NEPARAMETARSKA STATISTIKA II

STATISTIČKI PRAKTIKUM 2

9. VJEŽBE

4. Mann - Whitney - Wilcoxonov test za medijane dviju populacija

Imamo dva međusobno nezavisna slučajna uzorka $X_1, \ldots, X_{n_1}$ i $Y_1, \ldots, Y_{n_2}$ redom iz distribucija F i G.
Želimo provjeriti hipotezu da su populacije jednake tj.

$$H_0 : \quad F = G.$$

Rangovi

Neka je $n := n_1 + n_2$. Označimo $Z_1 := X_1, \ldots, Z_{n_1} := X_{n_1}$, $Z_{n_1+1} = Y_1, \ldots, Z_n = Y_{n_2}$. Neka su

$$Z_{(1)} \leq Z_{(2)} \leq \ldots \leq Z_{(n)}$$

uređene statistike. Definiramo nove slučajne varijable $R_1, R_2, \ldots, R_n$.

Rangovi

Ako je $X_1 = Z_{(k_1)}$, onda je $R_1 = k_1$.

$\vdots$

Ako je $X_j = Z_{(k_j)}$, onda je $R_j = k_j$.

$\vdots$

Ako je $X_{n_1} = Z_{(k_{n_1})}$, onda je $R_{n_1} = k_{n_1}$
Ako je $Y_1 = Z_{(k_{n_1+1})}$, onda je $R_{n_1+1} = k_{n_1+1}$.

$\vdots$

Ako je $Y_{n_2} = Z_{(k_n)}$, onda je $R_n = k_n$.

Wilcoxonova statistika

Wilcoxonovu statistiku definiramo kao zbroj rangova drugog uzorka

$$W := R_{n_1+1} + R_{n_1+2} + \ldots + R_n.$$

Neka su sada $S_1 \leq S_2 \leq \ldots \leq S_{n_2}$ uređene statistike rangova $R_{n_1+1}, \ldots, R_n$ slučajnih varijabli $Y_1, \ldots, Y_n$.
Neka su sada $1 \leq s_1 < s_2 < \ldots < s_{n_2} \leq n$ vrijedi

$$\mathbb{P}_{H_0}(S_1 = s_1, \ldots, S_{n_2} = s_{n_2}) = \binom{n}{n_2}^{-1}.$$

Mann - Whitneyjeva statistika

Za $i \in \{1, 2, \ldots, n_1\}$, $j \in \{1, 2, \ldots, n_2\}$ definiramo slučajnu varijablu

$$U_{ij} := \mathbf{1}_{(X_i < Y_j)}.$$

Mann - Whitneyjeva statistika je

$$U = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} U_{ij}.$$

Dakle, U predstavlja broj parova (X_i, Y_j) za koje vrijedi $X_i < Y_j$. Vrijedi

$$U = W - \frac{1}{2} n_2 (n_2 + 1).$$

Što vrijedi uz pretpostavku H_0

Uz pretpostavku H_0 vrijedi

$$\mathbb{E}_{H_0} U = \frac{n_1 n_2}{2},$$

$$\mathbf{Var}_{H_0} U = \frac{1}{12} n_1 n_2 (n+1).$$

Nadalje, uz pretpostavku H_0 vrijedi iz CGT-a

$$\frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{1}{12} n_1 n_2 (n+1)}} \xrightarrow{d} N(0,1), \quad \min\{n_1, n_2\} \to \infty.$$

Što ako su neke vrijednosti iste?

Ako su neke vrijednosti iste određuju se *podijeljeni rangovi*:
Za vrijednost od X_i, je R_i aritmetička sredina svih rangova k za koje je $Z_{(k)} = X_i$, tj.

$$R_i^* := \frac{1}{\#\{k : X_i = Z_{(k)}\}} \sum_{k=1}^{n} k \mathbf{1}_{(X_i = Z_{(k)})}.$$

Analogno se definiraju rangovi za Y_i. Sada se definiraju $S_1^*, \ldots S_{n_2}^*$ uređene statistike od $R_{n_1+1}^*, \ldots, R_n^*$. Prirodno se poopćuje Wilcoxonova statistika

$$W^* = S_1^* + \ldots + S_{n_2}^* = R_{n_1+1}^* + \ldots + R_n^*$$

Što ako su neke vrijednosti iste?

Za $i \in \{1, 2, \ldots, n_1\}$, $j \in \{1, 2, \ldots, n_2\}$ definiramo slučajnu varijablu

$$U_{ij}^* := \mathbf{1}_{(X_i = Y_j)}.$$

Mann - Whitneyjeva statistika je

$$U^* = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} (U_{ij} + \frac{1}{2} U_{ij}^*).$$

Dakle, U predstavlja broj parova (X_i, Y_j) za koje vrijedi $X_i < Y_j$ i polovicu broja parova (X_i, Y_j) za koje vrijedi $X_i = Y_j$.
Ponovo vrijedi

$$U^* = W^* - \frac{1}{2} n_2 (n_2 + 1).$$

Slično vrijedi asimptotska normalnost.

Alternativne hipoteze

Jedna od mogućih alternativnih hipoteza je

$$H_1 : \quad F \neq G.$$

Druga mogućnost je da testiramo da je neka stohastički veća od druge. Za slučajnu varijablu X kažemo da je **stohastički veća** od slučajne varijable Y ako za svaki $t \in \mathbb{R}$ vrijedi

$$\mathbb{P}(X \geq t) \geq \mathbb{P}(Y \geq t) \;\Leftrightarrow\; F_Y(t) \geq F_X(t).$$

Stoga bi mogli testirati i

$$H_1 : \quad F < G \quad \text{ili} \quad H_1 : \quad F > G.$$

Možemo testirati je li neka pojava (distribuirana sa F) stohastički veća ili manja od druge (distribuirane sa G). Ako je $F > G$ očekujemo da Wilcoxonova statistika bude veća, u suprotnom manja.
Statistike W (W^*) i U (U^*) su ekvivalentne. Njihova distribucija se može izračunati uz pretpostavku H_0 (za zadane podatke), što se i radi za manje brojeve, a za veće se koristi asimptotska normalnost. Ovisno o alternativnoj hipotezi imamo dvostrano ili jednostrano testiranje.

Mann - Whitney - Wilcoxonov test u R-u

Neka je `x` vektor podataka dobiven kao slučajna realizacija od F i vektor `y` vektor podataka dobiven kao slučajna realizacija od G nezavisno od `x`.
Alternativne hipoteze $F \neq G$, $F < G$, $F > G$ testiramo na sljedeći način:

```
> wilcox.test(x,y)
> wilcox.test(x,y,al="gre")
> wilcox.test(x,y,al="less")
```

Zadatak 4.

U datoteci `smrt.txt` zabilježeni su podaci s gradskih groblja u Zagrebu o dobi umrlih i njihovom spolu. Je li dob koju dožive žene *veća* od dobi koju dožive muškarci? U ovom slučaju testirat ćemo tvrdnju da se pod *veća* podrazumijeva *stohastički veća*.
Svoje rezultate potkrijepite i grafičkim prikazom.

5. Wilcoxova statistika rangova s predznacima

Pretpostavimo da imamo $2n$ opažanja, po dva od svakog od n subjekata. Primjerice, imamo n parova bolesnika s istom bolešću, pri čemu bolesnici u paru imaju iste simptome bolesti. Za svaki par na jednom članu primijenimo tretman novim lijekom, a drugog tretiramo na stari način.

Opažanja kod člana i-tog para testne skupine označimo s X_i, a opažanja kod člana kontrolne skupine Y_i. Zanima nas postoji li razlika između ove dvije skupine.

Pretpostavke

Npr. želimo testirati pokazuju li pacijenti tretirani novim lijekom brži oporavak. Pretpostavka je da smo od populacije parova na slučajan način odabrali njih n i da smo među njima slučajno odabrali kojeg ćemo liječiti novim lijekom, a kojeg starom metodom. Tada su

$$(X_1, Y_1), \ldots, (X_n, Y_n),$$

nezavisni jednako distribuirani vektori, gdje su X_i distribuirani sa F_X, a Y_i sa F_Y.
Promatramo razlike $Z_i = Y_i - X_i$, koje su nezavisne i jednako distribuirane.

Postupak

1. Neka je R_i rang od $|Z_i|$.
2. Tada je Wilcoxova statistika

$$V = \sum_{k=0}^{n} \mathbf{1}_{(Z_i>0)} R_i.$$

$$H_0 : \quad F_X = F_Y$$

Nulta hipoteza je ekvivalentna hipotezi da je medijan distribucije razlika $\{Z_i\}$ 0.

Pokazuje se (uz pretpostavku da su distribucije neprekidne) da je

$$\mathbb{E}_{H_0} V = \frac{n(n+1)}{4}, \quad \mathbf{Var}_{H_0} V = \frac{n(n+1)(2n+1)}{24}$$

Stoga kad je n velik možemo koristiti asimptotsku normalnost (CGT). Možemo i egzaktno izračunat distribuciju V-a za male n.

Što ako distribucije nisu neprekidne

1. Neka je R_i^* podijeljen rang od $|Z_i|$ (promatramo samo vrijednosti $|Z_k|$ koje su različite od nule).
2. Tada je Wilcoxova statistika

$$V^* = \sum_{k=0}^{n} \mathbf{1}_{(Z_i>0)} R_i^*.$$

$$\mathbb{E}_{H_0} V^* = \frac{n(n+1) - d_0(d_0+1)}{4},$$

$$\mathbf{Var}_{H_0} V^* = \frac{n(n+1)(2n+1) - d_0(d_0+1)(2d_0+1)}{24} - \frac{\sum_{k=1}^{g} d_k(d_k^2-1)}{48}.$$

Dakle za velike n koristimo asimptotsku normalnost.

Zadatak 5.

Promatramo dva kolegija *Programiranje 2* i Strukture podataka i algoritme. Prvi kolegij je nužni prethodnik za drugi. Promatramo osobe koje su dobile 4 iz programiranja. Njima vježbe drže dva asistenta X i Y. Ispitujemo jesu li asistenti jednako uspješni u držanju vježbi.
Odabrali smo po jedan par studenata koji su iz Programiranja 2 imali ocjenu dovoljan i izvrstan, i po 2 para s ocjenama dobar i vrlo dobar. Te smo očitali rezultate s njihova 1. kolokvija iz SPA.
Dobili smo sljedeće rezultate

```
> x=c(15,12,24,11.5,13,22.5)
> y=c(17,12,22,13,12,16)
```

6. Kruskal-Wallis test

- generalizacija Mann-Whitney-Wilcoxonovog testa za dva uzorka
- jednako jak kao i ANOVA, u slučaju nenormalnosti podataka ili outliera i jači
- promatramo $k \geq 2$ nezavisnih uzoraka i želimo usporediti pripadne distribucije, ali bez pretpostavki o samoj familiji distribuciji.

Promatramo hipoteze
$H_0 : M_1 = M_2 = ... = M_k$
$H_1 :$ ne H_0

Procedura

Promatramo k uzoraka: $X_1^{(j)}, ..., X_{n_j}^{(j)}, j = 1, .., k,\ n = \sum_{i=1}^{k} n_i$.

- Svih k uzoraka uredimo u rastući niz $Z_1 < Z_2 < ... < Z_n$
- Odredimo rangove $R_i^{(j)}$ za svaki $X_i^{(j)}$ unutar **cjelokupnog** uređenog uzorka
- Odredimo sumu i prosjek rangova po uzorcima

$$R^{(j)} = \sum_{i=1}^{n_j} R_i^{(j)}, \qquad \bar{R}^{(j)} = \frac{R^{(j)}}{n_j}$$

(Uočimo da je $R = \sum_{j=1}^{k} R^{(j)} = \frac{n(n+1)}{2}$ i $\bar{R} = \frac{n+1}{2}$)

- Izračunamo Kruskal-Wallis statistiku

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{k} n_j (\bar{R}^{(j)} - \bar{R})^2 =$$

$$= \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{(R^{(j)})^2}{n_j} - 3(n+1)$$

- Točnu distribuciju statistike H je teško odrediti (ovisi o vrijednostima $k, n_1, ..., n_k$) i egzaktne vrijednosti su poznate samo za mali broj slučajeva. Za velike duljine uzoraka distribucija se može aproksimirati s $\chi^2(k-1)$
- Kritično područje testa je $K = \{H > \chi^2_\alpha(k-1)\}$

Zadatak 6. - Ozon

Podaci `airquality` sadrže dnevna mjerenja kvalitete zraka u New Yorku od svibnja do rujna 1973.g. Gustoća ozona u zraku dana je u stupcu `Ozone`. Na nivou značajnosti od 5% provjerite razlikuje li se dnevna razina ozona po mjesecima.

Koristimo naredbu `kruskal.test` na podacima koji su spremljeni u *listu*.

```
> attach(airquality)
> lista=list(Ozone[Month==5],Ozone[Month==6],
> Ozone[Month==7],Ozone[Month==8],Ozone[Month==9])
> kruskal.test(lista)
```

ili

```
> kruskal.test(Ozone~Month,data=airquality)
```

Postupanje s nedostupnim podacima (`NA`) odredimo naredbom `na.action`.

7. Friedmanov test

- Generalizacija Wilcoxonovog testa rangova za dva uzorka
- Niz uređenih k-torki je nezavisan (moguća zavisnost među k-torkama)
- Pogodan model za promatranje jedne *jedinke* kroz više nivoa *tretmana* (tj. jednak broj mjerenja za svaki uzorak)

Procedura

Promatramo k uzoraka: $X_1^{(j)}, ..., X_n^{(j)}, j = 1, .., k$ iste duljine.

- Uredimo svaki od uzoraka zasebno, $X_{(1)}^{(j)}, ..., X_{(n)}^{(j)}, j = 1, .., k$
- Odredimo rangove $R_i^{(j)}$ za svaki $X_i^{(j)}$ unutar **pojedinog** uređenog uzorka
- Odredimo sumu rangova po uzorcima

$$R^{(j)} = \sum_{i=1}^{n} R_i^{(j)}$$

te Friedmanovu statistiku

$$S = \frac{12}{nk(k+1)} \sum_{j=1}^{k} (R^{(j)} - \frac{n(k+1)}{2})^2 =$$

$$= \frac{12}{nk(k+1)} \sum_{j=1}^{k} (R^{(j)})^2 - 3n(k+1)$$

- Kritično područje testa je $K = \{H > a_\alpha\}$ gdje je a_α kvantil razdiobe statistike S
- Egzaktna distribucija statistike S izračunata je za male n i k, za veće vrijednosti koristi se aproksimacija $\chi^2(k)$ distribucijom.

Zadatak 7. - Benzin

U datoteci `gasoline.txt` nalaze se podaci o prijeđenoj kilometraži četiri modela automobila za tri marke benzina. Na razini značajnosti 2% testirajte postoji li značajna razlika u trima markama benzina.
Koristimo naredbu `friedman.test` u R-u.