UNIVERZA V LJUBLJANI
FAKULTETA ZA MATEMATIKO IN FIZIKO

Finančna Matematika – 1. stopnja

Damjan Manevski

Test Mann-Whitney

Delo diplomskega seminarja

Mentor: izr. prof. dr. Maja Pohar Perme

Ljubljana, 2016

Kazalo

Test Mann-Whitney

Povzetek

Diplomska naloga preučuje problem, ko je potrebno oceniti koliko se razlikujeta dve neodvisni slučajni spremenljivki X in Y. Zato je v nalogi podrobneje predstavljen test Mann-Whitney, ki se pogosto uporablja kot neparametrična alternativa testa t. Kot stopnja prekrivanja se uporabi $\theta = P(X < Y) + \frac{1}{2}P(X = Y)$, ki je povezana z Mann-Whitneyjevo statistiko U. Naloga razloži intervale zaupanja za θ in je omejena samo na peto Newcombovo metodo, kot najbolj pogosto uporabljeno metodo. Naloga podrobno razloži razvoj metode in določi primere, ko metoda spodleti. Na koncu je predstavljena povezava s krivuljo ROC.

Mann-Whitney test

Abstract

The thesis paper tackles the problem of characterizing the degree of separation of two independent random variables X and Y. Under these circumstances, the Mann-Whitney test is commonly used as a non-parametric alternative for the t-test. The following measure: $\theta = P(X < Y) + \frac{1}{2}P(X = Y)$ is used as an effect size, and the paper presents the link with the Mann-Whitney U statistic. Furthermore, the derivation of the confidence intervals for the fifth Newcombe method is presented and simulations for the same method are made. The paper evaluates situations in which the method fails, and it also shows the relation between θ and the area under the recieving operating characteristic curve.

Math. Subj. Class. (2010): Statistics, Probability theory and stochastic processes

Ključne besede: test Mann-Whitney, interval zaupanja, stopnja prekrivanja, krivulja ROC

Keywords: Mann-Whitney test, confidence interval, effect size, ROC curve

1. Definicija

1.1. **Uvod.** Test Mann-Whitney spada v skupini statističnih neparametričnih testov, ki temeljijo na manj predpostavkah kot večina testov, ki smo jih spoznali na dodiplomskem študiju. V bistvu, ta test poskuša zavrniti ničelno domnevo, da imata dva neodvisna vzorca enako porazdelitev. Test lahko obravnavamo kot alternativo testa t za dva vzroca, ki pa predpostavi, da vzorca prihajata iz normalno porazdeljenih populacij.

Začetki testa segajo v leto 1914, ko je nemški znanstvenik Gustav Deuchler objavil definicijo prvotne statistike (Kruskal, 1957). Naslednji napredek je prišel leta 1945, ko je ameriški kemik Frank Wilcoxon, nezadovoljen z rezultati dotedanjih statističnih testov nad svojimi podatki, razvil nov pristop, ki je bil objavljen v istem letu. Namreč, ko je študiral test t in Fisherjevo analizo variance, je opazil kako velik vpliv so imele osameljce na testni statistiki. Menil je, da je to glavni razlog za neuspeh testov in kmalu je svojo alternativo objavil. Bolj rigorozna definicija je prišla leta 1947, ko sta ameriški matematik Henry Mann in njegov študent Donald Whitney objavila članek, kjer sta predstavila podrobnejšo analizo tega testa. Do teh ugotovitev sta prišla, ko sta analizirala porazdelitev plač med letoma 1940 in 1944 in zaradi njihovih prispevkov je bil test imenovan po njiju. Test Mann-Whitney je eden izmed prvih razvitih neparametričnih testov in v naslednjih 25-letih so različni matematiki popolnili luknje definicije ter predpostavk testa (Salsburg, 2001).

V tem poglavju si bomo pogledali potrebne predpostavke tega testa in postopek statističnega testiranja. Potem v drugem poglavju bomo definirali stopnjo prekrivanja (ang. *effect size*) in intervale zaupanja ter predstavili rezultate simulacij. V zadnjem poglavju bomo v kratkem predstavili krivulje ROC in njihovo povezavo s testom Mann-Whitney.

1.2. **Primerjava s testom t.** Velikokrat v raziskavah hočemo primerjati srednja vrednost (npr. povprečje) dveh neodvisnih vzorcev in ugotoviti, ali sta značilno različna, ali še več, ali vzorca prihajata iz različno porazdeljenih populacij. V takih primerih se najpogosteje uporabljata test t in test Mann-Whitney. Glavna razlika med njima je predpostavka o normalnosti. Test t namreč predpostavi, da vzorca prihajata iz normalno porazdeljenih populacij, medtem ko test Mann-Whitney nima nobenih predpostavk o porazdelitvi populacij ali o posameznih parametrih (kar je lastnost ne samo tega testa, temveč vseh neparametričnih testov). Tako vidimo, da je mogoče smiselno uporabiti test Mann-Whitney, ko je vsaj eden izmed vzorcih asimetričen, ali ko je velikost vzorcev majhna. V obeh primerih je nevarno predpostaviti normalnost, saj nas lahko privede do napačnega zaključka. Drugi primer, ko se izkaže, da je test Mann-Whitney bolj primeren, je takrat ko vzorca vsebujeta več osameljcev. Izkaže se, da je test t zelo občutljiv pri takih vrednostih, medtem ko se test Mann-Whitney temu izogne, kar bomo prikazali v tem poglavju. Za primer lahko pogledamo naslednji vzorec $(4, 36, 38, 39, 44, 45)$ in vpliv prvega člana na vzorčno povprečje ali standardno napako. Razlogi, ki smo jih omenili zgoraj, so tudi glavni zgodovinski razlogi zakaj je prišlo do ogromne razvitosti neparametričnih testov (Salsburg, 2001).

1.3. **Predpostavke testa Mann-Whitney.**

- Vzorca sta naključna.
- Opazovanja obeh vzorcev so med seboj neodvisna, in posledično sta vzorca neodvisna.

- Merska lestvica je za vrednosti vzorcev vsaj urejenostna (ordinalna) lestvica. Vzorce lahko tako uvrstimo po velikosti in za poljubne dve vrednosti ugotovimo katera vrednost je večja, katera manjša, in ali sta vrednosti enaki. To je šibkejša predpostavka kot pri testu t, ki predpostavi, da imata vzorca za mersko lestvico vsaj razmično lestvico. Ta je vsebovana v urejenostni lestvici, vendar z dodatno predpostavko, da so lestvične enote enake (na primer, da je razlika med 1 in 2 enako vrednotena kot razliko med 19 in 20).

1.4. **Ideja testa.** Naslednji zgled bo pokazal osnovni postopek testa. Podatki so vzeti iz [6]. Imamo dva vzorca, vsak velikosti 9:

Vzorec A: 14,9, 11,3, 13,2, 16,6, 17,0, 14,1, 15,4, 13,0, 16,9.
Vzorec B: 15,2, 19,8, 14,9, 18,3, 16,2, 18,9, 12,2, 15,3, 19,4.

Postopek se začne tako, da uvrstimo vsa opazovanja v naraščajočem vrstnem redu in jih shranimo kot nov seznam. Vsakemu opazovanju seznama priredimo rang, ki ustreza njegovemu položaju. Prirejene vrednosti so od 1 (za najmanjše opazovanje) do $n_A + n_B$ (za največje opazovanje). Za naš primer izgleda tako:

Vrednost	11,3	12,2	13	13,2	14,1	14,9	14,9	15,2	15,3	15,4	16,2	16,6	16,9	17	18,3	18,9	19,4	19,8
Rang	1	2	3	4	5	6,5	6,5	8	9	10	11	12	13	14	15	16	17	18

Tabela 1

Ko imamo več opazovanj z isto vrednostjo, jim dodelimo isti rang, ki je povprečje potencialnih rangov. Imenujemo jih vezani rangi. V našem primeru se je to zgodilo za vrednost 14,9.

Zakaj smo se na tak nacin lotili? Intuitivno, če primerjamo povprečji rangov posameznih vzorcev, in se ti povprečji dosti razlikujeta, mogoče lahko sklepamo da imamo različno porazdeljene populacije. Zato ta postopek definiramo kot statistični test in poiščemo smiselno testno statistiko.

1.5. **Formalna definicija.** Imamo vzorec velikosti n: X_1,...,X_n, kjer so sl. sprem. neodvisne in enako porazdeljene s porazdelitveno funkcijo F. Imamo drug vzorec velikosti m: Y_1,Y_2,...,Y_m, kjer so sl. sprem. neodvisne in enako porazdeljene s porazdelitveno funkcijo G. X_i in Y_j sta neodvisna za poljubna $i = 1, ..., n, j = 1, ..., m$.

Za ničelno domnevo bomo vedno privzeli, da sta oba vzorca enako porazdeljena, torej: $H_0 : F \overset{(d)}{=} G$. Ta domneva implicira tudi naslednjo ničelno domnevo: $H_0 : P(X < Y) = P(X > Y) = \frac{1}{2}$. Za alternativno domnevo imamo dve možnosti:

- Če je test obojestranski: $H_1 : F \neq G$ ali $H_1 : P(X < Y) \neq P(X > Y)$, odvisno od H_0.
- Če je test enostranski: $H_1 : P(X < Y) < P(X > Y)$ ali $H_1 : P(X < Y) > P(X > Y)$, odvisno od tega, na kateri strani testiramo.

Zdaj želimo poiskati smiselno testno statistiko s pomočjo rangov, ki smo jih definirali v prejšnjem zgledu. Intuitivno, če obstaja precejšna razlika med povprečji populacij, večina nižjih rangov bo pripadala enemu vzorcu, medtem ko bo večina višjih rangov pripadala drugemu. Zato si bomo pomagali z vsotami rangov posameznih vzorcev. Definirajmo:

$R_X :=$ vsota rangov iz prvega vzorca.
$R_Y :=$ vsota rangov iz drugega vzorca.

Pri tem ni treba izračunati obeh vrednosti, saj so vsa števila med 1 in $n+m$ vsebovana v R_X in R_Y, kar pomeni: $R_X + R_Y = \frac{(n+m)(n+m+1)}{2}$.

Naš test lahko testiramo s temi testnimi statistikami. Obstaja izrek, ki pravi, da lahko za testno statistiko uporabimo poljubno monotono transformacijo teh statistik [9]. V praksi običajno uporabimo:

$$U_X := R_X - \frac{n(n+1)}{2}, U_Y := R_Y - \frac{m(m+1)}{2}, U := min\{U_X, U_Y\}$$

Razlog zakaj smo jih tako definirali bomo razložili v naslednjem poglavju. Kot testno statistiko uporabimo U_X, U_Y ali U, odvisno od tega, ali je test enostranski ali obojestranski.

Ko enkrat imamo definirane testne statistike, si poglejmo, kako poiskati kritično vrednost.

1.5.1. *Zgled:* Vzet iz [12]. Podani so 4 podatki: $1, 3, 4, 6$. Razporedimo jih v dve skupini po dva elementa. Če predpostavimo, da velja H_0, je vsaka razporeditev enako verjetna. Vzemimo eno tako razporeditev:

Treatment	Control
1(1)	6(4)
3(2)	4(3)

Imamo 2 skupini, Treatment in Control. V prvi sta elementa 1 in 3, v drugi 4 in 6. V oklepajih so zapisani pripadajoči rangi. Vidimo, da: $R_T = 3, R_C = 7$, oziroma $U_T = 0, U_C = 4$.

Zanima nas, ali je razlika med R_T in R_C dovolj velika, da bi lahko zavrnili H_0, oz. domnevo, da imata Treatment in Control isto porazdelitev. Poglejmo, kolikšna je verjetnost posamezne razporeditve. Vemo, da lahko pogojno na H_0, eksplicitno izračunamo porazdelitev U_T in U_C, saj je vsaka porazdelitev rangov enako verjetna, ko velja H_0. Torej:

Ranks	U
{1,2}	0
{1,3}	1
{1,4}	2
{2,3}	2
{2,4}	3
{3,4}	4

u	0	1	2	3	4
$P(U=u)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{6}$

Tu je U definiran kot razlika med vsoto rangov vzorca in $\frac{2\cdot3}{2}$. Torej lahko za dano stopnjo značilnosti α, izračunamo kritično vrednost. V našem zgledu, ko je $\alpha = 0{,}05$, ne moremo določiti kritične vrednosti, ker so posamezne verjetnosti visoke. Vendar se z večanjem velikosti vzorcev posamezne verjetnosti zmanjšujejo in dobimo ustrezna območja zavrnitve. V našem primeru, če povečamo $\alpha = \frac{1}{6}$, je območje zavrnitve $\{3\}$, če je $\alpha = \frac{1}{3}$, je območje zavrnitve $\{3{,}4\}$ itn.

Podobno kot v tem zgledu, lahko izračunamo kritično vrednost za poljubna dva vzorca velikosti m in n, pri dani stopnji značilnosti α. Če H_0 drži, je vsaka razporeditev rangov naših $m+n$ opazovanj enako verjetna. Za te posamezne razporeditve lahko izračunamo vsoto rangov in posledično porazdelitev U-ja ter kritično vrednost. Pomembno je omeniti, da nikjer v zgledu nismo predpostavili, da sta opazovanja vzorca nekaterih verjetnostnih porazdelitev.

Kritične vrednosti ponavadi ne računamo kot v tem zgledu, temveč so tabelirane za dane m, n in α, podobno kot tabele, ki jih poznamo za t-porazdelitev in normalno porazdelitev.

Ko enkrat poznamo kritične vrednosti, lahko določimo območja zavrnitve. Odvisno od tega, ali vzamemo enostranski ali obojestranski test, pri stopnji značilnosti α, ničelno domnevo zavrnemo ko:

- $H_1 : P(X < Y) \neq P(X > Y)$ $\qquad U \leq U_\alpha$
- $H_1 : P(X < Y) < P(X > Y)$ $\qquad U_Y \leq U_{2\alpha}$
- $H_1 : P(X < Y) > P(X > Y)$ $\qquad U_X \leq U_{2\alpha}$

Pri tem sta $U_\alpha, U_{2\alpha}$ ustrezne kritične vrednosti. To sta največji vrednosti, za kateri pripadajoča stopnja značilnosti je manjša ali enaka α, oz. 2α.

1.5.2. *Zgled:* Uporabimo ista vzorca kot pri prejšnjem zgledu:

Vzorec X: 14,9, 11,3, 13,2, 16,6, 17,0, 14,1, 15,4, 13,0, 16,9.
Vzorec Y: 15,2, 19,8, 14,9, 18,3, 16,2, 18,9, 12,2, 15,3, 19,4.

Range smo že pokazali v Tabeli 1. Določimo domneve:

$H_0 : P(X < Y) = P(X > Y)$
$H_1 : P(X < Y) > P(X > Y)$

Izberemo $\alpha = 0{,}05$. Kritična vrednost za $m = n = 9$ je $U_{0,10} = 21$. $R_X = 68{,}5$, in zato je $U_X = 68{,}5 - \frac{9 \cdot 9}{2} = 28$. Ker je $28 > 21$, ne zavrnemo H_0 pri stopnji značilnosti $\alpha = 0{,}05$.

1.6. **Moč testa.** Eden izmed razlogov zakaj neparametrični testi niso najpogosteje uporabljeni testi je njihova moč, ki je asimptotično nižja kot pri testih, ki temeljijo na normalni porazdelitvi. Pri našem testu, je moč najbolj odvisna od velikosti vzorcev. Moč testa lahko primerjamo z močjo pri testu t le v primeru, ko sta populaciji normalno porazdeljeni. Izkaže se, da ima test Mann-Whitney asimptotično relativno učinkovitost (ang. *asymptotic relative efficiency*), ki v najslabšem primeru znaša 0,864 glede na test t. To pomeni, da če bi hoteli dobiti isto moč kot pri testu t moramo imeti $\frac{1}{0{,}864}$-krat večji vzorec kot ga imamo. Več o tem lahko najdemo v [4].

Obstajajo tudi primeri, ko predpostavke testa t niso izpolnjene, vendar je njegova uporaba koristna. Velikokrat se zgodi, da je število različnih vrednosti (razredov) vzorcev fiksno in ga vnaprej poznamo (npr. v anketah pogosto izbiramo med urejenimi odgovori po priljubljenosti). Izkaže se, da če so razredi pametno izbrani, lahko s testom t odkrijemo statistično značilne razlike v primerih, ko nam to ne uspe s testom Mann-Whitney [2]. Več o tem lahko najdemo v [16], kjer sta avtorja podrobno analizirala rezultate obeh testov za 5-razredne vzorce (ang. *5-point Likert scale*). Njun splošni zaključek je bil, da je v večini primerov razlika med močmi testov trivialna, vendar obstajajo primeri, ko moramo biti pozorni, kateri test bomo uporabili[1].

[1]V članku sta avtorja naredila simulacije, kjer sta konstruirala 14 različnih porazdelitev. Za različne kombinacije velikosti vzorcev, sta primerjala kateri test pogosteje pravilno odloča, kolikšna je moč testov in kolikšna je napaka prve vrste, ter sta objavila tabele, v katerih se vidi, kdaj se kateri test boljše obnaša. Poleg ugotovitve, da je razlika med močmi pogosto trivialna, sta ugotovila tudi, da je napaka prve vrste pri obeh testih dosti blizu zaželjene vrednosti in da če obstaja razlika

2. θ in intervali zaupanja

Z dosedanjo definicijo testa Mann-Whitney smo si pogledali pristop s katerim lahko dokažemo, da vzorca prihajata iz različnih populacij. Z vsakim testom dobimo zavrnitveno območje in na podlagi tega zavrnemo ali ne zavrnemo ničelno domnevo. Pogosto pa želimo oceniti kolikšno je odstopanje med populacijami. Medtem ko nam vrednost p pove, ali obstaja razlika med populacijami, je sama vrednost dokaj problematična, saj poda le binarni odgovor, poleg tega pa ne moremo govoriti o pomenu njenih posameznih vrednosti. Vemo tudi, da je zelo odvisna od velikosti vzorcev in za dovolj velikih vzorcev nam test odkrije skoraj vsako razliko med populacijami, razen če se popolnoma ujameta [14]. Zaradi teh razlogov bomo poiskali mero odstopanj naših populacij. Definirajmo:

$$\theta := P(X < Y) + \frac{1}{2} \cdot P(X = Y)$$

Rečemo ji stopnja prekrivanja. θ je sprejeta kot splošna vrednost, ki nam grobo pokaže razliko populacij [7]. Lahko zavzame vrednosti med 0 in 1, kar pomeni, da imamo univerzalno mero, ne glede na podatke. Če H_0 velja ($H_0 : F \overset{(d)}{=} G$), je $\theta = 0{,}5$. Obrat ne velja vedno.

V naslednjem zgledu si bomo pogledali, zakaj je θ primerna mera.

2.0.1. *Zgled.* Privzet iz [7]. Privzemimo, da sta $X \sim N(0{,}1)$ in $Y \sim N(\delta{,}1)$, za $\delta \in (0, \infty)$. Velja:

$\delta = E(Y) - E(X)$

$\theta = P(X < Y) + \frac{1}{2} \cdot P(X = Y) = P(X - Y \leq 0) = \Phi(\frac{\delta}{\sqrt{2}})$, kjer smo upoštevali, da je $X - Y \sim N(-\delta, 2)$, kar lahko izpeljemo s pomočjo karakterističnih funkcij.

V tabeli na Sliki 1 smo prikazali kakšen je odnos med θ in δ v obeh smereh. Izberemo θ kot mero z lepšimi lastnostmi, ker je manj odvisna od predpostavk porazdelitev v primerjavi z δ, oz. razliko matematičnih upanj [7]. θ je tudi bolj prikladna v primeru, ko narašča proti 1, saj vemo, da je tedaj δ velika, ampak ne moremo izbrati točno določene vrednosti. Omenimo še, da pri porazdelitvah, ki niso normalna, δ ni najbolj smiselna mera. Zato se bomo v nadaljevanju osredotočili le na θ.

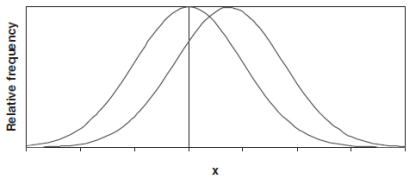


Figure 1. Relative frequency curves for Gaussian distributions N(0,1) and N(δ, 1) with $\delta = 0.7416$ corresponding to $\theta = 0.7$.

Table I. Correspondence between proposed measure θ and standardized difference δ in the homoscedastic Gaussian case.

θ	δ	δ	θ
0.50	0.000	0.00	0.500
0.55	0.178	0.25	0.570
0.60	0.358	0.50	0.638
0.65	0.545	0.75	0.702
0.70	0.742	1.00	0.760
0.75	0.954	1.25	0.812
0.80	1.190	1.50	0.856
0.85	1.466	1.75	0.892
0.90	1.812	2.00	0.921
0.95	2.326	2.50	0.961
0.99	3.290	3.00	0.983
0.999	4.370	3.50	0.993
		4.00	0.998

Slika 1. Graf in tabela sta privzeta iz [7].

med populacijami, jo bosta oba testa odkrila s približno isto verjetnostjo [2]. Več informacij glede rezultatov lahko najdemo v [16], povzetek tega pa lahko najdemo v [2].

2.1. **Ocena za θ.** Ker v praksi porazdelitvi X in Y praviloma nista podani, poskušamo oceniti θ na podlagi vzorcev. Definirajmo:

$$\hat{\theta} := \frac{1}{m \cdot n} \cdot \sum_{i=1}^{n} \sum_{j=1}^{m} U_{ij}$$

$$U_{ij} := \begin{cases} 1 & , \text{če } X_i < Y_j \\ \frac{1}{2} & , \text{če } X_i = Y_j \\ 0 & , \text{sicer} \end{cases}$$

θ je najbolj pogosto ocenjena z $\hat{\theta}$ [7]. Glavni razlog je njena povezanost s statistiko U, oziroma ta izrek:

Izrek 2.1.

$$\hat{\theta} = \frac{U_Y}{mn}$$

Dokaz. Naslednji dokaz je privzet in prilagojen iz [12], in velja samo, ko vzorca nimata vezanih rangov. Definirajmo:

$$V_{ij} := \begin{cases} 1 & \text{če } X_{(i)} < Y_{(j)} \\ 0 & \text{sicer} \end{cases}$$

Kjer smo z $X_{(i)}$ in $Y_{(j)}$ označili vrstilne statistike. Velja:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} U_{ij} = \sum_{i=1}^{n} \sum_{j=1}^{m} V_{ij},$$

saj so V_{ij} prerazporeditev U_{ij}. Vsoto na desni lahko izpeljemo tudi na drugi način:

$$\begin{aligned} \sum_{i=1}^{n} \sum_{j=1}^{m} V_{ij} = &(\text{število X-ov, ki so manjši kot } Y_{(1)}) \\ &+ (\text{število X-ov, ki so manjši kot } Y_{(2)}) + \\ &+ \cdots + (\text{število X-ov, ki so manjši kot } Y_{(m)}). \end{aligned}$$

Če z R_{ky} označimo rang statistike $Y_{(k)}$, je takrat število X-ov, manjših kot $Y_{(1)}$, enako $R_{1y} - 1$, stevilo X-ov, manjših kot $Y_{(2)}$, pa enako $R_{2y} - 2$, itn. Iz tega sledi:

$$\begin{aligned} \sum_{i=1}^{n} \sum_{j=1}^{m} V_{ij} &= (R_{1y} - 1) + (R_{2y} - 2) + (R_{my} - m) = \\ &= \sum_{i=1}^{m} R_{iy} - \sum_{i=1}^{m} i = R_Y - \frac{m(m+1)}{2} = U_Y \end{aligned}$$

Torej:

$$\hat{\theta} = \frac{U_Y}{mn}.$$

□

Povezava med Mann-Whitneyjevo statistiko U in oceno za θ je glavni razlog, zakaj v praksi uporabljamo U kot testno statistiko in ne vsoto rangov vzorcev.

2.2. **Intervali zaupanja.** V tem podrazdelku se bomo posvetili izračunu intervalov zaupanja, saj se pogosto v praksi poleg točkovne ocene poda tudi intervalno oceno. Intervali zaupanja so zanimiv problem za naš test, saj je eksplicitni izračun pogosto zelo zapleten, in zato običajno uporabljamo asimptotske metode. Zanimivost problema je tudi v tem, da ne obstaja ena metoda, za katero lahko rečemo, da je najboljša v večini primerov. Izračun intervalov zaupanja bo tudi glavni fokus te diplomske naloge.

Metode za izračun intervala zaupanja razdelimo v dve skupini:

- eksaktne metode (ang. *tail-area-based methods*),
- asimptotske metode.

Newcombe, v prvem delu svojega članka [7], je v podrobnosti razložil težave z eksaktnimi metodami. Kot glavni razlog je poudaril računsko zahtevnost teh metod, saj moramo pri izračunu obravnavati vse možne vrstne rede X-ov in Y-ov ter verjetnost teh dogodkov. Opozoril je, da postanejo te metode zelo počasne za vzorce velikosti deset ali več. V drugem delu članka [8] je predstavil osem različnih asimptotskih metod računanja intervalov zaupanja in naredil podrobno analizo kako se vsaka metoda obnaša pri različnih pogojih. V nalogi se bomo omejili na peto Newcombovo metodo, ki jo je predlagal kot najboljša izmed vseh obravnavanih. Zato jo sedaj predstavimo.

Kot prej, vzemimo vzorec velikosti n: X_1,...,X_n, kjer so sl. sprem. neodvisne in enako porazdeljene s porazdelitveno funkcijo F in drug vzorec velikosti m: Y_1,Y_2,...,Y_m, kjer so sl. sprem. neodvisne in enako porazdeljene s porazdelitveno funkcijo G. X_i in Y_j sta neodvisna za poljubna $i = 1, ..., n, j = 1, ..., m$.

$$\theta = P(X < Y) + \frac{1}{2} \cdot P(X = Y)$$

Definirajmo cenilko za θ:

$$\hat{\theta} = \frac{1}{m \cdot n} \cdot \sum_{i=1}^{n} \sum_{j=1}^{m} U_{ij}.$$

Definirajmo še cenilko za varianco $\hat{\theta}$:

$$V(\theta) := \frac{\theta \cdot (1-\theta) \cdot [1 + n_1 \cdot (1-\theta)/(2-\theta) + m_1 \cdot \theta/(1+\theta)]}{m \cdot n}, \tag{1}$$

kjer sta $m_1 = n_1 = \frac{1}{2}(m+n) - 1$.

$(1-\alpha)\%$ interval zaupanja za θ definiramo kot $[\theta_1, \theta_2]$, kjer sta θ_1 in θ_2 ustrezne ničle kvartilne enačbe:

$$(\theta - \hat{\theta})^2 = \Phi\left(\frac{1-\alpha}{2}\right)^2 \cdot V(\theta) \tag{2}$$

Metoda, tako kot ostale metode v članku [8], je bila razvita s predpostavko, da sta F in G zvezni porazdelitvi, torej v primeru, ko nastanejo vezani rangi z verjetnostjo 0. Vseeno, bomo dopuščali vezane range, saj se podatki hranijo v diskretni obliki [8].

2.2.1. *Izpeljava ocene variance in intervala zaupanja.* Formulo za varianco bomo dokazali, oziroma predstavili bomo kako je dobila svojo končno obliko, saj je več avtorjev prispevalo k njeni obliki. Dopolnili bomo tudi praznine, ki so jih avtorji v svojih člankih izpustili. Pri izpeljavi se bomo omejili samo na zvezno porazdeljene X in Y.

Prvi prispevek je naredil Noether, ki je v svoji knjigi [9] izpeljal formulo za varianco:

$$Var\left(\sum_{i=1}^{n}\sum_{j=1}^{m}U_{ij}\right) = mnVar(U_{12}) + \sum_{\substack{i=1\\j=1}}^{m,n}\sum_{\substack{k=1\\l=1}}^{m,n}Cov(U_{ij},U_{kl}) =$$

$$= mnVar(U_{12}) + \sum_{\substack{i=1\\j=1}}^{m,n}\sum_{\substack{l=1\\l\neq j}}^{n}Cov(U_{ij},U_{il}) + \sum_{\substack{i=1\\j=1}}^{m,n}\sum_{\substack{k=1\\k\neq i}}^{m}Cov(U_{ij},U_{kj}) +$$

$$\sum_{\substack{i=1\\j=1}}^{m,n}\sum_{\substack{k=1\\l=1\\k\neq i\\l\neq j}}^{m,n}Cov(U_{ij},U_{kl}) \tag{3}$$

U_{12} je Bernoulijeva slučajna spremenljivka s parametrom θ, torej $Var(U_{12}) = \theta(1-\theta)$. Zadnja kovarianca v enačbi (3) je enaka 0, ker je (U_{ij}, U_{kl}) porojen iz dveh neodvisnih parov slučajnih spremenljivk. Preostala dva sumanda razpišemo in dobimo:

$$Cov(U_{ij},U_{il}) = E(U_{ij}U_{il}) - \theta^2 = P(X_i < Y_j, X_i < Y_l) - \theta^2,$$

$$Cov(U_{ij},U_{kj}) = E(U_{ij}U_{kj}) - \theta^2 = P(X_i < Y_j, X_k < Y_j) - \theta^2.$$

Ko upoštevamo te vrednosti v formuli in pokrajšamo, dobimo:

$$Var(\hat{\theta}) = \frac{1}{m^2n^2}Var\left(\sum_{i=1}^{n}\sum_{j=1}^{m}U_{ij}\right) =$$

$$= mnVar(U_{12}) + mn(n-1)Cov(U_{12},U_{13}) + mn(m-1)Cov(U_{12},U_{32})$$

$$= \frac{1}{mn}\cdot(\theta\cdot(1-\theta) + (n-1)(Q_1-\theta^2) + (m-1)(Q_2-\theta^2)),$$

kjer sta $Q_1 := P(X_i < Y_j, X_i < Y_l)$ in $Q_2 := P(X_i < Y_j, X_k < Y_j)$.

Vidimo, da je $Var(\hat{\theta})$ odvisna od Q_1 in Q_2, ki pa sta odvisna od porazdelitev F in G. Ker jih v praksi ponavadi ne poznamo, bomo pri ocenjevanju variance za $\hat{\theta}$ morali oceniti tudi Q_1 in Q_2. Hanley in McNeil [3] trdita, da sta Q_1 in Q_2 odvisna predvsem od θ in manj od porazdelitev F in G. Preverila sta odvisnost variance od θ za različne pare porazdelitev (normalne, gama, eksponentne) in ugotovila, da ko sta X in Y eksponentno porazdeljena, je standardna napaka malo večja kot v ostalih primerih. Zato so ocenili Q_1 in Q_2 na podlagi predpostavke, da sta F in G eksponetno porazdeljena in s tem je varianca $\hat{\theta}$ kvečjemu precenjena.

Torej, predpostavimo, da: $F \sim Exp(\lambda)$, $G \sim Exp(\gamma)$. Velja:

$$\theta = P(X < Y) = \int_0^\infty dy \int_0^y \lambda e^{-\lambda x} \gamma e^{-\gamma y} dx = \int_0^\infty \gamma e^{-\gamma y}(1 - e^{-\lambda y}) dy$$

$$\theta = 1 + \frac{\gamma}{\gamma + \lambda}(0 - 1) = \frac{\lambda}{\lambda + \gamma} = \frac{1}{1 + \frac{\gamma}{\lambda}} \tag{4}$$

Na podoben način izrazimo Q_1:

$$Q_1 = P(X < Y_1, Y_2) = \int_0^\infty dx \int_x^\infty dy_1 \int_x^\infty \lambda e^{-\lambda x} \gamma^2 e^{-\gamma(y_1 + y_2)} dy_2 =$$

$$= \gamma\lambda \int_0^\infty e^{-x(\lambda+\gamma)} dx \int_x^\infty e^{-\gamma y_1} dy_1 =$$

$$= \lambda \int_0^\infty e^{-x(2\gamma+\lambda)} dx = \frac{\lambda}{\lambda + 2\gamma} = \frac{1}{1 + 2\frac{\gamma}{\lambda}}$$

Ko upoštevamo povezavo iz (4), dobimo:

$$Q_1 = \frac{1}{1 + 2(\frac{1}{\theta} - 1)} = \frac{\theta}{2 - \theta}$$

Na isti način izpeljemo:

$$Q_2 = 1 - \frac{2\lambda}{\lambda + \gamma} + \frac{\lambda}{\lambda + 2\gamma},$$

oziroma:

$$Q_2 = \frac{2\theta^2}{1 + \theta}.$$

Ko upoštevamo dobljene rezultate za Q_1 in Q_2, dobimo oceno za varianco $\hat{\theta}$:

$$V(\theta) = \frac{\theta \cdot (1 - \theta) \cdot [1 + (n - 1) \cdot (1 - \theta)/(2 - \theta) + (m - 1) \cdot \theta/(1 + \theta)]}{m \cdot n} \tag{5}$$

Sedaj pa poglejmo izpeljavo intervala zaupanja. Uporabili bomo idejo Wilsona [15], ki je isti pristop uporabil za ocenjevanje deleža. Predpostavimo, da je $\hat{\theta}$ normalno porazdeljena (ideja pride iz tega, da je $\hat{\theta}$ asimptotsko normalno porazdeljena [12]), in kot varianco uporabimo formulo (5). Zapišemo:

$$P\left(|\hat{\theta} - \theta| < \Phi\left(1 - \frac{\alpha}{2}\right)\sqrt{V(\theta)}\right) = 1 - \alpha$$

$$P\left((\hat{\theta} - \theta)^2 < \Phi\left(1 - \frac{\alpha}{2}\right)^2 V(\theta)\right) = 1 - \alpha$$

Če preoblikujemo zgornjo enačbo in poiščemo meje intervala θ_1 in θ_2, dobimo:

$$P(\theta_1 < \theta < \theta_2) = 1 - \alpha,$$

pri čemer moramo rešiti kvartilno enačbo, da dobimo θ_1 in θ_2:

$$(\hat{\theta} - \theta)^2 = \Phi(1 - \frac{\alpha}{2})^2 V(\theta)$$

Med štirimi koreni enačbe izberemo tiste, ki so realne in se nahajajo v intervalu [0,1]. Če najdemo dve taki ničli, ju uredimo po velikosti in razglasimo za meje intervala. Če dobimo samo eno tako ničlo, jo označimo θ_1, in izberemo interval $[\theta_1, 1]$, ali $[0, \theta_1]$, odvisno od tega, ali je θ_1 manjši ali strogo večji kot $\hat{\theta}$. V naših simulacijah nismo našli primera za tri ali več ničel, ki izpolnujejo začetne pogoje.

Kot zadnjo izboljšavo je Newcombe predlagal, da v formuli za varianco (5) namesto $(m-1)$ in $(n-1)$ uporabimo $m_1 = n_1 = \frac{1}{2}(m+n) - 1$. S tem dobimo našo oceno za varianco (1).

Če povzamemo peto Newcombovo metodo:

- Definiramo $V(\theta) := \frac{\theta \cdot (1-\theta) \cdot [1 + n_1(1-\theta)/(2-\theta) + m_1\theta/(1+\theta)]}{mn}$, kjer sta $m_1 = n_1 = \frac{1}{2}(m+n) - 1$.
- Poiščemo korene enačbe $(\hat{\theta} - \theta)^2 = \Phi\left(1 - \frac{\alpha}{2}\right)^2 V(\theta)$.
- Izberemo tiste korene, ki ustrezajo našim pogojem in jih označimo kot meje intervala zaupanja.

2.3. **Simulacije.** V tem podrazdelku bomo pogledali kakšne intervale zaupanja nam generira peto Newcombovo metodo. Za vnaprej dane porazdelitve bomo k–krat vzorčili iz teh porazdelitev in analizirali dobljenih intervalov. Kot glavno mero učinkovitosti metode bomo uporabili verjetnost pokritja (ang. *coverage probability*), ki predstavlja delež vseh ponovitev, za katere je θ vsebovana v izračunanem intervalu zaupanja. Torej, če predpostavimo, da metoda pravilno generira 95% intervale zaupanja, bomo pričakovali, da bo verjetnost pokritja blizu 0,95. Metodo bomo najprej testirali za porazdelitve za katere že imamo rezultate in potem bomo poiskali primere, ko metoda poda slabe rezultate

Uporabili bomo isto metodologijo kot pri [8]. Vzeli bomo vzorce velikosti:

(1) $m = n = 5$
(2) $m = 5, n = 20$
(3) $m = 20, n = 5$
(4) $m = n = 20$
(5) $m = 20, n = 100$
(6) $m = 100, n = 20$
(7) $m = n = 100$

Vnaprej bomo predpostavili vrednost θ, in na osnovi tega bomo poiskali parametre porazdelitev. Vzeli bomo naslednje vrednosti: $\theta \in \{0{,}5, 0{,}6, 0{,}7, 0{,}8, 0{,}9, 0{,}95, 0{,}99\}$ (v članku je testiral še za 0,999, vendar tega nismo vključili). Omejili smo se na $\theta \geq 0{,}5$, saj je pri $\theta < 0{,}5$ popolnoma simetrično.

Večina parov porazdelitev bo določena tako, da je prva porazdelitev F fiksna, medtem ko bo druga porazdelitev G imela znano obliko, vendar se bodo njeni parametri spreminjali v odvisnosti od θ.

Newcombe je tako testiral za šest parov porazdelitev, pri čemer so prvih pet bili normalno porazdeljeni, vsak par z različno kombinacijo parametrov, medtem ko je bil zadnji par binomsko porazdeljen. Ponovili bomo simulacije za prvi par in pogledali, ali dobimo dovolj podobne rezultate, oz. ali pravilno izvajamo naše simulacije. Zatem bomo metodo testirali na porazdelitvah, ki niso vključene v članku.

Za fiksno θ in velikosti vzorcev bomo navadno generirali 100000 vzorcev iz vnaprej izbranih porazdelitev. V nekaterih primerih bo število ponovitev manjše, zaradi časovne zahtevnosti, kar bo ustrezno navedeno.

Vse simulacije so bile izvedene v programski opremi R, različica 3.1.2.

2.3.1. *Normalne porazdelitve.* Predpostavimo: $F \sim N(0,1)$ in $G \sim N(\mu,1)$, oziroma $X_i \overset{(d)}{=} N(0,1)$, $Y_j \overset{(d)}{=} N(\mu,1)$, za vsak $i = 1,2,...,n$ in $j = 1,2,...,m$. Kot smo videli v zgledu 2.0.1, velja:

$$\theta = \Phi\left(\frac{\mu}{\sqrt{2}}\right)$$

Ker je θ vnaprej izbran in Φ strogo naraščajoča funkcija (torej obstaja inverz), lahko izpeljemo:

$$\mu = \sqrt{2} \cdot \Phi^{-1}(\theta).$$

V Tabeli 2 smo predstavili rezultate, ko 10000–krat generiramo vzorce iz F in G, pri $\theta = 0{,}8$:

m	n	average_left_endpoint	average_right_endpoint	average_width	min_width	max_width	cov_probability
5	5	0,44211	0,93878	0,49666	0,35730	0,58850	0,9608
20	5	0,52008	0,92553	0,40545	0,23930	0,49046	0,9620
5	20	0,52179	0,92661	0,40483	0,23930	0,49046	0,9631
20	20	0,62876	0,89809	0,26933	0,11064	0,33654	0,9553
100	20	0,67141	0,88132	0,20992	0,11259	0,26605	0,9580
20	100	0,67220	0,88184	0,20964	0,12460	0,26617	0,9580
100	100	0,73061	0,85294	0,12233	0,08807	0,14764	0,9558

TABELA 2

V tabeli smo za različne velikosti vzorcev podali naslednje vrednosti: povprečna leva in desna meja intervalov, povprečna širina intervalov, minimalna in maksimalna širina intervalov ter verjetnost pokritja. Stolpcev je toliko predvsem zato, ker želimo čim bolj nazorno prikazati, kako izgledajo intervali zaupanja in ali smo dobili smiselne rezultate. Poleg teh smo preverili tudi minimalne in maksimalne leve in desne meje, vendar smo jih zaradi preglednosti izpustili iz tabele. Opazimo, da je verjetnost pokritja stabilna za vse kombinacije vzorcev. Lahko sklepamo, da se širina intervalov zmanjšuje, ko se velikosti vzorcev povečujejo, kar je pričakovano (podobno, povprečna leva in desna meja se približujeta proti 0,8, ko povečujemo velikost vzorcev). Ena izmed glavnih prednosti te metode je to, da ne generira intervalov širine 0 [8], kar se je tudi pokazalo v tabeli. Takšni intervali so neuporabni, saj nam ne dajo nobene informacije o variabilnosti θ.

V Tabeli 2 lahko opazimo, da so vse verjetnosti pokritja večje ali enake 0,95. Vendar zanima nas kaj se zgodi, ko je neka verjetnost pokritja manj kot 0,95, oziroma ali je to odstopanje posledica naključja v simulacijah. Zato bomo poiskali neko spodnjo mejo, pod katero naša verjetnost pokritja ne bi smela biti. To bomo naredili tako, da bomo izračunali pripadajoči interval zaupanja.

2.3.2. *Interval zaupanja verjetnosti pokritja.* $X_i :=$ zavzame vrednost 1, če je pravi θ vsebovan v izračunanem intervalu zaupanja, 0 sicer. $X_i \sim \text{Ber}(0{,}95)$, za $i = 1,2,...,k$, kjer je k število ponovitev. Torej upoštevamo predpostavko, da naša metoda generira interval zaupanja, ki vsebuje pravo θ z verjetnostjo 0,95.

$$\sum_{i=1}^{k} X_i \sim Bin(k, 0{,}95) \approx N(k \cdot 0{,}95; k \cdot 0{,}95 \cdot 0{,}05)$$

Torej: $\bar{X} \approx \mathrm{N}(k \cdot 0{,}95, \frac{0{,}95 \cdot 0{,}05}{k})$.

(1-α)% interval zaupanja za verjetnost pokritja:

$$0{,}95 \pm \Phi^{-1}\left(\frac{1-\alpha}{2}\right)\sqrt{\frac{0{,}95 \cdot 0{,}05}{k}},$$

oziroma, pri $k = 100000$ je 99% interval zaupanja: $[0{,}9482; 0{,}95177]$.

Torej, pri predpostavki, da metoda pravilno generira intervale zaupanja, je verjetnost, da bo pokritje na intervalu $[0{,}9482; 0{,}95177]$ enaka 0,99. Vrednosti pokritij, ki bodo manjša od 0,9482, ne moremo pripisati naključju, temveč lahko z verjetnostjo 0,99 trdimo, da v tem primeru pokritje ni enako 0,95. To seveda še ne pomeni, da je metoda nesprejemljiva, vendar smo za boljšo preglednost v naslednjih tabelah z eno zvezdico označili tiste vrednosti, ki so med 0,92 in 0,9482, z dvemi zvezdicami pa tiste, ki so manjše kot 0,92.

Sedaj bomo poskusili narediti simulacije, za katere že imamo rezultate v članku. V Tabeli 3 smo predstavili naše rezultate, v Tabeli 4 pa rezultate, ki so vzeti iz članka, pri čemer smo izbrali: $m = n = 20$, $\theta = 0{,}8$ in naredili 100000 ponovitev.

average_width	min_width	max_width	cov_probability
0,26905	0,10133	0,33707	0,95607

Tabela 3

Method	Average width	Minimum width	Maximum width	Coverage probability
5	0.269100	0.101332	0.337024	0.95590

Tabela 4. Privzeta iz [8].

Vidimo, da se rezultati ne razlikujeta veliko. Vendar, da bi bili prepričani v našem postopku, lahko spet izračunamo 99% interval zaupanja za vsako vrednost iz Tabele 4. Dobimo naslednje intervale: $[0{,}2673; 0{,}2709]$ za povprečno širino intervala, $[0{,}09956; 0{,}10311]$ za minimalno širino intervala, $[0{,}33525; 0{,}33880]$ za maksimalno širino intervala, ter $[0{,}95413; 0{,}95768]$ za verjetnost pokritja. Vidimo, da je vsaka vrednost iz Tabele 3 vsebovana v pripadajočem intervalu zaupanja in s tem potrdimo, da smo z našimi simulacijami uspeli ponoviti rezultati avtorja.

V Tabeli 5 smo za različne θ in $m = n = 20$ izračunali iste vrednosti kot v Tabeli 4. Simulacije so bile ponovljene 10000–krat.

theta	average_width	min_width	max_width	cov_probability
0,50	0,33151	0,19065	0,33702	0,9528
0,60	0,32488	0,22295	0,33702	0,9524
0,70	0,30515	0,14783	0,33702	0,9526
0,80	0,26948	0,10133	0,33696	0,9573
0,90	0,21007	0,10133	0,32849	0,9689
0,95	0,16658	0,10133	0,28015	0,9690
0,99	0,11770	0,10133	0,21368	0,9754

Tabela 5

Vidimo, da je za vse θ verjetnost pokritja več kot 0,95. Za konec, prikažimo simulacijo, kjer upoštevamo vse predlagane kombinacije vzorcev in to ponovimo 100000:

theta	average_width	min_width	max_width	cov_probability
0,50	0,36026	0,25015	0,37099	0,94814*
0,60	0,35460	0,25011	0,37097	0,95157
0,70	0,33594	0,21855	0,37057	0,95714
0,80	0,30232	0,19392	0,36743	0,96029
0,90	0,24821	0,16869	0,34888	0,96186
0,95	0,21107	0,15908	0,32429	0,96957
0,99	0,16131	0,13389	0,25524	0,97100

Tabela 6

Glede na to, da je samo ena povprečna verjetnost pokritja manj kot 0,95, lahko rečemo, da smo zadovoljni s temi rezultati.

Dosedanje simulacije so bile narejene za normalne porazdelitve. V tem primeru imamo vedno na voljo test t, zato bomo testirali metodo za porazdelitve, pri katerih so predpostavke testa t močno kršene. En tak primer je Cauchyjevo porazdelitev, ki nima momentov in zato jo ne moremo obravnavati s parametričnimi metodami. Drug primer pa je ko imamo zelo asimetrične porazdelitve.

2.3.3. *$X \sim$ Cauchy(0,1), $Y \sim$ Cauchy(μ, 1).* Najprej si poglejmo, kako najti μ. Cauchyjevo porazdelitev s parametri a,b ima porazdelitveno funkcijo:

$$F(x; a, b) = \frac{1}{\pi} \cdot \arctan\left(\frac{x-a}{b}\right) + \frac{1}{2}$$

in gostoto:

$$f(x; a, b) = \frac{1}{\pi \cdot b \cdot (1 + (\frac{x-a}{b})^2)}.$$

Da bi izračunali θ, nas zanima $P(X < Y) = P(X - Y < 0)$. Izkaže se, da je $X - Y$ porazdeljen tudi Cauchy, to pa dokažemo s pomočjo karakterističnih funkcij.

Naj bosta $C_1 \sim$ Cauchy(a_1,b_1), $C_2 \sim$ Cauchy(a_2, b_2) in med seboj neodvisna. Katakeristična funkcija ima obliko:

$$\varphi_{C_1}(t) = E[e^{iC_1 t}] = e^{ia_1 t - b_1|t|}$$

Torej:

$$\varphi_{C_1 - C_2}(t) = E[e^{i(C_1 - C_2)t}] = E[e^{iC_1 t} \cdot e^{iC_2(-t)}] = e^{ia_1 t - b_1|t|} \cdot e^{-ia_2 t - b_2|t|} =$$

$$= e^{i(a_1 - a_2)t - (b_1 + b_2)|t|}$$

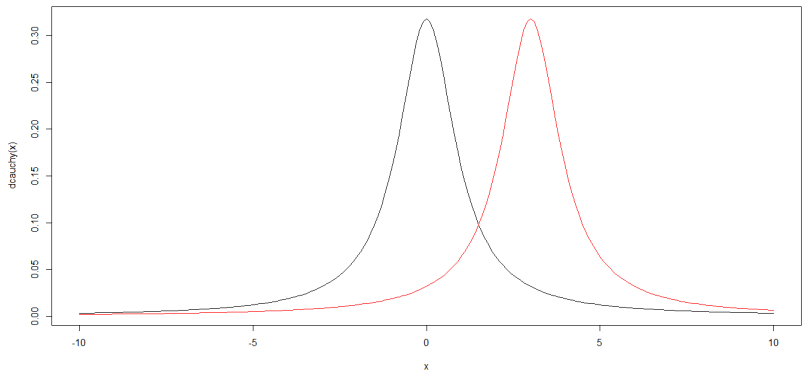
Vidimo: $C_1 - C_2 \sim$ Cauchy$(a_1 - a_2, b_1 + b_2)$, oziroma v našem primeru: $X - Y \sim$ Cauchy$(-\mu, 2)$. Torej upoštevamo, da $x = 0$, $a = -\mu$ in $b = 2$, in dobimo:

$$\theta = F(0; -\mu, 2) = \frac{1}{\pi} \cdot \arctan\left(\frac{\mu}{2}\right) + \frac{1}{2}$$

Če malo preuredimo: $\tan[\pi(\theta - \frac{1}{2})] = \frac{\mu}{2}$ in na koncu dobimo:

$$\mu = 2 \cdot \tan \pi(\theta - \frac{1}{2}).$$

Torej za dano θ lahko določimo parameter μ. Poglejmo si kako zgledata gostoti verjetnosti, ko je $\mu = 3$:



Slika 2. X ∼ Cauchy(0,1), Y ∼ Cauchy(3, 1)

V naslednjih simulacijah se bomo omejili samo na tri pare velikosti vzorcev, saj naša metoda ublaži vpliv vzorcev z različnimi velikostmi (m_1 in n_1 v formuli za varianco). Generirali bomo vzorce za: $m = n = 5$, $m = n = 20$, $m = n = 100$.

m in n	theta	average_theta_hat	average_width	cov_probability	m in n	theta	average_theta_hat	average_width	cov_probability
5	0,50	0,5004	0,5608	0,9443*	20	0,50	0,4999	0,3315	0,9511
5	0,60	0,5996	0,5534	0,9571	20	0,60	0,5998	0,3249	0,9480*
5	0,70	0,6991	0,5312	0,9541	20	0,70	0,6998	0,3047	0,9431*
5	0,80	0,8001	0,4914	0,9540	20	0,80	0,8002	0,2676	0,9297*
5	0,90	0,8997	0,4337	0,9186**	20	0,90	0,9000	0,2070	0,9406*
5	0,95	0,9500	0,3974	0,9664	20	0,95	0,9502	0,1620	0,9237*
5	0,99	0,9900	0,3655	0,9501	20	0,99	0,9899	0,1150	0,8223**

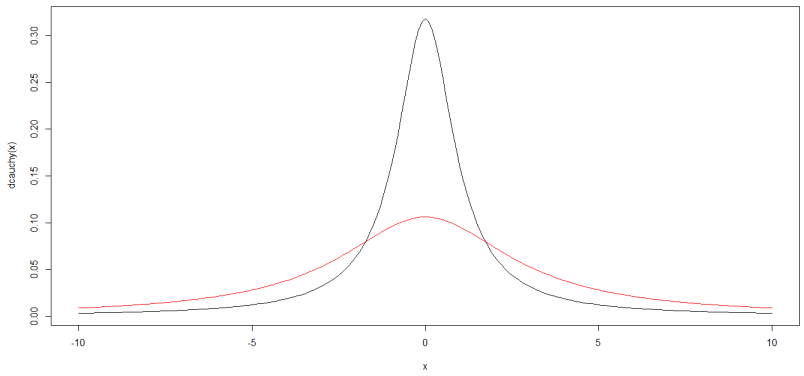
Tabela 7

m in n	theta	average_theta_hat	average_width	cov_probability
100	0,50	0,4999	0,15737	0,9496
100	0,60	0,5999	0,15369	0,9482*
100	0,70	0,7002	0,14231	0,9412*
100	0,80	0,8001	0,12210	0,9265*
100	0,90	0,9000	0,08945	0,9011**
100	0,95	0,9500	0,06446	0,8867**
100	0,99	0,9900	0,03240	0,9177**

Tabela 8

V tabelah 7 in 8 smo prikazali velikost vzorcev, privzet θ, ocenjen θ, povprečno širino in verjetnost pokritja. Vidimo, da so ocene θ dosti blizu prave vrednosti in da se povprečna širina intervalov zmanjšuje, ko θ in velikost vzorcev naraščata - kar bomo videli tudi za ostale pare Cauchyjevo porazdeljenih sl. spremenljivk. Če primerjamo te rezultate s tistimi, ko sta bila X in Y normalno porazdeljena, lahko opazimo, da smo našli primere, ko je verjetnost pokritja zelo daleč od pričakovane vrednosti 0,95 (npr. za $m = n = 20$ in $\theta = 0{,}99$ je verjetnost pokritja 0,82).

2.3.4. *$X \sim$ Cauchy(0,1), $Y \sim$ Cauchy(0, 3).* Gostoti verjetnosti izgledata tako:



Slika 3. X $\sim$ Cauchy(0,1), Y $\sim$ Cauchy(0, 3)

θ je fiksiran na 0,5. Poglejmo kako se metoda obnaša sedaj:

m	n	average_theta_hat	cov_probability
5	5	0,50100	0,94029*
20	20	0,50021	0,94160*
100	100	0,49985	0,93922*

Tabela 9

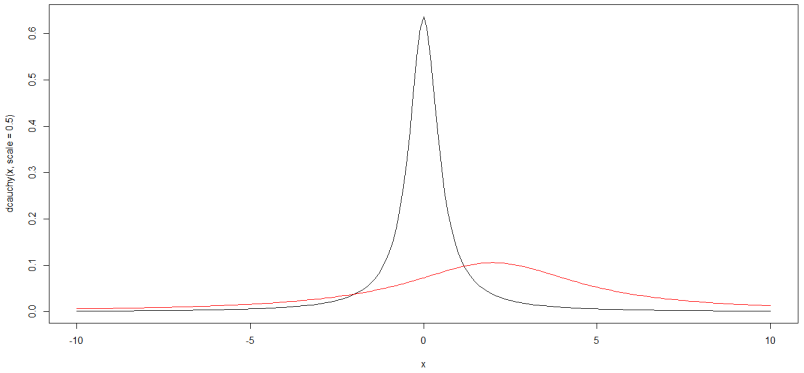
Čeprav je verjetnost pokritja manj kot 0,9482, lahko rečemo da je dovolj blizu 0,95.

S tem smo hoteli preveriti, kakšne rezultate dobimo, ko povečujemo variabilnost spremenljivk. Poglejmo kaj se zgodi, ko temu dodamo še premik srednje vrednosti.

2.3.5. *$X \sim$ Cauchy(0; 0,5), $Y \sim$ Cauchy(μ, 3).* Sedaj spet premikamo μ v odvisnosti od θ s to razliko, da je sedaj enačba takšna:

$$\mu = 3{,}5 \cdot tan(\pi \cdot (\theta - \frac{1}{2})).$$

Pogljemo, kako izgledajo gostote, za $\mu = 2$:



Slika 4. X $\sim$ Cauchy(0; 0,5), Y $\sim$ Cauchy(2, 3)

In rezultate:

theta	5.cov_probability	20.cov_probability	100.cov_probability
0,50	0,9356*	0,9309*	0,9299*
0,60	0,9280*	0,9272*	0,9262*
0,70	0,9507	0,9255*	0,9181**
0,80	0,9474*	0,9074**	0,9048**
0,90	0,9188**	0,9373*	0,8835**
0,95	0,9710	0,9232*	0,8828**
0,99	0,9508	0,8181**	0,9206*

TABELA 10

V tabeli 10 smo prikazali verjetnosti pokritja za vzorce velikosti $5, 20$ in 100, pri različnih θ. Dobili smo najnižje verjetnosti pokritja med vsemi simulacijami za Cauchyjevo porazdeljene spremenljivke. Vidimo, da so za vzorce velikosti 20 in 100 verjetnosti pokritja manjše od 0,93, medtem ko so za vzorce velikosti pet rezultati dokaj raznoliki.

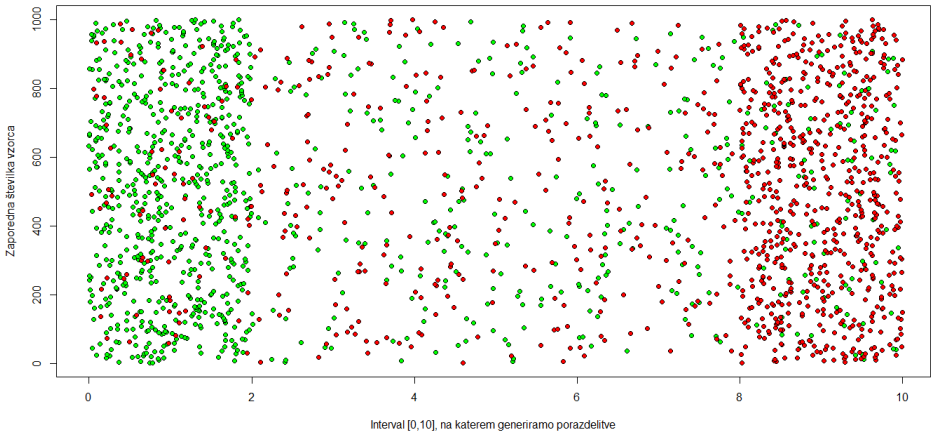
Za konec, naredimo simulacije za bolj zapletene asimetrične porazdelitve.

2.3.6. *F in G asimetrična kombinacija enakomernih porazdelitev.* Naj bosta:

$$F(p) = \begin{cases} Unif([0,2]) & \text{1 - p} \\ Unif([0,10]) & \text{p} \end{cases}$$

$$G(p) = \begin{cases} Unif([8,10]) & \text{1 - p} \\ Unif([0,10]) & \text{p} \end{cases}$$

Pri čemer je p $\in$ [0,1]. Naj bosta $X_i(p) \overset{(d)}{=} F(p)$ in $Y_j(p) \overset{(d)}{=} G(p)$. Vsak $X_i(p)$ generiramo z verjetnostjo $(1-p)$ iz porazdelitve $Unif([0,2])$ in z verjetnostjo p iz porazdelitve $Unif([0,10])$, ekvivalentno za $Y_j(p)$. Poglejmo, kako izgledata porazdelitvi za $p = 0{,}45$:



SLIKA 5. Generirana vzorca iz F in G, ko je $p = 0{,}45$

Zanima nas θ, oziroma $P(X < Y)$.

$$\theta = P(X < Y) = p \cdot p \cdot \frac{1}{2} + (1-p) \cdot p \cdot \frac{9}{10} + p \cdot (1-p) \cdot \frac{9}{10} + (1-p)^2 \cdot 1 =$$

$$= -\frac{3}{10}p^2 - \frac{p}{5} + 1$$

Za določen θ, je torej p ničla enačbe $\frac{3}{10}p^2 + \frac{p}{5} + \theta - 1 = 0$. V vseh primerih je ena ničla negativna (torej odpade), druga pa je med 0 in 1, torej jo vedno izberemo. Tabela 11 pokaže še zvezo med p in θ.

theta	0,50	0,60	0,70	0,80	0,90	0,95	0,99
p	1,00	0,83	0,63	0,45	0,24	0,12	0,02

Tabela 11

Rezultati:

m in n	theta	average_theta_hat	cov_probability	m in n	theta	average_theta_hat	cov_probability
5	0,50	0,50272	0,9444*	20	0,50	0,50086	0,9513
5	0,60	0,59840	0,9594	20	0,60	0,60189	0,9548
5	0,70	0,70071	0,9541	20	0,70	0,69913	0,9536
5	0,80	0,79959	0,9615	20	0,80	0,79982	0,9507
5	0,90	0,89904	0,9369*	20	0,90	0,90012	0,9584
5	0,95	0,95184	0,9618	20	0,95	0,95002	0,9457*
5	0,99	0,98968	0,9354*	20	0,99	0,99004	0,9101**

Tabela 12

m in n	theta	average_theta_hat	cov_probability
100	0,50	0,49992	0,9486
100	0,60	0,59976	0,9520
100	0,70	0,70021	0,9489
100	0,80	0,80047	0,9518
100	0,90	0,90010	0,9485
100	0,95	0,94993	0,9418*
100	0,99	0,99002	0,9387*

Tabela 13

V tabelah 12 in 13 smo za vzorce velikosti $5, 20$ in 100 prikazali θ, ocenjena $\hat{\theta}$ in verjetnosti pokritja, pri čemer smo naredili 10000 ponovitev. Lahko opazimo, da so ocene θ dosti blizu prave vrednosti in da so verjetnosti pokritja visoki, razen v primeru, ko je $\theta = 0{,}99$.

S tem smo končali simulacije za peto Newcombovo metodo. Lahko sklepamo, da je v večini testiranih primerov metoda stabilna, vendar ostaja še prostora za izboljšavo, saj smo videli, da ko sta F in G Cauchyjevo porazdeljena smo našli nekaj primerov, ko je verjetnost pokritja dosti oddaljena od 0,95.

3. Krivulje ROC in povezava s θ

Krivulja ROC (ang. *receiver operating characteristic curve*) je statistično orodje, ki ga lahko uporabimo, ko želimo primerjati dve porazdelitvi in opaziti morebitne razlike med njima. Za nekatere statistične teste, jo lahko uporabimo pri izbiri kritične vrednosti in pri primerjavi rezultatov različnih raziskav. Poleg tega lahko z njo prikažemo različne pojave na isti lestvici. Za to diplomsko nalogo je zanimiva zato, ker obstaja med njo in θ povezava, kar zelo olajša računanje posameznih vrednosti krivulje. Krivulje ROC so se najprej razvile med drugo svetovno vojno za potrebe vojske, kasneje pa se njena uporaba razširila v medicini, radiologiji in računalništvu.

V tem poglavju so vsi primeri iz medicine, saj besedilo temelji na osnovi knjige [11], v kateri je avtorica razložila krivulje z vidika medicine.

3.1. **Uvod.** Naj bosta D in Y Bernoullijevi sl. spremenljivki, pri čemer nam D pove pravo vrednost nekega pojava, ki ga raziskujemo, Y pa določi oceno za D na podlagi nekega testa (npr. v medicini D lahko zavzema vrednost 1, če smo odkrili bolezen nekega pacienta in 0 sicer, medtem ko je Y rezultat diagnostičnega testa, s katerim smo čim bolj natančno poskusili določiti pravo stanje). Vendar se v praksi lahko zgodi, da nam Y poda napačni odgovor, zato želimo takšne napake čim bolj minimizirati. Zato tu definirajmo vsa možna stanja vektorja (D, Y): resnični pozitivni, resnični negativni, lažni pozitivni in lažni negativni (ang. *true positive, true negative, false positive, false negative*). To lahko prikazemo v tabeli:

	D = 0	D = 1
Y = 0	True negative	False negative
Y = 1	False positive	True positive

Test ima lahko 2 dve vrsti napak, lažno pozitivno in lažno negativno napako. Definirajmo naslednje verjetnosti s katerimi bomo te napaki ovrednotili:

- FPF = $P(Y = 1|D = 0)$ (ang. false positive fraction),
- TPF = $P(Y = 1|D = 1)$ (ang. true positive fraction),

kjer je FPF = verjetnost lažne pozitivne napake in 1 - TPF = verjetnost lažne negativne napake. Z (FPF, TPF) imamo definirane verjetnosti, s katerimi nastanejo te napake.

Če bi imeli popoln test, velja FPF = 0, TPF = 1. Po drugi strani, če bi imeli neuporaben test, oz. če posamezna vrednost Y ne pomaga določiti prave vrednosti D, velja TPF = FPF.

To testiranje lahko obravnavamo kot osnovni statistični test. Če torej privzamemo, da je $(D = 0)$ ničelna domneva, $(D = 1)$ alternativna domneva, FPF stopnja značilnosti in TPF moč testa, smo definirali klasičen statistični test.

Namesto da ločujemo med tema dvema napakama, lahko uporabimo:

$$P(D \neq Y) = P(D = 1) \cdot (1 - TPF) + P(D = 0) \cdot FPF$$

Vendar ima ta mera nekaj slabosti. Kot prvo sta obe napaki 1 - TPF in FPF različna in imata lahko različne posledice. Npr, v medicini, če ne odkrijemo bolezni, čeprav je prisotna (torej lažna negativna napaka), lahko to povzroči smrt. Na drugi strani lažne pozitivne napake ponavadi niso toliko nevarne. Bolniki, ki nimajo

bolezni, so podvrženi dodatnemu zdravljenju, vendar imajo dolgoročno majhne posledice. Drugi razlog, zakaj je boljše uporabljati (FPF, TPF), je v tem, da ima pogostost te bolezni vpliv na $P(D \neq Y)$. Na primer, če imamo trivialen test, ki vedno odgovori negativno (torej $P(Y = 0) = 1$), in je bolezen zelo redka ($P(D = 1)$ je blizu 0), bo tudi $P(D \neq Y)$ blizu 0, vendar bo test neuporaben, ker ne bomo nikoli prepoznali bolezni.

Na podoben način kot smo definirali (FPF, TPT) lahko določimo druge mere, ki bi nam pomagale v uporabi testa. Npr. z $P(D = 1|Y = 1)$ in $P(D = 0|Y = 0)$ merimo kolikšna je verjetnost, da je neka bolezen res prisotna, če nam diagnostični test to odgovori. Več o teh merah lahko najdemo v [11].

3.2. **Definicija.** Sedaj pa si poglejmo, ko Y ni Bernoullijeva, temveč zvezna, ali diskretna (s končno mnogo vrednostmi). Ker je D še zmeraj Bernoullijeva, bomo potrebovali neko kritično vrednost za Y, s katero bomo sklepali, o kakšnem stanju gre.

Brez škode za splošnost naj bosta Y in D povezani tako, da se za večje vrednosti Y poveča verjetnost, da smo v stanju ($D = 1$). Če Y nima te lastnosti, najprej poskusimo poiskati neko transformacijo, ki jo pretvori v tako obliko, in potem nadaljujemo. Običajno je dovolj, da pomnožimo z -1.

Izbira kritične vrednosti je odvisna od tega, v kakšnem odnosu želimo, da sta naši napaki (lažna pozitivna napaka se zgodi, ko je Y večji kot kritična vrednost in $D = 0$; obratno velja za lažno negativno napako). Za različne kritične vrednosti sta ti napaki v drugačnem odnosu, in prav to nam predstavi krivuljo ROC.

Označimo kritično vrednost s c in privzemimo, da je Y zvezna. Iz sl. spremenljivke Y ločimo dva odgovora:

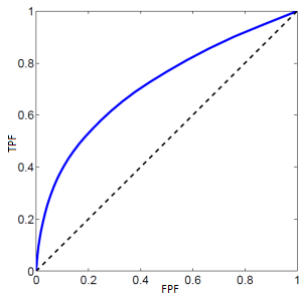
- pozitivni, če $Y \geq c$,
- negativni, če $Y < c$.

Definirajmo še ustrezna TPF in FPF:

$$\text{TPF(c)} = P(Y \geq c|D = 1)$$
$$\text{FPF(c)} = P(Y \geq c|D = 0)$$

Krivuljo ROC definiramo kot množico vseh parov (FPF(c), TPF(c)), ki jih lahko dobimo za različne kritične vrednosti, oziroma:

$$ROC(\cdot) = \{(FPF(c), TPF(c)), c \in (-\infty, \infty)\}$$



Slika 6. Primer krivulje ROC, označena z modro barvo. Slika privzeta iz [17]

Vidimo, da:

- ko $c \to \infty$, $\lim_{c\to\infty} FPF(c) = 0$ in $\lim_{c\to\infty}$ TPF(c) = 0,
- ko $c \to -\infty$, $\lim_{c\to-\infty}$ FPF(c) = 1 in $\lim_{c\to-\infty}$ TPF(c) = 1.

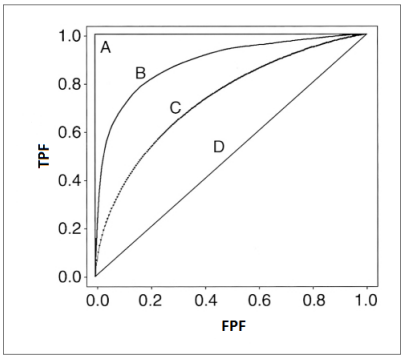
Opazimo, da se krivulja nahaja v prostoru (0,1)×(0,1). Drug način zapisa krivulje ROC je naslednji:

$$ROC(\cdot) = \{(t, ROC(t)), t \in (0,1)\},$$

kjer je ROC funkcija, ki slika iz t v TPF(c), in je c kritična vrednost, ki ustreza pogoju $FPF(c) = t$.

Kot prej, test je neuporaben, ko vrednost Y nam ne pomaga v določitvi stanja D. Takrat je za poljuben c TPF(c) = FPF(c), torej ROC je identična funkcija na intervalu (0,1). Na drugi strani je test popoln, če obstaja neka kritična vrednost c, za katero velja TPF(c) = 1 in FPF(c) = 0. Takrat privzamemo, da poteka krivulja ROC po levem in zgornjem robu enotnega kvadrata.

V večini primerov, je krivulja ROC nekje med krivuljo za popoln in krivuljo za neuporaben test. Bolj efektivni testi imajo krivuljo ROC, ki je blizu levega zgornjega kota. Na primer, če imamo dva testa B in C in pri dani vrednost FPF, ima B višjo TPF kot TPF testa C, rečemo, da je B boljši test kot C. Podobno, če izberemo kritične vrednosti c_B in c_C, za katere $TPF_B(c_B) = TPF_C(c_C)$, potem je $FPF_B(c_B) \leq FPF_C(c_C)$.



Slika 7. Prikažana sta zgoraj navedena primera krivulj ROC: A – popoln test, D – neuporaben test, ter B in C test, kjer je očitno B boljši kot C. Slika privzeta iz [10]

.

3.3. **Uporaba ROC krivulj.** Pogledali smo, kako nam krivulja ROC prikaže vse možne pare (FPF, TPF) in si s tem pomagamo, ko primerjamo več testov. Vendar se pogosto v praksi zgodi, da ne moremo narisati celotne krivulje, oziroma da imamo podatke samo za eno ali več kritičnih vrednosti. S tem postane primerjava več raziskav še težja, zlasti če je imel vsak test različno kritično vrednost.

S krivuljami ROC lahko se lotimo pri izbiri najustreznejše kritične vrednosti nekega testa. Izkaže se, da lahko problem prevedemo na minimizacijski problem, ki nam da optimalno vrednost. Več o tem, lahko najdemo v (Pepe, 2004).

Poleg tega, lahko krivulje ROC uporabimo tudi kot orodje, ko primerjamo dve porazdelitvi in želimo zaznati razlike med njima. Torej, naj bosta X_1 in X_2 dve neodvisni slučajni spremenljivki s porazdelitvami F in G. Slučajno spremenljivko D

definiramo tako, da je enaka 1, če je posamezna vrednost iz X_1, in 0, če je vrednost iz X_2. Definirajmo še:

$Y|(D=1) \overset{(d)}{=} X_1$, in
$Y|(D=0) \overset{(d)}{=} X_2$.

Torej, če poznamo porazdelitvi F in G ter za dano kritično vrednost c, lahko izračunamo posamezen par (FPF(c), TPF(c)), in posledično celotno krivuljo ROC. Tu si bomo pogledali analogijo med krivuljami ROC in θ. Vendar najprej definirajmo pomožni izrek.

Izrek 3.1. *Naj bosta X_1 in X_2 neodvisni sl. spremenljivki, ki porodita krivuljo ROC, na isti način kot zgoraj. Definirajmo naslednje preživetvene funkcije: $S_{X_1}(y) = P(Y \geq y|D=1)$ in $S_{X_2}(y) = P(Y \geq y|D=0)$. Potem, za $t \in (0,1)$:*

$$ROC(t) = S_{X_1}(S_{X_2}^{-1}(t))$$

Dokaz lahko najdemo v [11], kjer je navedenih še nekaj zanimih lastnosti krivulj ROC, kot na primer invariantnost krivulje ROC za monotone transformacije Y-a in definicija razmerja verjetij ter uporaba Neyman-Pearsonove leme.

3.4. **AUROC.** Pogosta praksa je, da namesto, da bi predstavili celotno krivuljo ROC, podamo le eno vrednost. Podobno kot pričakovana vrednost in varianca, ki pojasnujeta porazdelitev neke slučajne spremenljivke, želimo najiti mero, ki bo povzela neko krivuljo ROC. Najbolj znana takšna mera je AUC (ang. *area under the ROC curve*). Definirajmo:

$$AUC = \int_0^1 ROC(t)dt$$

Hitro opazimo, da AUC zavzame vrednost 1, če je test popoln in vrednost $\frac{1}{2}$, če je test neuporaben. Če za testa A in B in za $\forall t \in (0,1)$ velja:

$$ROC_A(t) \leq ROC_B(t),$$

potem je $AUC_A \leq AUC_B$. Obrat ne drži vedno.

Izrek 3.2. *Velja:*

$$AUC = P(X < Y) + \frac{1}{2}P(X = Y) = \theta,$$

kjer sta X in Y neodvisni slučajni spremenljivki, iz katerih je bila porojena prvotna krivulja ROC.

Dokaz. Dokazali bomo le v primeru, ko sta X in Y zvezna.

$$AUC = \int_0^1 ROC(t)dt = \int_0^1 S_Y(S_X^{-1}(t))dt = \int_{-\infty}^{\infty} S_Y(y)dS_X(y)$$

Tu smo upoštevali izrek 3.1 in substitucijo iz t v $y = S_X^{-1}(t)$. Če z f_X označimo gostoto spremenljivke X in upoštevamo neodvisnost X in Y, dobimo:

$$AUC = \int_{-\infty}^{\infty} P(Y > y)f_X(y)dy = \int_{-\infty}^{\infty} P(Y > X, X = y)dy = P(Y > X)$$

□

Zgornji izrek je zelo pomemben, tudi zato, ker nam pomaga pri ocenjevanju AUC s pomočjo $\hat{\theta}$, ki smo jo uporabili prej. Na isti način kot prej, lahko dobimo $(1-\alpha)\%$ interval zaupanja za AUC:

$$[AUC_1, AUC_2],$$

kjer sta AUC_1, AUC_2 ustrezne ničle enačbe:

$$(AUC - \widehat{AUC})^2 = \Phi\left(\frac{1-\alpha}{2}\right)^2 \cdot V(AUC),$$

in je $\widehat{AUC}$ enak $\hat{\theta}$.

Več o povezavi med AUC in θ lahko najdemo v (Bamber, 1975) in (Hanley, McNeil, 1982), kjer avtorji podrobneje obravnavajo to povezanost in predstavijo različne cenilke za varianco. Bolj podrobno razlago ocenjevanja (FPF, TPF) in ROC krivuljo v splošnem, najdemo v [11].

4. Zaključek

V diplomski nalogi smo prikazali ideje in predpostavke testa Mann-Whitney, predstavili razlike glede na test t in ga definirali kot statističen test. Uporabili smo θ kot stopnjo prekrivanja dveh slučajnih spremenljivk in pogledali kakšna je povezava med njeno cenilko $\hat{\theta}$ in Mann-Whitneyjevo statistiko U. Razložili smo probleme intervala zaupanja in različne pristope, ki jih lahko privzamemo.

Naloga se je osredotočila na peto Newcombovo metodo, ki je bila obsežno raziskana v članku [8] in predlagana kot najboljša izmed testiranih metod. Pokazali smo tudi celotno izpeljavo metode in jo testirali za porazdelitve, ki niso bile obravnavane v članku in za katere so predpostavke testa t neizpolnjene. Ugotovili smo, da ko sta X in Y Cauchyjevo porazdeljena, je verjetnost pokritja zelo nižja od dobljenih rezultatov za ostale pare porazdelitev, ki so predstavljeni v članku. Če bi hoteli poiskati razloge zakaj je peta Newcombova metoda spodletela, bi morali pogledati v samo izpeljavo intervala zaupanja. Kot glavni razlog lahko izpostavimo ocene za Q_1 in Q_2, kjer smo predpostavili, da sta X in Y eksponentno porazdeljena. Ta predpostavka ne ustreza oblikam Cauchyjevih porazdelitev, zato bi morali te vrednosti oceniti na drugačen način. Eden takšnih pristopov je ocenitev Q_1 in Q_2 direktno iz podatkov. Več o tem lahko najdemo v [8], kjer je Newcombe v svoji prvi metodi predstavil ocene s pomočjo U_{ij}. Poleg tega, smo si tudi pogledali kako se metoda obnaša, ko sta X in Y definirana kot mešanico enakomernih porazdelitev in v večini primerov smo dobili zadovoljive rezultate, čeprav je bila predpostavka o eksponentni porazdelitvi Q_1 in Q_2 kršena. Na koncu smo na kratko predstavili krivulje ROC, njihovo uporabo in povezavo med AUC in θ, kar nam omogoča računanje intervalov zaupanja za AUC.

Na podlagi vseh ugotovitev, v tej nalogi in navedenih literaturah, lahko sklepamo, da je ne glede na različne rezultate še vedno najbolje uporabljati peto Newcombovo metodo, čeprav ostaja še prostora za nadaljno raziskavo in izboljšavo metod za izračun intervala zaupanja za test Mann-Whitney.

Slovar strokovnih izrazov

analysis of variance analiza variance
asymptotic relative efficiency asimptotično relativno učinkovitost
coverage probability verjetnost pokritja
effect size stopnja prekrivanja
false negative lažni negativni
false positive lažni pozitivni
five-point Likert scale 5-razreden vzorec
interval estimate intervalna ocena
interval scale razmična lestvica
nonparametric test neparametričen test
ordinal scale urejenostna lestvica
outlier osameljec
point estimate točkovna ocena
rank rang
roc curve roc krivulja
scale of measurement merska lestvica
tied rank vezan rang
true negative resnični negativni
true positive resnični pozitivni

Literatura

[1] D. Bamber, *The Area above the Ordinal Dominance Graph and the Area below the Receiver Operating Characteristic Graph*, Journal of Mathematical Psychology **12** (1975) 387-415.

[2] J. Frost, v: The Minitab Blog, [ogled 6. 6. 2016], dostopno na `http://blog.minitab.com/blog/adventures-in-statistics/best-way-to-analyze-likert-item-data:-two-sample-t-test-versus-mann-whitney`.

[3] J. A. Hanley, B. J. McNeil, *The meaning and use of the area under a receiver operating characteristic (ROC) curve*, Radiology **143** (1982) 29-36.

[4] M. Hollander, D. A. Wolfe, *Nonparametric Statistical Methods*, **3**, John Wiley & Sons Inc., New York, 2014.

[5] W. H. Kruskal *Historical Notes on the Wilcoxon Unpaired Two-Sample Test*, Journal of the American Statistical Association **279** (1957) 356-60.

[6] I. Miller, M. Miller, *John E. Freund's Mathematical Statistics with Applications*, **7**, Upper Saddle River, New Jersey, 2004.

[7] R. G. Newcombe, *Confidence Intervals for an effect size measure based on the Mann-Whitney statistics. Part 1: General Issues and tail-area-based methods*, Statistics in Medicine **25** (2006) 543-557.

[8] R. G. Newcombe, *Confidence Intervals for an effect size measure based on the Mann-Whitney statistics. Part 2: asymptotic methods and evaluation*, Statistics in Medicine **25** (2006) 559-573.

[9] G. E. Noether, *Elements of Nonparametric statistics*, **1**, John Wiley & Sons Inc., New York, 1967.

[10] S. H. Park, J. M. Goo, C. H. Jo, *Receiver Operating Characteristic (ROC) Curve: Practical Review for Radiologists*, Korean Journal of Radiology **5** 2004 11-18.

[11] M. S. Pepe, *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford statistical science series **31**, Oxford University Press, ,2004.

[12] J. A. Rice, *Mathematical Statistics and Data Analysis*, **6**, Thomson/Brooks/Cole, Belmont, USA, 2007.

[13] D. Salsburg, *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. W.H. Freeman, New York, 2001.

[14] G. M. Sullivan, R. Feinn, *Using Effect Size - or Why the P Value Is Not Enough*, Journal of Graduate Medical Education **4** (2012) 279-282.
[15] E. B. Wilson, *Probable Inference, the Law of Succession, and Statistical Inference*, Journal of the American Statistical Association, **22** (1927) 209-212.
[16] J. C. F. De Winter in D. Dodou, *Five-point Likert items: t test versus Mann-Whitney-Wilcoxon*, Practical Assessment, Research and Evaluation **15** (2010) 1-12.
[17] *Computational Story Lab*, [ogled 20. 7. 2016], dostopno na `https://www.uvm.edu/storylab/2013/02/11/who-will-your-friends-be-next-week-the-link-prediction-problem/`.