# NEPARAMETARSKA STATISTIKA

STATISTIČKI PRAKTIKUM 2

7. VJEŽBE

# Zašto neparametarska statistika?

- provodimo statističku analizu podataka bez rigoroznih pretpostavki na distribuciju populacije
- korisna kada je pripadnost određenoj distribuciji teško provjeriti ili je distribucija nepoznata
- kod uzoraka male duljine
- izbjegavamo neopravdanu pretpostavku normalnosti

# Jakost neparametarske statistike

U situaciji kada je parametarski model za uzorak nedostupan, neparametarske metode su idealne. Parametarski testovi su općenito jači od neparametarskih (iako ne značajno), ali njihova snaga drastično pada ukoliko pretpostavke o populacijskom modelu nisu zadovoljene.

# Procedura

U većini slučajeva potrebne su minimalne pretpostavke o modelu (jednakost varijanci/distribucije među populacijama, neprekidnost distribucije i sl.), ali kako je riječ o jednostavnim pretpostavkama manja je mogućnost grešaka u zaključivanju.

Većina neparametarskih metoda se bazira na *uređajnoj statistici*; umjesto samih vrijednosti mjerenja promatramo njihove *rangove* (poziciju u uređenom uzorku).

# Rangovi

Za slučajni uzorak $X_1, \ldots, X_n$ promatramo uređajnu statistiku $X_{(1)}, \ldots, X_{(n)}$ (takvu permutaciju slučajnog uzorka da vrijedi $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$)

Rangovi su niz slučajnih varijabli duljine $n$: $R_1, \ldots, R_n$ takvi da vrijedi:

- $R_i =$ rang podatka $X_i$ u uređenom uzorku.

# 1.Procjena pouzdanih intervala

Za uzorke velike duljine korištenjem CGT-a možemo dobiti asimptotske pouzdane intervale za parametre pripadne distribucije. No kako asimptotika ovisi o brzini konvergencije prema normalnoj distribuciji, ova metoda nije pouzdana za uzorke male duljine i specifične distribucije.

Ukoliko želimo dobiti egazktne pouzdane intervale za male uzorke potrebne su određene pretpostavke na distribuciju populacije koje nije uvijek jednostavno provjeriti ili koje su neopravdane.

Također, zanimaju nas i metode za procjenu pouzdanih intervala za vrijednosti koje nisu parametri određene vjerojatnosne distribucije.

# Pouzdani interval za medijan

*Medijan* $M$ je često promatrana vrijednost u neparametarskom okruženju, uz pretpostavku neprekidnosti distribucije $F$ vrijedi

$$F(M) = \frac{1}{2}.$$

Za uzorak $X_1, .., X_n$ iz distribucije $F$, vjerojatnost da je svaki element veći (odnosno manji) od medijana je $\frac{1}{2}$. Tada je

$$N^- = \#\{i : X_i < M\} \sim B\left(n, \frac{1}{2}\right).$$

# Procedura

Za $\alpha \in \langle 0, 1 \rangle$ odredimo $(1-\alpha) \cdot 100\%$ - pouzdani interval za $M$

- Odaberemo $a, b \in \{0, \ldots, n\}$ t.d.

$$\mathbb{P}(N^- \leq a) = \mathbb{P}(N^- \geq b) = \frac{\alpha}{2}.$$

- Ukoliko nije moguće postići jednakost, odaberemo vrijednost $a$ t.d. je gornja vjerojatnost najbliža $\frac{\alpha}{2}$ i $b = n + 1 - a$.
- Za uređeni uzorak $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ vrijedi

$$\mathbb{P}(X_{(a)} < M < X_{(b)}) = (\approx) 1 - \alpha$$

.

# Zadatak 1 - Srčani puls

U datoteci `heartrate.txt` nalaze se izmjerene promjene pulsa grupe pacijenata nakon primanja određenog lijeka.

1. Provjerite dolaze li podaci iz normalne razdiobe.
2. Konstruirajte 98% pouzdani interval za medijan.

# Pouzdani intervali dobiveni bootstrap metodom

Ideja: uz dani uzorak $X_1, \dots, X_n$ bez dodatnih pretpostavki na populaciju iz koje dolazi, sve poznate informacije o distribuciji sadržane su u uzorku. Uzorak tretiramo kao populaciju i ponavljamo *resampling* iz tog uzorka. Na temelju tih novih uzoraka procjenjujemo vrijednosti promatrane statistike (parametra $\theta$).

U `R`-u:

```
> boot.out=boot(data, statistic,...)
> boot.ci(boot.out, conf = 0.95, type = "all", ...)
```

gdje je `type` tip bootstrap intervalne procjene. Jedna od metoda je "`perc`" koja $(1-\alpha)\cdot 100$ - pouzdani interval za $\theta$ određuje kao

$$\left[\hat{\theta}_{\alpha/2}, \hat{\theta}_{1-\alpha/2}\right].$$

# 2.Binomni test

Test jednakosti medijana, neparametarska verzija t-testa za jednakost očekivanja.
Promotrimo uzorak $X_1, .., X_n$ i hipoteze

$$H_0 : \quad M = m_0$$

$$H_1 : \quad M > m_0$$

Uredimo uzorak $X_{(1)}, .., X_{(n)}$ i promotrimo broj elemenata uzorka većih od $m_0$,

$$N^+ = \#\{i : X_{(i)} > m_0\}.$$

Ukoliko je hipoteza $H_0$ točna statistika $N^+$ ima binomnu distribuciju $B(n, \frac{1}{2})$.
*Napomena:* Elemente uzorka koji su jednaki $m_0$ izbacimo iz uzorka.

# Procedura

Imamo realizaciju uzorka $x_1, \ldots, x_n$ i razinu značajnosti $\alpha \in \langle 0, 1 \rangle$

1. Odredimo $n^+$, broj elemenata većih od $m_0$.
2. Izračunamo p-vrijednost

$$\gamma = \mathbb{P}(N^+ \geq n^+ | H_0) = \sum_{i=\lfloor n^+ \rfloor}^{n} \binom{n}{i} \left(\frac{1}{2}\right)^n$$

3. Kritično područje za test na razini značajnosti $\alpha$ je $K = \{\gamma < \alpha\}$

Moguće alternativne hipoteze

- $H_1 : \quad M < m_0$

  $$\gamma = \mathbb{P}(N^- \leq n^- | H_0),$$
  $$K = \{\gamma > \alpha\}$$

- $H_1 : \quad M \neq m_0$

  $$\gamma_i = \mathbb{P}(N^{+(-)} > (<) n^{+(-)} | H_0)/2,$$
  $$K_i = \{\gamma \geq (\leq) \alpha/2\},$$
  $$i = 1, 2$$

# Zadatak 2.

Dani su rezultati mjerenja visine devetero osnovnoškolske djece

$$1.51\ 1.35\ 1.69\ 1.48\ 1.29\ 1.27\ 1.54\ 1.39\ 1.45$$

Na razini značajnosti od 5% testirajte je li medijan veći od 1.4.

1. Test iskodirajte sami.
2. Koristite naredbu `binom.test` u R-u.

# 3. Wilcoxonov test za jednakost medijana

Promatramo uzorak $X_1, \ldots, X_n$ iz populacije sa simetričnom neprekidnom distribucijom. Želimo testirati hipotezu:

$$H_0: \quad M = m_0$$

$$H_1: \quad M \neq m_0$$

Wilcoxonov test je neparametarska alternativa jednostranom t-testu.

# Procedura

1. Odredimo $Z_i = |X_i - m_0|$, uredimo uzorak $Z$ te za svaki $Z_i$ odredimo njegov rang $R_i$;

$$R_i = j \text{ ako je } Z_{(j)} = Z_i$$

2. Za svaki od $Z_i$-eva provjerimo je li pripadni $X_i$ veći ili manji od $m_0$
3. Sa $W^+$ ($W^-$) označimo sumu rangova $Z$-ova čiji su pripadni $X$-evi veći (manji) od $m_0$
4. Ukoliko je nulta hipoteza točna statistika $W$ ima *Wilcoxonovu distribuciju*.

U R-u koristimo naredbu

```
> wilcox.test(x, y = NULL, alternative = c("two.sided",
"less", "greater"), paired = FALSE, exact = NULL,
conf.level = 0.95, ...)
```

Varijabla `exact` određuje računamo li p-vrijednosti preko Wilcoxonove distribucije ili njene normalne aproksimacije. Kako je

$$\mathbb{E}W^+ = \frac{n(n+1)}{4}$$
$$\mathrm{Var}W^+ = \frac{n(n+1)(2n+1)}{24} < \infty$$

Wilcoxonova distribucija zadovoljava CGT.

# Zadatak 3.

U datoteci `rent.txt` nalaze se podaci o mjesečnoj cijeni najma (u $) za 25 slučajno odabranih kućanstava u Bostonu. Testirajte na razini značajnosti od 5% hipotezu da je prosječni mjesečni najam veći od 750$

# Napomena

Wilcoxonov test odbacuje jednake vrijednosti unutar uzorka. Kako veličina uzorka značajno utječe na p-vrijednost, ne bismo smjeli odbaciti više od 10% elemenata uzorka. U slučaju da je to nužno treba razmotriti alternativu testu, npr. s korekcijama.

# 3. 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 $Y_1, \ldots, Y_{n_2}$.
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 $i \in \{1, 2, \dots, n_1\}$, $j \in \{1, 2, \dots, n_2\}$ 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

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

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

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.

# Wilcoxonova statistika rangova s predznacima

Pretpostavimo da imamo $2n$ opažanja, po dva od svakog $n$ subjekata. Primjerice, imamo $n$ parova bolesnika s sitom 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 tretiranog novim lijekom označimo s $X_i$, a opažanja kod člana tretiranog starim načinom $Y_i$. Zanima postoji li razlika između ove dvije skupine.

# Pretpostavke

Želimo testirati pokazuju li pacijenti tretirani novim lijekom da su se prije oporavili. 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

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

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

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

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

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

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

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

Dakle za velike $n$ koristimo asimptotsku normalnost.

# Zadatak 5.

Promatramo dva kolegija *Programiranje 2* i *Strukture podataka i algoritmi*. Prvi kolegij je nužni prethodnik za drugi. Njima vježbe drže dva asistenta $X$ i $Y$.
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)
```

Ispitajte jesu li asistenti jednako uspješni u držanju vježbi.

# 6. Kruskal-Wallis test

- generalizacija Mann-Whitney-Wilcoxonovog testa za dva uzorka na više uzoraka
- 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 distribucija.

Promatramo hipoteze

$H_0 : M_1 = M_2 = ... = M_k$

$H_1 :$ ne $H_0$

# Procedura

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

- Svih $k$ uzoraka uredimo u rastući niz $Z_1 < Z_2 < \ldots < Z_n$
- 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.

# 6. Friedmanov test

- Generalizacija Wilcoxonovog testa rangova za dva uzorka na više uzoraka
- 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: $X_1^{(i)}, ..., X_n^{(i)}$, $i = 1, .., k$ iste duljine.

- Uredimo svaki od uzoraka zasebno, $X_{(1)}^{(i)}, ..., X_{(n_i)}^{(i)}$, $i = 1, .., k$
- 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 $a_\alpha$ 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.