Neparametarski testovi

Dermišek, Antonija

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Antonija Dermišek

NEPARAMETARSKI TESTOVI

Diplomski rad

Voditeljica rada:
doc. dr. sc. Snježana Lubura Strunjak

Zagreb, rujan 2020.

Ovaj diplomski rad obranjen je dana ____________________ pred ispitnim povjerenstvom u sastavu:

1. ____________________________, predsjednik
2. ____________________________, član
3. ____________________________, član

Povjerenstvo je rad ocijenilo ocjenom _______________.

Potpisi članova povjerenstva:

1. ____________________________
2. ____________________________
3. ____________________________

„Budi junak, hrabar i radi! Ne boj se i ne plaši se, jer će Jahve, Bog, moj Bog biti s tobom! Neće te napustiti niti te ostaviti dok ne svršiš sav posao za službu oko Jahvina Doma." 1.Ljet 28,21

Ovaj diplomski rad posvećujem svojoj obitelji, zaručniku, prijateljima i kolegama koji su mi pomogli ostvariti ovo postignuće te Majci Milosrđa koja me svojim zagovorom pratila zajedno sa svojim Sinom. Zahvaljujem im na podršci koju su mi pružili tijekom cijelog studija. Posebno zahvaljujem mentorici doc.dr.sc. Snježani Luburi Strunjak, na čija sam predavanja voljela ići, što mi je pomogla svojim savjetima pri izradi ovog diplomskog rada te uvijek imala strpljenja i vremena. Hvala joj što mi je omogućila lijep završetak priče zvan „fakultet".

Sadržaj

Uvod

U mnogo slučajeva moramo pretpostaviti da populacija jako nalikuje obliku normalne distribucije ili da su uzorci neovisni. Tako je i kada imamo dvije populacije i kada se zna da su varijance jednake.
Budući da je mnogo situacija u kojima se te pretpostavke ne mogu ispuniti, statističari su razvili alternativne tehnike temeljene na ne tako strogim pretpostavkama koje su poznate kao **neparametarski testovi.** [3]
U mnogim neparametarskim testovima računanje je lagano te su zbog toga popularni.
Navedeni testovi ne postavljaju nikakve zahtjeve u pogledu parametara populacije iz koje su uzorci izvučeni, a uvjeti koje zahtijevaju blaži su od onih kod parametarskih testova.

Prednosti neparametarskih testova su što ne postavljaju zahtjeve u obliku distribucije mjere u populaciji pa se mogu primijeniti u velikom broju situacija, mogu se primjeniti na rangove i na normalne podatke te male uzorke. Ako su ispunjeni svi uvjeti za primjenu parametarskog testa primjenom neparametarskog testa, gube se neke informacije sadržane u podacima.

Od poglavlja 1.1 do 1.3 predstavit ćemo the sign test(test predznaka) kao neparametarsku alternativu testovima koji se tiču obilježja i razlike između sredina uparenih podataka.

Kroz poglavlja1.4, 1.5 i 1.6 proučit ćemo metode temeljene na zbrojevima ranga koji služe kao neparametarske alternative two-sample t testu i one-way analizi varijance. Rangovi se često koriste u neparametarskim testovima što ih čini primjenjivima za uporabu s ordinalnim podacima.

U poglavljima1.7, 1.8 i 1.9 proučit ćemo kako testirati slučajnost uzorka nakon odabira podataka te u poglavlju 1.10 predstaviti neparametarski test u pogledu odnosa između uparenih podataka.

0.1 Osnovne definicije [4]

U statistici pod skupom podataka podrazumijevamo vrijednosti dobivene mjerenjem (ili opažanjem) nekog *statističkog obilježja* (ili *varijable*) promatrane skupine objekata ili osoba. Varijabla može biti jednodimenzionalna ili dvodimenzionalna.
Grupa objekata ili osoba koju promatramo, odnosno za koju izučavamo odabrano statističko obilježje zove se **populacija**.
Često nije moguće popisati sve vrijednosti izučavanoga statističkoga obilježja pa u tom slučaju odabiremo (reprezentativni) **uzorak** iz populacije i iz njega popisujemo vrijednosti statističkog obilježja.
Razlikujemo skupove podataka dobivene mjerenjem (opažanjem) odabranog statističkog obilježja na populaciji (*populacijski podaci*) od onih dobivenih na uzorku iz populacije (*uzorački podaci*).
Cilj statističke analize je da se na osnovi podataka iz uzorka donesu određeni zaključci o populacijskoj razdiobi promatranog statističkog obilježja.
Podatke možemo podijeliti po tipu vrijednosti opažanog statističkog obilježja, odnosno varijable. Razlikujemo *numeričke* i *kategorijalne* varijable.
Numeričke varijable dijele se na:

- *diskretne*, obično predstavljaju neko prebrojavanje,
- *neprekidne*, obično predstavljaju rezultat mjerenje neke fizikalne ili novčane veličine

Kategorijalne varijable dijele se na:

- *dihotomne*, imaju samo dva razreda, (npr. spol)
- *nominalni*, razredi su neuređeni, (npr. uzrok štete)
- *ordinalne*, vrijednosti su im uređene, (npr. školske ocjene)

Definicija 0.1.1. ***Frekvencija*** *ili* ***učestalost*** *vrijednosti varijable je broj pojavljivanja te vrijednosti u skupu podataka,a njena* ***relativna frekvencija*** *je omjer frekvencije i ukupnog broja podataka.*

Neka su

$$x_1, x_2, ..., x_n \tag{1}$$

n vrijednosti varijable X koje čine skup podataka. Ako je X numerička varijabla tada je (1) niz brojeva.

Definicija 0.1.2. *Neka je X numerička varijabla.* ***Aritmetička sredina*** *brojeva (1) je broj*

$$\overline{x} = \frac{1}{n}(x_1 + x_2 + ... + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{2}$$

Ako X poprima samo par različitih vrijednosti

$$a_1, a_2, ..., a_n \tag{3}$$

koje se u nizu (1) pojavljuju više puta (vrijednost a_1 s frekvencijom f_1, vrijednost a_2 s frekvencijom f_2 itd.) tada se formula (2) za aritmetičku sredinu može zapisati u sažetom obliku:

$$\overline{x} = \frac{1}{n}(f_1a_1 + f_2a_2 + ... + f_ka_k) = \frac{1}{n}\sum_{j=1}^{k} f_ja_j. \tag{4}$$

Primijetite da je $n = f_1 + f_2 + ... + f_k$.
Neka je X numerička ili ordinalna varijabla. Tada je njene vrijednosti (1) moguće urediti:

$$x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}. \tag{5}$$

Definicija 0.1.3. ***Medijan*** *skupa podataka (1) je vrijednost od X za koju vrijedi da je* 50% *svih podataka u skupu manje od ili jednako toj vrijednosti i* 50% *svih podataka je veće od nje ili jednako njoj. Vrijedi sljedeće:*

- *Ako je broj podataka n, neparan broj, tj.* $n = 2k - 1$ *tada je medijan,* $m = x_{(k)}$
- *Ako je broj podataka n, tj. paran broj,* $n = 2k$ *tada je medijan,* $m = \frac{x_{(k)}+x_{(k+1)}}{2}$

Definicija 0.1.4. ***Mod*** *je vrijednost obilježja X koja se u skupu podataka (1) pojavljuje najviše puta, dakle, ima najveću frekvenciju.*

Definicija 0.1.5. ***Standardna devijacija*** *je srednje kvadratno odstupanje podataka od njihove aritmetičke sredine. Dana je formulom:*

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{6}$$

odnosno:

$$s = \sqrt{\frac{1}{n-1}\sum_{j=1}^{k} f_j(a_j - \overline{x})^2} \tag{7}$$

u slučaju da se k različitih vrijednosti (3) od X u (1) ponavljaju. Broj

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2, \ \textit{odnosno} \ s^2 = \frac{1}{n-1}\sum_{j=1}^{k} f_j(a_j - \overline{x})^2 \tag{8}$$

*zovemo **varijanca** skupa podataka .*

Vjerojatnosni prostor predstavlja matematički model za slučajni pokus.

Definicija 0.1.6. *Slučajni pokus je pokus koji ima više mogućih ishoda. Ishode slučajnog pokusa zovemo događajima.*

Označimo sa Ω skup elementarnih dogadaja ω_1, ω_2 itd. Ω zovemo *prostor elementarnih događaja*. Budući da se svaki događaj sastoji od elementarnih događaja koji su povoljni za njega, događaji su podskupovi od Ω. Specijalno, Ω i prazan skup $\emptyset$ su događaji. Ω zovemo sigurnim, a $\emptyset$ nemogućim događajem.

Definicija 0.1.7. *[5] Neka je Ω neprazan skup. Familija podskupova $\mathcal{F}$ od Ω se zove σ-algebra (ili σ-algebra događaja), ako vrijede sljedeća tri svojstva:*

i. $\Omega \in \mathcal{F}$,

ii. Ako je $A \in \mathcal{F}$, onda je i $A^C \in \mathcal{F}$ (zatvorenost na komplement),

iii. Ako su $A_j \in \mathcal{F}$, $j \in \mathbb{N}$ onda je i $\bigcup_{j=1}^{\infty} \in \mathcal{F}$ (zatvorenost na prebrojive unije).

Uređen par $(\Omega, \mathcal{F})$ zove se izmjeriv prostor.

Definicija 0.1.8. *Neka je $\mathcal{F}$ σ-algebra događaja na prostoru elementarnih događaja Ω. Preslikavanje $\mathbb{P}$ svakom događaju A iz $\mathcal{F}$ pridružuje realan broj $\mathbb{P}(A)$ tako da vrijedi:*

(P_1) $0 \leq \mathbb{P}(A) \leq 1$ *za sve događaje* $A \in \mathcal{F}$

(P_2) $\mathbb{P}(\Omega) = 1$,

(P_3) Za sve po parovima disjuktne događaje $A_1, A_2, ...$ iz $\mathcal{F}$ vrijedi:

$$\mathbb{P}(A_1 \cup A_2 \cup ...) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + ...,$$

*zovemo **vjerojatnost** na $(\Omega, \mathcal{F})$. Broj $\mathbb{P}(A)$ zovemo **vjerojatnost događaja** A. U slučaju kada je Ω prebrojiv skup, $\mathcal{F} = \mathcal{P}(\Omega)$. Uređena trojka $(\Omega, \mathcal{F}, \mathbb{P})$ zove se **vjerojatnosni prostor.***

Definicija 0.1.9. ***Slučajna varijabla*** *je funkcija koja svakom elementarnom događaju pridružuje broj. Slučajna varijabla je diskretna ako je ImX prebrojiv skup gdje sa ImX označavamo sliku slučajne varijable X. Diskretnoj slučajno varijabli X pridružujemo funkciju* $f_X : \mathbb{R} \to \mathbb{R}$ *definiranu sa:*

$$f_X(x) := \mathbb{P}(X = x),$$

koju zovemo ***funkcijom vjerojatnosti*** *od X ili* ***funkcijom gustoće*** *razdiobe od X.*

Općenito ako je $X(\Omega) = a_1, a_2, ..., a_n, ...$ i $P(X = a_i) = p_i, i = 1, 2, 3...$ onda pišemo u **tablicu distribucije**:

$$X \sim \begin{pmatrix} a_1 & a_2 & ... & a_n & .. \\ p_1 & p_2 & ... & p_n & ... \end{pmatrix}$$

Nadalje, vrijedi:

(G_1) $f_X(x) \geq 0$ za sve x,

(G_2) $\sum_{x \in ImX} f_X(x)$,

Ako neka realna funkcija s diskretnom slikom ima svojstva G_1, G_2, tada kažemo da je ona *funkcija gustoće neke diskretne vjerojatnosne razdiobe.*

Definicija 0.1.10. *Svakoj slučajnoj varijabli pridružujemo* ***funkciju distribucije*** F_X *koja se definira kao funkcija* $F_X : \mathbb{R} \to \mathbb{R}$,

$$F_X(x) := \mathbb{P}(X \leq x).$$

Za diskretnu slučajnu varijablu X s gustoćom razdiobe f_X *vrijedi:*

$$F_X(x) = \sum_{y \in Imx : y \leq x} f_X(y) \tag{9}$$

Definicija 0.1.11. *Slučajna varijabla X je* ***neprekidna*** *ako vrijedi:*

1. *ImX je interval u* $\mathbb{R}$,

2. *Skup* $a \leq X \leq b$ *je događaj za sve realne brojeve* $a < b$.

3. *Postoji funkcija* $f_X : \mathbb{R} \to \mathbb{R}$ *takva da je za sve* $a < b$,

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)dx$$

Funkciju f_X *zovemo funkcijom gustoće razdiobe od X ili gustoćom razdiobe od X.*

Za gustoću vrijedi:

(G_1) $f_X(x) \geq 0$ za sve x,

(G_2) $\int_{-\infty}^{\infty} f_X(x)dx = 1$,

Ako neka realna funkcija kojoj je slika interval realnih brojeva ima svojstva G_1, G_2, tada kažemo da je to *funkcija gustoće neke neprekidne vjerojatnosne razdiobe.*

Definicija 0.1.12. ***Matematičko očekivanje*** *slučajne varijable X interpretira se kao srednja (očekivana) vrijednost od X. Definira se kao broj* $\mathbb{E}[X]$*:*

$$\mathbb{E}[X] := \sum_{x \in ImX} x f_X(x) \qquad \textit{ako je X diskretna}$$

$$\mathbb{E}[X] := \int_{-\infty}^{\infty} x f_X(x) \qquad \textit{ako je X neprekidna}$$

Definicija 0.1.13. ***Varijanca*** *slučajne varijable je mjera raspršenja njenih vrijednosti od matematičkog očekivanja. Preciznije, varijanca od X je srednje kvadratno odstupanje X od* $\mathbb{E}[X]$*:*

$$\mathbb{V}ar[X] := \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \tag{10}$$

Standardna devijacija od X je drugi korijen varijance:

$$\sigma(X) := \sqrt{\mathbb{V}ar[X]}$$

Definicija 0.1.14. *Neka je X broj uspjeha u nizu od n nezavisnih pokusa, pri čemu je vjerojatnost uspjeha u svakom od pokusa jednaka p. Tada vrijedi:*

$$P(X = k) = \binom{n}{k} \cdot p^k (1 - p)^{n-k}, k \in 0, 1, ..., n$$

Uz oznaku $q = 1 - p$*:*

$$X \sim \begin{pmatrix} 0 & 1 & ... & n & .. \\ \binom{n}{0} \cdot p^0 q^n & \binom{n}{1} \cdot p^1 q^{n-1} & ... & \binom{n}{n} \cdot p^n q^0 & ... \end{pmatrix}$$

X zovemo **binomna slučajna varijabla** s parametrima n i p i označavamo s $X \sim B(n, p)$. Posebno za n=1 je **Bernoullijeva** slučajna varijabla.

Definicija 0.1.15. *Kažemo da diskretna slučajna varijabla ima* ***Poissonovu distribuciju*** *s parametrom λ, $\lambda > 0$, i pišemo $X \sim P(\lambda)$, ako je njena funkcija vjerojatnosti oblika:*

$$f_X(x) = \frac{\lambda^x}{x!} e^{-\lambda} \text{ za } x \in ImX = 0, 1...$$

Vrijedi:

$$\mathbb{E}[X] = \mathbb{V}ar[X] = \lambda$$

Definicija 0.1.16. *Kažemo da slučana varijabla X ima* ***gama distribuciju*** *s parametrima $\alpha > 0$ i $\lambda > 0$ i pišemo $X \sim \Gamma(\alpha, \frac{1}{\lambda})$ ako je strogo pozitivna $ImX = (0, +\infty)$ i gustoća razdiobe je:*

$$f_X(x) = \begin{cases} \dfrac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} & \text{za } x > 0 \\ 0 & \text{inače,} \end{cases}$$

gdje je $\Gamma(\alpha) = \int_0^{+\infty} t^{\alpha-1} e^{-t} dt$ Γ-funkcija. Vrijedi:

$$\mathbb{E}[X] = \frac{\alpha}{\lambda}, \mathbb{V}ar[X] = \frac{\alpha}{\lambda^2}$$

Definicija 0.1.17. *Ako je $X \sim \Gamma(1, \frac{1}{\lambda})$, tada kažemo da X ima* ***eksponencijalnu distribuciju*** *s parametrom $\lambda > 0$ i pišemo $X \sim Exp(\lambda)$. Gustoća razdiobe i funkcija distribucije od X, te matematičko očekivanje i varijanca su:*

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{za } x > 0 \\ 0 & \text{inače,} \end{cases} \quad F_X(x) = \begin{cases} 1 - e^{-\lambda x} & \text{za } x > 0 \\ 0 & \text{inače,} \end{cases}$$

$$\mathbb{E}[X] = \frac{1}{\lambda}, \mathbb{V}ar[X] = \frac{1}{\lambda^2}$$

Definicija 0.1.18. *Ako je $X \sim \Gamma(\frac{n}{2})$ za n prirodan broj, tada kažemo da X ima* ***χ^2-razdiobu*** *s n stupnjeva slobode i pišemo $X \sim \chi^2(n)$. Vrijedi:*

$$\mathbb{E}[X] = n, \mathbb{V}ar[X] = 2n.$$

Definicija 0.1.19. *Kažemo da X ima* ***normalnu razdiobu*** *s parametrima μ i σ^2 i pišemo $X \sim N(\mu, \sigma^2)$, ako je $ImX = \mathbb{R}$ i gustoća joj je:*

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

gdje je $\mu = \mathbb{E}[X], \sigma^2 = \mathbb{V}ar[X]$.

Standardizirana verzija normalne varijable X,

$$Z = \frac{X - \mu}{\sigma}$$

je normalno distribuirana s očekivanjem 0 i varijancom 1. Kažemo da Z ima **jediničnu normalnu razdiobu**. Vrijednostima od Z izražavamo koliko je standardnih devijacija pripadna vrijednost X udaljena (i na koju stranu) od svoje očekivane vrijednosti μ.
Vrijednosti jedinične normalne razdiobe su tabelirane.

Teorem 0.1.20 (Centralni granični teorem)**.** *Neka je $X_1, X_2, ...$ niz nezavisnih jednako distribuiranih slučajnih varijabli s konačnim matematičkim očekivanjem μ i konačnom varijancom $\sigma^2 > 0$. Nadalje, neka je $\overline{X_n} := \frac{(X_1+X_2+...+X_n)}{n}$ za sve prirodne brojeve n. Tada za sve $a < b$ vrijedi:*

$$\lim_{n\to\infty} \mathbb{P}\left(a \leq \frac{\overline{X_n} - \mu}{\sigma}\sqrt{n} \leq b\right) = \Phi(b) - \Phi(a)$$

gdje je $\Phi(x)$ funkcija distribucije jedinične normalne razdiobe.

Drugim riječima, kažemo da niz slučajnih varijabli $\frac{(\overline{X_n}-\mu)\sqrt{n}}{\sigma}$ *konvergira po distribuciji* jediničnoj normalnoj razdiobi kada n teži u beskonačnost i pišemo

$$\frac{\overline{X_n} - \mu}{\sigma}\sqrt{n} \xrightarrow{D} N(0,1), n \to \infty.$$

Primjena teorema na binomnu razdiobu:
Neka je X binomna $B(n, p)$ slučajna varijabla. Tada znamo da se X može reprezentirati pomoću zbroja $\sum_{i=1}^{n} X_i$ od n nezavisnih jednako distribuiranih Bernoullijevih slučajnih varijabli $X_i (i = 1, 2, ..., n)$ s parametrom p. Budući da je $\mu = p$, $\sigma^2 = p(1 - p)$ varijable $X_i (i = 1, 2, ..., n)$ zadovoljavaju uvjete CGT-a.
Dakle,

$$X \sim N(np, np(1 - p))$$

za velike n. n je dovoljno velik ako je istovremeno $np \geq 5, n(1 - p) \geq 5$

Definicija 0.1.21. ***Statistička hipoteza*** *je pretpostavka o populacijskoj razdiobi promatrane varijable. Osnovna hipoteza koja se testira zove se nulhipoteza i označava se sa H_0. Uz nulhipotezu, postavlja se i njoj* ***alternativna*** *hipoteza koju označavamo sa H_1.*

Definicija 0.1.22. ***Statistički test*** *je pravilo podjele prostora vrijednosti uzoraka na dva podskupa: na područje vrijednosti uzoraka koji su konzistentni sa H_0, i na njegov komplement u kojemu se nalaze vrijednosti nekonzistentne sa H_0.*

Testovi kojima ćemo se baviti dizajnirani su za odgovaranje na pitanje: “Daju li podaci dovoljno dokaza za odbacivanje H_0?” Odluka o odbacivanju ili ne odbacivanju nulhipoteze donosi se na osnovi vrijednosti **testne statistike**.
Područje vrijednosti koje testna statistika poprima (dakle, njezina slika) dijeli se na područje vrijednosti koje su konzistentne sa H_0 i na područje nekonzistentno sa H_0. Područje testne statistike koje je nekozistentno sa H_0 zove se **kritično područje.** Dakle, ako se opažena vrijednost testne statistike nalazi u kritičnom području, H_0 se odbacuje (u korist H_1).

Razina značajnosti testa α je vjerojatnost odbacivanja H_0 ako je H_0 istinita hipoteza. Za slučaj kada odbacimo H_0 ako je H_0 istinito, kažemo da smo počinili *pogrešku prve vrste*. *Pogrešku druge vrste* činimo kada ne odbacujemo H_0, a H_1 je istinito.

Definicija 0.1.23. ***p-vrijednost*** *je je najmanja razina značajnosti uz koju bi H_0 bila odbačena u korist zadane alternative H_1, uz vrijednost testne statistike koja je opažena.*

Pomoću p-vrijednosti možemo zaključiti sljedeće:

$$\text{ako je } p \leq \alpha \tag{11}$$

odbacujemo H_0 u korist H_1 na nivou značajnosti α,

$$\text{ako je } p > \alpha, \tag{12}$$

ne odbacujemo H_0 u korist H_1 na nivou značajnosti α.

Poglavlje 1

Neparametarski testovi

1.1 Test predznaka (Sign test)

Testovi malih uzoraka koji se odnose na srednju vrijednost i razlike između sredina temelje se na pretpostavci da su populacije uzorka imale otprilike oblik normalne distribucije. Ako je ta pretpostavka neodrživa u praksi, ove se standardne testove može zamijeniti bilo kojim od nekoliko neparametarskih testova, među njima i **testom predznaka** (sign test) koji ćemo proučiti sljedećim poglavljima.

Mali uzorci obično nam ne mogu reći slijedi li uzorak populacije normalnu razdiobu. Uobičajeni pokazatelj odstupanja od normalnosti je pojava jedne ili više vrijednosti koje su raštrkane daleko od drugih selektivnih vrijednosti. Neparametarski postupci koji se koriste u ovom poglavlju vrijede pod prilično općim uvjetima. Svakako bi ih trebalo koristiti kad postoji sumnja u prikladnost pretpostavke normalne raspodjele, a čak se mogu koristiti i kada bi pretpostavka normalne raspodjele bila razumna.

Test jednog uzorka (The one-sample sign test) odnosi se na medijan populacije koja ima neprekidnu razdiobu. Koristi se za ispitivanje nulhipoteze gdje je medijan distribucije jednak nekoj zadanoj vrijednosti μ_0. Vjerojatnost dobivanja vrijednosti uzorka koja je manja od medijana i vjerojatnost dobivanja vrijednosti uzorka koja je veća od medijana je $\frac{1}{2}$. Ako su vrijednosti populacije simetrično raspoređene oko medijana, tada su medijan M i srednja vrijednost μ jednaki. Tek povremeno smo u situaciji koja nam omogućava pretpostaviti simetriju populacije.
Zapisujemo ovako:

$$H_0 : M = \mu_0 \quad \text{ili} \quad H_0 : p = \tfrac{1}{2}$$

gdje je μ_0 određena vrijednost medijana dana u pitanju-zadatku te p vjerojatnost uspjeha u

binomnoj razdiobi.
Za prvi slučaj kada je nulhipoteza jednaka nekoj vrijednosti iz uzorka μ_0 tada one vrijednosti u uzorku koje su veće medijana bilježimo znakom plus.
U drugom slučaju, $p = \frac{1}{2}$ predstavlja binomnu razdiobu s parametrom $\frac{1}{2}$, jer se pretpostavlja da je populacija simetrična. Oba ova slučaja su objašnjena u 2.koraku.

Prilikom rješavanja problema mogu se slijediti sljedeći generalni koraci: [1]

1. korak: U ovom koraku testiramo hipoteze
   - $H_1 : \mu \neq \mu_0$ ili $H_1 : p \neq \frac{1}{2}$ (dvostrani test)
   - $H_1 : \mu > \mu_0$ ili $H_1 : p > \frac{1}{2}$ (jednostrani test na gornju granicu)
   - $H_1 : \mu < \mu_0$ ili $H_1 : p < \frac{1}{2}$ (jednostrani test na donju granicu)
     na razini značajnosti : α.
2. korak : Nakon postavljanja hipoteza računamo test-statistiku.
   Podaci se bilježe kao plus i minus znakovi, a ne numeričke veličine, zbog čega se naziva test predznaka (sign test).
   - Ako je promatranje veće od μ_0 bilježimo ga znakom +
   - Ako je promatranje manje od μ_0 bilježimo ga znakom -
   - Ako je promatranje jednako μ_0, ignoriramo ga.

   Tada je: $n = n_+ + n_-$ gdje je:
   n_+= broj opažanja sa pozitivnim predznakom, a n_- sa negativnim predznakom.
   Vjerojatnost dobivanja vrijednosti uzorka manja od medijana te vjerojatnost dobivanja uzorka veća od medijana je uvijek jednaka $\frac{1}{2}$. Razlog tome je što se pretpostavlja da je populacija simetrična, a distribucija se pretpostavlja da je binomna s jednakim vjerojatnostima za pozitivne i negativne ishode, pa je $q = 1 - p = \frac{1}{2}$.
   Ako označimo sa R broj pozitivnih ili negativnih znakova tada je pod nulhipotezom u osnovi binomna distribucija sa parametrima n i $\frac{1}{2}$, $R \sim B(n, \frac{1}{2})$. Za izračun test statistike koristi se normalna aproksimacija binomne distribucije.
   Koristeći normalnu aproksimaciju binomne distribucije slijedi:

   $$R \sim N(np, np(1-p)) = N(\frac{n}{2}, \frac{n}{4})$$

   gdje je $p = \frac{1}{2}$.
   Za dvostrani test razmatramo $R = min(n_+, n_-)$ a za jednostrani test na gornju granicu razmatramo broj pozitivnih znakova.
   Primjenom korekcije zbog neprekidnosti dobivamo formule za računanje vjerojatnosti:

a) $P(R \geq r) = P(R > r - \frac{1}{2})$ [jednostrani test]

b) $P(R \leq r) = P(R < r + \frac{1}{2})$ [dvostrani test]

Test-statistika se računa po formuli:

$$z = \frac{r \pm 0.5 - \mu}{\sigma}$$

gdje je $Z \sim N(0,1)$, $\mu = \frac{n}{2}, \sigma^2 = \frac{n}{4}, r = min(n_+, n_-)$ te znak plus koristimo za dvostrani test, a znak minus za jednostrani test.

3. korak: U ovom koraku kojeg kraće nazivamo kriterij odbijanja, razmatramo za koji slučaj odbacujemo hipotezu H_0.
4. korak: U ovom koraku donosimo odluku.
5. korak: Zaključak.

Primjer 1.1.1. *Smatra se da upitnik koji se koristi u procjeni daje medijan bodova 50 u grupi koja pohađa određeni tečaj. 20 studenata drugog tečaja dobilo je bodove: 26, 46, 39, 58, 62, 41, 65, 49, 54, 50, 61, 38, 58, 35, 27, 34, 46, 51, 29, 40. Testirajte hipotezu da medijan nije 50 na nivou značajnosti od* 5%.*[1]*

Rješenje:

1. Testiramo hipoteze:

$$H_0 : M = 50$$
$$H_1 : M \neq 50$$
$$\alpha = 0.05 \Rightarrow \text{(dvostrani test)} \Rightarrow \frac{\alpha}{2} = 0.025$$

2. Test statistika

a)

$$n_+ = 7, n_- = 12 \Rightarrow n = n_+ + n_- = 19$$
$$p = \frac{1}{2}, q = \frac{1}{2}$$
$$r = min(n_+, n_-) = min(7, 12) = 7$$

b)

$$R \sim N(np, npq)$$
$$R \sim N(19 \cdot 0.5, 19 \cdot 0.5 \cdot 0.5)$$
$$R \sim N(9.5, 4.75) \Rightarrow \mu = 9.5, \sigma^2 = 4.75$$

c)
$$P(R \leq r) = P(R < r + \frac{1}{2}) = P(R < 7.5)$$

d)
$$Z = \frac{r + \frac{1}{2} - \mu}{\sigma} = \frac{7.5 - 9.5}{\sqrt{4.75}} = -0.92$$

3. Odbacujemo H_0 ako je $|z| > z_{\frac{\alpha}{2}} = 1.96$

4. Odluka: $|-0.92| < 1.96 \Rightarrow$ ne odbacujemo H_0.

5. Na nivou značajnosti od 5% ne odbacujemo H_0.
   Drugim riječima, možemo raditi s pretpostavkom da je medijan jednak 50.

Primjer 1.1.2. *Dana su mjerenja čvrstoće na lomljenje određene vrste 2-inčne pamučne vrpce u kilogramima:*

163	165	160	189	161	171	158	151	169	162	163	139
172	165	148	166	172	163	187	173				

Koristeći test predznaka testirajte nulhipotezu $M = 160$ *protiv alternativne hipoteze* $M > 160$ *na nivou značajanosti od 0.025.[1]*

Rješenje:

1. Testiramo hipoteze:
$$H_0 : M = 160$$
$$H_1 : M > 160$$
$$\alpha = 0.025 \Rightarrow (\text{ jednostrani test})$$

2. Računamo test statistiku

a)
$$n_+ = 15, n_- = 4 \Rightarrow n = 15 + 4 = 19$$
$$p = \tfrac{1}{2}, q = \tfrac{1}{2}$$
$$r = n_+ = 15$$

b)
$$R \sim N(np, npq)$$
$$R \sim N(19 \cdot 0.5, 19 \cdot 0.5 \cdot 0.5)$$
$$R \sim N(9.5, 4.75) \Rightarrow \mu = 9.5, \sigma^2 = 4.75$$

c)
$$P(R \geq r) = P(R > r - \frac{1}{2}) = P(R > 14.5)$$

d)
$$z = \frac{r - \frac{1}{2} - \mu}{\sigma} = \frac{14.5 - 9.5}{\sqrt{4.75}} = 2.29$$

3. Odbacujemo H_0 ako je $z > z_\alpha = 1.96$

4. Odluka:
   Kako je $2.29 > 1.96 \Rightarrow$ odbacujemo H_0.

5. Na nivou značajnosti od 0.025 odbacujemo H_0.
   Drugim riječima, možemo raditi s pretpostavkom da je medijan veći od 160.

Pogledajmo idući primjer. Riješit ćemo ga na vrlo sličan način kao prethodni, ali pomoću p-vrijednosti, bez korištenja aproksimacije binomne slučajne varijable normalnom. Prilikom računanja p-vrijednosti koristimo tablicu 2.1 jer test statistika z, koja je jednaka broju plus znakova, uz pretpostavku da nulhipoteza vrijedi, ima binomnu razdiobu s parametrima n i 0.5.

Primjer 1.1.3. *Promatrajući prosječnu mjesečnu potrošnju boje na slučajnom uzorku ličioca, njihova prosječna potrošnja bila je 35.5, 37.3, 35.0, 37.7, 35.2, 33.9, 30.0, 41.3, 43.2, 34.6, 41.1, 33.9, 40.9, 27.8 i 32.0. Koristeći se jednostranim testom predznaka testirajte nulhipotezu M=35.0 galona boje protiv alternativne hipoteze M <35.0 galona, na nivou značajnosti* α. *[3]*

Rješenje:
S obzirom da imamo jednu vrijednost u uzorku jednaku M, nju odbacujemo, pa je $n = 14$.

1. Testiramo hipoteze:
$$\begin{aligned} M &= 35.0 \\ M &< 35.0 \end{aligned}$$

2. Nivo značajnosti: α.

3. Kriterij: Test statistika z je broj plus znakova.

4. Račun: Mijenjajući sve vrijednosti veće od 35.0 s plus znakom te sve vrijednosti manje od 35.0 s minus znakom i odbacujući vrijednost jednaku 35.0 dobivamo:

$$+ + + + - - + + - + - + - -$$

Dakle $z = 8$.

5. Odluka:
S obzirom na to da su ispitivanja neovisna i da je vjerojatnost ista za svako ispitivanje, koristimo binomnu distribuciju koja se za male n (broj pokušaja) iščitava direktno iz tablice binomnih vjerojatnosti 2.1.
Stoga imamo za: $p = 0.5$, $n = 14$ i $z \leq 8$ je:
$z_0 = 0$, $z_1 = 0.001$, $z_2 = 0.006$, $z_3 = 0.022$, $z_4 = 0.061$, $z_5 = 0.122$, $z_6 = 0.183$, $z_7 = 0.209, z_8 = 0.183$, gdje oznaka z_k predstavlja $\mathbb{P}(Z = k)$, gdje Z ima binomnu $B(14, 0.5)$ razdiobu.
Prazna mjesta u tablici daju vrijednost 0, jer vjerojatnosti zaokružene na 3 decimale iznose 0.
Odakle slijedi vjerojatnost da je $z \leq 8$:
$0.001 + 0.006 + 0.022 + 0.061 + 0.122 + 0.183 + 0.209 + 0.183 = 0.787$.
Dakle p-vrijednost je 0.787.
Dakle za sve $\alpha < 0.787$ ne odbacujemo nulhipotezu po (12).

Zadatak 1.1.4. *Simuliran je uzorak duljine $n = 15$ iz eksponencijalne razdiobe s parametrom $\lambda = 2$ pomoću programa R. Dobiveni su sljedeći podaci (zaokruženi na tri decimale).*

0.225	0.135	0.074	0.190	0.981	0.480	1.107	0.261
0.340	0.477	0.542	0.018	0.536	0.027	0.065	

Testirajte nulhipotezu $M = \frac{1}{2}$ protiv alternativne da je $M > \frac{1}{2}$ na nivou značajnosti α.

Naredbom za nasumični generator za eksponencijalnu razdiobu *y*$< -$*rexp(15,2)* dobili smo gornje podatke koje smo ispisali naredbom *y*.
rexp(15,2) predstavlja nasumični generator eksponencijalne razdiobe duljine 15 i parametrom 2.

Kako nemamo nijednu vrijednost u uzorku jednaku M, ostaje $n = 15$

1. Testiramo hipoteze:
$$\begin{aligned} M &= \tfrac{1}{2} \\ M &> \tfrac{1}{2} \end{aligned}$$

2. Nivo značajnosti: α.

3. Kriterij: Test statistika z je broj plus znakova.

4. Račun: Mijenjajući sve vrijednosti veće od $\frac{1}{2}$ sa znakom plus te sve vrijednosti manje od $\frac{1}{2}$ sa znakom minus dobivamo:

$$- - - - + - + - - - + - + - -$$

   Dakle $z = 4$.

5. Odluka:
   Tablica 2.1 prikazuje za $n = 15$ i $p = 0.50$ vjerojatnost da $z \geq 4$:
   $0.042 + 0.092 + 0.153 + 0.196 + 0.196 + 0.153 + 0.092 + 0.042 + 0.014 + 0.003 = 0.983$.

   Dakle p-vrijednost je 0.983.
   Dakle za sve $\alpha < 0.983$ ne odbacujemo nulhipotezu po (12).

1.2 Test uparenog uzorka (The paired-sample sign test)

Test predznaka također se može koristiti kada radimo s uparenim podacima. Pretpostavimo da se za usporedbu uzimaju uzorci iz iste populacije. Važno je istaknuti da imamo par promatranja koja su povezana.

Svaki par vrijednosti uzorka zamjenjuje se znakom plus ako je prva vrijednost veća od druge, znakom minus ako je prva vrijednost manja od druge, a odbacuje se kada su dvije vrijednosti jednake. Test predznaka za situaciju uparenih uzoraka ispituje ima li populacija razlika između parova medijan koji je jednak nuli.

Primjer 1.2.1. *Procjena za 9 pacijenata je dana u tablici:*

Pacijent	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Tretman A	*36.3*	*48.4*	*40.2*	*54.7*	*28.7*	*42.8*	*36.1*	*39.0*	*36*
Tretman B	*35.1*	*46.8*	*37.3*	*50.6*	*29.1*	*41.0*	*35.3*	*39.1*	*36*

Koristite test predznaka kako biste utvrdili predstavljaju li podaci dovoljno dokaza koji ukazuju na to da tretmani nisu jednako učinkoviti koristeći nivo značajnosti, $\alpha = 0.05$.*[1]*

Rješenje:

1. Testiramo hipoteze:
$$H_0 : M = 0 \quad H_1 : M \neq 0$$
$$H_0 : p = \tfrac{1}{2} \quad H_1 : p \neq \tfrac{1}{2}$$
   $\alpha = 0.05 \Rightarrow$ (dvostrani test) $\Rightarrow \frac{\alpha}{2} = 0.025$.
   Gdje je M oznaka za medijan razlike.

2. Računamo test statistiku

	1	2	3	4	5	6	7	8	9
razlika	1.2	1.6	2.9	4.1	-0.4	1.8	0.8	-0.1	0
predznak	+	+	+	+	-	+	+	-	ignorirati

a)

$$n_+ = 6, n_- = 2 \Rightarrow n = n_+ + n_- = 8$$
$$r = min(n_+, n_-) = min(2, 6) = 2$$
$$p = \tfrac{1}{2}, q = \tfrac{1}{2}$$

b)

$$R \sim N(np, npq)$$
$$R \sim N(8 \cdot 0.5, 8 \cdot 0.5 \cdot 0.5)$$
$$R \sim N(4, 2) \Rightarrow \mu = 4, \sigma^2 = 2$$

c)

$$P(R \leq r) = P(R < r + \frac{1}{2}) = P(R < 2.5)$$

d)

$$Z = \frac{r + \frac{1}{2} - \mu}{\sigma} = \frac{2.5 - 4}{\sqrt{2}} = -1.06$$

3. Odbacujemo H_0 ako je $|z| > z_{\frac{\alpha}{2}} = 1.96$
4. Odluka: $|-1.06| < 1.96 \Rightarrow$ ne odbacujemo H_0
5. Kako ne odbacujemo H_0, možemo pretpostaviti da su tretmani jednaki.

Primjer 1.2.2. *Prijevoznička tvrtka prodala je svoju flotu od osam starih kamiona zamijenivši ih sa osam novih kamiona kako bi smanjila broj mehaničkih kvarova koji su se dogodili. Slijedi broj kvarova koji su se dogodili tijekom tri mjeseca prije i tri mjeseca nakon nabave novih kamiona. U sljedećim parovima, prva vrijednost je broj kvarova starog kamiona, a druga vrijednost je broj kvarova nasumično uparenih novih kamiona.*

4 i 1	*2 i 2*
3 i 0	*4 i 3*
3 i 2	*0 i 1*
1 i 2	*2 i 1*

Koristite test uparenog uzorka za testiranje nulhipoteze da su novi kamioni jednako učinkoviti kao i stari kamioni, nasuprot alternativne hipoteze da su učinkovitiji, u smislu da imaju manje mehaničkih kvarova. [3]

Rješenje:
S obzirom da jedan od parova (2 i 2) sadrže iste vrijednosti, njih ignoriramo te je veličina uzorka $n = 8 - 1 = 7$.

1. Testiramo hipoteze:
   $H_0 : M_1 = M_2$ gdje je $M_1 = M_2$ prosječni broj mehaničkih kvarova za 3 mjeseca kod starih i novih kamiona.
   $H_1 : M_1 - M_2 > 0$
2. Na nivou značajnosti: α.
3. Statistika testa: z je broj znakova +
4. Račun: zamjenjujući svaki par vrijednosti sa + ako je prva vrijednost veća od druge te sa - ako je prva vrijednost manja od druge dobivamo:

   $$+ + + - + - +$$

   Dakle $z = 5$.
5. Odluka:

   Koristeći se tablicom 2.1 za n=7 i p =0.50 vjerojatnost $z \geq 5$ je 0.164+0.555+0.008=0.227. p-vrijednost je dakle 0.227. Dakle za sve $\alpha < 0.227$, nećemo odbaciti nulhipotezu.

Zadatak 1.2.3. *Za simulaciju uzorka duljine 100 iz binomne razdiobe s parametrima n=4 i p=0.5, za poduzorak duljine 30 dobivamo sljedeće podatke za opažanje A i opažanje B:*

Opažanja	*broj pojavljivanja opažanja*									
Opažanje A	2	2	2	2	4	1	3	3	1	2
Opažanje B	2	2	2	2	0	4	1	1	3	2
Opažanje A	2	2	4	2	2	2	3	1	2	2
Opažanje B	2	2	0	2	2	2	1	4	2	2
Opažanje A	2	1	2	3	1	4	2	3	1	3
Opažanje B	2	3	2	1	3	0	2	1	3	1

Možemo li na nivou značajnosti α tvrditi da je opažanje B učestalije od opažanja A?

Rješenje:

Naredbom *binom4*$< -$ *rbinom(100,4,0.5)* dobili smo gornje podatke koje smo ispisali naredbom *binom4[1:30]*.
rbinom(100,4,0.5) predstavlja nasumični generator binomne distribucije sa

parametrima redom:
n,broj opažanja, *size*, broj pokušaja te *prob*, vjerojatnosti uspjeha za svaki pokušaj.
U našem zadatku $n = 100$,size=4, $p = 0.5$.
Naredbom *rbinom[1:30]* ispisujemo poduzorak duljine 30.

S obzirom na to da petnaest parova (5 i 5) sadrže iste vrijednosti, njih ignoriramo te je veličina uzorka $n = 30 - 15 = 15$.

1. Testiramo hipoteze:

$$H_0 : M_1 = M_2$$
$$H_1 : M_1 - M_2 < 0$$

2. Nivo značajnosti: α.
3. Statistika testa: z je broj znakova +
4. Račun: zamjenjujući svaki par vrijednosti sa + ako je prva vrijednost veća od druge te sa znakom - ako je prva manja od druge dobivamo:

$$+ - + + - + + - - + - + + - +$$

   Dakle $z = 9$.

5. Odluka:

   Koristeći se tablicom 2.1 za $n = 15$ i $p = 0.5$ vjerojatnost da je $z \leq 9$ je:

$$0 + 0 + 0.003 + 0.014 + 0.042 + 0.092 + 0.153 + 0.196 + 0.196 + 0.153 = 0.849.$$

   Prve dvije nule su vrijednosti za z_0 i z_1 koje su objašnjene u primjeru 1.1.3.

   Dakle za sve $\alpha < 0.849$, ne odbacujemo H_0.

1.3 Test predznaka (veliki uzorci) [3]

Kada su np i $n(1 - p)$ oboje veći od 5 tada možemo koristiti normalnu aproksimaciju binomne distribucije. U tom slučaju se za test statistiku testa predznaka uzima

$$z = \frac{x - np_0}{\sqrt{np_0(1 - p_0)}} \tag{1.1}$$

gdje je $p_0 = 0.50$, x broj plus znakova objašnjeno u poglavlju 1.1 u 2.koraku.
Za jednostranu alternativnu hipotezu $p < p_0$ odbacujemo nulhipotezu ako je $z \leq -z_\alpha$,
kada je $p > p_0$ odbacujemo nulhipotezu ako je $z \geq z_\alpha$,
a za $p \neq p_0$ odbacujemo nulhipotezu ako je $z \leq -z_{\frac{\alpha}{2}}$ ili $z \geq z_{\frac{\alpha}{2}}$.
Gdje je p oznaka kao u poglavlju 1.1.

Primjer 1.3.1. *U nastavku su mjerenja dubine oceana na određenom području u fatomima: 46.4, 48.3, 51.9, 38.8, 46.5, 45.6, 52.1, 41.0, 54.2, 44.9, 52.3, 43.6, 48.7, 42.2 i 44.9. Koristeći se testom predznaka sa nivoom značajnosti* $\alpha = 0.05$ *testirajte nulhipotezu* $M = 43.0$ *protiv alternativne hipoteze* $M \neq 43.0$*.[3]*

Rješenje:

1. Testiramo hipoteze:
$$\begin{aligned} H_0 &: \quad M = 43.0 \\ H_1 &: \quad M \neq 43.0 \end{aligned}$$

2. Na nivou značajnosti: α.

3. Kriterij: $z_{\frac{\alpha}{2}}$ odgovara ulazu od $0.5000 - 0.025 = 0.475$ u tablici 2.2. Najbliži ulaz je 0.4750 što odgovara $z = 1.96$, pišemo $z_{0.025} = 1.96 = z_{\frac{\alpha}{2}}$.
Odbacujemo nulhipotezu ako je $z \leq -1.96$ ili $z \geq 1.96$ gdje z računamo po formuli (1.1).

4. Račun: zamjenujući svaku vrijednost veću od 43.0 sa znakom plus, a manju sa znakom minus dobivamo:
$$+ + + - + + + - + + + + + + - +$$
odakle slijedi da je x = 12 i uvrštavanjem u formulu (1.1) dobivamo:
$$z = \frac{12 - 15(0.50)}{\sqrt{15(0.50)(0.50)}} = 2.32$$

5. Odluka: Kako je $z = 2.32 > 1.96$ odbacujemo nulhipotezu.
Drugim riječima, srednja dubina oceana na navedenom području nije 43.0 fatoma kao što je prije izvješteno. Da smo iskoristili korekciju zbog neprekidnosti dobili bi $z = 2.07$ i odluka je ista.

Zadatak 1.3.2. *Koristeći se simulacijom poissonove razdiobe u programskom jeziku R duljine 100, a* $\lambda = 50$ *za poduzorak od 20 podataka dobili smo sljedeće vrijednosti:*

50 41 55 62 51 43 47 49 47 41 44 46 48 50 54 44 38 50 49 53

Na nivou značajnosti od $\alpha = 0.05$ *testirajte nulhipotezu* $M = 50$ *protiv alternativne* $M > 50$.

S obzirom da imamo tri vrijednosti koje su jednake 50 njih ignoriramo, pa naša veličina uzorka sada je jednaka $n = 17$.

Rješenje: Pripadni kod u programskom jeziku *R* je:

poiz< −rpois(100,50)
poiz[1:20]

gdje funkcija *rpois(n,lambda)* predstavlja nasumični generator Poissonove distribucije s parametrom lambda te brojem nasumičnih vrijednosti, n, za vratiti.
U našem primjeru $n = 100, \lambda = 50$.
Naredba *poiz[1:20]* ispisuje 20 vrijednosti od nasumičnih vrijednosti.

1. Testiramo hipoteze:
$$H_0 : \quad M = 50$$
$$H_1 : \quad M \geq 50$$

2. Na nivou značajnosti: $\alpha = 0.05$

3. Kriterij: Odbacujemo nulhipotezu ako je $z \geq 1.645$.

4. Račun: zamjenujući svaku vrijednost veću od 50 sa znakom plus, a manju sa znakom minus dobivamo:
$$- + + + - - - - - - - - - + - - - +$$
odakle slijedi da je x = 5 i uvrštavanjem u formulu (1.1) dobivamo:
$$z = \frac{5 - 17(0.50)}{\sqrt{17(0.50)(0.50)}} = -1.7$$

5. Odluka: Kako je $z = -1.7 < 1.645$ ne odbacujemo nulhipotezu.
Drugim riječima, medijan uzorka je 50.

1.4 U test[3]

U ovom odjeljku predstavit ćemo neparametarsku alternativu jednostranom t-testu za razlike u očekivanjima.

Naziva se **U test, Wilcoxon** ili **Mann-Whitney test**, po statističarima koji su doprinijeli njegovom razvoju. Različiti nazivi odnose se na različite metode ili organiziranje izračuna, ali postupci su logički ekvivalentni. Moći ćemo testirati nulhipotezu da dva uzorka potječu iz identične populacije, a da ne pretpostavljamo da testirane populacije imaju normalne raspodjele. U stvari, test zahtijeva samo da testirana populacija ima neprekidnu razdiobu populacije iz koje je uzet uzorak.

Pretpostavimo da nasumično odaberete uzorke iz svake populacije koje se uspoređuju. U test, kako je prvotno definirano, temelji se na broju puta kada su vrijednosti uzorka iz jedne populacije premašene vrijednostima uzorka iz druge populacije.
Da ilustriramo kako se na taj način izvodi U test, pretpostavimo da želimo usporediti veličinu zrna pijeska dobivenih s dva različita mjesta na Mjesecu na temelju sljedećih promjera mjerenih u milimetrima ([3]):

Lokacija 1	0.37,	0.70,	0.75,	0.30,	0.45,	0.16,	0.62,	0.73,	0.33	
Lokacija 2	0.86,	0.55,	0.80,	0.42,	0.97,	0.84,	0.24,	0.51,	0.92,	0.69

Aritmetičke sredine za ta dva uzorka su 0.49 i 0.68 i njihova je razlika velika, no preostaje nam vidjeti je li značajna. Kako bismo izveli U test, podatke spojimo u jedan uzorak na način da ih poredamo po veličini (od najmanjeg k najvećem) i zabilježimo iz kojeg uzorka je koji podatak. Na taj način dobivamo:

0.16	0.24	0.30	0.33	0.37	0.42	0.45	0.51	0.55	0.62
1	2	1	1	1	2	1	2	2	1
0.69	0.70	0.73	0.75	0.80	0.84	0.86	0.92	0.97	
2	1	1	1	2	2	2	2	2	

pri čemu smo za svaku vrijednost naznačili dolazi li s lokacije 1 ili lokacije 2.
Dodjeljujući podacima rangove 1,2,3, ..., i 19 vidimo da vrijednosti prvog uzorka (lokacija 1) zauzimaju rangove 1, 3, 4, 5, 7, 10, 12, 13 i 14 dok vrijednosti drugog uzorka zauzimaju rangove 2, 6, 8, 9, 11, 15, 16, 17, 18 i 19.

Ovdje nemamo iste vrijednosti jedne za drugom. U slučaju da imamo, dodijelili bismo im srednju vrijednost rangova koje zauzimaju. Na primjer, ako su treća i četvrta vrijednost jednake, dodijelili bismo svakoj rang $\frac{3+4}{2} = 3.5$ a ako su deveta, deseta i jedanaesta vrijednost jednake, dodijelili bismo svakoj rang $\frac{9+10+11}{3} = 10$.

Ako postoji osjetna razlika između aritmetičkih sredina dviju populacija, većina nižih rangova vjerojatno će biti dodijeljena vrijednostima jednog uzorka, dok će većina viših rangova biti dodijeljena vrijednostima drugog uzorka.

Test nulhipoteze da dva uzorka potječu iz identične populacije mogu se stoga temeljiti na W_1, zbroju rangova vrijednosti prvog uzorka ili na W_2, zbroju rangova vrijednosti drugog uzorka. U praksi nije važno koji uzorak nazivamo uzorkom 1, a koji nazivamo uzorkom 2, niti hoćemo li test temeljiti na W_1 ili W_2. Kada su veličine uzorka nejednake, uobičajena je praksa da je n_1 manji od dva.

Ako su veličine uzorka n_1 i n_2, zbroj W_1i W_2 jednostavno je zbroj prvih $n_1 + n_2$ cjelobrojnih brojeva koji je poznat kao jednakost $\frac{(n_1+n_2)(n_1+n_2+1)}{2}$.

Ova formula nam omogućava da pronađemo W_2 ako znamo W_1 i obrnuto. Za našu ilustraciju imamo: $W_1 = 1+3+4+5+7+10+12+13+14 = 69$ i s obzirom da je suma prvih 19 pozitivnih cijelih brojeva $\frac{19 \cdot 20}{2} = 190$ slijedi da je $W_2 = 190 - 69 = 121$ (vrijednost se može provjeriti zbrajajući 2, 6, 8, 9, 11, 15, 16, 17, 18 i 19). Kada je upotreba zbroja rangova prvi put predložena kao neparametarska alternativa dvostranom t testu, odluka se temeljila na W_1 ili W_2. Trenutno se odluka temelji na bilo kojoj povezanoj statistici

$$\boxed{U_1 = W_1 - \frac{n_1(n_1+1)}{2} \text{ ili } U_2 = W_2 - \frac{n_2(n_2+1)}{2}} \tag{1.2}$$

ili na statistici U, koja je uvijek jednaka $\min\{U_1, U_2\}$.
Rezultirajući testovi ekvivalentni su onima koji se temelje na W_1 ili W_2, ali oni imaju prednost što se čitaju lako iz tablice kritičnih vrijednosti. Ne samo da U_1 i U_2 uzimaju vrijednosti u intervalu od 0 do $n_1 n_2$ zapravo, njihov zbroj je uvijek jednak $n_1 n_2$ te su njihove distribucije uzoraka simetrične oko $\frac{n_1 n_2}{2}$. Upotreba U, koja je uvijek jednaka manjoj od vrijednosti U_1 i U_2, ima dodatnu prednost što je rezultirajući jednostran test i time je lakše tabelirati.

U skladu s tim, testiramo nulhipotezu da dva uzorka potječu iz identične populacije nasuprot alternativnoj hipotezi da dvije populacije imaju nejednake sredine sa sljedećim kriterijem:
Odbaciti nulhipotezu ako $U \leq U'_\alpha$ *gdje je* U'_α *dan u Tablici* 2.5 za $n_1 \leq 15, n_2 \leq 15$ *i* $\alpha = 0.05$ *ili* $\alpha = 0.01$.
U konstrukciji tablice 2.5 U'_α najveća je vrijednost U za koju je vjerojatnost $\mathbb{P}(U \leq U'_\alpha)$ manja ili jednaka α, a prazni razmaci ukazuju da se nulhipoteza ne može odbaciti na danoj razini značajnosti, bez obzira na vrijednost koju dobivamo za U.

Opsežnije tablice mogu se naći u priručnicima statističkih tablica, ali kada su i n_1 i n_2 veći od 8, obično se smatra razumnim koristiti test velikog uzorka (large-sample test) koji ćemo razmotriti u idućem odjeljku. Uočili su da se nulhipoteza odbacuje za male vrijednosti U. U svakom slučaju, $U = U_1$ ili $U = U_2$. Kada je U jednako U_1 tada uzorak jedan daje male vrijednosti. Slično, kada je U jednak U_2 tada uzorak dva daje male vrijednosti.

Primjer 1.4.1. *Koristeći se prethodnim podacima za veličinu zrna te upotrebljujući U test s nivoom značajnosti od 0.05, testirajte nulhipotezu da dva uzorka dolaze iz iste populacije nasuprot alternativne da dvije populacije imaju nejednake sredine. [3]*

Rješenje:

1. Testiramo hipoteze:

$$\begin{aligned} H_0 &: \text{ populacije su jednake} \\ H_1 &: \quad \mu_1 \neq \mu_2 \end{aligned}$$

2. Na nivou značajnosti: $\alpha = 0.05$.

3. Kriterij: Odbacujemo nulhipotezu ako je $U \leq 20$ što je vrijednost U'_α za $n_1 = 9$, $n_2 = 10$ i $\alpha = 0.05$ iščitano iz tablice 2.5.

4. Račun:
   Pokazujući da je $W_1 = 69$ i $W_2 = 121$ imamo:
   $$U_1 = 69 - \tfrac{9 \cdot 10}{2} = 24 \quad U_2 = 121 - \tfrac{10 \cdot 11}{2} = 66$$
   i zbog toga $U = 24$. Primjetimo $U_1 + U_2 = 24 + 66 = 90$ što je jednako $n_1 n_2 = 9 \cdot 10$.

5. Odluka:
   Kako je $U = 24$ veće od 20 ne odbacujemo nulhipotezu.
   Drugim riječima, ne možemo zaključiti da je stvarna razlika u veličini zrna pijeska s dvije lokacije na Mjesecu. Možemo pretpostaviti da su populacije jednake.

Test koji smo ovdje opisali također se može koristiti kada je alternativa $\mu_1 < \mu_2$ ili $\mu_1 > \mu_2$. No budući da je postupak kompliciraniji, u tom slučaju raspravljat ćemo samo za veliki uzorak u sljedećem dijelu.

Zadatak 1.4.2. *Koristeći se programom R za simulaciju eksponencijalne razdiobe sa parametrom $\lambda = 3$ te veličinama uzorka redom 5 i 3 dobivamo sljedeće vrijednosti:*

Model 1	1.92,	0.10,	0.53,	0.20,	0.10		
Model 2	0.48,	0.10,	0.17,	0.25,	0.46,	0.30	0.08

Koristeći U test baziran na tablici 2.5 na nivou značajnosti od 0.01 testirajte nulhipotezu da dva uzorka dolaze iz iste populacije, nasuprot alternativne da imaju nejednake sredine.

Rješenje: Pripadni kod u programskom jeziku *R* je:

```
uTest< −rexp(5,3)
uTest
UTest2< −rexp(7,3)
UTest2
```

1. Testiramo hipoteze:
   $$\begin{aligned} H_0 &: \text{ populacije su jednake} \\ H_1 &: \quad \mu_1 \neq \mu_2 \end{aligned}$$

2. Na nivou značajnosti: α.

3. Kriterij: Odbacujemo nulhipotezu ako je $U \leq 1$ što je vrijednost U'_α za $n_1 = 5$, $n_2 = 7$ i $\alpha = 0.01$ iščitano iz tablice 2.5.

4. Račun:

rang	1	2	3	4	5	6	7	8	9	10	11	12
vrijednosti	0.08	0.1	0.1	0.1	0.17	0.2	0.25	0.3	0.46	0.48	0.53	1.92
uzorak	2.	1.	1.	2.	2.	1.	2.	2.	2.	2.	1.	1.

S obzirom da vrijednost 0.1 zauzima rangove 2, 3 i 4 tada oni postaju:

$$\frac{2+3+4}{3} = 3$$

Slijedi:

rang	1	3	3	3	5	6	7	8	9	10	11	12
vrijednosti	0.08	0.1	0.1	0.1	0.17	0.2	0.25	0.3	0.46	0.48	0.53	1.92
uzorak	2.	1.	1.	2.	2.	1.	2.	2.	2.	2.	1.	1.

$$W_1 = 3+3+6+11+12 = 35$$
$$W_2 = 1+3+5+7+8+9+10 = 43$$

$$U_1 = 35 - \tfrac{5\cdot 6}{2} = 20 \quad U_2 = 43 - \tfrac{7\cdot 8}{2} = 15$$

i zbog toga $U = 15$.

5. Odluka:
Kako je $U = 15$,što je veće od 1, ne odbacujemo nulhipotezu.
Drugim riječima, možemo pretpostaviti da su populacije jednake.

1.5 U test (veliki uzorci) [3]

U-test velikog uzorka može se temeljiti na U_1 ili U_2 kako je prikazano u prethodnom poglavlju 1.4 .

Rezultirajući test bit će ekvivalentan tako da nije važno koji uzorak ćemo identificirati kao uzorak 1. U sljedećem opisu koristit ćemo statistiku U_1. Pod nulhipotezu da dva uzorka potječu iz identične populacije, može se pokazati da su srednja vrijednost i standardna devijacija raspodjele uzorka U_1 jednake:

$$\boxed{\mu_{U_1} = \frac{n_1 n_2}{2} \text{ i } \sigma_{U_1} = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} \tag{1.3}$$

Dalje, ako su n_1 i n_2 veće od 8, raspodjela uzorka U_1 može se aproksimirati približno normalnoj krivulji, zato temeljimo test nulhipoteze da dva uzorka potječu iz identične populacije na statistici

$$\boxed{z = \frac{U_1 - \mu_{U_1}}{\sigma_{U_1}}} \tag{1.4}$$

koja ima približno standardnu normalnu distribuciju.

Ako je alternativa hipoteza $\mu_1 \neq \mu_2$, odbacujemo nulhipotezu za $z \leq -z_{\frac{\alpha}{2}}$ ili $z \geq z_{\frac{\alpha}{2}}$.
Ako je alternativna hipoteza $\mu_1 < \mu_2$, odbacujemo nulhipotezu za $z \leq -z_\alpha$, budući da male vrijednosti U_1 odgovaraju malim vrijednostima zbroja ranga W_1.
Ako je alternativna hipoteza $\mu_1 > \mu_2$, odbacujemo nulhipotezu za $z \geq z_\alpha$.
Pronalaženje rangova kada su n_1 i n_2 veliki prije je bio kompliciran zadatak, no danas uz tehnologiju to nije tako.

Primjer 1.5.1. *Sljedeće su dobitci mase (u kilogramima) mladih purana koji se hrane s dvije različite hrane, ali se drže pod jednakim uvjetima:*

Hrana 1:	16.3,	10.1,	10.7,	13.5,	14.9,	11.8,	14.3	10.2,
	12.0,	14.7,	23.6,	15.1,	14.5,	18.4,	13.2,	14.0
Hrana 2:	21.3,	23.8,	15.4,	19.6,	12.0,	13.9,	18.8,	19.2,
	15.3,	20.1,	14.8,	18.9,	20.7,	21.1,	15.8,	16.2

Koristeći se U testom za velike uzorke sa nivoom značajnosti od 0.01, testirajte nulhipotezu da su dvije testirane populacije identične protiv alternativne hipoteze da u drugoj prehrani u prosjeku dolazi do većeg dobitka u težina. [3]

Rješenje:

1. Testiramo hipoteze:

$$\begin{aligned} H_0 &: \quad \text{populacije su jednake} \\ H_1 &: \quad \mu_1 < \mu_2 \end{aligned}$$

2. Na nivou značajnosti $\alpha = 0.01$

3. Kriterij: odbacujemo nulhipotezu ako je $z \leq -2.33$

4. Račun: Raspoređivanje podataka prema veličini koju dobijemo

rang	1	2	3	4	5	6	7	8	9	10
povećanje težine vrsta hrane	10.1 1	10.2 1	10.7 1	11.8 1	12.0 1	12.0 2	13.2 1	13.5 1	13.9 2	14.0 1
rang	11	12	13	14	15	16	17	18	19	20
povećanje težine vrsta hrane	14.3 1	14.5 1	14.7 1	14.8 2	14.9 1	15.1 1	15.3 2	15.4 2	15.8 2	16.2 2
rang	21	22	23	24	25	26	27	28	29	30
povećanje težine vrsta hrane	16.3 1	18.4 1	18.8 2	18.9 2	19.2 2	19.6 2	20.1 2	20.7 2	21.1 2	21.3 2
rang	31	32								
povećanje težine vrsta hrane	23.6 1	23.8 2								

Ispod povećanja težine nalaze se uzorci iz kojih su došli. Primijetite da postoji veza od 12.0 kilograma koja uključuje po jednu puretinu iz svake prehrane. To su peta i šesta pozicija na listi, pa je svakoj dodijeljen rang:

$$\frac{5+6}{2} = 5.5$$

Zbroj rang pozicija koje zauzimaju purani prvog uzorka:

$$W_1 = 1+2+3+4+5.5+7+8+10+11+12+13+15+16+21+22+31 = 181.5$$

i

$$U_1 = 181.5 - \frac{16 \cdot 17}{2} = 45.5$$

Stoga:

$$\mu_{U_1} = \frac{16 \cdot 16}{2} = 128 \text{ i } \sigma_{U_1} = \sqrt{\frac{16 \cdot 16 \cdot 33}{12}} \approx 26.53$$

slijedi:

$$z = \frac{45.5 - 128}{26.53} \approx -3.11$$

5. Odluka: Kako je $z = -3.11$, što je manje od -2.33, odbacujemo nulhipotezu. Drugim riječima, zaključujemo da druga prehrana u prosjeku donosi veći dobitak na težini.

Zadatak 1.5.2. *Koristeći se programom R i simulacijom uzorka binomne distribucije te pozivom iste funkcije dva puta za redom gdje promatramo 100 opažanja sa 100 pokušaja i vjerojatnosti 0.5 dobivamo sljedeće podatke:*

Poziv 1:	53	57	42	46	48	58	50	44	43	45
Poziv 2:	49	52	53	51	49	54	47	36	51	45

Upotrijebite U-test velikog uzorka na razini značajnosti 0.01 da vidimo je li uspješniji Poziv 1 nego Poziv 2.

Rješenje:
Pripadni kod u programskom jeziku *R* je:

$$binU < -rbinom(100{,}100{,}0.5)$$
$$binU[1:10]$$
$$binU2 < -rbinom(100{,}100{,}0.5)$$
$$binU2[1:10]$$

gdje su naredbe i funkcije opisane u zadatku 1.2.3.

1. Testiramo hipoteze:
$$\begin{aligned} H_0 &: \text{ populacije su jednake} \\ H_1 &: \quad \mu_1 > \mu_2 \end{aligned}$$

2. Na nivou značajnosti α.

3. Kriterij: odbacujemo nulhipotezu ako je $z \geq 2.33$

4. Račun: Raspoređivanje podataka prema veličini koju dobijemo

rang	1	2	3	4	5	6	7	8	9	10
broj	36	42	43	44	45	45	46	47	48	49
Uzorak	2.	1.	1.	1.	1.	2.	1.	2.	1.	2.
rang	11	12	13	14	15	16	17	18	19	20
vrijednost	49	50	51	51	52	53	53	54	57	58
uzorak	2.	1.	2.	2.	2.	1.	2.	2.	1.	1.

Primijetite da imamo iste brojeve na različitima pozicijama. To su 45 (pozicija: 5 i 6), 49 (pozicija: 10 i 11), 51 (pozicija: 13 i 14), te 53 (pozicija: 16 i 17) pa im dodjeljujemo rangove:

$$\tfrac{5+6}{2} = 5.5 \quad \tfrac{10+11}{2} = 10.5 \quad \tfrac{13+14}{2} = 13.5 \quad \tfrac{16+17}{2} = 16.5$$

Zbroj rang pozicija koje zauzimaju brojevi prvog uzorka:

$$W_1 = 2 + 3 + 4 + 5.5 + 7 + 9 + 12 + 16.5 + 19 + 20 = 98$$

i

$$U_1 = 98 - \frac{10 \cdot 11}{2} = 43$$

Stoga:

$$\mu_{U_1} = \frac{10 \cdot 10}{2} = 50 \text{ i } \sigma_{U_1} = \sqrt{\frac{10 \cdot 10 \cdot 21}{12}} \approx 13.23$$

slijedi:

$$z = \frac{43 - 50}{13.23} \approx -0.53$$

5. Odluka: Kako je $z = -0.53$, što je manje od 2.33, ne odbacujemo nulhipotezu.

1.6 H test [3]

H test je zbrojni-rang test koji služi za testiranje nulhipoteze da k neovisnih slučajnih uzoraka dolazi iz identične populacije u odnosu na alternativnu hipotezu da sredine ovih populacija nisu sve jednake. H test zamjenjuje standardni test jednofaktorske analize varijance i on ne mora imati pretpostavku normalnosti (koja je u standardnom testu pretpostavka, da slučajni uzorak uzimamo iz normalno distribuirane populacije). Kao u U testu, podaci su poredani zajedno od najmanjeg do najvećeg kao da čine jedan uzorak.

Ako je R_1 zbroj rangova dodijeljenih n_i vrijednostima i-tog uzorka i $n = n_1 + n_2 + ... + n_k$, H test temelji se na statistici

$$\boxed{H = \frac{12}{n \cdot (n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)} \tag{1.5}$$

Ako je nulhipoteza točna i svaki uzorak ima najmanje pet opažanja, raspodjela uzorka H ima približno hi-kvadratnu raspodjelu sa $k - 1$ stupnjom slobode.

Posljedično tome, odbacujemo nulhipotezu da su populacije uzoraka identične i prihvaćamo alternativnu hipotezu da sredine za ove populacije nisu sve jednake, ako je vrijednost koju dobijemo za H veća ili jednaka X_α^2 za $k - 1$ stupanj slobode.

Primjer 1.6.1. *Slijede konačni rezultati na uzorku učenika koji uče njemački pomoću tri različite metode:*

Metoda 1:	94,	87,	91,	74,	87,	97	
Metoda 2:	85,	82,	79,	84,	61,	72,	80
Metoda 3:	89,	67,	72,	76,	69		

Upotrijebite H test sa nivoom značajnosti 0.05 za testiranje nulhipoteze da su tri uzorkovane populacije identične protiv alternativne hipoteze da njihove sredine nisu jednake.[3]

Rješenje:

1. Testiramo hipoteze:

$$\begin{aligned} H_0 &: \quad \text{populacije su jednake} \\ H_1 &: \quad \text{sredine populacija nisu jednake} \end{aligned}$$

2. Na nivou značajnosti: $\alpha = 0.05$

3. Kriterij: Odbacujemo nulhipotezu ako je $H \geq 5.991$. Gdje 5.991 je vrijednost od $X^2_{0.05}$ za $3 - 1 = 2$ stupnja slobode iščitano iz tablice 2.3.

4. Račun: Raspoređivanje podataka prema veličini koju dobijemo:

vrijednosti	61	67	69	72	72	74	76	79	80	82	84	85	87	87	89	91	94	97
metoda	2	3	3	2	3	1	3	2	2	2	2	2	1	1	3	1	1	1

Dodjeljivanje podataka ovim redoslijedom 1, 2, 3, ..., 18 pronalazimo da vrijednosti prvog uzorka zauzimaju sljedeće pozicije:
$6, 13, 14, 16, 17, 18$ dok drugi uzorak:$1, 4.5, 8, 9, 10, 11, 12$, a treći: $2, 3, 4.5, 7, 15$.
(jer 87 pripada istom uzorku pa im dodijelimo rangove 13 i 14.)

Stoga:

$$R_1 = 6 + 13 + 14 + 16 + 17 + 18 = 84$$

$$R_2 = 1 + 4.5 + 8 + 9 + 10 + 11 + 12 = 55.5$$

$$R_3 = 2 + 3 + 4.5 + 7 + 15 = 31.5$$

i iz 1.5 slijedi

$$H = \frac{12}{18 \cdot 19}\left(\frac{84^2}{6} + \frac{55.5^2}{7} + \frac{31.5^2}{5}\right) - 3 \cdot 19 \approx 6.67$$

5. Odluka:
Kako je $H = 6.67$ što je veće od 5.99, odbacujemo nulhipotezu.
Drugim riječima, zaključujemo da tri metode poučavanja njemačkog jezika nisu jednako učinkovite.

Zadatak 1.6.2. *Koristeći se programski jezikom R za simulaciju uzorka eksponencijalne razdiobe dobivamo sljedeće podatke za $\lambda = 5$:*

Uzorak 1:	0.10	0.59	0.17						
Uzorak 2:	0.05	0.35	0.05	0.02	0.29	0.07	0.20	0.03	0.18
Uzorak 3:	0.15	0.01	0.08	0.36	0.23	0.10	0.23		

Upotrijebite H-test na razini značajnosti 0.05 za testiranje nulhipoteze da su tri populacije jednake protiv alternativne hipoteze da im sredine nisu jednake.

Rješenje: Pripadni kod u programskom jeziku *R* za dobivanje podataka je:

$$\begin{array}{c} pr1 < -rexp(3,5) \\ pr1 \\ pr2 < -rexp(9,5) \\ pr2 \\ pr3 < -rexp(7,5) \\ pr3 \end{array}$$

gdje su naredbe i funkcije opisane u zadatku 1.1.4.

1. Testiramo hipoteze:

$$\begin{array}{rc} H_0: & \text{populacije su jednake} \\ H_1: & \text{sredine populacija nisu jednake} \end{array}$$

2. Nivo značajnosti: $\alpha = 0.05$

3. Kriterij: Odbacujemo nulhipotezu ako je $H \geq 5.991$. Vrijednost $X^2_{0.05}$ za $3 - 1 = 2$ stupnja slobode.

4. Račun: Raspoređivanje podataka prema veličini koju dobijemo:

rang	1	2	3	4	5	6	7	8	9	
vrijednosti	0.01	0.02	0.03	0.05	0.05	0.07	0.08	0.1	0.1	
uzorak	3	2	2	2	2	2	3	1	3	
rang	10	11	12	13	14	15	16	17	18	19
vrijednosti	0.15	0.17	0.18	0.2	0.23	0.23	0.29	0.35	0.36	0.59
uzorak	3	1	2	2	3	3	2	2	3	1

Dodjeljivanje podataka ovim redoslijedom 1, 2, 3, ..., 19 pronalazimo da vrijednosti prvog uzorka zauzimaju sljedeće pozicije: 8.5, 11, 19

dok drugi uzorak: $2, 3, 4, 5, 6, 12, 13, 16, 17$
i treći: $1, 7, 8.5, 10, 14, 15, 18$.
S obzirom da 0.1 pripada različitim uzorcima (uzorak 1 i 3) pa im dodjeljujemo rangove $\frac{8+9}{2} = 8.5$. Dok 0.05 i 0.23 pripadaju istom uzorku pa im rang prepišemo.

Stoga:

$$R_1 = 8.5 + 11 + 19 = 38.5$$
$$R_2 = 2 + 3 + 4 + 5 + 6 + 12 + 13 + 16 + 17 = 78$$
$$R_3 = 1 + 7 + 8.5 + 10 + 14 + 15 + 18 = 73.5$$

i iz 1.5 slijedi

$$H = \frac{12}{19 \cdot 20}\left(\frac{38.5^2}{3} + \frac{78^2}{9} + \frac{73.5^2}{7}\right) - 3 \cdot 20 \approx 1.321$$

5. odluka:
Kako je $H = 1.321$, što je manje od 5.991, ne odbacujemo nulhipotezu.
Drugim riječima, zaključujemo da su populacije jednake.

1.7 Test homogenosti niza (Runs test) [3]

Sve metode zaključivanja o kojima smo raspravljali temelje se na pretpostavci da su uzorci slučajni, ali postoji mnogo primjena u kojima je teško odlučiti je li ta pretpostavka opravdana. To je istina, osobito, kad imamo malo kontrole ili nemamo kontrolu nad odabirom podataka, kao što je to slučaj, na primjer, kada se oslanjamo na sve dostupne zapise za pravljenje dugoročnih predviđanja vremena, kada koristimo dostupne podatke za procjenu stope smrtnosti od neke bolesti ili kada koristimo zapise o prodaji za posljednjih nekoliko mjeseci da bismo predvidjeli prodaju robnih kuća. Niti jedan od ovih podataka ne predstavlja slučajni uzorak u strogom smislu.

Postoji nekoliko metoda ocjenjivanja nasumičnosti uzorka na temelju redoslijeda kojim su observacije opažene (prikupljene). One nam omogućavaju da odlučimo, nakon što smo prikupili podatke da li uzorci koji izgledaju sumnjivo nenasumično mogu biti pripisani slučajnosti. Dakle, ako imamo uzorak koji izgleda nenasumično, a testovi pokažu da uzorak je nasumičan, tada se sumnja na nenasumičnost pripisuje slučajnosti. Tehnika koju ćemo opisati ovdje i u sljedeća dva poglavlja temelji se na teoriji homogenosti niza (runs).

Niz (runs) je niz identičnih slova (ili drugih vrsta simbola) koja slijede i kojima prethode različita ili nikakva slova.
Za ilustraciju razmotrit ćemo sljedeće uređenje:

H- zdrav i D-bolestan, grabovih stabala koji su posađeni prije dugo godina duž ceste:

$$\underline{HHHH}\ \underline{DDD}\ \underline{HHHHHHH}\ \underline{DD}\ \underline{HH}\ \underline{DDDD}$$

koristeći podcrtavanje za kombiniranje slova koja čine nizove utvrdili smo da postoji prvi niz od 4 slova H, potom niz od 3 slova D, onda niz od 7 slova H, zatim niz od 2 slova D, niz od 2 slova H te na kraju niz od 4 slova D. Ukupni broj nizova koji se pojavljuju u rasporedu ove vrste često je dobar pokazatelj mogućeg nedostatka slučajnosti. Ako ima premalo nizova, možda ćemo posumnjati u definirano grupiranje ili klasteriranje. Ako ima previše nizova mogli bismo posumnjati u neku vrstu ponovljenog izmjeničnog ili cikličkog obrasca. U prethodnom primjeru čini se da postoji određeno skupljanje - čini se da oboljela stabla dolaze u skupinama, ali ostaje da se vidi je li to značajno ili se može pripisati slučajnosti.
Ako je n_1 slova jedne vrste i n_2 slova druge vrste i u nizova ovakvu odluku temeljimo na sljedećem kriteriju:
Odbacujemo nulhipotezu slučajnosti ako $u \leq u'_{\frac{\alpha}{2}}$ *ili* $u \geq u_{\frac{\alpha}{2}}$ *gdje su* $u'_{\frac{\alpha}{2}}$ *i* $u_{\frac{\alpha}{2}}$ *dani u tablicama 2.6 i 2.7 za* $n_1 \leq 15, n_2 \leq 15, \alpha = 0.05$ *ili* $\alpha = 0.01$.
U konstrukciji tablica 2.6 i 2.7, $u'_{\frac{\alpha}{2}}$ je najveća vrijednost od u za koju je vjerojatnost da je $u \leq u'_{\frac{\alpha}{2}}$ manja ili jednaka $\frac{\alpha}{2}$, dok $u_{\frac{\alpha}{2}}$ je najmanja vrijednost od u za koju je vjerojatnost da je $u \geq u_{\frac{\alpha}{2}}$ je manja ili jednaka $\frac{\alpha}{2}$, a prazni razmaci označavaju da se nulhipoteza slučajnosti ne može odbaciti za vrijednosti u tom slijedu raspodjele uzorkovanja u, bez obzira na vrijednost koju dobivamo za u. Kada su n_1 i n_2 oboje barem 10, općenito smatrat ćemo razumljivo koristiti test velikog uzorka koji će biti opisan u idućem odjeljku.

Primjer 1.7.1. *Koristeći zdrava i bolesna drveća koja smo spomenuli, upotrijebite u test sa nivoom značajnosti od 0.05 testirajte nulhipotezu o slučajnosti protiv alternativne hipoteze da raspored nije slučajan. [3]*

Rješenje:

1. Testiramo hipoteze:
$$\begin{aligned} H_0 &: \text{ raspored je slučajan} \\ H_1 &: \text{ raspored nije slučajan} \end{aligned}$$

2. Na nivou značajnosti α.

3. Kriterij: Kako je $n_1 = 13$, $n_2 = 9$ i $\alpha = 0.05$ imamo iz tablice2.6 da je $u'_{0.025} = 6$ i $u_{0.025} = 17$ prema tome, odbacujemo nulhipotezu ako je $u \leq 6$ ili $u \geq 17$.

4. Račun: $u = 6$ što je vidljivo iz podataka.

5. Odluka: Kako je $u = 6$ što je manje ili jednako 6, odbacujemo nulhipotezu. Drugim riječima, zaključujemo da redoslijed zdravih i bolesnih stabala nije slučajan. Doista, čini se da bolesna drva dolaze u klasterima.

1.8 Test homogenosti niza (Runs test) (veliki uzorci) [3]

Pod nulhipotezu da je n_1 broj slova jedne vrste i n_2 broj slova druge vrste koja su slučajno raspoređena, može se pokazati da su srednja vrijednost i standardno odstupanje od u(ukupni broj nizova (runs)):

$$\boxed{\mu_u = \frac{2n_1n_2}{n_1+n_2} + 1 \text{ i } \sigma_u = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}} \tag{1.6}$$

ako su n_1 i n_2 oba veća ili jedaka 10, test statistika u može se aproksimirati normalnom razdiobom.
Stoga test nulhipoteze slučajnosti temeljimo na statistici:

$$\boxed{z = \frac{u - \mu_u}{\sigma_u}} \tag{1.7}$$

koji ima standardnu normalnu raspodjelu.
Ako je alternativna hipoteza da raspored nije slučajan,
odbacujemo nulhipotezu za $z \leq -z_{\frac{\alpha}{2}}$ ili $z \geq z_{\frac{\alpha}{2}}$.
Ako je alternativna hipoteza da postoji klasteriranje ili trend(tendencija),
odbacujemo nulhipotezu za $z \leq -z_\alpha$
i ako je alternativna hipoteza da postoji alternativni ili ciklički obrazac,
odbacujemo nulhipotezu za $z \geq z_\alpha$.

Primjer 1.8.1. *Slijedi raspored muškaraca, M i žena, W koji su u liniji za kupnju karata za rock koncert:*

$$\begin{matrix} M & W & M & W & M & M & M & W & M & W & M & M & M & W & W & M \\ M & M & M & W & W & M & W & M & M & M & W & M & M & M & W & W \\ W & M & W & M & M & M & W & M & W & M & M & M & M & W & W & M \end{matrix}$$

Testirajte na nivou značajanosti od 0.05 da je raspored osoba slučajan koristeći se testom homogenosti niza. [3]

Rješenje:

1. Testiramo hipoteze:
$$H_0: \quad \text{raspored je slučajan}$$
$$H_1: \quad \text{raspored nije slučajan}$$

2. Na nivou značajnosti: α.

3. Kriterij: odbacujemo nulhipotezu ako je $z \leq -1.96$ ili $z \geq 1.96$.

4. Račun:
Kako je $n_1 = 30, n_2 = 18, u = 27$ te uvrštavanjem u formulu 1.6 slijedi:
$$\mu_u = \frac{2 \cdot 30 \cdot 18}{30 + 18} + 1 = 23.5$$
$$\sigma_u = \sqrt{\frac{2 \cdot 30 \cdot 18(2 \cdot 30 \cdot 18 - 30 - 18)}{(30 + 18)^2(30 + 18 - 1)}} \approx 3.21$$
i zbog toga
$$z = \frac{27 - 23.5}{3.21} \approx 1.09$$

5. Odluka: Kako je $z = 1.09$ što je između -1.96 i 1.96, ne odbacujemo nulhipotezu. Drugim riječima, nema stvarnih dokaza o nedostatku slučajnosti.

Zadatak 1.8.2. *Koristeći se programom R i simulacijom uzorka binomne distribucije gdje promatramo 100 opažanja sa jednim pokušajem i vjerojatnosti 0.6 te za poduzorak duljine 45 dobivamo sljedeće rezultate:*

1	1	1	1	0	0	0	1	0	1	1	1	0	0	1
1	1	0	0	0	1	0	1	0	1	1	1	1	1	1
1	0	1	0	1	1	1	1	0	1	0	1	1	1	1

Testirajte na nivou značajnosti od 0.01 da je raspored slučajan.

Rješenje: Pripradni kod u programskom jeziku *R* je:

```
binom<-rbinom(100,1,0.6)
binom[1:45]
```

gdje su funkcije i naredbe opisane u zadatku 1.2.3.

1. Testiramo hipoteze:
$$H_0: \quad \text{raspored je slučajan}$$
$$H_1: \quad \text{raspored nije slučajan}$$

2. Na nivou značajnosti: α.

3. Kriterij: odbacujemo nulhipotezu ako je $z \leq -2.578$ ili $z \geq 2.578$

4. Račun:
   Kako je $n_1 = 30, n_2 = 15, u = 21$ te uvrštavanjem u formulu 1.6 slijedi:

$$\mu_u = \frac{2 \cdot 30 \cdot 15}{30 + 15} + 1 = 21$$

$$\sigma_u = \sqrt{\frac{2 \cdot 30 \cdot 15(2 \cdot 30 \cdot 15 - 30 - 15)}{(30 + 15)^2(30 + 15 - 1)}} \approx 2.94$$

   i zbog toga

$$z = \frac{21 - 21}{2.94} \approx 0$$

5. Odluka: Kako je $z = 0$ između -2.578 i 2.578, ne odbacujemo nulhipotezu.
   Drugim riječima, nema stvarnih dokaza o nedostatku slučajnosti.

1.9 Test homogenosti niza (runs) iznad i ispod medijana [3]

Bilo koji uzorak koji se sastoji od numeričkih mjerenja ili opažanja može se tretirati slično korištenjem slova a i b koja predstavljaju vrijednosti koje padaju iznad i ispod medijana uzorka. Brojevi jednaki medijanu su izostavljeni. Rezultirajući niz a-ova i b-ova koji predstavljaju podatke u izvornom redoslijedu može se testirati za slučajnosti na temelju ukupnog broja niza slova a i slova b, tj. ukupnog broja niza ispod i iznad medijana. Ovisno o veličini n_1 i n_2 koristimo tablice 2.6 i 2.7 ili test velikog uzorka odjeljka 1.8.

Primjer 1.9.1. *Na uzastopna putovanja između dva grada, bus je prevozio*

24	19	32	28	21	23	26	17	20	28	30	24
13	35	26	21	19	29	27	18	26	14	21	23

putnika. Koristeći ukupni broj nizova iznad i ispod medijana i razine značajnosti $\alpha = 0.01$ *odlučite je li razumno tretirati podatke kao da čine slučajni uzorak. [3]*

Rješenje: Kako je medijan 23.5 imamo sljedeći redoslijed vrijednosti ispod i iznad medijana:

$$a \quad b \quad a \quad a \quad b \quad b \quad a \quad b \quad b \quad a \quad a \quad a \quad b \quad a \quad a \quad b \quad b \quad a \quad a \quad b \quad a \quad b \quad b \quad b$$

1. Testiramo hipoteze:
$$\begin{aligned} H_0 &: \text{ raspored je slučajan} \\ H_1 &: \text{ raspored nije slučajan} \end{aligned}$$

2. Na nivou značajnosti: α.

3. Kriterij: Kako je $n_1 = 12, n_2 = 12, \alpha = 0.01$ imamo da je $u'_{0.005} = 6$ i $u_{0.005} = 20$ iz tablice 2.7. Stoga nulhipoteza se odbacuje ako je $u \leq 6$ ili $u \geq 20$.

4. Račun: Kako se može vidjeti iz uređenja slova a i b, $u = 14$ nizova.

5. Odluka: Kako je $u = 14$ što je između 6 i 20 ne odbacujemo nulhipotezu.
   Drugim rječima, ne postoje stvarni dokazi koji bi sugerirali da se s podacima ne može postupati kao da čine slučajni uzorak.

1.10 Korelacija ranga

[2]
Korelacija je mjera linearne zavisnosti dviju serija podataka $x_1, x_2, ...x_n$ i $y_1, y_2, .., y_n$.
Drugim riječima, ako su točke $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ grupirane oko regresijskog pravca, regresijski pravac je pravac s jednadžbom $y = ax + b$, gdje koeficijente a i b računamo po formuli:

$$a = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} \qquad b = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - (\sum x_i^2)}$$

gdje su $\sum_{i=1}^n x_i^2, \sum_{i=1}^n x_i, n, \sum_{i=1}^n y_i$ koeficijenti sustava koji su poznati jer se dobiju iz mjernih podataka,

onda govorimo da su podatci **korelirani (linearno korelirani)**. Na osnovu toga govori se da su pripadne veličine x, y korelirane.

Razina koreliranosti mjeri se **koeficijentom korelacije**

$$r := \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2} \cdot \sqrt{n\sum y_i^2 - (\sum y_i)^2}} \tag{1.8}$$

[3]

U *linearno regresijskoj analizi* pretpostavljamo da su x-evi poznate konstante i da za svaki x slučajna varijabla y ima određenu distribuciju (kao što je prikazano na slici 1.1)

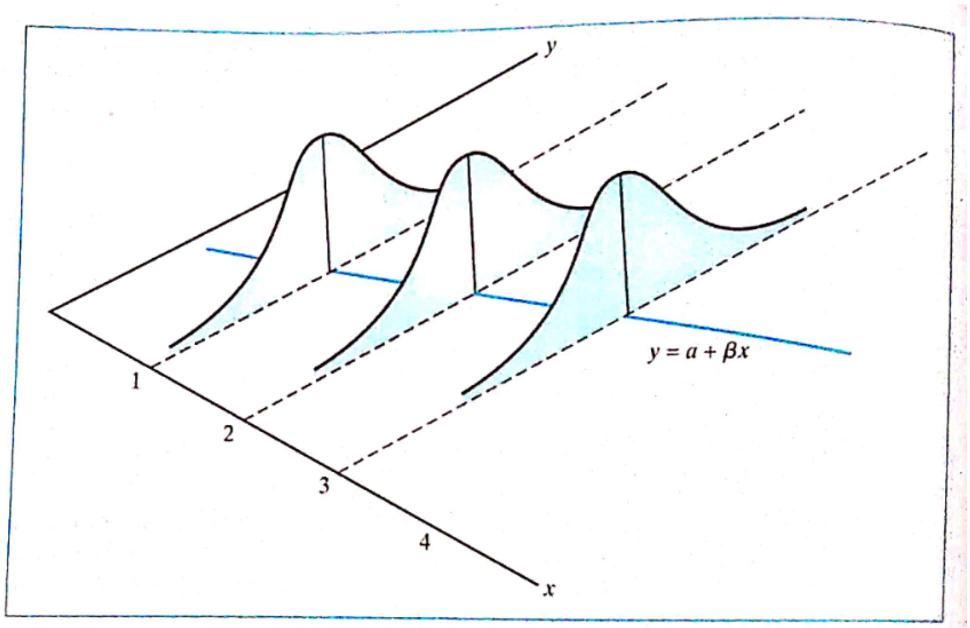
sa aritmetičkom sredinom $\alpha + \beta x$.
U *normalnoj regresijskoj analizi* pretpostavljamo da su ove distribucije normalne s istom standardnom devijacijom σ. Drugim riječima, distribucije prikazane na slici 1.1 su normalne krivulje s aritmetičkim sredinama na krivulji $\alpha + \beta x$ i istom standardnom devijacijom σ.

Kako bismo testirali nulhipotezu nekoreliranosti koristit ćemo iste pretpostavke kao gore.

Kad su ove pretpostavke zadovoljene u razumnom stupnju aproksimacije odbacujemo nulhipotezu nekoreliranosti ako

$$r \leq -r_{\frac{\alpha}{2}} \text{ ili } r \geq r_{\frac{\alpha}{2}}$$

gdje vrijednost $r_{\frac{\alpha}{2}}$ možemo pronaći u tablici 2.4 za $\alpha = 0.05$ i $\alpha = 0.01$.
Ako se nulhipoteza može odbaciti, kažemo da postoji značajna korelacija, u suprotonom kažemo da vrijednost od r nije statistički značajna.



Slika 1.1: Distribucija od y za dane vrijednosti od x

[3]

Budući da se test značajnosti za r temelji na vrlo strogim pretpostavkama, ponekad koristimo neparametarsku alternativu koja se može primijeniti pod mnogo općenitijim uvjetima.

Ovaj test nulhipoteze nekoreliranosti je baziran na koeficijentu rang-korelacije često zvanom **Spearmanov koeficijent korelacije ranga** ili koeficijent rang-korelacije kojeg označujemo s r_s.
Izračunavanje koeficijenta korelacije rangiranja za dani skup od n parova x-ova i y-ona zahtijeva nekoliko koraka. Prvo rangiramo x-eve između sebe od najmanjeg do najvećeg. Potom y na isti način. Nakon toga pronađemo razliku d između rangova i uvrstimo u formulu:

koeficijent korelacije ranga:

$$\boxed{r_s = 1 - \frac{6(\sum d^2)}{n(n^2-1)}} \tag{1.9}$$

Kada postoje rangovi istih vrijednosti u promatranju, postupamo kao i prije, svakom dodijelimo sredinu rangova koje zajedno zauzimaju.

Primjer 1.10.1. *Slijede brojevi sati učenja za ispit 10 studenata i njihove ocjene:*

broj sati učenja	*ocjena na ispitu*
x	y
9	56
5	44
11	79
13	72
10	70
5	54
18	94
15	85
2	33
8	65

Izračunajte r_s. [3]

Rješenje: Rangirajući x-ove među sobom od najmanjih do najvećih, a također i y, dobivamo poredak prikazan u prva dva stupca sljedeće tablice:

Poredak x-eva	poredak y-a	d	d^2
5	4	1.0	1.00
2.5	2	0.5	0.25
7	8	-1.0	1.00
8	7	1.0	1.00
6	6	0.0	0.00
2.5	3	-0.5	0.25
10	10	0.0	0.00
9	9	0.0	0.00
1	1	0.0	0.00
4	5	-1.0	1.00
		Ukupno	4.50

Određivanje d i njegovog kvadrata i uvrštavanjem u formulu za r_s (1.9) imamo:

$$r_s = 1 - \frac{6(4.50)}{10(10^2 - 1)} \approx 0.97$$

Možemo vidjeti da se r_s lako izračuna ručno pa se ponekad koristi umjesto r (1.8) kad nije dostupan kalkulator. Kada ne postoje veze, r_s je zapravo koeficijent korelacije r izračunat za dva skupa rangova,a kad postoje veze, može postojati mala razlika. Koristeći rangove umjesto originalnih podataka možemo izgubiti neke informacije, ali to se obično kompenzira računskom lakoćom izračunavanja koeficijenta korelacije. Kada r_s koristimo za testiranje nulhipoteze da nema povezanosti dviju varijabli x i y, ne moramo donositi nikakve pretpostavke o prirodi testiranih populacija. Po nulhipotezi nekoreliranosti, zaista nulhipoteza da su x i y nasumično spareni, uzoračka distribucija od r_s ima srednju vrijednost 0 i standardnu devijaciju:

$$\sigma_{r_s} = \frac{1}{\sqrt{n-1}}$$

Budući da se raspodjela uzoraka može aproksimirati s normalnom raspodjelom čak i za relativno male vrijednosti n, test nulhipoteze o nepostojanju korelacije temeljimo na statistici:

$$\boxed{z = \frac{r_S - 0}{\frac{1}{\sqrt{n-1}}} = r_s\sqrt{n-1}} \tag{1.10}$$

Primjer 1.10.2. *S obzirom na prethodni primjer gdje smo imali $n = 10$, a $r_s = 0.97$, testirajte postojanje korelacije između podataka na razini značajnosti 0.01 [3]*

Rješenje:

1. Testiramo hipoteze:
$$H_0 : \quad \rho = 0$$
$$H_1 : \quad \rho \neq 0$$

2. Na nivou značajnosti: α.

3. Kriterij: odbacujemo nulhipotezu ako je $z \leq -2.575$ ili $z \geq 2.575$ gdje je
$$z = r_s \sqrt{n-1}$$

4. Račun: Za $n = 10$ i $r_s = 0.97$ imamo: $z = 0.97\sqrt{10-1} \approx 2.91$

5. Odluka: Kako je $z = 2.91$ veće od 2.575 odbacujemo nulhipotezu.
   Drugim riječima, zaključujemo da postoji veza između vremena ispitivanja i ocjena u populaciji uzorka.

Poglavlje 2

Tablice

Binomne vjerojatnosti												
n	**x**	**p**										
		0.05	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**0.95**
7	0	0.698	0.478	0.210	0.082	0.028	0.008	0.002				
	1	0.257	0.372	0.367	0.247	0.131	0.055	0.017	0.004			
	2	0.041	0.124	0.275	0.318	0.261	0.164	0.077	0.025	0.004		
	3	0.004	0.023	0.115	0.227	0.290	0.273	0.194	0.097	0.029	0.003	
	4		0.003	0.029	0.097	0.194	0.273	0.290	0.227	0.115	0.023	0.004
	5			0.004	0.025	0.077	0.164	0.261	0.318	0.275	0.124	0.041
	6				0.004	0.017	0.055	0.131	0.247	0.367	0.372	0.257
	7					0.002	0.008	0.028	0.082	0.210	0.478	0.698
14	0	0.488	0.229	0.044	0.007	0.001						
	1	0.359	0.356	0.154	0.041	0.007	0.001					
	2	0.123	0.257	0.250	0.113	0.032	0.006	0.001				
	3	0.026	0.114	0.250	0.194	0.085	0.022	0.003				
	4	0.004	0.035	0.172	0.229	0.155	0.061	0.014	0.001			
	5		0.008	0.086	0.196	0.207	0.122	0.041	0.007			
	6		0.001	0.032	0.126	0.207	0.183	0.092	0.023	0.002		
	7			0.009	0.062	0.157	0.209	0.157	0.062	0.009		
	8			0.002	0.023	0.092	0.183	0.207	0.126	0.032	0.001	
	9				0.007	0.041	0.122	0.207	0.196	0.086	0.008	
	10				0.001	0.014	0.061	0.155	0.229	0.172	0.035	0.004
	11					0.003	0.022	0.085	0.194	0.250	0.114	0.026
	12					0.001	0.006	0.032	0.113	0.250	0.257	0.123
	13						0.001	0.007	0.041	0.154	0.356	0.359
	14							0.001	0.007	0.044	0.229	0.488
15	0	0.463	0.206	0.035	0.005							
	1	0.366	0.343	0.132	0.031	0.005						
	2	0.135	0.267	0.231	0.092	0.022	0.003					
	3	0.031	0.129	0.250	0.170	0.063	0.014	0.002				
	4	0.005	0.043	0.188	0.219	0.127	0.042	0.007	0.001			
	5	0.001	0.010	0.103	0.206	0.186	0.092	0.024	0.003			
	6		0.002	0.043	0.147	0.207	0.153	0.061	0.012	0.001		
	7			0.014	0.081	0.177	0.196	0.118	0.035	0.003		
	8			0.003	0.035	0.118	0.196	0.177	0.081	0.014		
	9			0.001	0.012	0.061	0.153	0.207	0.147	0.043	0.002	
	10				0.003	0.024	0.092	0.186	0.206	0.103	0.010	0.001
	11				0.001	0.007	0.042	0.127	0.219	0.188	0.043	0.005
	12					0.002	0.014	0.063	0.170	0.250	0.129	0.031
	13						0.003	0.022	0.092	0.231	0.267	0.135
	14							0.005	0.031	0.132	0.343	0.366
	15								0.005	0.035	0.206	0.463

Tablica 2.1: Binomne vjerojatnosti

[3]

Područja normalne krivulje										
z	**.00**	**.01**	**.02**	**.03**	**.04**	**.05**	**.06**	**.07**	**.08**	**.09**
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4967	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Tablica 2.2: Područja normalne krivulje

[3]

Vrijednosti od χ^2			
s.s.	$\chi^2_{0.05}$	$\chi^2_{0.01}$	**s.s.**
1	3.841	6.635	1
2	5.991	9.210	2
3	7.815	11.345	3
4	9.488	13.277	4
5	11.070	15.086	5
6	12.592	16.812	6
7	14.067	18.475	7
8	15.507	20.090	8
9	16.919	21.666	9
10	18.307	23.209	10
11	19.675	24.725	11
12	21.026	26.217	12
13	22.362	27.688	13
14	23.685	29.141	14
15	24.996	30.578	15
16	26.296	32.000	16
17	27.587	33.409	17
18	28.869	34.805	18
19	30.144	36.191	19
20	31.410	37.566	20
21	32.671	38.932	21
22	33.924	40.289	22
23	35.172	41.638	23
24	36.415	42.980	24
25	37.652	44.314	25
26	38.885	45.642	26
27	40.113	46.963	27
28	41.337	48.278	28
29	42.557	49.588	29
30	43.773	50.892	30

Tablica 2.3: Vrijednosti od χ^2

[3]

Kritične vrijednosti od *r*					
n	$r_{0.025}$	$r_{0.005}$	**n**	$r_{0.025}$	$r_{0.005}$
3	0.997		18	0.468	0.590
4	0.950	0.999	19	0.456	0.575
5	0.878	0.959	20	0.444	0.561
6	0.811	0.917	21	0.433	0.549
7	0.754	0.875	22	0.423	0.537
8	0.707	0.834	27	0.381	0.487
9	0.666	0.798	32	0.349	0.449
10	0.632	0.765	37	0.325	0.418
11	0.602	0.735	42	0.304	0.393
12	0.576	0.708	47	0.288	0.372
13	0.553	0.684	52	0.273	0.354
14	0.532	0.661	62	0.250	0.325
15	0.514	0.641	72	0.232	0.302
16	0.497	0.623	82	0.217	0.283
17	0.482	0.606	92	0.205	0.267

Tablica 2.4: Kritične vrijednosti od r

[3]

Kritične vrijednosti od *U*														
n_2	**Vrijednosti od $U'_{0.05}$**													
n_1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
2							0	0	0	0	1	1	1	1
3				0	1	1	2	2	3	3	4	4	5	5
4			0	1	2	3	4	4	5	6	7	8	9	10
5		0	1	2	3	5	6	7	8	9	11	12	13	14
6		1	2	3	5	6	8	10	11	13	14	16	17	19
7		1	3	5	6	8	10	12	14	16	18	20	22	24
8	0	2	4	6	8	10	13	15	17	19	22	24	26	29
9	0	2	4	7	10	12	15	17	20	23	26	28	31	34
10	0	3	5	8	11	14	17	20	23	26	29	30	36	39
11	0	3	6	9	13	16	19	23	26	30	33	37	40	44
12	1	4	7	11	14	18	22	26	29	33	37	41	45	49
13	1	4	8	12	16	20	24	28	30	37	41	45	50	54
14	1	5	9	13	17	22	26	31	36	40	45	50	55	59
15	1	5	10	14	19	24	29	34	39	44	49	54	59	64
n_2	**Vrijednosti od $U'_{0.01}$**													
n_1		**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
3								0	0	0	1	1	1	2
4					0	0	1	1	2	2	3	3	4	5
5				0	1	1	2	3	4	5	6	7	7	8
6			0	1	2	3	4	5	6	7	9	10	11	12
7			0	1	3	4	6	7	9	10	12	13	15	16
8			1	2	4	6	7	9	11	13	15	17	18	20
9		0	1	3	5	7	9	11	13	16	18	20	22	24
10		0	2	4	6	9	11	13	16	18	21	24	26	29
11		0	2	5	7	10	13	16	18	21	24	27	30	33
12		1	3	6	9	12	15	18	21	24	27	31	34	37
13		1	3	7	10	13	17	20	24	27	31	34	38	42
14		1	4	7	11	15	18	22	26	30	34	38	42	46
15		2	5	8	12	16	20	24	29	33	37	42	46	51

Tablica 2.5: Kritične vrijednosti od *U*

[3]

Kritične vrijednosti od u

n_2 / n_1	Vrijednosti od $u_{0.025}$												
		4	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
4			9	9									
5		9	10	10	11	11							
6		9	10	11	12	12	13	13	13	13			
7			11	12	13	13	14	14	14	14	15	15	15
8			11	12	13	14	14	15	15	16	16	16	16
9				13	14	14	15	16	16	16	17	17	18
10				13	14	15	16	16	17	17	18	18	18
11				13	14	15	16	17	17	18	19	19	19
12				13	14	16	16	17	18	19	19	20	20
13					15	16	17	18	19	19	20	20	21
14					15	16	17	18	19	20	20	21	22
15					15	16	18	18	19	20	21	22	22

n_2 / n_1	Vrijednosti od $u'_{0.025}$													
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
2											2	2	2	2
3					2	2	2	2	2	2	2	2	2	3
4				2	2	2	3	3	3	3	3	3	3	3
5			2	2	3	3	3	3	3	4	4	4	4	4
6		2	2	3	3	3	3	4	4	4	4	5	5	5
7		2	2	3	3	3	4	4	5	5	5	5	5	6
8		2	3	3	3	4	4	5	5	5	6	6	6	6
9		2	3	3	4	4	5	5	5	6	6	6	7	7
10		2	3	3	4	5	5	5	6	6	7	7	7	7
11		2	3	4	4	5	5	6	6	7	7	7	8	8
12	2	2	3	4	4	5	6	6	7	7	7	8	8	8
13	2	2	3	4	5	5	6	6	7	7	8	8	9	9
14	2	2	3	4	5	5	6	7	7	8	8	9	9	9
15	2	3	3	4	5	6	6	7	7	8	8	9	9	10

Tablica 2.6: Kritične vrijednosti od u

[3]

Kritične vrijednosti od *u* (Nastavak)												
n_2 / n_1	**Vrijednosti od $u_{0.005}$**											
		5	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
5			11									
6		11	12	13	13							
7			13	13	14	15	15	15				
8			13	14	15	15	16	16	17	17	17	
9				15	15	16	17	17	18	18	18	19
10				15	16	17	17	18	19	19	19	20
11				15	16	17	18	19	19	20	20	21
12					17	18	19	19	20	21	21	22
13					17	18	19	20	21	21	22	22
14					17	18	19	20	21	22	23	23
15						19	20	21	22	22	23	24

n_2 / n_1	**Vrijednosti od $u'_{0.005}$**											
	3	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
3										2	2	2	2
4						2	2	2	2	2	2	2	3
5				2	2	2	2	3	3	3	3	3	3
6			2	2	2	3	3	3	3	3	3	4	4
7			2	2	3	3	3	3	4	4	4	4	4
8		2	2	3	3	3	3	4	4	4	5	5	5
9		2	2	3	3	3	4	4	5	5	5	5	6
10		2	3	3	3	4	4	5	5	5	5	6	6
11		2	3	3	4	4	5	5	5	6	6	6	7
12	2	2	3	3	4	4	5	5	6	6	6	7	7
13	2	2	3	3	4	5	5	5	6	6	7	7	7
14	2	2	3	4	4	5	5	6	6	7	7	7	8
15	2	3	3	4	4	5	6	6	7	7	7	8	8

Tablica 2.7: Kritične vrijednosti od *u* (Nastavak)

[3]

Bibliografija

[1] Raeesa Ganey, *Lecture A26 - Non-parametric- Sign Test*, 2016 (pristupljeno 31.7. 2020.), https://www.youtube.com/watch?v=ZbjsXS8oKfo/.

[2] Ivica Gusić, *Metoda najmanjih kvadrata i linearna korelacija*, 2008/09 (pristupljeno 25.8.2020.), http://matematika.fkit.hr/statistika.html.

[3] Benjamin M. Perles John E. Freund, *Statistics: a first course*, Pearson Prentice Hall, 2004.

[4] Huzak M., *Vjerojatnost i matematička statistika*, 2006 (pristupljeno 27.8.2020.), http://aktuari.math.pmf.unizg.hr/docs/vms.pdf.

[5] Vondraček Z. Sandrić N., *Vjerojatnost*, 2019 (pristupljeno 4.9.2020.), https://web.math.pmf.unizg.hr/nastava/vjer/files/vjer_predavanja.pdf.

Sažetak

U ovom diplomskom radu prikazane su određene metode neparametarskih testova i simulacija podataka u programskom jeziku *R* za neke distribucije.

Neparametarski testovi koriste se kada nam ništa nije poznato o distribuciji populacije, kada imamo uzorke male duljine, kada statističke parametre nije moguće izračunati zbog prirode obilježja te kada su vrijednosti obilježja rangirane.

Također, u radu su prikazane osnovne definicije i pojmovi koji se koriste u neparametarskim testovima, a sistematičnom podjelom testova opisane su njihove mogućnosti, primjeri te provođenje postupka za zadatke čiji su podaci izgenerirani u programskom jeziku *R*.

Summary

In this thesis certain methods of non-parametric tests and data simulations in programming language *R* for some distributions are shown.
Non-parametric tests are used when we don't know anything about the distribution of a population, when samples are small, when statistical parameters can not be calculated due to the nature of characteristics and when the values of characteristics are ranked.
Basic definitions and terms which are being used in non-parametric tests are also shown in this thesis, while their possibilities, examples and implementations in tasks whose data are generated in programming language *R* are described with systematic division of the tests.

Životopis

Rođena sam 11. listopada 1993. godine u Zagrebu. Završila sam Osnovnu školu braće Radić u Kloštru Ivaniću, Srednju školu Ivan Švear u Ivanić Gradu, smjer: opća gimnazija. Zatim sam upisala Prirodoslovno-matematički fakultet u Zagrebu gdje sam završila prediplomski sveučilišni studij Matematika.

Nakon toga sam upisala Diplomski studij Matematika i informatika; smjer: nastavnički, tijekom kojeg sam počela raditi kao programer u tvrtci Vestigo. Uz to sam voditeljica instrukcija u svojoj župi te potpredsjednica Caritasa.