme dva nezávislé výběry $X_1, \ldots,$ X_{n_1} a $Y_1, \ldots, Y_{n_2}$. V této souvislosti budeme používat označení $N = n_1 + n_2$ pro součet velikostí obou výběrů a $Z_1, \ldots, Z_N$ pro výběr sdružený z těchto výběrů, přesněji

$$Z_i = \begin{cases} X_i, & \text{pro} \quad i = 1, \ldots, n_1 \\ Y_i, & \text{pro} \quad i = n_1 + 1, \ldots, n. \end{cases}$$

Pod zápisem $X_n \xrightarrow{P} X$ budeme rozumět, že posloupnost náhodných veličin X_n konverguje k náhodné veličině X v pravděpodobnosti. Podobně, $X_n \xrightarrow{s.j.} X$ a $X_n \xrightarrow{\mathcal{D}} X$ budou značit konvergenci skoro jistě, respektive v distribuci. Pro indikátor jevu A budeme používat označeni $1_{\{A\}}$. $N(\mu,\sigma^2)$ bude zástupný symbol pro náhodnou veličinu s normálním rozdělením se střední hodnotou μ a rozptylem σ^2. Podobně, $Exp(\lambda)$ bude označovat exponenciální rozdělení se střední hodnotou $1/\lambda$. Symbolem $\mathbb{R}$ budeme značit obor reálných čísel, symbolem $\mathbb{N}$ obor přirozených čísel. Další značení je vysvětleno dále v textu.

Kapitola 1

U-statistiky

1.1 Jednovýběrové U-statistiky

U-statistiky jsou velice častým typem statistik, se kterým se setkáváme v oblasti parametrické i neparametrické statistiky. Nejznámějšími U-statistikami jsou pravděpodobně výběrový průměr a výběrový rozptyl. V této kapitole definujeme pojem U-statistik a popíšeme některé jejich vlastnosti. Zaměříme se především na limitní chování U-statistik. Od jednovýběrových statistik přejdeme k zobecnění na dva výběry. Zmíníme se také o V-statistikách, které tvoří velmi blízkou třídu ke třídě U-statistik a stručně popíšeme jejich vzájemný vztah.

1.1.1 Definice a základní vlastnosti

Předpokládejme, že máme výběr $X_1, \ldots, X_n$ nezávislých reálných náhodných veličin s rozdělením charakterizovaným distribuční funkcí F.[1] Označme $\mathcal{F}$ třídu distribučních funkcí a uvažujme funkcionál $\theta(F)$ s hodnotami na reálné přímce, jehož definičním oborem je $\mathcal{F}$.

Definice 1.1. Řekneme, že $\theta = \theta(F)$ je *regulární funkcionál nad* $\mathcal{F}$ pokud pro všechny distribuční funkce $F \in \mathcal{F}$ existuje nestranný odhad $\theta(F)$. Jinými

[1] Pro jednoduchost budeme předpokládat, že $X_1, \ldots, X_n$ jsou náhodné veličiny s hodnotami na reálné přímce. Definice U-statistik se však často rozšiřuje i na případ reálných náhodných vektorů či na obecnější měřitelné prostory. Platnost velké části tvrzení o U-statistikách, které uvedeme v této kapitole, přitom zůstane zachována.

slovy, pokud pro všechna $F \in \mathcal{F}$, může být $\theta(F)$ reprezentováno jako

$$\theta(F) = \mathrm{E}_F h(X_1, \ldots, X_m) = \int \ldots \int h(x_1, \ldots, x_m) dF(x_1) \ldots dF(x_m) \quad (1.1)$$

pro nějakou funkci $h = h(x_1, \ldots, x_m)$, kterou budeme nazývat *jádro* funkcionálu $\theta(F)$. Pokud podmínka (1.1) platí, ekvivalentně se také říká, že $\theta(F)$ je *odhadnutelný*.

Definice 1.2. Pro libovolné jádro $h(x_1, \ldots, x_m)$ se příslušná *U-statistika* pro odhad $\theta(F)$ na základě výběru $X_1, \ldots, X_n$ velikosti $n \geq m$ definuje jako průměr jádra $h(x_1, \ldots, x_m)$ přes všechny permutace pozorování, tedy

$$U_n = U_n(h) = \frac{(n-m)!}{n!} \sum_{\mathbf{P}_{m,n}} h(X_{i_1}, \ldots, X_{i_m}). \quad (1.2)$$

kde součet prochází množinu $\mathbf{P}_{m,n}$ všech $n!/(n-m)!$ permutací $(i_1, \ldots, i_m)$ velikosti m z prvků množiny $\{1, \ldots, n\}$. Pokud je jádro h symetrické, v smyslu invariantní vzhledem k permutacím svých argumentů, U_n má ekvivalentní tvar

$$U_n = U_n(h) = \binom{n}{m}^{-1} \sum_{\mathbf{C}_{m,n}} h(X_{i_1}, \ldots, X_{i_m}), \quad (1.3)$$

kde součet prochází množinu $\mathbf{C}_{m,n}$ všech $\binom{n}{m}$ kombinací m čísel $i_1 < \cdots < i_m$, vybraných z množiny $\{1, \ldots, n\}$.

Poznámka 1.1. Každou nesymetrickou h lze v rovnicích (1.1) a (1.2) nahradit symetrickou h' takovou, že:

$$h'(x_1, \ldots, x_m) = \frac{1}{m!} \sum_{\mathbf{P}_m} h(x_{i_1}, \ldots, x_{i_m}),$$

kde $\sum_{\mathbf{P}_m}$ značí součet přes všech $m!$ permutací $(i_1, \ldots, i_m)$ prvků množiny $\{1, \ldots, m\}$.

Nestrannost a vlastnost optimality Pokud $\theta(F) = \mathrm{E}_F h(X_1, \ldots, X_m)$ existuje pro všechny distribuční funkce $F \in \mathcal{F}$, potom jednou z vlastností příslušné U-statistiky U_n s jádrem h je, že je *nestranným odhadem* $\theta(F)$. Navíc, pokud $\mathcal{F}$ splňuje určité podmínky je dokonce *nejlepším nestranným odhadem* $\theta(F)$. Taková situace nastává například když $\mathcal{F}$ obsahuje všechna

rozdělení s konečným nosičem nebo všechny absolutně spojité distribuční funkce F. Příslušná U-statistika s jádrem h tvaru (1.3) má potom minimální rozptyl mezi všemi nestrannými odhady θ. Je zde souvislost s Lehmannovou-Scheffeho větou (viz např. [Anděl 2002], str. 135). U-statistiku lze totiž vyjádřit jako funkci pořádkové statistiky. Proto v situaci, kdy je pořádková statistika navíc úplnou suficientní statistikou a U_n má konečný rozptyl, je podle této věty jediným nejlepším nestranným odhadem θ.

Poznámka k historii U-statistik Autorem první studie o U-statistikách byl Halmos a byla publikována v roce 1946. Halmos jako první identifikoval tuto třídu statistik jako třídu nestranných odhadů s minimálním rozptylem. V roce 1948 Hoeffding zavedl pojmenování U-statistiky, protože mají vlastnost nestrannosti, anglicky *unbiasedness*. Hoeffding rovněž ukázal, jak lze spočítat rozptyl U-statistiky a demonstroval asymptotickou normalitu této třídy statistik. (viz [Hoeffding 1948]).

1.1.2 Příklady

Výběrový průměr Zřejmě nejčastěji užívanou U-statistikou je *výběrový průměr* $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Je odhadem střední hodnoty rozdělení charakterizovaného distribuční funkcí F:

$$\mu(F) = \int x\, dF(x).$$

V tomto případě má jádro tvar

$$h(x) = x.$$

Výběrový rozptyl Další, rovněž velice často užívanou, U-statistikou je *výběrový rozptyl* $S_n^2 = \frac{1}{n-1}\sum_{i-1}^{n}(X_i - \bar{X}_n)^2$. Je odhadem rozptylu rozdělení s distribuční funkcí F:

$$\sigma^2(F) = \int (x - \mu(F))^2\, dF(x).$$

Dále je možné rozptyl vyjádřit následovně:

$$\begin{aligned}
\sigma^2(F) &= \int \left(x - \int y\, dF(y)\right)^2 dF(x) \\
&= \int \left[x^2 - 2x \int y\, dF(y) + \left(\int y\, dF(y)\right)^2\right] dF(x) \\
&= \left(\frac{1}{2} + \frac{1}{2}\right) \int x^2 dF(x) - \int \left(\int xy dF(y)\right) dF(x)
\end{aligned}$$

$$= \int \left(\int \frac{x^2}{2} - xy + \frac{y^2}{2} dF(y) \right) dF(x)$$
$$= \int \left(\int \frac{(x-y)^2}{2} dF(y) \right) dF(x).$$

To vede k volbě symetrického jádra tvaru

$$h(x,y) = \frac{(x-y)^2}{2}.$$

Po úpravách dostaneme, že příslušná U-statistika

$$U_n = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{(X_i - X_j)^2}{2}$$

je skutečně rovna výběrovému rozptylu S_n^2.

1.1.3 Rozptyl

Předpokládejme, že rozptyl jádra $h(x_1, \ldots, x_m)$ je konečný, tj.

$$\operatorname{Var} h(X_1, \ldots, X_m) < \infty.$$

Abychom mohli určit rozptyl U-statistiky U_n, definujeme nejdříve posloupnost funkcí přidružených k jádru h.
Pro $c = 0, 1, \ldots, m$ označme

$$h_c(x_1, \ldots, x_c) = \mathrm{E}\, h(x_1, \ldots, x_c, X_{c+1}, \ldots, X_m). \tag{1.4}$$

To znamená, že $h_c(X_1, \ldots, X_c)$ je podmíněná střední hodnota $h(X_1, \ldots, X_m)$ při daných $X_1, \ldots, X_c$. Potom $h_0 = \mathrm{E}\, h(X_1, \ldots, X_m)$ a $h_m(x_1, \ldots, x_m) = h(x_1, \ldots, x_m)$.
Dále, pro $c = 0, 1, \ldots, m$ označme

$$\sigma_c^2 = \operatorname{Var} h_c(X_1, \ldots, X_c). \tag{1.5}$$

Definice 1.3. V případě, že $\sigma_1^2 = 0$, řekneme, že U_n je *degenerovaná U-statistika* a příslušné jádro je *degenerované jádro*. V opačné situaci říkáme, že U_n a příslušné jádro jsou *nedegenerované*.
Číslo k, které splňuje $\sigma_j^2 = 0$ pro všechna $j < k$ a $\sigma_k^2 > 0$ se nazývá *řád degenerovanosti*.

Rozptyl U_n je daný následujícím vztahem:

$$\operatorname{Var} U_n = \operatorname{Var}\left(\binom{n}{m}^{-1} \sum_{\{i_1,\ldots,i_m\}\in\mathbf{C}_{m,n}} h(X_{i_1},\ldots,X_{i_m})\right)$$
$$= \binom{n}{m}^{-2} \sum_{\{i_1,\ldots,i_m\}\in\mathbf{C}_{m,n}} \sum_{\{j_1,\ldots,j_m\}\in\mathbf{C}_{m,n}} \operatorname{Cov}\left(h(X_{i_1},\ldots,X_{i_m}), h(X_{j_1},\ldots,X_{j_m})\right).$$

Dá se ukázat, že pro libovolnou dvojici kombinací $\{i_1,\ldots,i_m\},\{j_1,\ldots,j_m\} \in \mathbf{C}_{m,n}$ je kovariance $\operatorname{Cov}\left(h(X_{i_1},\ldots,X_{i_m}), h(X_{j_1},\ldots,X_{j_m})\right)$ rovna σ_c^2 (které je definované vztahem (1.5)), kde c označuje počet indexů, které jsou společné pro obě kombinace $\{i_1,\ldots,i_m\}$ a $\{j_1,\ldots,j_m\}$.

Z toho se dá dále odvodit, že rozptyl U-statistiky U_n je daný výrazem uvedeným v následující větě.

Věta 1.1. *Nechť U_n je U-statistika tvaru (1.3). Potom*

$$\operatorname{Var} U_n = \binom{n}{m}^{-1} \sum_{c=1}^{m} \binom{m}{c}\binom{n-m}{m-c}\sigma_c^2. \tag{1.6}$$

Důkaz. Viz např. [Lee 1990], str. 12–13

1.1.4 Hoeffdingova dekompozice

Reprezentace U-statistik, kterou zde uvedeme, lze považovat za užitečný nástroj při studiu asymptotických vlastností U-statistik založených na výběru nezávislých stejně rozdělených náhodných veličin. Je zobecněním rozšířené projekční techniky používané v mnohých oblastech neparametrické statistiky. Předtím než zavedeme zmíněnou reprezentaci, definujeme nejdříve projekční jádra $h^{(0)},\ldots,h^{(m)}$.

Nechť H_x označuje distribuční funkci rozdělení koncentrovaného v bodě x. Jádra $h^{(c)}$ pro $c = 1\ldots,m$ pak definujeme jako

$$h^{(c)}(x_1,\ldots,x_c) = \int\cdots\int h(t_1,\ldots,t_m)\prod_{i=1}^{c}(dH_{x_i}(t_i) - dF(t_i))\prod_{j=c+1}^{m} dF(t_j). \tag{1.7}$$

Věta 1.2. *Pro $j = 1,\ldots,m$ nechť $U_n^{(j)}$ je U-statistika založena na jádře $h^{(j)}$, které je definované vztahem (1.7). Potom*

$$U_n = \operatorname{E} U_n + \sum_{j=1}^{m} \binom{m}{j} U_n^{(j)}. \tag{1.8}$$

Důkaz. Viz např. [Lee 1990], str. 25–26

Tato reprezentace je známá pod názvem *Hoeffdingova dekompozice* nebo také *H-dekompozice*. Její užitečnost je především v tom, že $U_n^{(j)}$ v Hoeffdingově dekompozici jsou všechny nekorelované statistiky. První člen je součtem nezávislých stejně rozdělených náhodných veličin a dalších $m-1$ členů jsou degenerované U-statistiky. Navíc, řády rozptylů (vzhledem k n) U-statistik $U_n^{(j)}$ klesají spolu s rostoucím j.

1.1.5 Asymptotické vlastnosti

Nedegenerované U-statistiky Asymptotické chování nedegenerovaných U-statistik je srovnatelné s asymptotickým chováním součtů nezávislých náhodných veličin. To je zjevné z Hoeffdingovy dekompozice U_n, která byla popsána v předchozí části podkapitoly. Pokud je člen $mU_n^{(1)} = m\sum_{i=1}^{n} h^{(1)}(X_i)$ nenulový, pro velké hodnoty n dominuje součtu U-statistik (danému vztahem (1.8)). Následující větu dokázal Hoeffding v jeho základním článku o U-statistikách (viz [Hoeffding 1948]).

Věta 1.3. *Předpokládejme, že* $\mathrm{E}\,(h(X_1,\ldots,X_m))^2 < \infty$ *a že* $\sigma_1^2 > 0$. *Potom* $\sqrt{n}(U_n - \theta)$ *má asymptoticky normální rozdělení s nulovou střední hodnotou a rozptylem* $m^2\sigma_1^2$.

Důkaz. Viz např. [Lee 1990], str. 76.

Degenerovaný případ Limitní teorie degenerovaných U-statistik se liší od nedegenerovaného případu a je mnohem složitější. Podrobněji je popsána například v knize [Lee 1990]. Zde se zaměříme pouze na případ, kdy je řád degenerovanosti U-statistiky rovný 1.

Pokud má zkoumaná U-statistika řád degenerovanosti rovný 1, jinými slovy: pokud jsme v situaci v situaci, kdy $\sigma_1^2 = 0$ a zároveň $\sigma_2^2 > 0$, je první člen Hoeffdingovy dekompozice (1.8) rovný nule s pravděpodobností 1. Z dekompozice je pak možné vyvodit, že $n(U_n - \mathrm{E}\,U_n)$ a $\binom{m}{2} nU_n^{(2)}$ mají stejné asymptotické chování. Asymptotické rozdělení U_n v takové situaci popisuje následující věta.

Věta 1.4. *Pokud* $\mathrm{E}\,(h(X_1,\ldots,X_m))^2 < \infty$ *a* $\sigma_1^2 = 0 < \sigma_2^2$, *pak* $n(U_n - \theta)$ *konverguje v distribuci k* $Ym(m-1)/2$, *kde* Y *je náhodní veličina tvaru*

$$Y = \sum_{i=1}^{\infty} \lambda_j(\chi_{1j}^2 - 1),$$

kde $\chi^2_{11},\chi^2_{12},\ldots$ *jsou nezávislé náhodné veličiny s* χ^2_1*- rozdělením a* λ_j *jsou vlastní čísla integrální rovnice*

$$\int h^{(2)}(x_1,x_2)f(x_2)dF(x_2) = \lambda f(x_1),$$

pro $h^{(2)}$*, které jsme v předchozí části podkapitoly definovali rovnicí (1.7).*

Důkaz. Věta je převzata z knihy [Serfling 1980], str. 193–199.

Z věty plyne, že pokud chceme určit limitní rozdělení degenerované U-statistiky, musíme nejdříve určit vlastní čísla integrální rovnice. V případě, že neznáme rozdělení výběru, lze rozdělení aproximovat například pomocí tzv. *resampling metod*.

Silný zákon velkých čísel pro U-statistiky Další důležitou vlastností U-statistik je, že jsou nejen nestrannými, ale i *silně konzistentními* odhady příslušných parametrů, jinými slovy: U-statistiky konvergují k odhadovanému parametru, skoro jistě. Následující věta je zobecněním silného zákona velkých čísel pro případ U-statistik. Platí pro nedegenerované i pro degenerované U-statistiky.

Věta 1.5. *Předpokládejme, že* $\mathrm{E}\,|h(X_1,\ldots,X_m)| < \infty$*. Potom* U_n *konverguje k* $\theta = \mathrm{E}\,h(X_1,\ldots,X_m)$*, skoro jistě.*

Důkaz. Věta je převzata z knihy [Lee 1990], str. 122.

1.2 Zobecněné U-statistiky

Pojem U-statistik je možné zobecnit i na více než jeden výběr. Pro takovou třídu statistik se používá označení *zobecněné U-statistiky*. Protože dále v textu budeme studovat především situaci, kde se srovnávají dva náhodné výběry, omezíme se v této podkapitole pouze na popis dvouvýběrových zobecněných U-statistik.

1.2.1 Definice a základní vlastnosti

Předpokládejme, že máme výběry $X_1,\ldots,X_{n_1}$ a $Y_1,\ldots,Y_{n_2}$ nezávislých reálných náhodných veličin z rozdělení charakterizovaného distribuční funkcí F, respektive G. Analogicky k jednovýběrovému případu, uvažujme funkcionál $\theta(F,G)$ s hodnotami na reálné přímce, jehož definičním oborem $\mathcal{F}\times\mathcal{G}$.

Definice 1.4. Řekneme, že $\theta = \theta(F,G)$ je *regulární funkcionál* nad $\mathcal{F} \times \mathcal{G}$ pokud pro všechna $(F,G) \in \mathcal{F} \times \mathcal{G}$ je možné reprezentovat $\theta(F,G)$ jako

$$\begin{aligned}\theta(F,G) &= \mathrm{E}_{F,G} h(X_1,\ldots,X_{m_1},Y_1,\ldots,Y_{m_2}) \\ &= \int \cdots \int h(x_1,\ldots,x_{m_1};y_1,\ldots,y_{m_2}) dF(x_1)\ldots dF(x_{m_1}) dG(y_1)\ldots dG(y_{m_2})\end{aligned}$$

pro nějakou funkci $h = h(x_1,\ldots,x_{m_1};y_1,\ldots,y_{m_2})$, kterou, stejně jako v jednovýběrovém případě, budeme nazývat *jádro*.

Poznámka 1.2. Bez újmy na obecnosti lze předpokládat, že h je symetrická vzhledem k permutacím uvnitř obou skupin argumentů. Nesymetrickou h lze nahradit symetrickou funkcí

$$\begin{aligned}&h'(x_1,\ldots,x_{m_1};y_1,\ldots,y_{m_2}) \\ &\qquad = \frac{1}{m_1! m_2!} \sum_{\mathbf{P}_{m_1}} \sum_{\mathbf{P}_{m_2}} h(x_{i_1},\ldots,x_{i_{m_1}};y_{j_1},\ldots,y_{j_{m_2}}).\end{aligned}$$

kde $\sum_{\mathbf{P}_{m_1}}$ značí součet přes všech $m_1!$ permutací $(i_1,\ldots,i_{m_1})$ prvků množiny $\{1,\ldots,m_1\}$. Analogicky, $\sum_{\mathbf{P}_{m_2}}$ je součet přes všechny permutace $(j_1,\ldots,j_{m_2})$ prvků množiny $\{1,\ldots,m_2\}$.

Definice 1.5. Předpokládejme, že $X_1,\ldots,X_{n_1}$ a $Y_1,\ldots,Y_{n_2}$ jsou dva nezávislé výběry skládající se z nezávislých stejně rozdělených náhodných veličin a nechť F a G jsou příslušné distribuční funkce. Předpokládejme dále, že X_i a Y_j jsou nezávislé pro všechna $i = 1,\ldots,m_1$ a $j = 1,\ldots,m_2$. Potom pro funkci $h(x_1,\ldots,x_{m_1};y_1,\ldots,y_{m_2})$, symetrickou uvnitř obou skupin svých argumentů $x_1,\ldots,x_{m_1}$ a $y_1,\ldots,y_{m_2}$, definujeme příslušnou *zobecněnou U-statistiku s jádrem h* vztahem

$$U_{n_1,n_2} = \binom{n_1}{m_1}^{-1} \binom{n_2}{m_2}^{-1} \sum_{\mathbf{C}_{m_1,n_1}} \sum_{\mathbf{C}_{m_2,n_2}} h(X_{i_1},\ldots,X_{i_{m_1}};Y_{j_1},\ldots,Y_{j_{m_2}}), \quad (1.9)$$

kde součty přecházejí přes množiny $\mathbf{C}_{m_1,n_1}$ a $\mathbf{C}_{m_1,n_1}$ všech kombinací m_1 prvků $1 \le i_1 < \cdots < i_{m_1} \le n_1$, respektive m_2 prvků $1 \le j_1 < \cdots < j_{m_2} \le n_2$.

Nestrannost Stejně jako v případě jednovýběrové U-statistiky, je i dvouvýběrová U-statistika U_{n_1,n_2} nestranným odhadem

$$\theta(F,G) = \mathrm{E}_{F,G} h(X_1,\ldots,X_{m_1};\, Y_1,\ldots,Y_{m_2})$$

za předpokladu, že střední hodnota $\mathrm{E}_{F,G} h(X_1,\ldots,X_{m_1};\, Y_1,\ldots,Y_{m_2})$ je konečné číslo.

1.2.2 Rozptyl

Předtím než uvedeme vztah pro rozptyl zobecněné U-statistiky, analogický vztahu pro jednovýběrovou U-statistiku, definujme nejdříve funkce $h_{c,d}$ analogické k funkcím h_c v jednovýběrovém případě (viz (1.4)):

$$\begin{aligned} h_{c,d} &= (x_1,\dots,x_c;y_1,\dots,y_d) \\ &= \mathrm{E}\, h(x_1,\dots,x_c,X_{c+1},\dots,X_{k_1};y_1,\dots,y_d,Y_{d+1},\dots,Y_{k_2}). \end{aligned} \tag{1.10}$$

Označme dále $\sigma^2_{c,d}$ rozptyl těchto funkcí:

$$\sigma^2_{c,d} = \mathrm{Var}\, h_{c,d}(X_1,\dots,X_c;Y_1,\dots,Y_d). \tag{1.11}$$

Užitím podobných argumentů jako v jednovýběrovém případě se dá ukázat následující věta.

Věta 1.6. *Nechť U_{n_1,n_2} zobecněná U-statistika tvaru (1.9).*
Potom

$$\mathrm{Var}\, U_{n_1,n_2} = \sum_{c=0}^{m_1}\sum_{d=0}^{m_2} \frac{\binom{m_1}{c}\binom{m_2}{d}\binom{n_1-m_1}{m_1-c}\binom{n_2-m_2}{m_2-d}}{\binom{n_1}{m_1}\binom{n_2}{m_2}}\sigma^2_{c,d}. \tag{1.12}$$

Důkaz. Lze najít například v knize [Lee 1990].

1.2.3 Hoeffdingova dekompozice

Pro účely odvození Hoeffdingovy dekompozice pro dvouvýběrové U-statistiky, uveďme analogii funkcí $h^{(c)}$ definovaných vztahem (1.7).
Nechť opět H_x označuje distribuční funkci rozdělení koncentrovaného v bodě x. Funkce $h^{(c,d)}$ pro $c = 0,\dots,m_1$ a $d = 0,\dots,m_2$ pak definujeme vztahem

$$\begin{aligned} &h^{(c,d)}(x_1,\dots,x_c;y_1,\dots,y_d) \\ &= \int\cdots\int h(s_1,\dots,s_{m_1};t_1,\dots,t_{m_2}) \prod_{i=1}^{c}(dH_{x_i}(s_i) - dF(s_i)) \prod_{j=c+1}^{m_1} dF_1(s_j) \\ &\qquad \times \prod_{k=1}^{d}(dH_{y_k}(t_k) - dG(t_k)) \prod_{l=d+1}^{m_2} dF_2(t_l). \end{aligned} \tag{1.13}$$

Věta 1.7. *Každou zobecněnou U-statistiku je možné reprezentovat jako součet*

$$U_{n_1,n_2} = \sum_{c=0}^{m_1} \sum_{d=0}^{m_2} \binom{m_1}{c} \binom{m_2}{d} U_{n_1,n_2}^{(c,d)},$$

kde $U_{n_1,n_2}^{(c,d)}$ je zobecněná U-statistika založena na jádře $h^{(c,d)}$ a je definována vztahem

$$U_{n_1,n_2}^{(c,d)} = \binom{n_1}{c}^{-1} \binom{n_2}{d}^{-1} \sum_{\mathbf{C}_{c,n_1}} \sum_{\mathbf{C}_{d,n_2}} h^{(c,d)}(X_{i_1}, \ldots, X_{i_c}; Y_{j_1}, \ldots, Y_{j_d}).$$

pro $c = 0, \ldots, m_1$ a $d = 0, \ldots, m_2$.

Důkaz. Věta je převzata z knihy [Lee 1990], str. 40.

Analogicky k případu jednovýběrové U-statistiky dále platí, že zobecněné U-statistiky $U_{n_1,n_2}^{(c,d)}$ jsou nekorelované.

1.2.4 Asymptotické vlastnosti

Při výše uvedeném značení, každá U-statistika s $m = 2$ může být ekvivalentně zapsaná jako

$$U_{n_1,n_2} = \mathrm{E}\, U_{n_1,n_2} + m_1 U_{n_1,n_2}^{(1,0)} + m_2 U_{n_1,n_2}^{(0,1)} + R_{n_1,n_2}$$

kde

$$U_{n_1,n_2}^{(1,0)} = \frac{1}{n_1} \sum_{i=1}^{n_1} h^{(1,0)}(X_i) \qquad U_{n_1,n_2}^{(0,1)} = \frac{1}{n_2} \sum_{j=1}^{n_2} h^{(0,1)}(Y_j)$$

a $\mathrm{Var}\, R_{n_1,n_2} = o(N^{-1})$ (viz [Lee 1990], str. 140), pro $N = n_1 + n_2$. Statistiky $\sqrt{n_1} U_{n_1,n_2}^{(1,0)}$ a $\sqrt{n_2} U_{n_1,n_2}^{(0,1)}$ obě konvergují k normálním rozdělením se středními hodnotami rovnými 0 a rozptyly $\sigma_{1,0}^2$, respektive $\sigma_{0,1}^2$ a jsou navzájem nezávislé. Platí následující věta:

Věta 1.8. *Nechť U_{n_1,n_2} je zobecněná U-statistika založená na dvou nezávislých výběrech $X_1, \ldots, X_{n_1}$ a $Y_1, \ldots, Y_{n_2}$ s jádrem h, které splňuje*

$$\mathrm{E}\, \left(h(X_1, \ldots, X_{m_1}; Y_1, \ldots, Y_{m_2})\right)^2 < \infty. \tag{1.14}$$

Předpokládejme, že platí

$$\sigma_{1,0}^2 = \mathrm{Var}\, h^{(1,0)}(X_1) > 0 \tag{1.15}$$

a

$$\sigma_{0,1}^2 = \operatorname{Var} h^{(0,1)}(Y_1) > 0, \tag{1.16}$$

kde funkce $h^{(1,0)}$ *a* $h^{(0,1)}$ *jsou definovány vztahem (1.13). Dále předpokládejme, že*

$$\frac{n_1}{N} \to p, \tag{1.17}$$

pro konstantu $p \in (0,1)$.
Potom

$$\sqrt{N}(U_{n_1,n_2} - \theta) \xrightarrow{\mathcal{D}} N\left(0,\sigma_{as}^2\right)$$

pro $N \to \infty$, *kde*

$$\sigma_{as}^2 = \frac{1}{p} m_1^2 \sigma_{1,0}^2 + \frac{1}{1-p} m_2^2 \sigma_{0,1}^2. \tag{1.18}$$

Důkaz. Věta je převzata z knihy [Lee 1990], str. 141. Důkaz je naznačen v úvaze, která předchází vyslovení věty.

Silný zákon velkých čísel pro dvouvýběrové U-statistiky Silný zákon velkých čísel patří mezi ty vlastnosti jednovýběrových U-statistik, které nelze přímočaře zobecnit na dvouvýběrový případ. Platí však následující věta

Věta 1.9. *Nechť* $\mathrm{E}\,|h(X_1,\ldots,X_{m_1};Y_1,\ldots,Y_{m_2})| < \infty$ a $n_1 = n_2$. *Potom* U_{n_1,n_2} *konverguje k* $\theta = \mathrm{E}\,h(X_1,\ldots,X_{m_1};Y_1,\ldots,Y_{m_2})$, *skoro jistě.*

Důkaz. Věta je převzata z práce [Janssen 1988], str. 9

Poznámka 1.3. Bylo též ukázáno (viz [Janssen 1988], str. 9), že pro výběry nestejné velikosti má dvouvýběrová U-statistika U_{n_1,n_2} vlastnost silné konzistence, pokud

$$\mathrm{E}\left(|h(X_1,\ldots,X_{m_1};Y_1,\ldots,Y_{m_2})| \log^+ |h(X_1,\ldots,X_{m_1};Y_1,\ldots,Y_{m_2})|\right) < \infty$$

pro $n_1,n_2 \to \infty$.

1.3 V-statistiky

V-statistiky jsou velice blízkou třídou ke třídě U-statistik. Ve srovnání s U-statistikami, obsahují V-statistiky navíc členy $h(X_{i_1},\ldots,X_{i_m})$, kde index i_j může být roven indexu i_k pro nějaké $j \neq k$. Normovací konstanta je proto jiná: n^m místo $\binom{n}{m}$, kde n je velikost výběru a m označuje počet argumentů jádra h. V souvislosti s U-statistikami se s V-statistikami setkáváme například při aplikaci metody bootstrap s vracením.

1.3.1 Jednovýběrové V-statistiky

Definice 1.6. Pro symetrickou funkci $h(x_1,\dots,x_m)$ se příslušná *V-statistika* na základě výběru $X_1,\dots,X_n$ velikosti $n \geq m$ definuje jako

$$V_n = \frac{1}{n^m} \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} h(X_{i_1},\dots,X_{i_m}).$$

Ekvivalentně lze V-statistiky vyjádřit také jako střední hodnotu z jádra h vzhledem k empirickému rozdělení:

$$V_n = \int \cdots \int h(x_1,\dots,x_m) dF_n(x_1) \dots dF_n(x_m) = \theta(F_n),$$

kde F_n označuje empirickou distribuční funkci na základě výběru $X_1,\dots,X_n$:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i \leq x\}}.$$

Je tedy možné se na V_n dívat také jako na empirický protějšek $\theta(F)$.

Za určitých podmínek má U-statistika i V-statistika, založená na stejném jádře, stejné limitní rozdělení. Tato situace je popsána v následující větě:

Věta 1.10. *Předpokládejme, že máme náhodný výběr $X_1,\dots,X_n$ z rozdělení s distribuční funkcí F a že jádro $h(x_1,\dots,x_m)$ splňuje podmínku*

$$\max_{1 \leq k \leq m} \left[\frac{1}{k^m} \sum_{i_1=1}^{k} \cdots \sum_{i_m=1}^{k} h(X_{i_1},\dots,X_{i_m}) \right] < \infty.$$

Potom pro všechna n je $\mathrm{E}\, V_n < \infty$ a platí

$$V_n - U_n = O(n^{-1})$$

s pravděpodobností 1.[2]

Důkaz. Věta je převzata z článku [Bönner a Kirschner 1977].

[2]Poznamenejme, že $x_n = O(a_n)$, kde x_n a a_n jsou posloupnosti reálných čísel, používáme pro označení situace, kdy existuje konstanta $C \in \mathbb{R}$ taková, že $|x_n| \leq C a_n \forall n \in \mathbb{N}$.

1.3.2 Zobecněné V-statistiky

Stejným způsobem, jakým byl zobecněn pojem U-statistik na dva výběry, lze zobecnit i pojem V-statistik. Pro dva navzájem nezávislé náhodné výběry $X_1,\dots,X_{n_1}$ a $Y_1,\dots,Y_{n_2}$ tak dostaneme *zobecněnou V-statistiku*

$$V_{n_1,n_2} = \frac{1}{n_1^{m_1} n_2^{m_2}} \sum_{i_1=1}^{n_1} \cdots \sum_{i_{m_1}=1}^{n_1} \sum_{j_1=1}^{n_2} \cdots \sum_{j_{m_2}=1}^{n_2} h(X_{i_1},\dots,X_{i_{m_1}};Y_{j_1},\dots,Y_{j_{m_2}}).$$

kde $h(x_1,\dots,x_{m_1};y_1,\dots,y_{m_2})$ je symetrická uvnitř obou skupin svých argumentů.

Předpokládejme, že máme zobecněnou U-statistiku založenou na dvou výběrech $X_1,\dots,X_{n_1}$ a $Y_1,\dots,Y_{n_2}$ a zobecněnou V-statistiku se stejným jádrem, založenou na stejných výběrech. Pak platí (viz [Puri a Sen 1971], str. 65), že

$$|U_{n_1,n_2} - V_{n_1,n_2}| = O_P(n_{min}^{-1}), \quad \text{pro } n_{min} = \min\{n_1,n_2\}.^3$$

[3]Poznamenejme dále, že $X_n = O_P(a_n)$, kde X_n je posloupnost náhodných veličin a a_n posloupnost reálných čísel, používáme pro označení situace, kdy je $\lim_{C\to\infty} \sup_{n\in\mathbb{N}} \mathrm{P}(|X_n| > Ca_n) = 0$.

Kapitola 2

Obecný tvar testové statistiky

V literatuře je možné najít mnoho velmi rozmanitých testů, kde se testová statistika dá vyjádřit jako U-statistika nebo funkce jedné či několika U-statistik. V této kapitole se zaměříme na problém dvouvýběrových testů. Jedním z nejznámějších dvouvýběrových testů založených na U-statistice je Mannův-Whiteyův (respektive ekvivalentní dvouvýběrový Wilcoxonův test). Existuje i mnoho testů, kde souvislost s U-statistikami není na první pohled zřejmá. Jedná se například o testy založené na empirických charakteristických funkcích, které byly popsány v článcích [Meintanis 2005] a [Hušková a Meintanis 2006].

Testováním hypotézy, že rozdělení dvou náhodných veličin jsou stejná, proti alternativě, že se nějakým způsobem liší, se zabývá parametrická i neparametrická statistika. Zde se zaměříme na neparametrické testy a situaci obecné alternativy. Pro tuto situaci se používají dvouvýběrové U-statistiky typu (1.9) s vlastností $m = m_1 = m_2$.

2.1 Obecný tvar testu

Uvažujme dva náhodné výběry $X_1, \ldots, X_{n_1}$ a $Y_1, \ldots, Y_{n_2}$ nezávislých stejně rozdělených náhodných veličin charakterizované distribuční funkcí $F(x) = P(X_i \leq x)$ pro $i = 1, \ldots, n_1$, respektive $G(y) = P(Y_j \leq y)$ pro $j = 1, \ldots, n_2$. Budeme studovat obecný problém, kdy testujeme nulovou hypotézu

$$H_0 : \quad F \equiv G \tag{2.1}$$

proti obecné alternativě

$$H_1 : \quad F \not\equiv G. \tag{2.2}$$

Základem testu bude zobecněná dvouvýběrová statistika, která srovnává po-

zorování z obou výběrů pomocí jádra $h(x_1,\ldots,x_m;y_1,\ldots,y_m)$, tedy U-statistika tvaru

$$U_{n_1,n_2} = \binom{n_1}{m}^{-1}\binom{n_2}{m}^{-1}\sum_{\mathbf{C}_{m,n_1}}\sum_{\mathbf{C}_{m,n_2}} h(X_{i_1},\ldots,X_{i_m};Y_{j_1},\ldots,Y_{j_m}). \qquad (2.3)$$

Budeme předpokládat, že střední hodnota

$$\mathrm{E}\, h(X_1,\ldots,X_m;Y_1,\ldots,Y_m) = \mathrm{E}\, U_{n_1,n_2}$$

je za platnosti nulové hypotézy rovna 0. Připomeňme jen, že každou U-statistiku se známou střední hodnotou θ_0 za platnosti nulové hypotézy lze převést na tento případ jednoduchou úpravou jádra (místo $h(x_1,\ldots,x_m;y_1,\ldots,y_m)$ se zvolí jádro $h'(x_1,\ldots,x_m;y_1,\ldots,y_m) = h(x_1,\ldots,x_m;y_1,\ldots,y_m) - \theta_0$). Navíc budeme předpokládat, že za platnosti alternativy nabývá $\mathrm{E}\, U_{n_1,n_2}$ hodnoty různé od nuly. Kombinace těchto dvou předpokladů poskytuje tu vlastnost, že se nulová hypotéza bude zamítat pro velké hodnoty $|U_{n_1,n_2}|$.

Pro určení kritického oboru testu založeného na U-statistice (2.3) se zpravidla používá aproximace známým limitním rozdělením nebo některá z tzv. *resampling metod* (například metoda bootstrap nebo permutační testy). V některých případech (například u Mannova-Whitneyova testu pro malé hodnoty n_1 a n_2 – viz podkapitola 3.1) se kritické hodnoty rozdělení dají odvodit, nebo najít v tabulkách.

Jak bylo popsáno v kapitole 2, U-statistika (2.3) má za předpokladu, že hodnoty $\sigma^2_{1,0}$ a $\sigma^2_{0,1}$ jsou obě větší než 0 a podíl n_1/N konverguje ke konstantě $p \in (0,1)$, normální rozdělení se střední hodnotou $\mathrm{E}\, h(X_1,\ldots,X_m;Y_1,\ldots,Y_m)$ a rozptylem $\sigma^2_{as} = p^{-1}m^2\sigma^2_{1,0} + (1-p)^{-1}m^2\sigma^2_{0,1}$ (viz věta 1.8).

Pokud je známý rozptyl $\mathrm{Var}\, U_{n_1,n_2}$ nebo rozptyl σ^2_{as} za platnosti nulové hypotézy, kritérium pro zamítnutí nulové hypotézy bude

$$|U_{n_1,n_2}| \ \geq u_{1-\alpha/2}\sqrt{\mathrm{Var}\, U_{n_1,n_2}},$$

respektive

$$|U_{n_1,n_2}| \ \geq u_{1-\alpha/2}\sqrt{\frac{\sigma^2_{as}}{N}},$$

kde $u_{1-\alpha/2}$ označuje $(1-\alpha/2)$-kvantil normovaného normálního rozdělení pro danou hladinu významnosti α.

Pokud rozptyl není známý, použije se vhodný odhad pro rozptyl $\mathrm{Var}\, U_{n_1,n_2}$, respektive pro rozptyl limitního rozdělení σ^2_{as}. Problematika odhadů rozptylu pro dvouvýběrové U-statistiky tvaru (2.3) je blíže popsaná v podkapitole 2.3. Při použití odhadu $\widehat{\sigma}^2_{n_1,n_2}$ pro hodnotu σ^2_{as}, který je konzistentní za platnosti nulové hypotézy i za platnosti alternativy, bude mít náhodná veličina

$U_{n_1,n_2}/\widehat{\sigma}_{n_1,n_2}$ za platnosti nulové hypotézy normální rozdělení se střední hodnotou 0 a rozptylem 1. Kritérium pro zamítnutí nulové hypotézy bude mít v tomto případě tvar

$$|U_{n_1,n_2}| \geq u_{1-\alpha/2}\sqrt{\frac{\widehat{\sigma}^2_{n_1,n_2}}{N}}.$$

V případě, kdy nejsou splněny podmínky věty o normálním rozdělení dvouvýběrových U-statistik, je situace složitější a liší se případ od případu. Reprezentantem této třídy dvouvýběrových testů je test založený na vzdálenosti empirických distribučních funkcí (viz podkapitola 3.5) a patří sem i testy založené na empirických charakteristických funkcích (viz kapitola 4). V obou těchto případech má testová statistika $\frac{n_1 n_2}{N} U^{edf}_{n_1,n_2}$, respektive $\frac{n_1 n_2}{N} U^{ecf}_{n_1,n_2}(w)$, za platnosti nulové hypotézy degenerované rozdělení. U testu založeného na vzdálenosti empirických distribučních funkcí je limitní rozdělení známé a jeho kvantily lze najít v tabulkách. U testů založených na empirických charakteristických funkcích není známá explicitní forma limitního rozdělení. Za platnosti alternativy konverguje hodnota $\frac{n_1 n_2}{N}|U_{n_1,n_2}|$ v obou případech do nekonečna.

2.2 Speciální tvar testové statistiky – rozdíl dvou jednovýběrových U-statistik

Některé dvouvýběrové testy jsou založeny na srovnání hodnot určitého parametru pro rozdělení výběrů (viz [Ferger 2004]). Hodnotu tohoto parametru, pro rozdělení s distribuční funkcí H, budeme značit symbolem $\tau(H)$. Například může jít o srovnání středních hodnot, rozptylů, nebo dalších momentů. Předpokládá se přitom, že $\tau(H)$ je odhadnutelný parametr. Lze ho tedy vyjádřit jako

$$\tau(H) = \int \cdots \int h_\tau(x_1,\ldots,x_m) dH(x_1)\ldots dH(x_m)$$

pro nějakou funkci h_τ, která je symetrická vzhledem k permutacím svých argumentů.
Uvažujme tedy U-statistiku tvaru

$$D_{n_1,n_2} = \binom{n_1}{m}^{-1} \sum_{\mathbf{C}_{m,n_1}} h_\tau(X_{i_1},\ldots,X_{i_m}) - \binom{n_2}{m}^{-1} \sum_{\mathbf{C}_{m,n_2}} h_\tau(Y_{j_1},\ldots,Y_{j_m})$$
$$= T^X_{n_1} - T^Y_{n_2}, \quad (2.4)$$

kde $T^X_{n_1}$ označuje U-statistiku s jádrem h_τ, vypočtenou na základě výběru $X_1,\ldots,X_{n_1}$ a $T^Y_{n_2}$ označuje U-statistiku se stejným jádrem, vypočtenou na základě výběru $Y_1,\ldots,Y_{n_2}$. D_{n_1,n_2} je potom nestranným odhadem pro

$$\theta(F,G) = \tau(F) - \tau(G).$$

Pokud dále označíme

$$h_\theta(x_1,\ldots,x_m;y_1,\ldots,y_m) = h_\tau(x_1,\ldots,x_m) - h_\tau(y_1,\ldots,y_m),$$

vidíme, že se jedná o speciální tvar testové statistiky (2.3), s jádrem, které je antisymetrické vzhledem k vzájemné výměně obou skupin argumentů, tj.

$$h_\theta(x_1,\ldots,x_m;y_1,\ldots y_m) = -h_\theta(y_1,\ldots y_m;x_1,\ldots,x_m).$$

Poznámka 2.1. Všimněme si, že obě degenerované projekce U-statistiky (2.4), které určují limitní rozdělení v případě popsané ve větě 1.8, jsou zároveň degenerovanou projekcí $T^X_{n_1}$, respektive $T^Y_{n_2}$ vzhledem k příslušnému rozdělení. Pro lepší názornost uvedeme jejich tvar pro $m = 2$:

$$h^{(1,0)}_\theta(x) = h^1_\tau(x) = \mathrm{E}_F h_\tau(x,X_1) - \mathrm{E}\, h_\tau(X_1,X_2).$$

$$h^{(0,1)}_\theta(y) = h^1_\tau(y) = \mathrm{E}_G h_\tau(y,Y_1) - \mathrm{E}\, h_\tau(Y_1,Y_2).$$

Pro jiné hodnoty m se jádra odvodí analogicky: $h^{(1,0)}$ bude vždy záviset pouze na rozdělení s distribuční funkcí F a $h^{(0,1)}$ pouze na na rozdělení s distribuční funkcí G. Z výše uvedené úvahy plyne, že pokud jsou splněny podmínky (1.14)-(1.17), D_{n_1,n_2} bude mít normální rozdělení s asymptotickým rozptylem

$$\sigma^2_{as} = \frac{\sigma^2_F}{p} + \frac{\sigma^2_G}{1-p} \tag{2.5}$$

kde σ^2_F označuje asymptotický rozptyl jednovýběrové U-statistiky $T^X_{n_1}$, která je funkcí výběru $X_1,\ldots,X_{n_1}$ s distribuční funkcí F. Podobně, σ^2_G je asymptotický rozptyl U-statistiky $T^Y_{n_2}$, určené na základě výběru s rozdělením G.

2.3 Odhad rozptylu pro U-statistiky pomocí metody jackknife

V následující podkapitole popíšeme metodu jackknife odhadů rozptylu pro U-statistiky pro situaci, kdy dvouvýběrová U-statistika konverguje k normálnímu rozdělení.

2.3.1 Obecná dvouvýběrová U-statistika

Tuto metodu odhadu rozptylu pro jednovýběrové U-statistiky rozpracoval Arvesen v roce 1969 (viz [Arvesen 1969]), avšak v ekvivalentní formě ji odvodil Sen v roce 1960 (viz [Puri a Sen 1971], str. 59).

Zde uvedeme zobecnění na dvouvýběrové statistiky, které je příbuzné s Arvesenovym odvozením jackknife odhadu pro jednovýběrové U-statistiky.[1] Jedná se o odhad rozptylu pro situaci popsanou ve větě 1.8, tj. případ, kdy zobecněná U-statistika s jádrem $h(x_1,\dots,x_m;y_1,\dots,y_m)$ konverguje k normálnímu rozdělení se střední hodnotou $\theta = \mathrm{E}\, h(X_1,\dots,X_m;Y_1,\dots,Y_m)$ a rozptylem σ^2_{as}. Označme

$$S^{(1,0)}_{n_1,n_2} = \frac{1}{n_1}\sum_{k=1}^{n_1}\left(\binom{n_1-1}{m-1}^{-1}\binom{n_2}{m}^{-1}\sum h(X_k,X_{i_2},\dots,X_{i_m};Y_{j_1},\dots,Y_{j_m})\right)^2$$

kde druhá suma prochází všechna $1 \le i_2 < \cdots < i_m \le n_1$ s vlastností $i_c \neq k$, $c = 2,\dots,m$ a všechna $1 \le j_1 < \cdots < j_m \le n_2$. Dále označme

$$S^{(0,1)}_{n_1,n_2} = \frac{1}{n_2}\sum_{k=1}^{n_2}\left(\binom{n_1}{m}^{-1}\binom{n_2-1}{m-1}^{-1}\sum h(X_{i_1},\dots,X_{i_m};Y_k,Y_{j_2},\dots,Y_{j_m})\right)^2$$

kde druhá suma prochází všechna $1 \le i_1 < \cdots < i_m \le n_1$ a všechna $1 \le j_2 < \cdots < j_m \le n_2$ s vlastností $j_d \neq k$, $d = 2,\dots,m$.
Následující lemma obsahuje tvrzení o konzistenci této metody odhadu pro asymptotický rozptyl σ^2_{as}. Tvrzení platí za platnosti nulové hypotézy i za platnosti alternativy.

Lemma 2.1. *Předpokládejme, že $n_1/N \to p$ pro konstantu $p \in (0,1)$ a že je splněna podmínka*

$$\mathrm{E}\,(h(X_1,\dots,X_m;Y_1,\dots,Y_m))^2 < \infty.$$

Potom

$$\widehat{\sigma}^2_{n_1,n_2} = \frac{N}{n_1}m^2\left(S^{(1,0)}_{n_1,n_2} - U^2_{n_1,n_2}\right) + \frac{N}{n_2}m^2\left(S^{(0,1)}_{n_1,n_2} - U^2_{n_1,n_2}\right) \qquad (2.6)$$

je slabě konzistentním odhadem pro σ^2_{as}, tj. $\widehat{\sigma}^2_{n_1,n_2} \xrightarrow{P} \sigma^2_{as}$.

Důkaz. Lemma je převzato z článku [Ferger 2004].

[1] V knize [Puri a Sen 1971] (str. 66) je popsáno zobecnění Senovy metody jackknife odhadu rozptylu.

2.3.2 Rozdíl dvou jednovýběrových U-statistik

Z úvahy uvedené v Poznámce 2.1 plyne, že vhodným odhadem pro asymptotický rozptyl dvouvýběrové testové statistiky založené na rozdílu dvou jednovýběrových U-statistik je

$$\widehat{\sigma}^2_{n_1,n_2} = \frac{N}{n_1} m^2 \left(S^X_{n_1} - \left(T^X_{n_1}\right)^2 \right) + \frac{N}{n_2} m^2 \left(S^Y_{n_2} - \left(T^Y_{n_2}\right)^2 \right), \tag{2.7}$$

kde

$$S^X_{n_1} = \frac{1}{n_1} \sum_{k=1}^{n_1} \left(\binom{n_1 - 1}{m-1}^{-1} \sum h_\tau(X_k, X_{i_2}, \ldots, X_{i_m}) \right)^2$$

a

$$S^Y_{n_2} = \frac{1}{n_2} \sum_{k=1}^{n_2} \left(\binom{n_2 - 1}{m-1}^{-1} \sum h_\tau(Y_k, Y_{i_2}, \ldots, Y_{i_m}) \right)^2$$

jsou jednovýběrové jackknife odhady pro σ^2_F/m^2, respektive pro σ^2_G/m^2 (viz [Arvesen 1969] nebo [Lee 1990], str. 218). V obou případech součty procházejí všechna $1 \le i_2 < \cdots < i_m \le n_1$, respektive $1 \le i_2 < \cdots < i_m \le n_2$ s vlastností $i_c \neq k$.

Poznámka 2.2. Jedná se o odhad, který je identický s odhadem určeným na základě metody uvedené v předchozím odstavci. Stejný tvar odhadu bychom dostali po několika jednoduchých úpravách. Konzistence odhadu tak plyne z lemmatu 2.1.

Poznámka 2.3. V literatuře se lze setkat i s dalšími metodami odhadu rozptylu pro dvouvýběrové U-statistiky. Například u permutačních testů se používají odhady rozptylu, které jsou založené na sdruženém výběru (viz například [Sen 1967] nebo [Horváth a Hušková 2005]).

Kapitola 3

Konkrétní testy pro obecný dvouvýběrový problém

3.1 Mannův-Whitneyův test

Takzvaný dvouvýběrový Mannův-Whitneyův test je neparametrický test, který se požívá pro testování hypotézy (2.1) proti alternativě (2.2). Pro odvození vlastností testu se zpravidla předpokládá, že výběry $X_1, \ldots, X_{n_1}$ a $Y_1, \ldots, Y_{n_2}$ pocházejí ze spojitých rozdělení s distribuční funkcí F, respektive G.

3.1.1 Testová statistika

Mannův-Whitneyův test je založen na statistikách:

$$U_1 = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} h(X_i, Y_j)$$

a

$$U_2 = n_1 n_2 - \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} h(X_i, Y_j)$$

kde jádro h nabývá hodnoty

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

Poznámka 3.1. Mannův-Whitneyův test patří do kategorie testů založených na pořadích a je ekvivalentní s tzv. *dvouvýběrovým Wilcoxonovým testem.* Vzájemný vztah je vyjádřen rovnicemi

$$T_1 = n_1 n_2 + \frac{1}{2} n_1(n_1+1) - U_1$$
$$T_2 = n_1 n_2 + \frac{1}{2} n_2(n_2+1) - U_2$$

kde T_1 označuje součty pořadí hodnot $X_1,\ldots,X_{n_1}$ ve sdruženém výběru. Analogicky, T_2 je součet pořadí $Y_1,\ldots,Y_{n_2}$.

3.1.2 Určování kritických hodnot pro zamítnutí nulové hypotézy

Pro určení kritických hodnot Mannova-Whitneyova testu se používá několik metod. Pro malé hodnoty n_1 a n_2 lze kritické hodnoty určit přímo (viz [Jurečková 1981], str. 47). Využije se přitom vlastnost ekvivalence s dvouvýběrovým Wilcoxonovým testem. Pro všechny kombinace $s_1 < \cdots < s_{n_2}$ z čísel $1,\ldots,N$ se určí hodnota součtu $\sum_{i=1}^{n_2} s_i$ a získané hodnoty se vzestupně seřadí. Kritický obor pro test s hladinou α je pak tvořen M_{n_1,n_2} největšími hodnotami tohoto součtu, kde

$$M_{n_1,n_2} = \alpha \binom{N}{n_2}.$$

Pokud žádné M_{n_1,n_2} nesplňuje tuto podmínku, použije se pro určení kritické hodnoty metoda randomizace. Pro větší velikostí výběrů n_1 a n_2, pro něž by tento způsob byl početně příliš náročný, se hodnota $\min\{U_1,U_2\}$ srovnává s tabelovanými kritickými hodnotami. Pokud jsou oba výběry velikosti menší nebo rovné 40, lze kritické hodnoty najít například v [Likeš a Laga 1978], str. 374–407. Pokud pro daná n_1 a n_2 už nelze najít kritické hodnoty v tabulkách, používá se aproximace normálním rozdělením.
Příslušná zobecněná U-statistika má tvar

$$U^{MW}_{n_1,n_2} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} h(X_i,Y_j).$$

Za platnosti nulové hypotézy je střední hodnota $U^{MW}_{n_1,n_2}$ rovna $\mathrm{E}\, U^{MW}_{n_1,n_2} = \mathrm{P}(X < Y) = 1/2$. Dále se dá odvodit (viz např. [Anděl 2002]), že rozptyl za

platnosti nulové hypotézy je roven

$$\operatorname{Var} U_{n_1,n_2}^{MW} = \frac{N+1}{12n_1n_2}$$

a že statistika

$$\sqrt{\frac{12n_1n_2}{N+1}}\left(U_{n_1,n_2}^{MW} - \frac{1}{2}\right)$$

má asymptoticky normální rozdělení s nulovou střední hodnotou a jednotkovým rozptylem.
Nulovou hypotézu potom zamítneme pokud

$$U_{n_1,n_2}^{MW} \geq \frac{1}{2} + u_{1-\alpha/2}\sqrt{\frac{N+1}{12n_1n_2}}$$

$$\text{nebo} \quad U_{n_1,n_2}^{MW} \leq \frac{1}{2} - u_{1-\alpha/2}\sqrt{\frac{N+1}{12n_1n_2}}$$

kde $u_{1-\alpha/2}$ označuje $(1-\alpha/2)$-kvantil normovaného normálního rozdělení pro danou hladinu významnosti α.

3.1.3 Další charakteristiky testu

Z věty 1.8 dále plyne (viz [Lehmann 1951]), že U_{n_1,n_2}^{MW} má (za předpokladu, že n_1/N konverguje ke konstantě z intervalu (0,1)) normální rozdělení nejen za platnosti nulové hypotézy $F \equiv G$, ale i v případě, že $F \not\equiv G$. Zároveň konverguje v pravděpodobnosti k hodnotě

$$U_{n_1,n_2}^{MW} \xrightarrow{P} \mathrm{E}\, h(X_1,Y_1) = \mathrm{P}(X_1 < Y_1).$$

Původně byl Mannův-Whitneyův test navržen pro testování hypotézy $F \equiv G$ proti obecné alternativě $F \not\equiv G$. Je však citlivý zejména na tzv. *alternativu posunutí*, tj.

$$F(x) = G(x-\Delta), \quad \Delta \neq 0.$$

Proti alternativám s vlastností

$$F(x) < G(x) \qquad \forall x$$

je Mannův-Whitneyův *konzistentním* i *nestranným* testem (viz [Lehmann 1951]).
Proti alternativě posunutí je *lokálně nejsilnějším testem pro logistické rozdělení* (viz [Jurečková 1981], str. 46), tj. pro rozdělení s hustotou

$$f(x) = \frac{e^{-x}}{(1+e^{-x})^2} \qquad x \in \mathbb{R}.$$

3.2 Sukhatmův test

Jako příklad dvouvýběrového testu testu založeného na U-statistice můžeme uvést dále test, který odvodil Sukhatme (viz [Sukhatme 1957] nebo [Lee 1990]) a který se používá zejména pro testování hypotézy (2.1) proti alternativě (2.2) pro výběry z dvou absolutně spojitých rozdělení se stejným mediánem. Nejčastěji se používá v situacích, kdy se očekává, že za platnosti alternativy mají rozdělení různé rozptyly.

3.2.1 Testová statistika

Předpokládejme tedy, že F a G jsou dvě distribuční funkce a že obě mají medián rovný nule. Nechť $X_1, \ldots, X_{n_1}$ a $Y_1, \ldots, Y_{n_2}$ jsou dva náhodné výběry z F, respektive G. Testová statistika pro test hypotézy (2.1) proti alternativě (2.2) má v takovém případě tvar

$$U^S_{n_1,n_2} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} h(X_i,X_j),$$

kde

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

Za platnosti nulové hypotézy je střední hodnota testové statistiky $U^S_{n_1,n_2}$ rovna

$$\mathrm{E}\, U^S_{n_1,n_2} = \mathrm{E}\, h(X_1,Y_1) = \mathrm{P}(0 < X_1 < Y_1) + \mathrm{P}(Y_1 < X_1 < 0) = \frac{1}{4}.$$

3.2.2 Určování kritických hodnot pro zamítnutí nulové hypotézy

Věta 1.8 poskytuje aproximaci pro kritické hodnoty pomocí normálního rozdělení. Odvoďme nejprve platnost podmínek věty:
Za platnosti nulové hypotézy platí pro funkce $h^{(1,0)} = h^{(0,1)}$ následující vztahy:

$$h^{(1,0)}(x) = h^{(0,1)}(x) = \begin{cases} \frac{3}{4} - F(x), & \text{pro } x > 0 \\ F(x) - \frac{1}{4}, & \text{pro } x < 0. \end{cases}$$

Hodnoty $\sigma^2_{1,0}$ a $\sigma^2_{0,1}$ jsou potom obě rovny

$$\sigma^2_{1,0} = \sigma^2_{0,1} = \int_{-\infty}^{0} \left(F(x) - \frac{1}{2}\right)^2 dF(x) + \int_{0}^{\infty} \left(\frac{1}{2} - F(x)\right)^2 = \frac{1}{12} dF(x),$$

Za předpokladu (1.17) je rozptyl asymptotického rozdělení testové statistiky roven

$$\widehat{\sigma}^2_{n_1,n_2} = \frac{1}{12p(1-p)}$$

Podle věty 1.8 má statistika

$$\sqrt{12p(1-p)}\sqrt{N}\left(U^S_{n_1,n_2} - \frac{1}{4}\right)$$

za platnosti nulové hypotézy asymptoticky normální rozdělení se střední hodnotou rovnou 0 a rozptylem 1. Příslušný test je tedy asymptoticky tzv. *distribution free*.
Nulová hypotéza se potom zamítne pokud

$$U^S_{n_1,n_2} \geq \frac{1}{4} + u_{1-\alpha/2}\frac{1}{\sqrt{12p(1-p)}}\frac{1}{\sqrt{N}}$$

$$\text{nebo} \quad U^S_{n_1,n_2} \leq \frac{1}{4} - u_{1-\alpha/2}\frac{1}{\sqrt{12p(1-p)}}\frac{1}{\sqrt{N}}$$

kde $u_{1-\alpha/2}$ označuje $(1-\alpha/2)$-kvantil normovaného normálního rozdělení pro danou hladinu významnosti α.

3.2.3 Další charakteristiky testu

Podobně, jako u Mannova-Whitneyova testu má také testová statistika Sukhatmova testu asymptoticky normální rozdělení nejen za platnosti nulové hypotézy, ale i za platnosti alternativy. Dále platí (viz [Sukhatme 1957]), že

$$U^S_{n_1,n_2} \xrightarrow{P} \mathrm{E}\, h(X_1,Y_1) = \int_{0}^{\infty} F dG + \int_{-\infty}^{0} F dG$$

pro $n_1,n_2 \to \infty$. Zkoumaný test je tedy *konzistentní*.

Sukhatmův test byl navržen jako alternativa k parametrickému F-testu (viz např. [Anděl 2002], str. 77), u nějž se předpokládá, že oba výběry pocházejí z normálního rozdělení. Podle [Sukhatme 1957] poskytuje test založený na $U^S_{n_1,n_2}$ uspokojivé výsledky pro výběry z normálního rozdělení a může poskytovat velmi dobré výsledky pro výběry z některých jiných rozdělení, například z Laplaceova.

3.3 Test založený na rozdílu výběrových rozptylů

3.3.1 Testová statistika

Předpokládejme opět, že máme dva navzájem nezávislé náhodné výběry $X_1,\dots,X_{n_1}$ a $Y_1,\dots,Y_{n_2}$ s rozdělením charakterizovaným distribuční funkcí F, respektive G. Budeme studovat testovou statistiku, která je založena na rozdílu nestranných odhadů rozptylu:

$$D^{var}_{n_1,n_2} = \frac{1}{n_1-1}\sum_{i=1}^{n_1}\left(X_i - \bar{X}_{n_1}\right)^2 - \frac{1}{n_2-1}\sum_{j=1}^{n_2}\left(Y_j - \bar{Y}_{n_2}\right)^2,$$

kde $\bar{X}_{n_1}$, respektive $\bar{Y}_{n_2}$ značí výběrový průměr. V podkapitole 1.1.2 jsme odvodili, že výběrový rozptyl se dá vyjádřit jako U-statistika založená na jádře

$$h_\tau(x,y) = \frac{(x-y)^2}{2}$$

a testová statistika $D^{var}_{n_1,n_2}$ se dá vyjádřit ve tvaru:

$$D^{var}_{n_1,n_2} = \frac{2}{n_1(n_1-1)}\sum_{i_1=1}^{n_1}\sum_{i_2=i_1+1}^{n_1}\frac{(X_{i_1}-X_{i_2})^2}{2} - \frac{2}{n_2(n_2-1)}\sum_{j_1=1}^{n_2}\sum_{j_2=j_1+1}^{n_2}\frac{(Y_{j_1}-Y_{j_2})^2}{2}$$

Test založený na rozdílu výběrových rozptylů proto patří do třídy testů založených na rozdílu dvou U-statistik (viz podkapitola 2.2).

Za nulové hypotézy je střední hodnota statistiky $D^{var}_{n_1,n_2}$ rovna 0, za alternativy je rovna rozdílu rozptylů rozdělení výběrů. Rozptyl $\operatorname{Var} D^{var}_{n_1,n_2}$ závisí za nulové hypotézy i za alternativy na rozděleních výběrů. Zpravidla proto není jeho hodnota známá.

3.3.2 Asymptotické vlastnosti

Za předpokladu, že jsou splněny předpoklady věty 1.8, má statistika $D^{var}_{n_1,n_2}$ asymptoticky normální rozdělení.
Pro funkce $h^{(1,0)} = h^{(0,1)}$ platí následující vztahy:

$$h^{(1,0)}(x) = (x - \operatorname{E} X_1)^2 - \operatorname{Var} X_1,$$

$$h^{(0,1)}(y) = (y - \mathrm{E}\, Y_1)^2 - \mathrm{Var}\, Y_1.$$

Hodnoty $\sigma^2_{1,0}$ a $\sigma^2_{0,1}$ obě závisí na rozdělení výběru $X_1, \ldots, X_{n_1}$, respektive $Y_1, \ldots, Y_{n_2}$ a jsou dány vztahy

$$\sigma^2_{1,0} = \mathrm{E}\,(X_1 - \mathrm{E}\, X_1)^4 - (\mathrm{Var}\, X_1)^2\,,$$

$$\sigma^2_{0,1} = \mathrm{E}\,(Y_1 - \mathrm{E}\, Y_1)^4 - (\mathrm{Var}\, Y_1)^2\,.$$

Podmínky (1.14)-(1.16) jsou splněny pokud obě rozdělení výběrů mají tu vlastnost, že jejich čtvrtý centrální moment rozdělení je konečný a zároveň je ostře větší než druhá mocnina rozptylu rozdělení. Po přidání předpokladu (1.17), má potom náhodná veličina $\sqrt{N}(D^{var}_{n_1,n_2} - (\mathrm{Var}\, X_1 - \mathrm{Var}\, Y_1))$ asymptoticky normální rozdělení s rozptylem

$$\sigma^2_{as} = \frac{1}{p}\left(\mathrm{E}\,(X_1 - \mathrm{E}\, X_1)^4 - (\mathrm{Var}\, X_1)^2\right) + \frac{1}{1-p}\left(\mathrm{E}\,(Y_1 - \mathrm{E}\, Y_1)^4 - (\mathrm{Var}\, Y_1)^2\right).$$

Odtud plyne, že

$$D^{var}_{n_1,n_2} \xrightarrow{P} \mathrm{Var}\, X_1 - \mathrm{Var}\, Y_1$$

pro n_1,$n_2 \to \infty$. Test bude proto vhodný zejména v situacích, kdy budou srovnávány dva výběry, jejichž rozdělení mají různé rozptyly.

Za platnosti nulové hypotézy je $\mathrm{E}\, D^{var}_{n_1,n_2} = 0$ a rozptyl má tvar

$$\sigma^2_{as} = \mathrm{E}\,(X_1 - \mathrm{E}\, X_1)^4 - (\mathrm{Var}\, X_1)^2\,.$$

Nulová hypotéza o shodě rozdělení obou výběrů se zamítne pokud

$$\left|D^{var}_{n_1,n_2}\right| \geq \; u_{1-\alpha/2}\sqrt{\frac{\widehat{\sigma}^2_{n_1,n_2}}{N}},$$

kde $\widehat{\sigma}^2_{n_1,n_2}$ označuje odhad rozptylu $D^{var}_{n_1,n_2}$ a $u_{1-\alpha/2}$ je $(1-\alpha/2)$-kvantil normovaného normálního rozdělení pro danou hladinu významnosti α. Jackknife odhad rozptylu (viz podkapitola 2.3) bude mít v tomto případě tvar

$$\widehat{\sigma}^2_{n_1,n_2} = \frac{N}{n_1}\left(S^{(1,0)}_N - (D^{var}_{n_1,n_2})^2\right) + \frac{N}{n_2}\left(S^{(0,1)}_N - (D^{var}_{n_1,n_2})^2\right)$$

kde

$$S^{(1,0)}_N = \frac{1}{n_1}\sum_{k=1}^{n_1}\left(\frac{1}{n_1 - 1}\sum_{\substack{i=1\\ i\neq k}}^{n_1}\frac{(X_k - X_i)^2}{2}\right)^2,$$

$$S^{(0,1)}_N = \frac{1}{n_2}\sum_{k=1}^{n_2}\left(\frac{1}{n_2 - 1}\sum_{\substack{i=1\\ i\neq k}}^{n_2}\frac{(Y_k - Y_i)^2}{2}\right)^2.$$

3.3.3 Další charakteristiky testu

V práci [Dufour a Farhat 2002] byla zkoumaná permutační verze testu založeného na statistice $D^{var}_{n_1,n_2}$. Ve srovnání s ostatními zkoumanými testovými statistikami pro dvouvýběrový problém, měl permutační test založený na statistice $D^{var}_{n_1,n_2}$ dobrou sílu testu při testování odchylek od normality, zejména pokud šlo o výběry z rozdělení se stejnou střední hodnotou a různými rozptyly.

3.4 Test založený na rozdílu obecných měr variability

Pro náhodný výběr z rozdělení s distribuční funkcí H je výběrový rozptyl nestranným odhadem parametru $\tau(H) = \mathrm{E}_H(X_1 - X_2)^2/2$. To vede k úvaze o obecnějším tvaru odhadu variability výběru. Jednovýběrová U-statistika s jádrem

$$h_\tau(x,y) = |x - y|^p \tag{3.1}$$

pro zvolenou hodnotu $p > 0$, je nestranným odhadem $\tau(H) = \mathrm{E}_H |X_1 - X_2|^p$. Pro volbu $p = 1$ dostaneme parametr, který je literatuře známý pod názvem *Giniho průměrná diference*. Jako míra variability má tento parametr oproti rozptylu výhodu v tom, že pro jeho existenci není potřebné předpokládat konečnost druhých momentů rozdělení. Pro označení $\tau(H) = \mathrm{E}_H |X_1 - X_2|^p$ pro obecné hodnoty p se někdy používá výraz *zobecněná průměrná diference* (viz například [Yitzhaki 2003]).

3.4.1 Testová statistika

Na U-statistikách s jádrem tvaru (3.1) lze tedy založit test, který bude srovnávat tyto obecné míry variability pro rozdělení obou výběrů. Testová statistika bude mít následující tvar:

$$D^{(p)}_{n_1,n_2} = \frac{2}{n_1(n_1-1)} \sum_{i_1=1}^{n_1} \sum_{i_2=i_1+1}^{n_1} |X_{i_1} - X_{i_2}|^p \\ - \frac{2}{n_2(n_2-1)} \sum_{j_1=1}^{n_2} \sum_{j_2=j_1+1}^{n_2} |Y_{j_1} - Y_{j_2}|^p ,$$

kde $p > 0$ je hodnota parametru. Pro volbu $p = 2$ dostaneme dvojnásobek testové statistiky popsané v předchozí podkapitole.

3.4.2 Asymptotické vlastnosti

Tvar podmínek (1.14)-(1.17) pro U-statistiku $D^{(p)}_{n_1,n_2}$ je analogický tvaru podmínek pro testovou statistiku založenou na rozdílu výběrových rozptylů. Podmínka (1.14) je zřejmě splněna, pokud jsou konečné $2p$ momenty obou rozdělení:

$$\mathrm{E}\,|X_1|^{2p} < \infty \quad \text{a} \quad \mathrm{E}\,|Y_1|^{2p} < \infty.$$

Jádra $h^{(1,0)}(x)$ a $h^{(0,1)}(y)$ mají následující tvar:

$$h^{(1,0)}(x) = \mathrm{E}\,|x - X_1|^p - \mathrm{E}\,|X_1 - X_2|^p\,,$$

$$h^{(0,1)}(y) = \mathrm{E}\,|y - Y_1|^p - \mathrm{E}\,|Y_1 - Y_2|^p\,.$$

Pokud jsou pro pro ně splněny podmínky (1.15) a (1.16) a je splněna podmínka (1.17), má statistika $D^{(p)}_{n_1,n_2}$ asymptoticky normální rozdělení se střední hodnotou

$$\mathrm{E}\,D^{(p)}_{n_1,n_2} = \mathrm{E}\,|X_1 - X_2|^p - \mathrm{E}\,|Y_1 - Y_2|^p\,.$$

Dále platí:

$$D^{(p)}_{n_1,n_2} \xrightarrow{P} \mathrm{E}\,|X_1 - X_2|^p - \mathrm{E}\,|Y_1 - Y_2|^p\,,$$

pro n_1,$n_2 \to \infty$. Za platnosti nulové hypotézy je $\mathrm{E}\,D^{(p)}_{n_1,n_2} = 0$. Test bude proto citlivý zejména v situacích, kdy budou srovnávány dva výběry z rozdělení s různými mírami variability.

Rozptyl asymptotického rozdělení lze určit na základě vztahu (1.18), respektive (2.5). Za platnosti nulové hypotézy i za platnosti alternativy závisí tato hodnota na neznámých rozděleních výběrů. Jackknife odhad rozptylu se určí ze vztahu (2.7) analogicky jako pro test založený na rozdílu dvou výběrových průměrů.

3.5 Test založený na vzdálenosti distribučních funkcí

Test založený na vzdálenosti distribučních funkcí, který zde popíšeme, byl navržen v práci [Lehmann 1951]. Opět budeme předpokládat, že máme dva nezávislé výběry $X_1,\ldots,X_{n_1}$ a $Y_1,\ldots,Y_{n_2}$ z rozdělení s distribuční funkcí F, respektive G. I v tomto případě se pro odvození vlastností testu předpokládá, že F i G jsou distribuční funkce spojitých rozdělení.

3.5.1 Testová statistika

Test je založen na U-statistice s jádrem

$$h(X_1,X_2,Y_1,Y_2) = \begin{cases} \frac{1}{3} & \text{pokud } \min\{X_1,X_2\} > \max\{Y_1,Y_2\} \\ & \text{nebo } \max\{X_1,X_2\} > \min\{Y_1,Y_2\} \\ -\frac{1}{6} & \text{jinak} \end{cases} \tag{3.2}$$

Příslušná U-statistika

$$U^{edf}_{n_1,n_2} = \frac{1}{n_1(n_1-1)} \frac{1}{n_2(n_2-1)} \sum_{i_1=1}^{n_1} \sum_{i_2=i_1+1}^{n_1} \sum_{j_1=1}^{n_2} \sum_{j_2=j_1+1}^{n_2} h(X_{i_1},X_{i_2},Y_{i_1},Y_{j_2})$$

je nestranným odhadem funkcionálu (viz [Puri a Sen 1971], str. 65)

$$\theta(F,G) = \int_{-\infty}^{\infty} (F(x) - G(x))^2 d\left(\frac{F(x)+G(x)}{2}\right). \tag{3.3}$$

$\theta(F,G)$ lze považovat za funkci vzdálenosti mezi distribučními funkcemi F a G. Pokud $F \equiv G$, je hodnota $\theta(F,G)$ rovna 0.

3.5.2 Asymptotické vlastnosti

Situace za platnosti nulové hypotézy Asymptotické rozdělení za platnosti nulové hypotézy pro dvouvýběrovou statistiku $U^{edf}_{n_1,n_2}$ s jádrem (3.2) je popsáno v následující větě

Věta 3.1. *Předpokládejme, že $F \equiv G$ a $n_1/N \to p \in (0,1)$ pro $n_1,n_2 \to \infty$. Náhodná veličina*

$$\frac{n_1 n_2}{N} U^{edf}_{n_1,n_2}$$

má potom asymptotické rozdělení tvaru

$$Y = \sum_{i=1}^{\infty} \frac{1}{\pi^2 j^2} (\chi^2_{1j} - 1),$$

kde $\chi^2_{11},\chi^2_{12},\ldots$ jsou nezávislé náhodné veličiny s χ^2_1- rozdělením.

Důkaz. Věta je přímým důsledkem tvrzení z článku [Wegner 1956] a tvrzení o rozdělení testové statistiky dvouvýběrového Cramérova-von Misesova testu uvedeném v [Jurečková 1981], str. 64.

Vidíme, že asymptotické rozdělení $U^{edf}_{n_1,n_2}$ za nulové hypotézy je stejného typu jako bylo popsáno ve větě o limitním rozděleni degenerovaných U-statistik (věta 1.4). Pro určení kritických hodnot lze použít tabulky, které jsou zahrnuty například v článcích [Anderson a Darling 1952] nebo [Csörgő a Faraway 1996].

Poznámka 3.2. Asymptotické rozdělení statistiky $U^{edf}_{n_1,n_2}$, které je popsané ve výše uvedené větě, je stejné jako asymptotické rozdělení testové statistiky jednovýběrového Cramérova-von Misesova testu založeného na výběru $X_1,\ldots,X_n$:

$$V_n^{CV} = n\int \left(F_n(t) - F(t)\right)^2 dF(t)$$

kde F_n značí empirickou distribuční funkci výběru $X_1,\ldots,X_n$ a F je distribuční funkce rozdělení, s nímž je výběr srovnáván. Poznamenejme dále, že pro danou distribuční funkci F je V_n^{CV}/n V-statistika s jádrem tvaru (viz [Lee 1990], str. 160)

$$h(x,y) = \int (1_{\{x\le t\}} - F(t))(1_{\{y\le t\}} - F(t))dF(t).$$

Zároveň se jedná o stejné asymptotické rozdělení, které má i testová statistika pro dvouvýběrový Cramérův-von Misesův test (viz [Wegner 1956])

$$V^{CV}_{n_1,n_2} = \frac{n_1 n_2}{N}\int \left(F_{n_1}(t) - G_{n_2}(t)\right)^2 d\left(\frac{F_{n_1}(t) + G_{n_2}(t)}{2}\right)$$

kde $F_{n_1}(z)$ je empirická distribuční funkce výběru $X_1,\ldots,X_{n_1}$ a $G_{n_2}(z)$ je empirická distribuční funkce výběru $Y_1,\ldots,Y_{n_2}$.

Situace za platnosti alternativy Asymptotické chování testové statistiky za platnosti alternativy je popsáno v následující větě.

Věta 3.2. *Předpokládejme, že $n_1/N \to p \in (0,1)$ pro $n_1,n_2 \to \infty$. Potom statistika*

$$\sqrt{\frac{n_1 n_2}{N}}\left(U^{edf}_{n_1,n_2} - \mathrm{E}\, U^{edf}_{n_1,n_2}\right)$$

má asymptoticky normální rozdělení. Pokud vyloučíme distribuční funkce F a G, pro něž buď $F \equiv G$, $\int F dG = 0$ nebo $\int F dG = 1$, potom třída distribučních funkcí, pro které dostaneme nedegenerovanou testovou statistiku $U^{edf}_{n_1,n_2}$ zahrnuje všechny spojité distribuční funkce F a G, jež jsou v oblasti své variace striktně rostoucí[1].

[1] F je v oblasti své variace striktně rostoucí, pokud je splněna podmínka, že pro libovolnou dvojici čísel $x,y \in \mathbb{R}$, $x < y$ platí buď $F(x) < F(y)$ nebo $F(x) = F(y) = 0$ nebo $F(x) = F(y) = 1$. Pro distribuční funkci G se podmínka formuluje analogicky.

Důkaz. Viz [Wegner 1956]. Věta je důsledkem věty 1.8.

Pro libovolnou hladinu významnosti je test založený na $U^{edf}_{n_1,n_2}$ *konzistentní* (viz [Lehmann 1951]), za předpokladu, že $\min\{n_1,n_2\} \to \infty$. Jeho síla je totiž rostoucí funkcí vzdálenosti mezi rozděleními F a G, reprezentované výrazem (3.3).

3.5.3 Další charakteristiky testu

Za nevýhodu tohoto testu lze považovat fakt, že nemá vlastnost nestrannosti (viz [Wegner 1956]). Wegner dále srovnával test založený na $U^{edf}_{n_1,n_2}$ s Mannovým-Whitneyovým testem z hlediska síly testu. Ukázalo se, že oba testy měli srovnatelné vlastnosti.

Poznámka 3.3. Jak již bylo naznačeno výše, Lehmannem navržený test je příbuzný s Cramérovým-von Misesovým testem pro dva výběry (viz [Wegner 1956]). Testovou statistiku dvouvýběrového Cramérova-von Misesova testu dostaneme, pokud distribuční funkce F a G ve výrazu (3.3) nahradíme jejich empirickými odhady. Pro situaci, kdy jsou velikosti obou výběrů stejné, jsou oba testy ekvivalentní. Pro velká n_1 a n_2 a za předpokladu, že podíl n_1/n_2 konverguje ke kladné konstantě, jsou testy asymptoticky ekvivalentní. Proto se v literatuře (viz např. [Zajta a Pandikow 1977]) tento test objevuje také pod označením Cramérův-von Misesův-Lehmannův test.

Kapitola 4

Testy založené na empirických charakteristických funkcích

4.1 Testová statistika $V_{n_1,n_2}(w)$

V článku [Meintanis 2005] je navržena testová statistika pro obecný dvouvýběrový problém, která je založena na rozdílu mezi dvěma empirickými charakteristickými funkcemi. Předpokládejme opět, že máme dva navzájem nezávislé výběry nezávislých stejně rozdělených náhodných veličin $X_1, \ldots, X_{n_1}$ a $Y_1, \ldots, Y_{n_2}$. Uvažujme testovou statistiku

$$V_{n_1,n_2}(w) = \int_{-\infty}^{\infty} |\varphi_{n_1}(t) - \varphi_{n_2}(t)|^2 w(t) dt,$$

kde φ_{n_1} a φ_{n_2} označují empirické charakteristické funkce určené na základě výběrů $X_1, \ldots, X_{n_1}$, respektive $Y_1, \ldots, Y_{n_2}$. Jsou definovány vztahy

$$\varphi_{n_1}(t) = \frac{1}{n_1} \sum_{j=1}^{n_1} \exp(itX_j) \qquad \varphi_{n_2}(t) = \frac{1}{n_2} \sum_{j=1}^{n_2} \exp(itY_j),$$

Označme dále

$$h_w(\beta) = \int_{-\infty}^{\infty} \cos(\beta t) w(t) dt. \tag{4.1}$$

S využitím rovnosti $\exp(iz) = \cos(z) + i \sin(z)$ a součtových vzorců pro go-

niometrické funkce, se statistika $V_{n_1,n_2}(w)$ dá dále upravit na tvar

$$\begin{aligned} V_{n_1,n_2}(w) &= \int_{-\infty}^{\infty} \frac{1}{n_1^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} \cos(t(X_i - X_j)) w(t) dt \\ &\quad + \int_{-\infty}^{\infty} \frac{1}{n_2^2} \sum_{i=1}^{n_2} \sum_{i=1}^{n_2} \cos(t(Y_i - Y_j)) w(t) dt \\ &\quad - \int_{-\infty}^{\infty} \frac{2}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{i=1}^{n_2} \cos(t(X_i - Y_j)) w(t) dt \\ &= \frac{1}{n_1^2} \sum_{j=1}^{n_1} \sum_{k=1}^{n_1} h_w(X_j - X_k) + \frac{1}{n_2^2} \sum_{j=1}^{n_2} \sum_{k=1}^{n_2} h_w(Y_j - Y_k) \\ &\quad - \frac{2}{n_1 n_2} \sum_{j=1}^{n_1} \sum_{k=1}^{n_2} h_w(X_j - Y_k) \end{aligned} \quad (4.2)$$

Ze symetrie kosínové funkce plyne, že funkce $h'_w(x,y) = h_w(x-y)$ je symetrická vzhledem k záměně svých argumentů. Poznamenejme, že na V_{n_1,n_2} můžeme pohlížet jako na realizaci statistického funkcionálu

$$\begin{aligned} \theta(F,G) &= \int_{-\infty}^{\infty} |\varphi_1(t) - \varphi_2(t)|^2 w(t) dt \\ &= \int_{-\infty}^{\infty} \left| \int_{-\infty}^{\infty} \exp(ixt) dF(x) - \int_{-\infty}^{\infty} \exp(iyt) dG(y) \right|^2 w(t) dt \end{aligned} \quad (4.3)$$

kde φ_1, respektive φ_2 označuje charakteristickou funkci rozdělení s distribuční funkcí F, respektive G. Statistiku V_{n_1,n_2} lze tedy vyjádřit jako $V_{n_1,n_2} = \theta(F_{n_1},G_{n_2})$, kde F_{n_1} označuje empirickou distribuční funkci určenou na základě výběru $X_1,\ldots,X_{n_1}$ a G_{n_2} empirickou distribuční funkci výběru $Y_1,\ldots,Y_{n_2}$. Navíc, $\theta(F,G)$ je odhadnutelný (viz definice 1.4). Po analogické úpravě k úpravě $V_{n_1,n_2}(w)$ na vyjádření tvaru (4.2) dostaneme

$$\begin{aligned} \theta(F,G) &= \int_{\infty}^{\infty} \left(\int_{\infty}^{\infty} h_w(x-y) dF(x) \right) dF(y) \\ &\quad + \int_{\infty}^{\infty} \left(\int_{\infty}^{\infty} h_w(x-y) dG(x) \right) dG(y) \\ &\quad - 2 \int_{\infty}^{\infty} \left(\int_{\infty}^{\infty} h_w(x-y) dF(x) \right) dG(y). \end{aligned} \quad (4.4)$$

To se dá dále upravit na tvar

$$\theta(F,G) = \int_{\infty}^{\infty} \left(\int_{\infty}^{\infty} h_w(x-y) d\left(F(x) - G(x)\right) \right) d\left(F(y) - G(y)\right) \quad (4.5)$$

nebo také na složitější tvar

$$\theta(F,G) = \int_{\infty}^{\infty}\left(\int_{\infty}^{\infty}\left(\int_{\infty}^{\infty}\left(\int_{\infty}^{\infty} h_w(x-y) + h_w(u-v) - 2h_w(x-u)dF(x)\right)dF(y)\right)dG(u)\right)dG(v). \quad (4.6)$$

Volby váhových funkcí Pro některé volby váhových funkcí lze získat testovou statistiku, která má poměrně jednoduchou formu, výhodnou z početních důvodů. V článku [Meintanis 2005] je proto zvláštní pozornost věnována zejména těmto dvěma volbám váhových funkcí:

$$w_a^{(1)}(t) = \exp(-a|t|) \quad (4.7)$$

$$w_a^{(2)}(t) = \exp(-at^2) \quad (4.8)$$

kde $a > 0$ je parametr, který má za úlohu kontrolovat míru klesání hodnot váhových funkcí při rostoucím t. Pro vyšší hodnoty a klesají obě váhové funkce rychleji, pro hodnoty blízké 0 je klesání váhové funkce pomalejší.
Z rovnice 4.1 máme

$$\begin{aligned} h_a^{(1)}(\beta) &= \int_{-\infty}^{\infty}\cos(\beta t)\exp(-a|t|)dt = 2\int_0^{\infty}\cos(\beta t)\exp(-at)dt \\ &= 2\mathrm{Re}\left[\int_0^{\infty}\exp(-(a+i\beta)t)dt\right] = 2\mathrm{Re}\left[\frac{1}{a+i\beta}\right] = \frac{2a}{a^2+\beta^2} \end{aligned}$$

a

$$\begin{aligned} h_a^{(2)}(\beta) &= \int_{-\infty}^{\infty}\cos(\beta t)\exp(-at^2)dt = \mathrm{Re}\left[\int_{-\infty}^{\infty}\exp(-at^2 - i\beta t)dt\right] \\ &= \exp\left(-\frac{\beta^2}{4a}\right)\mathrm{Re}\left[\int_{-\infty}^{\infty}\exp -\sqrt{a}\left(t + \frac{i\beta t}{2\sqrt{a}}\right)^2 dt\right]. \end{aligned}$$

Pro zmíněné volby váhových funkcí pak dostaneme následující vyjádření testové statistiky $V_{n_1,n_2}(w)$:

$$V^{(1)}_{a,n_1,n_2} = 2a\frac{1}{n_1^2}\sum_{j=1}^{n_1}\sum_{k=1}^{n_1}\left[\frac{1}{a^2+(X_j-X_k)^2}\right] + 2a\frac{1}{n_2^2}\sum_{j=1}^{n_2}\sum_{k=1}^{n_2}\left[\frac{1}{a^2+(Y_j-Y_k)^2}\right] - 2a\frac{2}{n_1 n_2}\sum_{j=1}^{n_1}\sum_{k=1}^{n_2}\left[\frac{1}{a^2+(X_j-Y_k)^2}\right], \quad (4.9)$$

$$V^{(2)}_{a,n_1,n_2} = \sqrt{\frac{\pi}{a}}\frac{1}{n^2}\sum_{j=1}^{n_1}\sum_{k=1}^{n_1}\exp\left(-\frac{(X_j-X_k)^2}{4a^2}\right) + \sqrt{\frac{\pi}{a}}\frac{1}{n_2^2}\sum_{j=1}^{n_2}\sum_{k=1}^{n_2}\exp\left(-\frac{(Y_j-Y_k)^2}{4a^2}\right) - \sqrt{\frac{\pi}{a}}\frac{2}{n_1 n_2}\sum_{j=1}^{n_1}\sum_{k=1}^{n_2}\exp\left(-\frac{(X_j-Y_k)^2}{4a^2}\right). \quad (4.10)$$

4.1.1 Asymptotické vlastnosti

Limitní chování za platnosti nulové hypotézy Pro studium limitního chování zkoumané testové statistiky je výhodné vyjádření pomocí jádra degenerované projekce jádra $h_w(x-y)$ (viz podkapitola 1.2.3). Označme toto degenerované jádro symbolem $\tilde{h}(x,y)$:

$$\tilde{h}(z_1,z_2) = h_w(z_1-z_2) - \mathrm{E}\, h_w(z_1-Z_k) - \mathrm{E}\, h_w(Z_j-z_2) - \mathrm{E}\, h(Z_j-Z_k) \quad (4.11)$$

kde $Z_1, \ldots, Z_N$ označuje sdružený výběr.

Stejným postupem, jako bylo odvozeno v článku [Hušková a Meintanis 2006], lze odvodit následující dekompozici pro statistiku $V_{n_1,n_2}(w)$, za předpokladu platnosti nulové hypotézy H_0:

$$V_{n_1,n_2}(w) = A_1 + A_2 + A_3,$$

kde

$$A_1 = \frac{N}{n_1 n_2}\left(\frac{1}{n_1}\sum_{j=1}^{n_1}\sum_{\substack{k=1\\k\neq j}}^{n_1}\tilde{h}(X_j,X_k) + \frac{1}{n_2}\sum_{j=1}^{n_2}\sum_{\substack{k=1\\k\neq j}}^{n_2}\tilde{h}(Y_j,Y_k) - \frac{1}{N}\sum_{j=1}^{N}\sum_{\substack{k=1\\k\neq j}}^{N}\tilde{h}(Z_j,Z_k)\right) \quad (4.12)$$

$$A_2 = \frac{N}{n_1 n_2}\left(\int w(t)dt - \mathrm{E}\, h_w(Z_1 - Z_2)\right) \quad (4.13)$$

$$A_3 = -\frac{2}{n_1^2}\sum_{j=1}^{n_1}\left(\mathrm{E}\left(h_w(X_j - X_k)|X_j\right) - \mathrm{E}\, h_w(X_1 - X_2)\right) - \frac{2}{n_2^2}\sum_{j=1}^{n_2}\left(\mathrm{E}\left(h_w(Y_j - Y_k)|Y_j\right) - \mathrm{E}\, h(Y_1 - Y_2)\right). \quad (4.14)$$

Jak je vidět, A_1 je funkcí tří degenerovaných U-statistik. Všechny tři U-statistiky mají stejné jádro, jsou však založeny na různých skupinách dat. Za předpokladu nulové hypotézy je rozptyl prvního členu asymptoticky řádu

$$\mathrm{Var}\, A_1 \approx 2\left(\frac{N}{n_1 n_2}\right)^2 \mathrm{E}\,\tilde{h}^2(Z_1,Z_2),$$

kde $a_n \approx b_n$, pro posloupnosti reálných čísel a_n a b_n, v tomto kontextu znamená, že $a_n/b_n \to 1$ pro $n \to \infty$. Rozptyl třetího členu je asymptoticky

$$\mathrm{Var}\, A_3 \approx 4\left(\frac{1}{n_1^3} + \frac{1}{n_2^3}\right)^2 \mathrm{Var}\left(\mathrm{E}\left(h_w(Z_1 - Z_2)|Z_1\right)\right).$$

Z toho lze vyvodit, že člen A_3 nemá vliv na limitní rozdělení statistiky $V_{n_1,n_2}(w)$. Členy A_2 a $\sqrt{\mathrm{Var}\, A_1}$ jsou stejného řádu a tudíž oba ovlivňují limitní rozdělení.

Označme $\lambda_j, j = 1,2,\ldots$ vlastní čísla integrálního operátoru definovaného pomocí degenerovaného jádra $\tilde{h}(x,y)$ (viz věta 1.4). Rozptyl $\tilde{h}(Z_1,Z_2)$ je pak roven

$$\mathrm{E}\,\tilde{h}^2(Z_1,Z_2) = \sum_{j=1}^{\infty}\lambda_j^2.$$

Limitní rozdělení zkoumané třídy statistik je popsáno v následující větě.

Věta 4.1. *Předpokládejme, že $Z_1, \ldots, Z_N$ jsou nezávislé stejně rozdělené náhodné veličiny s distribuční funkcí F. Nechť w je symetrická nezáporná váhová funkce, splňující podmínku*

$$0 < \int w(t)dt < \infty.$$

Potom, pro $\min\{n_1,n_2\} \to \infty$ *je limitní chování* $\frac{n_1 n_2}{N} V_{n_1,n_2}(w)$ *stejné jako limitní chování náhodné veličiny*

$$Y = \left| \left(\int w(u)du - \mathrm{E}\, h_w(Z_1 - Z_2) \right) + \sum_{j=1}^{\infty} \lambda_j \left(\chi^2_{1j} - 1 \right) \right|$$

kde $\chi^2_{11}, \chi^2_{12}, \ldots$ *jsou nezávislé náhodné veličiny s* χ^2_1*- rozdělením.*

Důkaz. Věta je převzata z [Hušková 2006].

Limitní chování v obecném případě V obecné situaci , tj. nejen za platnosti nulové hypotézy, ale i za platnosti alternativy, konverguje statistika $V_{n_1,n_2}(w)$ v pravděpodobnosti k hodnotě $\theta(F,G)$ (viz [Meintanis 2005]), definované rovnicí (4.3):

$$V_{n_1,n_2}(w) \xrightarrow{P} \theta(F,G),$$

kde F a G označují distribuční funkce uvažovaných dvou výběrů. Poznamenejme, že $\theta(F,G)$ může být interpretováno jako míra vzdálenosti mezi distribucemi zkoumaných dvou výběrů. Pokud $F \equiv G$, je $\theta(F,G) = 0$, v opačném případě je ostře větší než 0. Z toho plyne, že za platnosti alternativy bude pro $n_1,\ n_2 \to \infty$ hodnota $\frac{n_1 n_2}{N} V_{n_1,n_2}$ konvergovat do nekonečna. Tento výsledek implikuje *konzistenci* testů založených na empirických charakteristických funkcích.

4.1.2 Permutační test založený na $V^{(1)}_{a,n_1,n_2}$ a $V^{(2)}_{a,n_1,n_2}$

V článku [Meintanis 2005] jsou dále uvedeny výsledky simulační studie, pomocí které byly srovnány testy založené na $V^{(1)}_{a,n_1,n_2}$ a $V^{(2)}_{a,n_1,n_2}$ (definované vztahy (4.9) a (4.10)), pro různé volby parametru a, s několika dalšími testy pro obecný dvouvýběrový problém, které byly studovány v práci [Dufour a Farhat 2002]. Použitá metoda se zakládala na kombinací permutačního principu a metody Monte Carlo.

Ukázalo se, že v zachování hladiny testu jsou pro spojitá rozdělení oba testy robustní vzhledem k a. Volby a blízké k 0 však pro spojitá rozdělení

poskytovaly lepší vlastnosti v oblasti síly testu. V téměř všech studovaných situacích dávali testy založené na empirických charakteristických funkcích, zejména test založený na $V^{(1)}_{a,n_1,n_2}$ lepší výsledky než ostatní testy.

V případě diskrétních rozdělení byly oba testy méně robustní v zachování hladiny testu. Nejlepší výsledky byly dosaženy pro $V^{(1)}_{a,n_1,n_2}$ a $V^{(2)}_{a,n_1,n_2}$ s většími hodnotami parametru a ($a = 1.0, 1.5, 2.0$). Pro tyto hodnoty měli testy aplikované na diskrétní rozdělení také větší sílu. V srovnání s ostatními uvažovanými testy pro diskrétní rozdělení měli oba uvažované testy založené na empirických charakteristických funkcích s většími hodnotami parametru a lepší výsledky než Kolmogorovův-Smirnovův test a byly srovnatelné s $\hat{L}_\infty$-testem[1], přičemž výhodnější vlastnosti měl v případě diskrétních rozdělení test založený na $V^{(2)}_{a,n_1,n_2}$.

4.2 U-statistika odvozená od $V_{n_1,n_2}(w)$

4.2.1 Testová statistika

Vraťme se teď k úvaze o statistickém funkcionálu $\theta(F,G)$ definovaném rovnicí (4.3). Jak již bylo zmíněno, $\theta(F,G)$ je odhadnutelný a na základě vyjádření (4.6) je vidět, že za určitých podmínek je možné nestranně odhadnout $\theta(F,G)$ pomocí součtu tří U-statistik. Součet těchto tří U-statistik je možné také vyjádřit jako jedinou zobecněnou dvouvýběrovou U-statistiku s jádrem

$$\begin{aligned} h(x_1,x_2,y_1,y_2) &= h_w(x_1 - x_2) + h_w(y_1 - y_2) \\ &\quad - \frac{1}{2}\left(h_w(x_1 - y_1) + h_w(x_1 - y_2) + h_w(x_2 - y_1) + h_w(x_2 - y_2)\right), \end{aligned}$$

kde $h_w(\beta)$ je definováno vztahem (4.1). Za nulové hypotézy (2.1) je

$$\theta(F,F) = \mathrm{E}_{F,F} h(X_1,X_2,Y_1,Y_2) = 0.$$

Příslušná U-statistika na základě výběrů $X_1,\ldots,X_{n_1}$ a $Y_1,\ldots,Y_{n_2}$ má potom tvar

$$U^{ecf}_{n_1,n_2}(w) = \frac{1}{n_1(n_1-1)} \frac{1}{n_2(n_2-1)} \sum_{i_1=1}^{n_1} \sum_{\substack{i_2=1 \\ i_2 \neq i_1}}^{n_1} \sum_{j_1=1}^{n_2} \sum_{\substack{j_2=1 \\ j_2 \neq j_1}}^{n_2} h(X_{i_1},X_{i_2},Y_{j_1},Y_{j_2}).$$

[1] Jedná se o test založený na L_∞- vzdálenosti mezi jádrovými odhady hustoty dvou výběrů. Test je podrobněji popsán v článku [Dufour a Farhat 2002].

Po úpravě dostaneme

$$U^{ecf}_{n_1,n_2}(w) = \frac{1}{n_1(n_1-1)} \sum_{i=1}^{n_1} \sum_{\substack{j=1 \\ j\neq i}}^{n_1} h_w(X_i - X_j) + \frac{1}{n_2(n_2-1)} \sum_{i=1}^{n_2} \sum_{\substack{j=1 \\ j\neq i}}^{n_2} h_w(Y_i - Y_j) - \frac{2}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} h_w(X_i - Y_j).$$

$U^{ecf}_{n_1,n_2}(w)$ je tedy lineárni kombinací tří U-statistik s jádrem h_w, analogickým k součtu (4.2).

4.2.2 Hoeffdingova dekompozice

Pro projekce jádra $h(x_1,x_2,y_1,y_2)$ (viz podkapitola 1.2.3) máme vztahy

$$h^{(1,0)}(x) = \mathrm{E}\, h_w(x - X_1) - \mathrm{E}\, h_w(X_1 - X_2) - \mathrm{E}\, h_w(x - Y_1) + \mathrm{E}\, h_w(X_1 - Y_1) \quad (4.15)$$

$$h^{(0,1)}(y) = \mathrm{E}\, h_w(y - Y_1) - \mathrm{E}\, h_w(Y_1 - Y_2) - \mathrm{E}\, h_w(y - X_1) + \mathrm{E}\, h_w(Y_1 - X_1) \quad (4.16)$$

Obě tato jádra jsou za předpokladu, že výběry pocházejí ze stejného rozdělení, identicky rovny 0. Další tří projekce jádra nejsou obecně rovny nule, ani za předpokladu $F \equiv G$

$$h^{(2,0)}(x_1,x_2) = h_w(x_1 - x_2) - \mathrm{E}\, h_w(x_1 - X_1) - \mathrm{E}\, h_w(X_1 - x_2) + \mathrm{E}\, h_w(X_1 - X_2)$$

$$h^{(0,2)}(y_1,y_2) = h_w(y_1 - y_2) - \mathrm{E}\, h_w(y_1 - Y_1) - \mathrm{E}\, h_w(Y_1 - y_2) + \mathrm{E}\, h_w(Y_1 - Y_2)$$

$$h^{(1,1)}(x,y) = \frac{1}{2}\mathrm{E}\, h_w(x - Y_1) + \frac{1}{2}\mathrm{E}\, h_w(y - X_1) - \frac{1}{2} h_w(x - y) - \frac{1}{2}\mathrm{E}\, h_w(X_1 - Y_1)$$

Dále máme

$$h^{(2,1)} \equiv 0 \qquad h^{(1,2)} \equiv 0$$

a

$$h^{(2,2)}(x_1,x_2,y_1,y_2) = \mathrm{E}\, h(x_1 - Y_1) - \mathrm{E}\, h(x_1 - X_1) + \mathrm{E}\, h(x_2 - Y_1) - \mathrm{E}\, h(x_2 - X_1) + \mathrm{E}\, h(y_1 - X_1) - \mathrm{E}\, h(y_1 - Y_1) + \mathrm{E}\, h(y_2 - X_1) - \mathrm{E}\, h(y_2 - Y_1).$$

4.2.3 Asymptotické vlastnosti

Limitní chování za platnosti nulové hypotézy Jak už bylo zmíněno, pokud platí $F \equiv G$, je

$$h^{(1,0)}(x) \equiv 0 \quad \text{a} \quad h^{(0,1)}(y) \equiv 0.$$

U-statistika $U^{ecf}_{n_1,n_2}(w)$ je proto degenerovaná. V takové situaci navíc platí $h^{(2,2)} \equiv 0$. V situaci, kdy jsou oba výběry ze stejného rozdělení, má U-statistika $U^{ecf}_{n_1,n_2}(w)$ podle věty o Hoeffdingově dekompozici pro zobecněné U-statistiky (věta 1.7) tvar

$$\begin{aligned} U^{ecf}_{n_1,n_2}(w) &= U^{(2,0)}_{n_1,n_2} + U^{(0,2)}_{n_1,n_2} + 4U^{(1,1)}_{n_1,n_2} \\ &= \frac{1}{n_1(n_1-1)} \sum_{i=1}^{n_1} \sum_{\substack{j=1 \\ j\neq i}}^{n_1} \tilde{h}(X_i,X_j) + \frac{1}{n_2(n_2-1)} \sum_{i=1}^{n_2} \sum_{\substack{j=1 \\ j\neq i}}^{n_2} \tilde{h}(Y_i,Y_j) \\ &\qquad - \frac{2}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \tilde{h}(X_i,Y_j), \end{aligned}$$

kde $U^{(2,0)}_{n_1,n_2}$, $U^{(0,2)}_{n_1,n_2}$ a $U^{(1,1)}_{n_1,n_2}$ jsou U-statistiky založeny na jádrech $h^{(2,0)}$, $h^{(0,2)}$, respektive $h^{(1,1)}$ a $\tilde{h}(x,y)$ označuje degenerovanou projekci jádra h_w vzhledem k rozdělení sdruženého výběru (viz (4.11)). Po úpravě dostaneme

$$\begin{aligned} U^{ecf}_{n_1,n_2}(w) = \frac{N-1}{n_1 n_2} \Bigg(& \frac{1}{n_1-1} \sum_{j=1}^{n_1} \sum_{\substack{k=1 \\ k\neq j}}^{n_1} \tilde{h}(X_j,X_k) + \frac{1}{n_2-1} \sum_{j=1}^{n_2} \sum_{\substack{k=1 \\ k\neq j}}^{n_2} \tilde{h}(Y_j,Y_k) \\ & - \frac{1}{N-1} \sum_{\substack{j=1 \\ k\neq j}}^{N} \sum_{k=1}^{N} \tilde{h}(Z_j,Z_k) \Bigg). \end{aligned}$$

Vidíme, že statistika $U^{ecf}_{n_1,n_2}(w)$ je asymptoticky ekvivalentní s A_1 v (4.12). Přesněji:

$$U^{ecf}_{n_1,n_2}(w) = \frac{N-1}{N} A_1 + \frac{N-1}{n_1 n_2} B_{n_1 n_2}$$

kde

$$\begin{aligned} B_{n_1,n_2} = & \frac{1}{n_1(n_1-1)} \sum_{j=1}^{n_1} \sum_{\substack{k=1 \\ k\neq j}}^{n_1} \tilde{h}(X_j,X_k) + \frac{1}{n_2(n_2-1)} \sum_{j=1}^{n_2} \sum_{\substack{k=1 \\ k\neq j}}^{n_2} \tilde{h}(Y_j,Y_k) \\ & - \frac{1}{N(N-1)} \sum_{\substack{j=1 \\ k\neq j}}^{N} \sum_{k=1}^{N} \tilde{h}(Z_j,Z_k). \end{aligned}$$

B_{n_1,n_2} je součtem tří jednovýběrových U-statistik se střední hodnotou rovnou 0. Podle silného zákona velkých čísel pro jednovýběrové U-statistiky (věta 1.5) konverguje B_{n_1,n_2} skoro jistě k 0, pro $n_1,\ n_2 \to \infty$. Dále, podíl $(N-1)/N \to 1$ pro $N \to \infty$. Důsledkem důkazu věty 4.1 je proto následující tvrzení:

Důsledek 4.2. *Nechť jsou splněny předpoklady věty 4.1. Potom, pro*

$$\min\{n_1,n_2\} \to \infty,$$

je limitní chování $\frac{n_1 n_2}{N} U^{ecf}_{n_1,n_2}(w)$ *stejné jako limitní chování náhodné veličiny*

$$Y = \sum_{j=1}^{\infty} \lambda_j \left(\chi^2_{1j} - 1\right),$$

kde $\chi^2_{11},\chi^2_{12},\ldots$ *jsou nezávislé náhodné veličiny s* χ^2_1*- rozdělením a a* λ_j *jsou vlastní čísla integrální rovnice*

$$\int \tilde{h}(x_1,x_2) f(x_2) dF(x_2) = \lambda f(x_1),$$

Důkaz. Tvrzení je důsledkem důkazu věty 4.1.

Poznámka 4.1. Podle věty 1.4 je rozdělení $\frac{n_1 n_2}{N} U^{ecf}_{n_1,n_2}(w)$ stejné jako rozdělení NC_N, kde C_N je degenerovaná jednovýběrová statistika s jádrem $\tilde{h}$:

$$C_N = \frac{1}{N(N-1)} \sum_{\substack{j=1 \\ k\neq j}}^{N} \sum_{k=1}^{N} \tilde{h}(Z_j,Z_k)$$

Limitní chování za platnosti alternativy Za předpokladu, že jsou splněny podmínky (1.14)-(1.17), lze na U-statistiku $U^{ecf}_{n_1,n_2}(w)$ aplikovat větu 1.8. Podle této věty bude mít statistika $\sqrt{N}(U^{ecf}_{n_1,n_2}(w) - \theta(\mathrm{F},\mathrm{G}))$ asymptoticky normální rozdělení. Podobně jako v případě $V_{n_1,n_2}(w)$ platí

$$U^{ecf}_{n_1,n_2}(w) \xrightarrow{P} \theta(F,G).$$

To implikuje *konzistenci* testu založeného na $U^{ecf}_{n_1,n_2}(w)$.

Poznámka 4.2. Statistiky $V_{n_1,n_2}(w)$ a $U^{ecf}_{n_1,n_2}(w)$ mají tedy velmi podobné asymptotické vlastnosti. Na rozdíl od $V_{n_1,n_2}(w)$ má však U-statistika $U^{ecf}_{n_1,n_2}(w)$ navíc tu výhodnou vlastnost, že střední hodnota $\mathrm{E}\left(\frac{n_1 n_2}{N} U^{ecf}_{n_1,n_2}(w)\right)$ je za nulové hypotézy vždy rovna 0, nezávisle na rozdělení výběrů.

Naproti tomu, pro $V_{n_1,n_2}(w)$ platí

$$\mathrm{E}\left(\frac{n_1 n_2}{N} V_{n_1,n_2}(w)\right) = \int_{-\infty}^{\infty} w(t)dt - \mathrm{E}_F h_w(Z_1,Z_2).$$

Tato hodnota závisí na rozdělení výběrů a není v obecném případě známá.

4.3 Příklad použití metody bootstrap pro statistiky $V_{n_1,n_2}(w)$ a $U^{ecf}_{n_1,n_2}(w)$

Pro ilustraci chování statistik založených na empirických charakteristických funkcích uvedeme příklad použití *metody bootstrap s vracením* pro testové statistiky $V_{n_1,n_2}(w)$ a $U^{ecf}_{n_1,n_2}(w)$. Metodu aplikujeme na simulovaná data. Omezíme sa pouze na speciální volbu váhové funkce (4.7). Pro statistiku $V_{n_1,n_2}(w)$ určenou na základě této speciální volby jsme zavedli označení $V^{(1)}_{a,n_1,n_2}$ (viz (4.9)). Statistika $U^{ecf}_{n_1,n_2}(w)$ pro tuto speciální volbu váhové funkce se určí analogicky. Budeme pro ni používat označení U^{ecf}_{a,n_1,n_2}.

Postup pro určování kritických hodnot Nejdříve se na základě výběrů $X_1,\ldots,X_{n_1}$ a $Y_1,\ldots,Y_{n_2}$ určí hodnota testových statistik $\frac{n_1 n_2}{N} U^{ecf}_{a,n_1,n_2}$ a $\frac{n_1 n_2}{N} V^{(1)}_{a,n_1,n_2}$ pro různé volby parametru a. Následuje simulace dvou bootstrapových výběrů $X^*_1,\ldots,X^*_{n_1}$ a $Y^*_1,\ldots,Y^*_{n_2}$. Oba výběry jsou nezávisle vybírány z empirického rozdělení sdruženého výběru $Z_1,\ldots,Z_N$. Na základě bootstrapových výběru se určí hodnoty bootstrapových verzí uvažovaných statistik, které budeme značit $\frac{n_1 n_2}{N} U^{ecf*}_{a,n_1,n_2}$ a $\frac{n_1 n_2}{N} V^{(1)*}_{a,n_1,n_2}$. Simulace bootstrapových výběrů se B-krát zopakuje. Pro každou statistiku se tak získá B hodnot, na základě kterých lze získat aproximaci pro podmíněné rozdělení zkoumaných statistik (podmíněně na sdruženém výběru $Z_1,\ldots,Z_N$). Kritické hodnoty pro zamítnutí nulové hypotézy se pak určí jako $(1-\alpha)$-kvantily tohoto rozdělení.[2]

Aplikace na simulovaná data V prostředí R byly simulovány dva náhodné výběry o velikostech $n_1 = n_2 = 40$. První výběr byl vybrán z normálního rozdělení s nulovou střední hodnotou a jednotkovým rozptylem. Druhý výběr pocházel z exponenciálního rozdělení se střední hodnotou rovnou 1. Pro srovnání těchto dvou výběrů byly použity testové statistiky $\frac{n_1 n_2}{N} U^{ecf}_{a,n_1,n_2}$ a $\frac{n_1 n_2}{N} V_{a,n_1,n_2}$ pro hodnoty $a = 0{,}25;\ 0{,}5;\ 1{,}0;\ 2{,}0;\ 3{,}0$. Dále bylo simulováno $B = 400$ bootstrapových výběrů se sdruženého rozdělení a na základě těchto

[2] Zdrojový kód programu pro určení kritických hodnot metodou bootstrap pro uvažované statistiky je v příloze.

výběrů byly pro každou uvažovanou testovou statistiku určeny kritické hodnoty $c^*_{1-\alpha}$ pro $\alpha = 0{,}01;\ 0{,}05;\ 0{,}10$.

	$a = 0{,}25$	$a = 0{,}5$	$a = 1{,}0$	$a = 2{,}0$	$a = 3{,}0$
$\frac{n_1 n_2}{N} U^{ecf}_{a,n_1,n_2}$	17,31094	12,68183	7,57126	3,37423	1,74432
$c^*_{0,99}$	11,15513	5,42106	2,04216	0,56865	0,37606
$c^*_{0,95}$	6,73301	2,32472	1,30922	0,43107	0,25144
$c^*_{0,90}$	4,03019	1,81141	0,78754	0,30818	0,15490

Tabulka 4.1: Srovnání hodnoty $\frac{n_1 n_2}{N} U^{ecf}_{a,n_1,n_2}$ pro výběr z $N(0{,}1)$ a výběr z $Exp(1)$ s kritickými hodnotami určenými pomocí metody bootstrap.

	$a = 0{,}25$	$a = 0{,}5$	$a = 1{,}0$	$a = 2{,}0$	$a = 3{,}0$
$\frac{n_1 n_2}{N} V^{(1)}_{a,n_1,n_2}$	23,68675	15,26991	8,46081	3,61034	1,83808
$c^*_{0,99}$	17,48672	8,05974	2,99650	0,86345	0,49643
$c^*_{0,95}$	13,03282	5,05351	2,20825	0,71862	0,36639
$c^*_{0,90}$	10,54680	4,52074	1,75768	0,58072	0,27776

Tabulka 4.2: Srovnání hodnoty $\frac{n_1 n_2}{N} V^{(1)}_{a,n_1,n_2}$ pro výběr z $N(0{,}1)$ a výběr z $Exp(1)$ s kritickými hodnotami určenými pomocí metody bootstrap.

Z tabulek 4.3 a 4.3 vidíme, že ve všech případech byla nulová hypotéza o shodě rozdělení zamítnuta na hladině $\alpha = 0{,}01$. V souladu s očekáváním jsou hodnoty $\frac{n_1 n_2}{N} V^{(1)}_{a,n_1,n_2}$ i příslušné simulované kritické hodnoty větší než hodnoty pro statistiku $\frac{n_1 n_2}{N} U^{ecf}_{a,n_1,n_2}$ se stejnou hodnotou parametru a.

Poznámka 4.3. Použili jsme metodu bootstrap, aniž bychom předtím teoreticky odvodili, že podmíněné rozdělení bootstrapových verzí uvažovaných testových statistik poskytuje dobrou aproximaci pro rozdělení původních statistik za platnosti nulové hypotézy. Metoda bootstrap pro jednovýběrové degenerované U-statistiky je popsána v článcích ([Arcones a Giné 1992] a [Hušková a Janssen 1993]). Podle poznámky 4.1 má U-statistika $U^{ecf}_{n_1,n_2}$ stejné limitní rozdělení jako degenerovaná jednovýběrová U-statistika. Nabízí se otázka, zda by oprávněnost použití metody bootstrap pro zkoumané testové statistiky nemohla být odvozena od této ekvivalence. Hlubší studium zmíněné problematiky jsme však vynechali, s ohledem na rozsah této práce.

Závěr

Na U-statistikách je založeno mnoho testů pro testování různých typů hypotéz. Užší zaměření této práce tvořili testy založené na U-statistikách pro obecný dvouvýběrový problém. U-statistiky mají výhodné vlastnosti pro testování například v tom, že jsou nestrannými odhady určitých parametrů rozdělení. Navíc, za určitých podmínek, mají podobné asymptotické chování jako součty nezávislých stejně rozdělených náhodných veličin. V takovém případě konverguje jejich $\sqrt{N}$-násobek v distribuci k asymptoticky normálnímu rozdělení. Patří sem například Mannův-Whitneyův test nebo Sukhatmův test.

Speciálním typem testové statistiky založené na dvouvýběrové U-statistice je rozdíl dvou jednovýběrových U-statistik. Testy tohoto typu například srovnávají odhady momentů dvou rozdělení. V této práci jsme popsali test založený na rozdílu výběrových rozptylů a zmínili jsme se také o testu založeném na obecnějším tvaru míry variability. Určili jsme podmínky, za kterých mají testové statistiky asymptoticky normální rozdělení a uvedli vhodný odhad pro jejich asymptotický rozptyl.

Některé dvouvýběrové U-statistiky mají tu vlastnost, že jejich $\frac{n_1 n_2}{N}$-násobek má za nulové hypotézy specifický tvar limitního rozdělení (jsou tzv. degenerované), zatímco za alternativy konverguje tato hodnota do nekonečna. Reprezentantem tohoto typu testu je test založený na vzdálenosti empirických distribučních funkcí a patří sem i testy založené na empirických charakteristických funkcích.

Testové statistiky založené na empirických funkcích mají tvar V-statistik, které tvoří příbuznou třídu ke třídě U-statistik. Studovali jsme vlastnosti U-statistik odvozených od zmíněných V-statistik. Oba typy statistik mají podobné asymptotické vlastnosti. U-statistiky odvozené od testů založených na empirických charakteristických funkcích mají navíc tu vlastnost, že za nulové hypotézy je jejich střední hodnota rovna 0, zatímco střední hodnota původní V-statistiky závisí na rozdělení výběrů. Pro ilustraci chování obou typů statistik jsme připojili krátký příklad aplikace testů na simulovaná data. Kritické hodnoty byly aproximovány pomocí metody bootstrap s vracením.

Příloha A

Zdrojový kód programu

Zdrojový kód programu pro určování kritických hodnot statistik V_{a,n_1,n_2} a U^{ecf}_{a,n_1,n_2} na základě metody bootstrap s vracením pro dva simulované výběry. Výpočty byly provedeny v prostředí R.

```
N1<-40        #velikost 1. vyberu
N2<-40        #velikost 2. vyberu
N<-N1+N2
vektora<-c(0.25,0.5,1,2,3)
#parametry vahove funkce-> konkretni podoba jadra
B<-400         #pocet opakovani bootstrapovych vyberu
alfa0<-0.01; alfa1<-0.05; alfa2<-0.1;
###
hch1<-function(y1,y2,a){# a je parameter
d<-(y1-y2); vysl<-2*a/(a^2+d^2);
return(vysl)}
UV4<-function(data,h,a){#U-statistika a V-statistika
vyb1<-data[1:N1]; vyb2<-data[N1+1:N];
pom1<-0; pom2<-0; pom3<-0;
for (i in 1:(N1-1)){for (j in (i+1):N1){
pom1<-pom1+h(vyb1[i],vyb1[j],a)}}
for (i in 1:(N2-1)){for (j in (i+1):N2){
pom2<-pom2+h(vyb2[i],vyb2[j],a)}}
for (i in 1:N1){for (j in 1:N2){
pom3<-pom3+h(vyb1[i],vyb2[j],a)}}
pom4<-0; pom5<-0;
for (i in 1:N1){#diagonaly
pom4<-pom4+h(vyb1[i],vyb1[i],a)}
for (i in 1:N2){pom5<-pom5+h(vyb2[i],vyb2[i],a)}
```

```
vU<-2*N1*N2*((pom1/(N1*(N1-1)))+(pom2/(N2*(N2-1)))
-(pom3/(N1*N2)))/N
vV<-N1*N2*((2*pom1+pom4)/(N1*N1)+(2*pom2+pom5)/(N2*N2)
-2*pom3/(N1*N2))/N
return(c(vU,vV))}
#####
vyber1<-rnorm(N1)      #simuluj 1. vyber
vyber2<-rexp(N2)       #simuluj 2. vyber
zvyber<-c(vyber1,vyber2)   #sdruzeny vyber
#####
for(j in 1:length(vektora)){#opakuj pro ruzne a
a<-vektora[j]
bootU1<-NULL; bootV1<-NULL;
for (m in 1:B){#Bx opakuj
actual<-sample(1:(N1+N2), replace=TRUE)
bootvyber<-zvyber[actual] #bootstrapovy vyber
buv1<-UV4(bootvyber,hch1,a);
bootU1<-c(bootU1,buv1[1]);
bootV1<-c(bootV1,buv1[2]);
}
sort1<-sort(bootU1);
sort1v<-sort(bootV1);
c10<-sort1[B*(1-alfa0)];  #pro B=nasobek 100
c11<-sort1[B*(1-alfa1)];
c12<-sort1[B*(1-alfa2)];
c10v<-sort1v[B*(1-alfa0)];
c11v<-sort1v[B*(1-alfa1)];
c12v<-sort1v[B*(1-alfa2)];
print("a="); print(a);
uv1<-UV4(zvyber,hch1,a);
print("U1="); print(uv1[1]);
print("V1="); print(uv1[2]);
print("U - krit. h. pro alpha=0,01"); print(c10);
print("U - krit. h. pro alpha=0,05"); print(c11);
print("U - krit. h. pro alpha=0,10"); print(c12);
print("V - krit. h. pro alpha=0,01"); print(c10v);
print("V - krit. h. pro alpha=0,05"); print(c11v);
print("V - krit. h. pro alpha=0,10"); print(c12v);
}
```

Literatura

Anderson, T. W. a Darling, D. A. (1952). Asymptotic theory of certain 'goodness of fit' criteria based on stochastic processes, *Annals of Mathematical Statistics* **23**: 193–213.

Anděl, J. (2002). Základy matematické statistiky. Preprint MFF UK.

Arcones, M. A. a Giné, E. (1992). On the bootstrap of U and V statistics, *The Annals of Statistics* **20**: 655–674.

Arvesen, J. N. (1969). Jackknifing U-statistics, *The Annals of Mathematical Statistics* **40**: 2076–2100.

Bönner, N. a Kirschner, H.-P. (1977). Note on conditions for weak convergence of von Mises' differentiable statistical functions, **5**: 405–407.

Csörgő, S. a Faraway, J. (1996). The exact and asymptotic distributions of Cramér-von Mises statistic, *Journal of the Royal Statistical Society, Series B (Methodological)* **58**: 221–234.

Dufour, J.-M. a Farhat, A. (2002). Exact nonparametric two-sample homogeneity tests, *Goodness-of-fit tests and model validity (Paris, 2000)*, Birkhäuser Boston, Boston, MA, pp. 435–448.

Ferger, D. (2004). Maximal asymptotic power and efficiency of two-sample tests based on generalized U-statistics, *Metrika* **60**: 33–57.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution, *The Annals of Mathematical Statistics* **19**: 293–325.

Horváth, L. a Hušková, M. (2005). Testing for changes using permutations of U-statistics, *Journal of Statistical Planning and Inference* **128**.

Hušková, M. (2006). Test procedures based on empirical characteristic functions, prezentace na konferenci Robust 2006.

Hušková, M. a Janssen, P. (1993). Consistency of the generalized bootstrap for degenerate U-statistics, *The Annals of Statistics* **21**: 1811–1823.

Hušková, M. a Meintanis, S. G. (2006). Change point analysis based on empirical characteristic functions, *Metrika* **63**: 145–168.

Janssen, P. (1988). *Generalized empirical distribution functions with statistical application*, Limburgs Universitair Centrum, Diepenbeek.

Jurečková, J. (1981). *Pořadové testy*, SPN, Praha.

Lee, A. J. (1990). *U-statistics. Theory and practice*, Marcel Dekker, New York.

Lehmann, E. L. (1951). Consistency and unbiasedness of certain nonparametric tests, *The Annals of Mathematical Statistics* **22**: 165–179.

Likeš, J. a Laga, J. (1978). *Základní statistické tabulky*, SNTL, Praha.

Meintanis, S. G. (2005). Permutation tests for homogeneity based on the empirical characteristic function, *Journal of Nonparametric Statistics* **17**: 583–592.

Puri, M. L. a Sen, P. K. (1971). *Nonparametric methods in multivariate analysis*, Wiley, New York.

Sen, P. K. (1967). On some multisample permutation tests based on a class of U-statistics, *Journal of the American Statistical Association* **62**: 1201–1213.

Serfling, R. I. (1980). *Approximation theorems of mathematical statistics*, Wiley, New York.

Sukhatme, B. V. (1957). On certain two-sample nonparametric tests for variances, *The Annals of Mathematical Statistics* **28**: 188–194.

Wegner, L. H. (1956). Properties of some two-sample tests based on a particular measure of discrepancy, *The Annals of Mathematical Statistics* **27**: 1006–1016.

Yitzhaki, S. (2003). Gini's mean difference: a superior measure of variability for non-normal distributions, *Metron* **61**(2): 285–316.

Zajta, A. J. a Pandikow, W. (1977). A table of selected percentiles of the Cramér-von Mises-Lehmann test, *Biometrika* **64**.