Sveučilište Josipa Jurja Strossmayera u Osijeku
Fakultet primijenjene matematike i informatike

Sveučilišni prijediplomski studij Matematika

# Testiranje statističkih hipoteza

Završni rad

Mentor:
**doc. dr. sc. Ivan Papić**

Student:
**Doris Jukić**

Osijek, 2024.

# Sadržaj

# 1 | Uvod

Koliko je neki lijek učinkovit? Kolika je vjerojatnost da će Biden izgubiti na idućim izborima? Možemo li reći da više golova po sezoni zabije Messi od Ronalda? Odgovor na sva ta pitanja možemo dobiti testiranjem statističkih hipoteza - ono ima vrlo važnu ulogu u današnjim istraživanjima, obradama podataka i raznim anketama.

Jedan od prvih "poznatih" primjera testiranja je *Trial of the Pyx* - tradicija Ujedinjenog Kraljevstva koja postoji i danas, no zapravo seže sve do kraja 11. stoljeća. Kako bi bili sigurni da se u kovanice stavlja točna količina zlata i srebra koja je propisana, nasumce su izabirali par kovanica i stavljali ih u metalnu kutiju zvanu *pyx*. Kasnije su tih par kovanica uspoređivali s očekivanom težinom i ukoliko se težina razlikovala, glavni majstor bio je kažnjen. Tu vidimo sve korake statističkog testiranja koje koristimo i danas : kovanice u kutiji su uzorak na kojem provodimo testiranja, nul hipoteza je da je težina tih kovanica jednaka propisanoj težini, a alternativna hipoteza je da je ta težina drugačija. Veliki matematičari kao što su J. Arbuthnot, P. S. Laplace, K. Pearson, R. Fisher, E. Pearson i J. Neyman napravili su temelje za testiranje kakvo postoji danas.

Kroz ovaj rad prvo ćemo se podsjetiti nekih bitnih definicija i pojmova. Zatim ćemo proći kroz razne testove kojima se danas koristimo u svrhu analize podataka i donošenja zaključaka. To su testovi o očekivanju, testovi o proporciji, testovi o distribuciji i test o kvantilima. To su testovi koje provodimo na temelju jednog uzorka. Isto tako, testiranje možemo provoditi na temelju dva uzorka, pri čemu oni mogu biti međusobno nezavisni i zavisni slučajni uzorci. Ponovno ćemo se susresti s testovima o očekivanju, proporciji i distribuciji, a ovdje će se pojaviti i testovi o lokacijskom pomaku i nezavisnosti. Za stvaranje primjera kroz rad koristit ćemo baze podataka s Kaggle-a, programski jezik za statističko računanje i grafiku R, verziju 4.1.2 iz 2021. godine te pakete *datasets* i *BSDA*.

# 2 | Osnovni pojmovi

Za početak potrebno se prisjetiti definicija iz vjerojatnosti i statistike radi lakšeg razumijevanja nastavka.

**Definicija 1.** *Neka je dan neprazan skup $\Omega$. Familija $\mathcal{F}$ podskupova skupa $\Omega$ je $\sigma$-algebra na $\Omega$ ako vrijedi:*

*i)* $\emptyset \in \mathcal{F}$,

*ii) ako je* $A \in \mathcal{F}$, *onda je i* $A^C \in \mathcal{F}$,

*iii) ako je dana prebrojiva familija skupova* $(A_n, n \in \mathbb{N}) \subseteq \mathcal{F}$, *onda* $\mathcal{F}$ *sadrži i njihovu uniju.*

Iz definicije $\sigma$-algebre možemo vidjeti da ona sadrži i prebrojive presjeke i razlike skupova iz $\mathcal{F}$.

**Definicija 2.** *Neka je $\Omega$ neprazan skup elementarnih događaja i $\mathcal{F}$ $\sigma$-algebra skupova na njemu. Funkciju $P : \mathcal{F} \to \mathbb{R}$ zovemo vjerojatnost na $\Omega$ ako zadovoljava:*

*A1)* $P(A) \geq 0$, *za sve* $A \in \mathcal{F}$,

*A2)* $P(\Omega) = 1$,

*A3) ako je dana prebrojiva familija međusobno disjunktnih skupova* $(A_i, i \in I) \subseteq \mathcal{F}, I \subseteq \mathbb{N}$, *tj.* $A_i \cap A_j = \emptyset$ *čim je* $i \neq j$ *tada vrijedi:*

$$P\left(\bigcup_{i \in I} A_i\right) = \sum_{i \in I} P(A_i).$$

Uređenu trojku $(\Omega, \mathcal{F}, P)$ zovemo vjerojatnosni prostor.

**Definicija 3.** *Neka je $(\Omega, \mathcal{F}, P)$ vjerojatnosni prostor. Funkcija $X : \Omega \to \mathbb{R}$ je slučajna varijabla na $\Omega$ ako je $X^{-1}(B) \in \mathcal{F}$ za sve $B \in \mathbb{R}$.*

Skup svih vrijednosti koje slučajna varijabla $X$ može poprimiti nazivamo slika slučajne varijable i taj skup označit ćemo s $\mathcal{R}(X) = \{x_i, i \in \mathbb{N}\}$. Dakle, vidimo da je funkcija slučajna varijabla ukoliko je praslika svakog skupa iz $\mathbb{R}$ u $\mathcal{F}$.

**Definicija 4.** *Neka je $X$ slučajna varijabla na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, P)$. Funkciju $F : \mathbb{R} \to [0,1]$ definiranu s $F(x) = P(\{\omega \in \Omega : X(\omega) \leq x\}) = P(\{X \leq x\})$ zovemo funkcija distribucije slučajne varijable $X$.*

Ovisno o funkciji distribucije razlikujemo diskretnu i neprekidnu slučajnu varijablu.

**Definicija 5.** *Neka je dan diskretan vjerojatnosni prostor* $(\Omega, \mathcal{F}, P)$. *Svaku funkciju* $X : \Omega \to \mathbb{R}$ *zvat ćemo diskretna slučajna varijabla. Ako označimo skup* $\mathcal{R}(X) = \{x_i, i \in I\}, I \subseteq N$ *i niz* $(p_i, i \in I)$ *gdje su* $p_i = P(\{\omega \in \Omega : X(\omega) = x_i\}) = P(\{X = x_i\})$, $i \in I$, *onda diskretnu slučajnu varijablu X možemo prikazati pomoću tablice distribucije:*

$$X = \begin{pmatrix} x_1 & x_2 & x_3 & \dots & x_n & \dots \\ p_1 & p_2 & p_3 & \dots & p_n & \dots \end{pmatrix}.$$

Upoznajmo se sada i s neprekidnom slučajnom varijablom.

**Definicija 6.** *Neka je dan vjerojatnosni prostor* $(\Omega, \mathcal{F}, P)$ *i funkcija* $X : \Omega \to \mathbb{R}$ *za koju vrijedi:*

- $\{\omega \in \Omega : X(\omega) \leq x\} = \{X \leq x\} \in \mathcal{F}, \forall x \in \mathbb{R}$,
- *postoji nenegativna realna funkcija realne varijable* $f$ *takva da:*

$$P(\{\omega \in \Omega : X(\omega) \leq x\}) = P(\{X \leq x\}) = \int_{-\infty}^{x} f(t)dt, \forall x \in \mathbb{R}.$$

*Funkciju X zovemo neprekidna slučajna varijabla na* $\Omega$, *a funkciju* $f$ *funkcija gustoće.*

Uz pojam slučajne varijable pojavljuje se i slučajni vektor.

**Definicija 7.** *Neka je* $(\Omega, \mathcal{F}, P)$ *vjerojatnosni prostor pridružen slučajnom pokusu. Funkciju* $(X_1, \dots, X_n)$ *koja svakom ishodu pokusa pridružuje uređenu n-torku realnih brojeva* $(x_1, \dots, x_n)$ *zovemo n-dimenzionalan slučajan vektor ako vrijedi :*

$$\{X_1 \leq x_1\} \cap \dots \cap \{X_n \leq x_n\} \in \mathcal{F}$$

*za svaki* $x_1, \dots, x_n \in \mathbb{R}$.

**Definicija 8.** *Statistički model* $\mathcal{F}$ *je familija funkcija distribucije slučajnog vektora koja se uzima u obzir pri zaključivanju u danom problemu.*

Kao poseban statistički model treba istaknuti parametarski statistički model u kojem je poznat tip distribucije do na nepoznati parametar. Kako se nepoznati parametar najčešće označava s $\theta$, funkciju distribucije označavamo s $F_\theta$. Parametarski statistički model je tada $P = \{F_\theta : \theta \in \Theta\}$ pri čemu je $\Theta \subset \mathbb{R}^k$ prostor parametara.

**Definicija 9.** *Slučajni vektor* $\mathbb{X} = (X_1, X_2, \dots, X_n)$ *zovemo slučajni uzorak, a njegovu realizaciju* $(x_1, x_2, \dots, x_n) \in \mathbb{R}^n$ *uzorak.*

**Definicija 10.** *Jednostavni slučajni uzorak (j.s.u.) iz distribucije zadane slučajnom varijablom X je uređena n-torka slučajnih varijabli* $(X_1, X_2, \dots X_n)$ *od kojih svaka ima istu distribuciju kao X i međusobno su nezavisne.*

Jedna od numeričkih karakteristika slučajne varijable koja će se često koristiti kroz ovaj rad je očekivanje. Očekivanje $\mu$ slučajne varijable $X$ definira se kao:

$$EX = \mu = \begin{cases} \sum_{i \in I} x_i p_i, & \text{ako je } X \text{ diskretna} \\ \int_{-\infty}^{\infty} x f(x) dx, & \text{ako je } X \text{ neprekidna} \end{cases},$$

pri čemu su $x_i$ i $p_i$ vrijednosti definirane u 5, a $f$ je realna funkcija realne varijable iz 6. Varijancu slučajne varijable $\sigma^2$ računamo kao $Var(X) = E(X - EX)^2 = EX - (EX)^2$, dok $\sqrt{Var(X)}$ nazivamo standardna devijacija slučajne varijable i označavamo ju sa $\sigma$. Slijedeće ćemo definirati normalnu i Bernoullijevu distribuciju.

**Definicija 11.** *Za neprekidnu slučajnu varijablu kažemo da ima Gaussovu ili normalnu distribuciju s parametrima $\mu$ i $\sigma^2$ ako je njezina funkcija gustoće dana izrazom*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x \in \mathbb{R},$$

*gdje su $\mu$ i $\sigma, \sigma > 0$ realni brojevi. Ako slučajna varijabla ima normalnu distribuciju s parametrima $\mu$ i $\sigma^2$ koristimo oznaku $X \sim \mathcal{N}(\mu, \sigma^2)$. Normalna distribucija s parametrima $\mu = 0$ i $\sigma^2 = 1$ zove se jedinična ili standardna normalna distribucija.*

**Definicija 12.** *p-kvantil slučajne varijable $X$ s funkcijom distribucije $F$ je $x_p \in \mathbb{R}$ takav da je*
*$P(X \leq x_p) \geq p$ i $P(X \geq x_p) \geq 1 - p$ za $p \in \langle 0, 1 \rangle$.*

Neki od poznatijih kvantila su 0.5-kvantil kojeg nazivamo medijan slučajne varijable $X$, 0.25-kvantil i 0.75-kvantil koje zovemo donji i gornji kvartil.

**Definicija 13.** *Za slučajnu varijablu kažemo da ima Bernoullijevu distribuciju s parametrom $p$ ako ona poprima točno dvije vrijednosti $\mathcal{R}(X) = 0, 1$. Njena tablica distribucije je oblika:*

$$X = \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix}, \ p \in (0,1), q = 1 - p.$$

# 3 | Statističke hipoteze

Statistička hipoteza je slutnja koja se odnosi na statistički model i predstavlja pretpostavku o populacijskoj razdiobi . Pojam „testiranja hipoteza" odnosi se na postupak donošenja odluke o prihvaćanju ili odbacivanju statističkih hipoteza na temelju rezultata eksperimenata. U statističkom modelu $\mathcal{P}$ statistička hipoteza $\mathcal{H}$ je njen podskup, tj. $\mathcal{H} \subset \mathcal{P}$. U parametarskim modelima hipoteza je izjava o vrijednosti nepoznatog parametra - „parametarska hipoteza" koju možemo identificirati s nekim podskupom prostora parametara.
Razlikujemo jednostavnu i složenu statističku hipotezu. Statistička hipoteza je jednostavna ako je njome jednoznačno određena populacijska razdioba, u suprotnom je složena. Problem koji nas zanima moramo prevesti u statističku hipotezu. Uvijek imamo dvije statističke hipoteze, tzv. nul statističku hipotezu $\mathcal{H}_0$ i alternativnu statističku hipotezu $\mathcal{H}_1$. Nul hipoteza se ne može potvrditi, prema tome u nju stavljamo ono što želimo odbaciti kako bismo mogli potvrditi alternativnu hipotezu. Hipoteze moraju biti disjunktne, ali u uniji činiti cijeli statistički model, tj.

$$\mathcal{H}_0 \cap \mathcal{H}_1 = \emptyset \ i \ \mathcal{P} = \mathcal{H}_0 \cup \mathcal{H}_1.$$

Hipoteze se uvijek postavljaju tako da prije provođenja testa vrijedi nul-hipoteza. Odluku u postupku testiranja hipoteza donosimo na temelju odabranog kriterija i realizacije uzorka. Odabrati kriterij znači definirati pravilo za odbacivanje nul-hipoteze. Obzirom kako se odluka temelji na realizaciji slučajnog uzorka $(X_1, X_2, \ldots X_n)$ statističkog modela, pravilo mora biti iskazano u terminu tog slučajnog vektora. U tu svrhu uvodimo pojam "statistika".

**Definicija 14.** *Neka je $(X_1, X_2, \ldots X_n)$ slučajni uzorak statističkog modela $\mathcal{P}$ na temelju kojega donosimo zaključke i neka je $t : \mathcal{R}(X_1, X_2, \ldots X_n) \to \mathbb{R}^k$ zadana funkcija. Slučajni vektor $T = t(X_1, X_2, \ldots X_n)$ zovemo statistika za taj statistički model.*

Primjetimo da se u slučaju kada je $k = 1$ radi o slučajnoj varijabli.

Statistički test koji koristimo prilikom testiranja statističke hipoteze kreiran je kako bismo korištenjem informacija iz prikupljenih podataka o realizacijama slučajne varijable donosili odluku o odbacivanju nul-hipoteze u korist alternativne hipoteze ili neodbacivanju nul-hipoteze.

## 3.1 Kritično područje

Pravilo po kojemu odbacujemo postavljenu hipotezu podijelit će skup svih mogućih realizacija slučajnog uzorka statističkog modela na dva disjunktna dijela.

Te dijelove označavamo s $C$ i $C^C$ , pri čemu je $C$ dio koji odgovara odbacivanju nul-hipoteze i kojega zovemo kritično područje.

**Definicija 15.** *Za skup $C \subset R(X_1, ..., X_n)$ kažemo da je kritično područje ako $(x_1, ..., x_n) \in C$ povlači odbacivanje nul-hipoteze.*

## 3.2 Pogreške prve i druge vrste

Sve odluke temeljene na uzorcima iz populacije nisu potpuno pouzdane pa tako ni zaključak testa ne mora biti potpuno pouzdan. Uočimo kako nul-hipoteza i alternativna hipoteza pri definiranju statističkog testa nisu ravnopravne, točnije nul-hipotezu ne možemo prihvatiti. Razlog tome je što pri odlučivanju u statističkom testu imamo toleranciju malih vjerojatnosti pogrešne odluke.
Odluka koja je donesena statističkim testom može biti ili pogrešna ili ispravna. Pri tome se mogu dogoditi dva tipa pogrešne odluke:

- pogreška I. tipa: odbaciti $\mathcal{H}_0$ ako je ona istinita, točnije ako je uzorak $\mathbf{x} = (x_1 \ldots, x_n)$ u kritičnom području, ali nul-hipoteza je istinita,
- pogreška II. tipa: ne odbaciti $\mathcal{H}_0$ ako je $\mathcal{H}_1$ istinita, točnije $\mathbf{x} = (x_1 \ldots, x_n)$ nije u kritičnom području, ali nul-hipoteza nije istinita.

Najzornije pogreške I. i II. tipa možemo uočiti kroz primjer suđenja. Postavljanje hipoteza možemo usporediti sa suđenjem: "nitko nije kriv dok mu se ne dokaže krivnja". U tom slučaju se hipoteze postavljaju na sljedeći način:

$$\mathcal{H}_0 : \textit{ optuženik nije kriv,}$$
$$\mathcal{H}_1 : \textit{ optuženik je kriv.}$$

Pogreška I.tipa je optužiti nevinu osobu, dok je pogreška II.tipa osloboditi osobu koja je stvarno kriva.
Vjerojatnosti pogreške I. i II. tipa ovise o stvarnoj distribuciji slučajne varijable na temelju koje testiramo hipoteze, tj. te su vjerojatnosti funkcije koje distribuciji iz $\mathcal{P}$ pridružuju broj, ali one nisu definirane na istoj domeni. Vjerojatnost pogreške I. tipa može se računati za svaku distribuciju iz $\mathcal{H}_0$, a vjerojatnost pogreške II. tipa računa se za svaku distribuciju iz $\mathcal{H}_1$. Slučajna varijabla o kojoj je ovdje riječ zove se test-statistika, a nju koristimo kako bi izračunali jakost testa.

## 3.3 Funkcija jakosti testa

Idealan bi statistički test bio kada bi se $\mathbf{x}$ nalazio u kritičnom području i $\mathcal{H}_0$ nije istinita, a istovremeno kada $\mathbf{x}$ nije u kritičnom području i $\mathcal{H}_0$ je istinita. Takav test gotovo je nemoguće konstruirati, ali možemo konstruirati test tako da postignemo što manju vjerojatnost pogreške I. i II. tipa. Vjerojatnost pogrešaka opisujemo funkcijom jakosti testa.
Funkcija jakosti testa, u oznaci $\pi : P \to [0, 1]$, definirana je za svaku distribuciju iz $\mathcal{P}$ (odnosno za svaku dozvoljenu vrijednost parametra ako je statistički model

parametarski) kao $\pi(\mathbb{F}) = P_{\mathbb{F}}\{(X_1, X_2, \ldots, X_n) \in C\}$, pri čemu $P_{\mathbb{F}}$ označava vjerojatnost pod pretpostavkom da slučajan uzorak $(X_1, \ldots, X_n)$ ima distribuciju $\mathbb{F}$, odnosno $\mathbb{F} \in \mathcal{P}$. Za parametarski statistički model funkcija jakosti testa definirana je kao $\pi(\theta) = P_\theta\{(X_1, \ldots, X_n) \in C\}, \theta \in \Theta$, gdje je $\Theta$ prostor parametara, a $P_\theta$ vjerojatnost pod pretpostavkom kako je $\theta$ stvarna vrijednost parametra.

Treba uočiti sljedeće:

- $\pi(\mathbb{F})$ (ili $\pi(\theta)$) je vjerojatnost odbacivanja nul-hipoteze,
- funkcija $\alpha : \mathcal{H}_0 \to [0,1]$, $\alpha(\mathbb{F}) = \pi(\mathbb{F}) = P_{\mathbb{F}}\{(X_1, \ldots, X_n) \in C\}$ daje vjerojatnost pogreške I.tipa,
- funkcija $\beta : \mathcal{H}_1 \to [0,1]$, $\beta(\mathbb{F}) = 1 - \pi(\mathbb{F}) = P_{\mathbb{F}}\{(X_1, \ldots, X_n) \in C\}$ daje vjerojatnost pogreške II.tipa.

U nastavku se radi jednostavnosti fokusiramo na parametarski model i parametarske hipoteze pri čemu $\pi$, $\alpha$ i $\beta$ promatramo kao funkcije nepoznatog parametra $\theta \in \Theta$.

## 3.4 Razina značajnosti

Prilikom definiranja testa, odnosno kritičnog područja, funkcije $\alpha$ i $\beta$ želimo učiniti što manjima. Definiramo:

- maksimalnu vjerojatnost pogreške I.tipa: $\alpha = \sup\limits_{\theta \in \Theta_0} \alpha(\theta) = \sup\limits_{\theta \in \Theta_0} P_\theta(\mathbb{X} \in C)$,
- maksimalnu vjerojatnost pogreške II.tipa: $\beta = \sup\limits_{\theta \in \Theta_1} \beta(\theta) = \sup\limits_{\theta \in \Theta_1} P_\theta(\mathbb{X} \notin C)$.

Vjerojatnost pogreške prvog i drugog tipa ovisi o stvarnoj distribuciji test-statistike. Želimo da su vjerojatnosti pogreške što manje. Vrlo je teško u klasičnom pristupu, odnosno Neyman-Persanovom pristupu, istodobno učiniti što manjima i $\alpha$ i $\beta$, stoga se najprije fiksira maksimalna dozvoljena vjerojatnost pogreške I.tipa $\alpha_0$. Vrijednosti koje se obično uzimaju za $\alpha_0$ su 0.01, 0.05 ili 0.1. Nakon što smo odredili $\alpha_0$ test definiramo tako da $\beta$ bude što je moguće manji, odnosno kritično područje biramo tako da $\alpha < \alpha_0$ . Vrijednost $\alpha_0$ nazivamo razina značajnosti testa ili nivo signifikantnosti testa te ju obično označavamo s $\alpha$. $\mathcal{H}_0$ odbacujemo ako se test-statistika $T$ realizira u kritičnom području, odnosno $T = t(\mathbb{X}) \in C$.

## 3.5 $p$-vrijednost

**Definicija 16.** *Najmanja razina značajnosti uz koju odbacujemo $\mathcal{H}_0$, odnosno vjerojatnost da test-statistika poprimi vrijednosti koje su uz pretpostavku da je $\mathcal{H}_0$ istinita manje ili jednako vjerojatne od opažene vrijednosti test-statistike zove se p-vrijednost.*

Pomoću $p$-vrijednosti zaključujemo sljedeće:

1° $p \leq \alpha$: odbacujemo $\mathcal{H}_0$ na razini značajnosti $\alpha$, 2° $p > \alpha$: ne odbacujemo $\mathcal{H}_0$ na razini značajnosti $\alpha$.

Dokaz protiv $\mathcal{H}_0$ je jači što je $p$-vrijednost manja. Uočimo kako su prilikom definiranja testa nul i alternativna hipoteza u neravnopravnom odnosu te moramo paziti kako ih definiramo. Primjerice, za nul hipotezu uglavnom uzimamo jednostavnu hipotezu i one hipoteze za koje želimo kontrolirati vjerojatnost pogreške I.tipa dok za alternativnu hipotezu biramo ono što želimo dokazati jer nju možemo prihvatiti.

# 4 | Testovi o očekivanju

## 4.1 Testovi o očekivanju normalno distribuirane populacije uz poznatu varijancu

Neka je $\mathbb{X} = (X_1, \ldots, X_n)$ je jednostavni slučajan uzorak koji je iz $\mathcal{N}(\mu, \sigma^2)$ distribucije pri čemu je $\mu \in \mathbb{R}$ nepoznati parametar, a $\sigma^2 > 0$ poznata konstanta. U nul-hipotezi pretpostavljamo da je očekivanje $\mu$ jednako već unaprijed zadanoj vrijednosti $\mu_0 \in \mathbb{R}$. Testovi se razlikuju s obzirom na oblik alternativne hipoteze. Razlikujemo dvostrane i jednostrane testove. Ovaj test je još poznat kao $z$-test jer test-statistika ima normalnu $(z)$ distribuciju.

### 4.1.1 Dvostrani test

Hipoteze testa su oblika:

$$\begin{aligned} \mathcal{H}_0 &: \mu = \mu_0, \\ \mathcal{H}_1 &: \mu \neq \mu_0. \end{aligned}$$

Potrebna nam je statistika čija se distribucija razlikuje u $\mathcal{H}_0$ i $\mathcal{H}_1$. U tu svrhu promotrimo

$$T(\mathbb{X}) = \frac{\overline{X_n} - \mu_0}{\sigma}\sqrt{n}.$$

U uvjetima $\mathcal{H}_0$ pripadna test-statistika imat će $\mathcal{N}(0,1)$ distribuciju, dok će u uvjetima $\mathcal{H}_1$ ponovno biti normalno distribuirana, ali očekivanje neće biti jednako nula te će ovisiti o $\mu$. Pretpostavljamo kako je $\mathcal{H}_0$ točna i očekujemo kako će razlika između uzoračkog prosjeka i očekivane vrijednosti biti vrlo mala zbog čega se očekuje kako će test-statistika pratiti standardnu normalnu distribuciju koja ima očekivanje 0 i varijancu 1, odnosno $\mathcal{N}(0,1)$ .

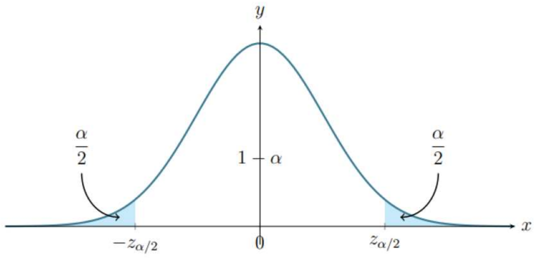
$$E(T) = \frac{E(\overline{X_n} - \mu_0)}{\sigma}\sqrt{n} = \frac{\mu - \mu_0}{\sigma}\sqrt{n}.$$

Ako nam ja ipak alternativna hipoteza točna, to tumačimo kao da postoji stvarna razlika između uzoračkog prosjeka i očekivane vrijednosti te će i test-statistika odražavati tu vrijednost. Test-statistika će i dalje pratiti normalnu distribuciju, ali joj očekivanje neće biti jednako nuli već će ovisiti o razlici uzoračkog prosjeka i očekivane vrijednosti.

$p$-vrijednost računamo prema sljedećoj formuli,

$$p = P(|T(\mathbb{X})| \geqslant |\hat{t}|) = 2 \cdot P_{\mu_0}(T(\mathbb{X}) \geqslant |\hat{t}|).$$

$\hat{t}$ u prethodnoj formuli označavarealizacije test-statistike te je on broj iz skupa realnih brojeva ($\hat{t} \in \mathbb{R}$). Kako bismo definirali test pronađimo kritično područje $C$ u obliku komplementa intervala simetričnog oko nula pri čemu kontroliramo pogrešku I.tipa. Neka je $C = \langle -\infty, -\frac{z_\alpha}{2} \rangle \cup \langle \frac{z_\alpha}{2}, \infty \rangle$ pri čemu je $\frac{z_\alpha}{2}$ jednak $(1 - \frac{\alpha}{2})$-kvantilu standardne normalne distribucije i $\alpha$ razina značajnosti.




Slika 4.1: Prikaz pronalaska $z_{\alpha/2}$

Primjerice, uzmimo da je $\alpha = 0.05$. Ako je $\mathcal{H}_0$ istinita, tada se s vjerojatnošću 0.95 test-statistika realizira u intervalu $[-\frac{z_\alpha}{2}, \frac{z_\alpha}{2}]$. Ako dođe do takve realizacije tada nemamo razloga odbaciti $\mathcal{H}_0$. Ako se dogodi da je realizacija test-statistike van tog intervala, vjerojatnost tog događaja je manja od 0.05 pod uvjetom da je $\mathcal{H}_0$ istinita. Prema tome, na razini značajnosti 0.05 odbacujemo $\mathcal{H}_0$ jer je realizacija jako malo vjerojatna u uvjetima $\mathcal{H}_0$.

### 4.1.2 Jednostrani test

Hipoteze testa su oblika:

a)

$$\mathcal{H}_0 : \mu = \mu_0,$$
$$\mathcal{H}_1 : \mu > \mu_0.$$

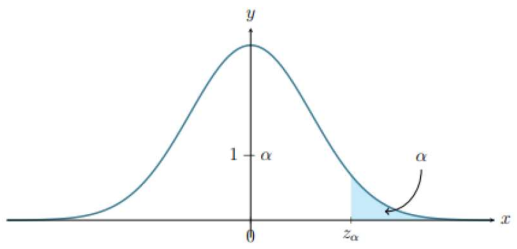
Test-statistika je ista, tj. $T(\mathbb{X}) = \dfrac{\overline{X_n} - \mu_0}{\sigma}\sqrt{n}$ i u uvjetima $\mathcal{H}_1$ je $\dfrac{\mu - \mu_0}{\sigma}\sqrt{n} > 0$. U uvjetima $\mathcal{H}_0$ test-statistika ima $\mathcal{N}(0,1)$ distribuciju, dok u uvjetima $\mathcal{H}_1$ ima ponovno normalnu distribuciju, ali s očekivanjem različitim od nula ($\mu - \mu_0 > 0$). Test-statistika ponovo prati standardnu normalnu distribuciju pod pretpostavkom kako je nulta hipoteza točna. Ako pretpostavimo kako vrijedi alternativna hipoteza tada očekujemo postojanje razlike između aritmetičke sredine uzorka $\overline{X_n}$ i očekivane vrijednosti $\mu_0$. Prethodno definirana razlika bit će veća od nule što implicira kako očekivanje test-statistike neće biti nula već upravo definirana razlika $\mu - \mu_0$ pri čemu varijanca i dalje ostaje jednaka

1 (test-statistika je standardizirana). Ako alternativna hipoteza vrijedi test-statistika će se pomaknuti od nule u smjeru $\mu - \mu_0$.

$p$-vrijednost računamo na sljedeći način,

$$p = P_{\mu_0}(T(\mathbb{X}) \geqslant \hat{t}).$$

Kritično područje je $C = [z_\alpha, \infty\rangle$, pri čemu je $z_\alpha$ $(1-\alpha)$-kvantil $\mathcal{N}(0,1)$ distribucije.



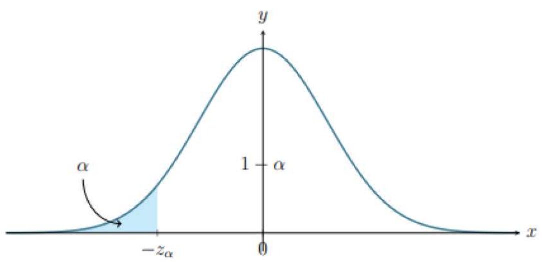

Slika 4.2: Prikaz pronalaska $z_\alpha$ za a) slučaj

b)

$$\mathcal{H}_0 : \mu = \mu_0,$$
$$\mathcal{H}_1 : \mu < \mu_0.$$

Test-statistika je ista kao u a), ali kritično područje se razlikuje i ono je oblika $C = \langle -\infty, z_\alpha]$.$p$-vrijednost računamo na sljedeći način,

$$p = P_{\mu_0}(T(\mathbb{X}) \leqslant \hat{t}).$$




Slika 4.3: Prikaz pronalaska $z_\alpha$ za b) slučaj

Za većinu testova, računalni program izbacuje kao rezultat $p$-vrijednost.

## 4.2 Test o očekivanju normalno distribuirane populacije uz nepoznatu varijancu

Neka je $(X_1, \ldots, X_n)$ jednostavni slučajni uzorak iz $\mathcal{N}(\mu, \sigma^2)$ distribucije pri čemu je $\sigma^2$ nepoznata, a $\mu \in \mathbb{R}$ nepoznati parametar od interesa. Koristimo test-

statistiku oblika
$$T(\mathbb{X}) = \frac{\overline{X_n} - \mu_0}{S_n}\sqrt{n}.$$

Možemo primjetiti kako je test statistika identična kao u prethodnom testu pri čemu je $\sigma$ iz prethodne test-statistike zamijenjen sa $S_n$, što označava promatranu (korigiranu) varijancu uzorka koja predstavlja procjenitelja nepoznate varijance $\sigma^2$.

### 4.2.1 Dvostrani test

Hipoteze su oblika:
$$\mathcal{H}_0 : \mu = \mu_0,$$
$$\mathcal{H}_1 : \mu \neq \mu_0.$$

U uvjetima $\mathcal{H}_0$ test-statistika ima Studentovu distribuciju s $n-1$ stupnjeva slobode, tj. $\mathcal{T}_{n-1}$ distribuciju. To tvrdimo prema Teoremu 8.4.3 u [1], kojeg navodimo u nastavku.

**Teorem 1.** *Ako $X_1, ..., X_n$ označava slučajni uzorak iz $\mathcal{N}(\mu, \sigma^2)$ tada*
$$\frac{\overline{X_n} - \mu_0}{S_n}\sqrt{n} \sim t(n-1).$$

$p$-vrijednost računamo prema sljedećoj formuli $p = P_{\mu_0}(|T(\mathbb{X})| \geqslant |\hat{t}|)$, pri čemu je $\hat{t}$ vrijednost test-statistike. Kritično područje bit će $C = \langle -\infty, -t_{n-1,\frac{\alpha}{2}}] \cup [t_{n-1,\frac{\alpha}{2}}, \infty\rangle$ pri čemu je $t_{n-1,\frac{\alpha}{2}}$ jednak $1-\frac{\alpha}{2}$ kvantilu Studentove distribucije s $n-1$ stupnjeva slobode.

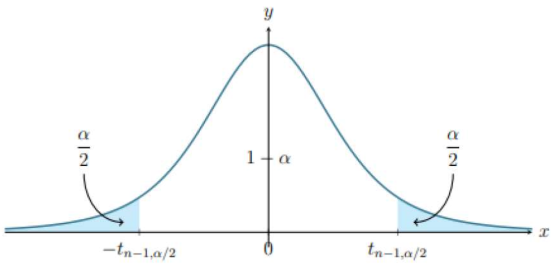
### 4.2.2 Jednostrani testovi

a)
$$\mathcal{H}_0 : \mu = \mu_0,$$
$$\mathcal{H}_1 : \mu > \mu_0.$$

$p$-vrijednost računamo prema sljedećoj formuli $p = P_{\mu_0}(T(\mathbb{X}) \geqslant \hat{t})$. Kritično područje jednako je $C = [t_{n-1,\alpha}, \infty\rangle$ pri čemu je $t_{n-1,\alpha}$ $(1-\alpha)$-kvantil $\mathcal{T}_{n-1}$ distribucije.

b)
$$\mathcal{H}_0 : \mu = \mu_0,$$
$$\mathcal{H}_1 : \mu < \mu_0.$$

$p$-vrijednost računamo prema sljedećoj formuli $p = P_{\mu_0}(T(\mathbb{X}) \leqslant \hat{t})$. Kritično područje jednako je $C = \langle -\infty, t_{n-1,\alpha}]$ pri čemu je $t_{n-1,\alpha}$ $(1-\alpha)$-kvantil $\mathcal{T}_{n-1}$ distribucije.




Slika 4.4: Prikaz pronalaska $t_{n-1,\alpha/2}$

**Primjer 1.** *Prosječna temperatura u lipnju u američkim gradovima iznosi* 71.6 *Fahrenheitovih stupnjeva*[1]. *U bazi airquality iz paketa datasets u R-u nalaze se podaci za temperature u New Yorku. Jesu li očekivane temperature u New Yorku u lipnju različite od prosječne temperature u lipnju u cijelim Sjedinjenim Američkim Državama?*

Radi se o dvostranom testu o očekivanju uz nepoznatu varijancu. Pretpostavljamo kako je temperatura u lipnju u New Yorku normalno distribuirana i neovisna (temperature mjerene na različite dane nisu povezane). Normalnost uzorka također možemo provjeriti vizualno (npr. histogram-grafički prikaz distribucije podataka) ili provesti Shapiro-Wilk test kako bi bili sigurni da naši podaci zadovoljavaju pretpostavku normalnosti. Za detaljniji opis testa pogledati [6], a nešto više o histogramu možemo pronaći u [3]. Hipoteze koje želimo testirati:

$$\mathcal{H}_0 : \mu = 71.6 \text{ (očekivana temperatura je ista)},$$
$$\mathcal{H}_1 : \mu \neq 71.6 \text{ (očekivana temperatura nije ista)}.$$

Realizacija test-statistike jednaka je:

$$\hat{t}(\mathbf{x}) = \frac{\overline{\mathbf{x}} - \mu_0}{s_n}\sqrt{n} = \frac{79.1 - 71.6}{6.60}\sqrt{30} = 6.23.$$

Kritično područje oblika je $C = \langle -\infty, -t_{n-1,\frac{\alpha}{2}}] \cup [t_{n-1,\frac{\alpha}{2}}, \infty\rangle$. U našem slučaju $t_{n-1,\frac{\alpha}{2}} = t_{29,\frac{0.05}{2}} = 2.05$, stoga je kritično područje $C = \langle -\infty, -2.05] \cup [2.05, \infty\rangle$ iz čega vidimo da se test-statistika realizira unutar kritičnog područja. Za $p$-vrijednost dobivamo $p = 8.597 \cdot 10^{-7} < 0.05$ pa na razini značajnosti 0.05 možemo odbaciti $H_0$ i prihvatiti $H_1$, odnosno, tvrdimo da se očekivana temperatura u lipnju u New Yorku razlikuje od prosječne temperature u cijelom SAD-u u lipnju.

## 4.3 Test o očekivanju za velike uzorke

Uzimajući u obzir raznolike baze podataka, svojstva tih podataka te varijabilnost među podacima, teško je točno definirati veliki uzorak. Kroz ovaj rad smatrat ćemo da ukoliko je broj opažanja $n$ veći od 30, uzorak smatramo velikim, no to nije općenito tako. Neka je $\mathbb{X}$ jednostavni slučajni uzorak iz neke distribucije, $EX_1 = \mu$ i $VarX_1 = \sigma^2$, pri čemu je $\sigma^2$ konačna. Za testiranje takvog uzorka koristimo $t$-test koji je asimptotski.

[1]Podatak sa stranice *https://www.tuiholidays.ie/*.

**Primjer 2.** *U tablici Abbey paketa BSDA nalaze se podaci dnevnog povrata cijene* (*u penijima*) *dionica Abbey National između 31.srpnja 1991. i 8.listopada 1991. Možemo li na razini značajnosti* 0.05 *tvrditi da je očekivani dnevni povrat u penijima veći od* 250*?*

Kako nas zanima je li očekivani dnevni povrat veći od 250 penija, potreban nam je jednostrani $t$-test čija je realizacija test-statistike, obzirom na promatrane podatke:

$$\hat{t}(\mathbf{x}) = \frac{\overline{\mathbf{x}} - \mu_0}{s_n}\sqrt{n} = \frac{299.96 - 250}{5.61}\sqrt{50} = 63.02.$$

Hipoteze koje testiramo su:

$$\begin{aligned} \mathcal{H}_0 &: \mu = 250, \\ \mathcal{H}_1 &: \mu > 250. \end{aligned} \tag{4.1}$$

Kritično područje stoga je jednako $C = [t_{n-1,\alpha}, \infty\rangle$. Kako je $n = 50$ i $\alpha = 0.05$, kritično područje je: $C = [t_{49,0.05}, \infty\rangle = [1.67, \infty\rangle$. Test-statistika realizira se unutar kritičnog područja iz čega zaključujemo da odbacujemo $\mathcal{H}_0$. Jednak zaključak dobivamo iz $p$-vrijednosti koja iznosi $p = 2.2 \cdot 10^{-16} < 0.05$ i prihvaćamo alternativnu hipotezu, odnosno možemo tvrditi da je očekivani dnevni povrat veći od 250 penija.

# 5 | Test o proporciji (vjerojatnosti događaja)

Neka je $\mathbb{X} = (X_1, \ldots, X_n)$ jednostavni slučajan uzorak koji ima Bernoullijevu distribuciju

$$\begin{pmatrix} 0 & 1 \\ 1-\theta & \theta \end{pmatrix},$$

pri čemu je $\theta \in \langle 0, 1 \rangle$ nepoznati parametar koji interpretiramo kao vjerojatnost ili proporciju od interesa. Ovim testom želimo testirati je li parametar $\theta$ jednak već nekoj unaprijed zadanoj vrijednosti $\theta_0$ uz razinu značajnosti $\alpha$. Pripadna test-statistika brojat će uspjehe i ona je oblika

$$T(\mathbb{X}) = \sum_{i=1}^{n} X_i.$$

U uvjetima istinitosti $\mathcal{H}_0$, odnosno kada je $\mathcal{H}_0 : \theta = \theta_0$, test-statistika ima binomnu distribuciju sa parametrima $n$ i $\theta_0$ (odnosno $T(\mathbb{X}) \sim \mathcal{B}(n, \theta_0)$). Priroda podataka nam govori o kojoj je distribuciji riječ, točnije broj uspjeha(u ovom slučaju naša test-statistika) u $n$ pokusa, pri čemu je uspjeh s vjerojatnošću $\theta_0$ prati binomnu distribuciju.

## 5.1 Dvostrani test

Hipoteze u dvostranom testu su oblika:

$$\begin{aligned} \mathcal{H}_0 &: \theta = \theta, \\ \mathcal{H}_1 &: \theta \neq \theta_0. \end{aligned}$$

Kritično područje je oblika $C = \{0, \ldots, \underline{k}\} \cup \{\overline{k}, \ldots, n\}$, pri čemu je $\underline{k} \in \{0, 1, 2, \ldots, n\}$ najveći broj takav da je

$$P_{\theta_0}(T(\mathbb{X}) \leq \underline{k}) = \sum_{k=0}^{\underline{k}} \binom{n}{k} \theta_0^k (1-\theta_0)^k \leq \frac{\alpha}{2},$$

a $\overline{k}$ je najmanji element od $\{0, 1, \ldots, n\}$ takav da je

$$P_{\theta_0}(T(\mathbb{X}) \geq \overline{k}) = \sum_{k=0}^{\overline{k}} \binom{n}{k} \theta_0^k (1-\theta_0)^k \leq \frac{\alpha}{2}.$$

Ako bi $\hat{t}$ bila realizacija test-statistike, onda će $p$-vrijednost biti vjerojatnost da $T$ bude jednako ili manje vjerojatna nego $\hat{t}$, tj.

$$p = \sum_{k \in \{0,1,\ldots,n\} : P_{\theta_0}(T=k) \leq P_{\theta_0}(T=\hat{t})} P_{\theta_0}(T = k).$$

**Primjer 3.** *Tablica PhDPublications.csv*[1] *sadrži podatke o doktorandima biokemije. Uz broj djece, broj objavljenih raddova i bračni status, također sadrži i podatke o spolu. Možemo li tvrditi da je udio muškaraca i žena različit?*

Obzirom da preispitujemo udio muškaraca i žena, radi se o testu o proporciji čije su hipoteze:

$$\mathcal{H}_0 : \theta = 0.5,$$
$$\mathcal{H}_1 : \theta \neq 0.5.$$

$\theta$ može biti ili udio žena ili udio u muškaraca u uzorku, no mi ćemo se odlučiti da je $\theta$ udio žena u uzorku. Od ukupno 915 promatranja, 421 su žene. Dakle, realizacija test-statistika bit će:

$$\hat{t}(\mathbf{x}) = \sum_{i=1}^{915} x_i = 421.$$

Za kritično područje $C = \{0, \ldots, \underline{k}\} \cup \{\overline{k}, \ldots, n\}$ trebaju nam $\underline{k}$ i $\overline{k}$. U našem slučaju:

- $\underline{k} \in \{0, 1, 2, \ldots, 915\}$ najveći broj takav da je

$$P_{0.5}(T(\mathbb{X}) \leq \underline{k}) = \sum_{k=0}^{\underline{k}} \binom{915}{k} (0.5)^k (0.5)^k \leq \frac{0.05}{2},$$

- $\overline{k} \in \{0, 1, \ldots, 915\}$ najmanji broj takav da je

$$P_{0.5}(T(\mathbb{X}) \geq \overline{k}) = \sum_{k=0}^{\overline{k}} \binom{915}{k} (0.5)^k (0.5)^k \leq \frac{0.05}{2}.$$

Iz prethodnog dobijemo da su $\underline{k} = 427$ $\overline{k} = 488$ pa je kritično područje $C = \{0, \ldots, 427\} \cup \{488, \ldots, 915\}$. Vidimo da se test-statistika realizira unutar kritičnog područja iz čega zaključujemo da možemo odbaciti $\mathcal{H}_0$. Isti zaključak dobijemo ukoliko pogledamo $p$-vrijednost koja iznosi $0.01725 < 0.05$. Dakle, odbacujemo $\mathcal{H}_0$ i možemo tvrditi da je udio muškaraca i žena različit.

[1] Stvorena od podataka iz knjige *Social Forces* autora J. S. Long, nalazi se u skoplu poglavlja *The Origins of Sex Differences in Science*.

## 5.2 Jednostrani testovi

Jednostrane testove koristimo kod ispitivanja je li proporcija veća ili manja od zadane vrijednosti, odnosno kada imamo specifična očekivanja o smjeru promjene, a testiranjem vidimo je li promjena u tom smjeru značajna. Hipoteze su slijedeće:

a)

$$\mathcal{H}_0 : \theta = \theta_0,$$
$$\mathcal{H}_1 : \theta > \theta_0,$$

te je kritično područje $C = \{k_1, \dots, n\}$ gdje je $k_1$ najmanji element iz $\{0, 1, \dots, n\}$ takav da vrijedi

$$P_{\theta_0}(T \geq k_1) = \sum_{k=k_1}^{n} \binom{n}{k} \theta_0^k (1-\theta_0)^k \leq \alpha.$$

b)

$$\mathcal{H}_0 : \theta = \theta_0$$
$$\mathcal{H}_1 : \theta < \theta_0,$$

gdje je kritično područje $C = \{0, 1, \dots, k_2\}$ za $k_2$ najveći element iz $\{0, 1, \dots, n\}$ takav da je

$$P_{\theta_0}(T \leq k_1) = \sum_{k=0}^{k_2} \binom{n}{k} \theta_0^k (1-\theta_0)^k \leq \alpha.$$

$P$-vrijednost za prvi slučaj jednaka je $p = P_{\theta_0}(T \geq \hat{t})$, dok je za drugi slučaj jednaka $p = P_{\theta_0}(T \leq \hat{t})$.

**Primjer 4.** *Ana šeta Budimpeštom i odlučila je da želi popiti dobru kavu. Prvi put se našla u Budimpešti i ne zna koji kafić je dobar. U blizini vidi Starbucks i odlučila je pogledati Google recenzije. Prolazeći na brzinu kroz ocjene vidjela je da ima jako puno ocjena* 1*. Iako joj je stvarno hladno, ne želi ući ukoliko je udio jedinica* 51 *posto ili više.*

Koristimo bazu podataka *Starbucks Reviews Dataset*[2] koja sadrži ocjene iz recenzija za kafić *Starbucks*, s naglaskom na to da smo nedostajuće podatke, odnosno recenzije bez ocjena uklonili iz promatranja. Ukupan broj promatranja tada je 705. Kako se radi o jednostranom testu o proporciji, potreban nam je broj uspjeha za test-statistiku. Kako je uspjeh ukoliko je ocjena 1, ukupan broj recenzija kojima je ocjena 1 je 451, stoga je realizacija test-statistika jednaka:

$$\hat{t}(\mathbf{x}) = \sum_{i=1}^{n} x_i = \sum_{i=1}^{705} x_i = 451.$$

[2]Baza podataka pronađena je na *https://www.kaggle.com/*.

Neka je $\theta$ udio jedinica u uzorku. Tada su hipoteze:

$$\mathcal{H}_0 : \theta = 0.51,$$
$$\mathcal{H}_1 : \theta > 0.51.$$

Kritično područje je $C = \{k_1, \ldots, n\}$, pri čemu je $k_1 \in \{0, 1, 2, \ldots, n\}$ najveći broj takav da

$$P_{0.51}(T \geq k_1) = \sum_{k=k_1}^{705} \binom{705}{k} (0.51)^k (0.49)^k \leq 0.05.$$

Vrijednosti tog broja je $k_1 = 382$ pa je kritično područje $C = \{382, \ldots, 705\}$ iz čega vidimo da se test-statistika realizira u kritičnom području. Analogan zaključak slijedi i pogledamo li $p$-vrijednost. Za $p$-vrijednost dobivamo $p = 2.559 \cdot 10^{-12} < 0.05$ pa na razini značajnosti 0.05 možemo odbaciti $H_0$ i prihvatiti $H_1$, odnosno udio jedinica veći je od 51 posto i Ana neće ući u kafić.

# 6 | Test o kvantilima

Neka je $\mathbb{X} = (X_1, \ldots, X_n)$ j.s.u. iz funkcije distribucije $F$ slučajne varijable $X$ sa funkcijom kvantila definiranom kao $Q(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}, p \in (0,1)$, te ako je $F$ strogo rastuća i neprekidna, onda je $Q = F^{-1}$, što ćemo u nastavku pretpostaviti. Također, neka je $X$ neprekidna slučajna varijabla. Želimo testirati da je vrijednost $Q(p_0)$ jednaka unaprijed zadanoj vrijednosti $x_0$ za zadani $p_0$, odnosno

$$\mathcal{H}_0 : Q(p_0) = x_0.$$

Hipotezu ekvivalentno možemo zapisati kao

$$\mathcal{H}_0 : F(x_0) = p_0 \Leftrightarrow \mathcal{H}_0 : P(X \leq x_0) = p_0$$

pa test samim time svodimo na testiranje hipoteze

$$\mathcal{H}_0 : \theta = \theta_0,$$

gdje je $\theta = P(X \leq x_0)$ nepoznati parametar i $\theta_0 = p_0$. Zbog toga što je test egzaktan to ga čini primjenjivim i na male uzorke. Slijedi primjer alternativne hipoteze za dvostrani test:

$$\mathcal{H}_1 : Q(p_0) \neq x_0 \Leftrightarrow \mathcal{H}_1 : F(x_0) \neq p_0 \Leftrightarrow \mathcal{H}_1 : \theta \neq p_0.$$

**Primjer 5.** *Dogecoin je kriptovaluta koju su 6. prosinca 2013. godine osmislili Billy Markus i Jackson Palmer. Stvarajući ovu "šaljivu" kriptovalutu htjeli su se narugati nepredvidivosti ulaganja u kriptovalate, no neki su ovo vidjeli kao perspektivno ulaganje tako da se u Dogecoin može uložiti i danas. Cijena ove kriptovalute čak je u 2021. godini doživjela porast, a raste i danas. Medijan cijene do 2021. godine bio je* 0.027 *dolara. Možemo li tvrditi da je 2021. godine medijan cijene Dogecoin-a bio veći od* 0.027*?*

Koristimo bazu podataka *DogeCoin Historical Price Data*[1]. Želimo li provjeriti je li medijan cijene veći od 0.027 dolara, moramo se koristiti testom o kvantilima s pretpostavkama:

$$\mathcal{H}_0 : medijan = 0.027 \Leftrightarrow \mathcal{H}_0 : Q(0.5) = 0.027 \Leftrightarrow F(0.027) = 0.5 \Leftrightarrow P(X \leq 0.027) = 0.5,$$
$$\mathcal{H}_1 : medijan > 0.027 \Leftrightarrow \mathcal{H}_1 : Q(0.5) > 0.027 \Leftrightarrow F(0.027) < 0.5 \Leftrightarrow P(X \leq 0.027) < 0.5.$$

Sada sa $\theta$ označimo $P(X \leq 0.027)$ te slijedi

$$\mathcal{H}_0 : \theta = 0.5,$$
$$\mathcal{H}_1 : \theta < 0.5.$$

[1] Baza podataka pronađena je na *https://www.kaggle.com/*

Time smo se sveli na jednostrani test o proporciji. Kako promatramo vjerojatnost događaja $X \leq 0.027$, u bazi podataka 27 podataka zadovoljava taj uvjet, dakle realizacija test-statistike ovog testa je:

$$\hat{t}(\mathbf{x}) = \sum_{i=1}^{n} x_i = \sum_{i=1}^{365} x_i = 27.$$

Za kritično područje potreban nam je $k_2$ najveći element iz $\{0, 1, \ldots, n\}$ takav da vrijedi

$$P_{0.5}(T \leq k_1) = \sum_{k=0}^{k_2} \binom{365}{k} (0.5)0^k (0.5)^k \leq 0.05 = 166$$

pa je kritično područje $C = \{0, 1, \ldots, 166\}$. Vidimo kako se test-statistika realizira unutar kritičnog područja zbog čega bi zaključili da odbacujemo $\mathcal{H}_0$. Isti zaključak slijedi i pogledamo li $p$-vrijednost $2.2 \cdot 10^{-16} < 0.05$. Dakle, možemo tvrditi da je medijan cijene kriptovalute *Dogecoin* u 2021. godini veći od 0.027 dolara.

# 7 | Testiranje hipoteza o distribuciji

Neka je $\mathbb{X} = (X_1, \ldots, X_n)$ j.s.u. iz nepoznate funkcije distribucije $F$. Testiranje hipoteza o distribuciji zapravo se radi o testiranju jednakosti distribucije $F$ s funkcijom distribucije $F_0$:

$$\mathcal{H}_0 : F(x) = F_0(x), \text{ za svaki } x \in \mathbb{R}.$$

## 7.1 $\chi^2$ test

Ovaj test koristimo u slučaju kada imamo distribuciju $F_0$ diskretne slučajne varijable čija je slika konačna, odnosno kada vrijedi

$$F_0 \sim \begin{pmatrix} y_1 & \cdots & y_m \\ p_1^0 & \cdots & p_m^0 \end{pmatrix}.$$

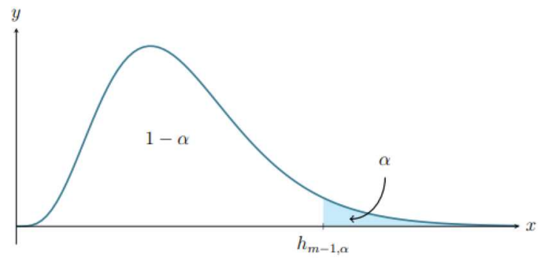
Sada ćemo definirati dvije veličine koje ćemo koristiti kod definiranja test-statistike:

- opažene ili empirijske frekvencije od $y_k$, $k = 1, \ldots, m$ definirane s $f_k = \sum_{i=1}^n I_{\{X_i = y_k\}}$, $k = 1, \ldots, m$.
- očekivane ili teorijske frekvencije od $y_k$, $k = 1, \ldots, m$ definirane s $np_k^0$, $k = 1, \ldots, m$.

Test-statistika je oblika

$$T(\mathbb{X}) = \sum_{k=1}^{m} \frac{(f_k - np_k^0)^2}{np_k^0}$$

koja u uvjetima istinitosti $\mathcal{H}_0$ asimptotski ima $\chi^2_{m-1}$ distribuciju. Test-statistika zapravo je suma kvadrata normalnih distribucija koje nisu nezavisne te $\chi^2_{m-1}$ distribuciju dobijemo limesom. $m$ ovdje označava veličinu slike, a ne cijelog uzorka. Možemo primjetiti da alternativnoj hipotezi $\mathcal{H}_1$ u prilog idu veće vrijednosti $T(\mathbb{X})$ pa je kritično područje oblika $C = [h_{m-1,\alpha}, \infty)$, gdje je $h_{m-1,\alpha} = Q_{\chi^2_{m-1}}(1-\alpha)$ $(1-\alpha)$-kvantil $\chi^2$ distribucije s $m-1$ stupnjeva slobode kao što je prikazano na idućoj slici:




Slika 7.1: Prikaz pronalaska $h_{m-1,\alpha}$

Napominje se da kod korištenja ovog testa treba voditi računa o idućem:

- ako je $F_0$ diskretna distribucija s beskonačnom slikom dosta opaženih frekvencija bit će 0 pa u tom slučaju realizacije bi se mogle kategorizirati tako da daju distribuciju s konačnom slikom pa nad tim uzorkom koristiti $\chi^2$ test
- test je asimptotski pa je bitno da aproksimacija distribucije test-statistike bude vrlo dobra (obično gledamo da očekivane frekvencije moraju biti barem 5). Za više detalja o testu pogledati Poglavlje 10 u [7].

**Primjer 6.** *U* 80 *bacanja nekog novčića, pismo se okrenulo* 53 *puta, a glava* 27 *puta. Možemo li na razini značajnosti* 0.05 *tvrditi da novčić nije pravilan?*

Neka je pad novčića na glavu događaj 0, a na pismo događaj 1. Ukoliko je novčić pravilan, glava i pismo realiziraju se s jednakom vjerojatnošću. Dakle, $P(0) = P(1) = 0.5$. Kako želimo testirati je li novčić pravilan, neka je $X$ distribucija pravilno izrađenog novčića. Tada su hipoteze:

$$\mathcal{H}_0 : X \sim \begin{pmatrix} 0 & 1 \\ 1/2 & 1/2 \end{pmatrix},$$

$$\mathcal{H}_1 : \text{ne } \mathcal{H}_0.$$

Neka je $f_1$ frekvencija padanja glave, a $f_2$ frekvencija pada pisma. Realizacija test-statistike jednaka je:

$$\begin{aligned}\hat{t}(\mathbf{x}) &= \sum_{k=1}^{m} \frac{(f_k - np_k^0)^2}{np_k^0} = \sum_{k=1}^{2} \frac{(f_k - np_k^0)^2}{np_k^0} = \frac{(f_1 - np_1^0)^2}{np_1^0} + \frac{(f_2 - np_2^0)^2}{np_2^0} \\ &= \frac{(27-40)^2}{40} + \frac{(53-40)^2}{40} = 8.45.\end{aligned}$$

Kritično područje oblika je $C = [h_{m-1,\alpha}, \infty)$ pri čemu je $h_{m-1,\alpha} = Q_{\chi^2_{m-1}}(1-\alpha) = Q_{\chi^2_1}(0.95) = 3.84$. Dakle, kritično područje je $C = [3.84, \infty)$. Zbog činjenice da se test-statistika realizira unutar kritičnog područja, kao i zbog činjenice da za $p$-vrijednost dobivamo $p = 0.00365 < 0.05$ odbacujemo $\mathcal{H}_0$ i prihvaćamo $\mathcal{H}_1$, odnosno možemo tvrditi da je novčić nepravilan.

## 7.2 Kolmogorov - Smirnovljev test

Kod uzoraka iz neprekidnih distribucija koristimo Kolmogorov-Smirnovljev (KS) test. Hipoteze su iduće:

$$\begin{aligned}\mathcal{H}_0 &: F(x) = F_0(x), \text{ za svaki } x \in \mathbb{R},\\ \mathcal{H}_1 &: \text{postoji } x \in \mathbb{R} \text{ takav da je } F(x) \neq F_0(x).\end{aligned}$$

Ovdje se test-statistika bazira na razlici između empirijske funkcije distribucije $\hat{F}_n(x)$ i pretpostavljene teorijske distribucije $F_0$. Test-statistiku definiramo kao:

$$D_n = \sup_{x\in\mathbb{R}} \left|\hat{F}(x) - F_0(x)\right|.$$

Kod uvjeta istinitosti hipoteze $H_0$ distribucija $D_n$ ovisi o veličini uzorka, ali ne o funkciji distribucije $F$. Za više detalja o testu pogledati Poglavlje 13 u [7].

**Primjer 7.** *Na razini značajnosti* 0.05 *testirajte dolazi li realizacija* $(0.09, 0.11, 0.26, 0.86, 0.83)$ *j.s.u.* $(X_1, X_2, X_3, X_4, X_5)$ *iz* $\mathcal{U}(0, 1)$ *distribucije.*

Uvedemo oznake:

- $F_0$ - funkcija distribucije $\mathcal{U}(0, 1)$,
- $F$ - funkcija distribucije iz koje dolazi zadan j.s.u.

Definiramo hipoteze koje želimo testirati:

$$\begin{aligned}\mathcal{H}_0 &: F = F_0 \text{ (uzorak dolazi iz } \mathcal{U}(0,1)),\\ \mathcal{H}_1 &: F \neq F_0 \text{ (uzorak ne dolazi iz } \mathcal{U}(0,1)).\end{aligned}$$

Za $p$-vrijednost dobivamo $p = 0.5073 > 0.05$ i ne odbacujemo $\mathcal{H}_0$. Kako je test egzaktan može se primjeniti i na ovako malom uzorku.

## 7.3 Testovi normalnosti

Kako se brojni testovi baziraju na pretpostavci da je distribucija uzorka normalna, nerijetko se pojavi ovakav problem kod hipoteza:

$$\begin{aligned}\mathcal{H}_0 &: \text{uzorak dolazi iz normalne distribucije,}\\ \mathcal{H}_1 &: \text{uzorak ne dolazi iz normalne distribucije.}\end{aligned}$$

Jedan od boljih testova normalnosti je Shapiro-Wilkov test. Za detaljniji opis testa pogledati [6].

**Primjer 8.** *U tablici Chesapea paketa BSDA nalaze se podaci mjerenja površinske slanosti rijeke Severn izvan obale Annapolisa, Maryland iz 1927. godine. Na razini značajnosti* 0.05 *ispitajte hipotezu o normalnoj distribuiranosti danog uzorka.*

Hipoteze koje testiramo:

$$\begin{aligned}\mathcal{H}_0 &: \text{površinska slanost ima normalnu distribuciju,}\\ \mathcal{H}_1 &: \text{površinska slanost nema normalnu distribuciju.}\end{aligned}$$

Dobivena $p$-vrijednost je $0.4973 > 0.05$ te ne odbacujemo $\mathcal{H}_0$.

# 8 | Zaključivanje temeljeno na dva uzorka

U slučaju promatranja dva uzorka, vrlo je bitno radi li se o nezavisnim ili vezanim slučajnim uzorcima.
Neka je $\mathbb{X} = (X_1, \ldots, X_{n_1})$ jednostavan slučajan uzorak iz neke distribucije i $\mathbb{Y} = (Y_1, \ldots, Y_{n_2})$ jednostavan slučajan uzorak iz neke distribucije. Kažemo da su ova dva uzorka nezavisna ako su sve $X_i, i = 1, \ldots, n_1$ i $Y_j, j = 1, \ldots, n_2$ međusobno nezavisne, $X_1, \ldots, X_{n_1}$ su jednako distribuirane, a $Y_1, \ldots, Y_{n_2}$ su jednako distribuirane.
Promatramo dva slučajna vektora $\mathbb{X}^{(1)} = (X_1^{(1)}, \ldots, X_n^{(1)})$ i $\mathbb{X}^{(2)} = (X_1^{(2)}, \ldots, X_n^{(2)})$. Kažemo da su to vezani (ili zavisni) slučajni uzorci ukoliko vrijedi da je $\left((X_1^{(1)}, X_1^{(2)}), \ldots, (X_n^{(1)}, X_n^{(2)})\right)$ dovdimenzionalan jednostavan slučajan uzorak, tj. u tom slučaju su $(X_1^{(1)}, X_1^{(2)}), \ldots, (X_n^{(1)}, X_n^{(2)})$ jednako distribuirani i nezavisni slučajni vektori. Važno je za napomenuti da su u tom slučaju i $\mathbb{X}^{(1)}$ i $\mathbb{X}^{(2)}$ jednostavni slučajni uzorci iz nekih distribucija, no to ne znači da su i $X_k^{(1)}$ i $X_k^{(2)}$ nezavisne.

## 8.1 Testovi za nezavisne slučajne uzorke

Uvedimo oznake i pretpostavke kojima ćemo se koristiti u slučaju nezavisnih uzoraka:

- $\mathbb{X} = (X_1, \ldots, X_{n_1})$ j.s.u. iz neke distribucije,
- $\mathbb{Y} = (Y_1, \ldots, Y_{n_2})$ j.s.u. iz neke distribucije,
- $\mathbb{X}$ i $\mathbb{Y}$ su nezavisni,
- $E[X_1] = \mu_1$, $E[Y_1] = \mu_2$.

Prvo ćemo se osvrnuti na testove o očekivanju za nezavisne slučajne uzorke. Razlika u odnosu na testiranje na jednom uzorku je ta što ovdje ne poznajemo niti $\mu_1$ niti $\mu_2$, dok smo u testiranjima s jednim uzorkom imali neku zadanu vrijednost s kojom smo uspoređivali.

### 8.1.1 Testovi o očekivanju za nezavisne slučajne uzorke s normalno distribuiranim populacijama i poznatim varijancama

Promatramo:

- $\mathbb{X} = (X_1, \ldots, X_n)$ jednostavan slučajan uzorak iz $\mathcal{N}(\mu_1, \sigma_1^2)$,
- $\mathbb{Y} = (Y_1, \ldots, Y_n)$ jednostavan slučajan uzorak iz $\mathcal{N}(\mu_2, \sigma_2^2)$,
- $\mathbb{X}$ i $\mathbb{Y}$ nezavisni,
- $\sigma_1^2 > 0$ i $\sigma_2^2 > 0$ poznati,
- $\mu_1 \in \mathbb{R}$ i $\mu_2 \in \mathbb{R}$ nepoznati parametri.

Test-statistika kojom se koristimo ovdje jest:

$$T(\mathbb{X}, \mathbb{Y}) = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

gdje su $\overline{X}_{n_1} = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i$, $\overline{Y}_{n_2} = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_i$.

Imamo $n$ slučajnih varijabli koje dolaze iz j.s.u. čija je distribucija normalna, odnosno $\mathcal{N}(\mu, \sigma^2)$. Kako sve varijable dolaze iz j.s.u. lako izračunamo kakvu aritmetička sredina ima distribuciju. Aritmetičku sredinu označit ćemo s $\overline{X_n}$ te idemo izračunati očekivanje.

$$\mathbf{E}[\overline{X_n}] = \mathbf{E}\left[\frac{1}{n} \sum_{i=1}^{n} (X_i)\right] = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{E}[X_i]) = \frac{1}{n} \cdot n \cdot \mu = \mu.$$

Sada još trebamo izračunati varijancu aritmetičke sredine,

$$Var(\overline{X_n}) = Var\left(\frac{1}{n} \sum_{i=1}^{n} (X_i)\right) = \frac{1}{n^2} \sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2} \cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}.$$

Time smo pokazali kako aritmetička sredina ima $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ distribuciju [tvrdnja preuzeta iz [2]].

Prema prethodnoj tvrdnji vrijedi:

- $\overline{X}_{n_1} \sim \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$,
- $\overline{Y}_{n_2} \sim \mathcal{N}\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$.

Činjenica da su $\overline{X}_{n_1}$ i $\overline{Y}_{n_2}$ međusobno nezavisne slijedi iz tog da su definirane kao sume međusobno nezavisnih varijabli. Razlika normalnih nezavisnih varijabli ponovno je normalno distribuirana, tj. $\overline{X}_{n_1} - \overline{Y}_{n_2} \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$. Stoga testove o očekivanju za dva nezavisna uzorka zapravo svodimo na isti $z$-test koji koristimo za testove o očekivanju za jedan uzorak te pri tome u alternativnoj hipotezi uspoređujemo s 0, odnosno:

- u slučaju dvostranog testa hipoteze su:

$$\mathcal{H}_0 : \mu_1 - \mu_2 = 0 \; (\textit{zapravo gledamo je li } \mu_1 = \mu_2),$$
$$\mathcal{H}_1 : \mu_1 - \mu_2 \neq 0 \; (\textit{zapravo gledamo je li } \mu_1 \neq \mu_2).$$

- U slučaju jednostranih testova hipoteze su:

a)

$$\mathcal{H}_0 : \mu_1 - \mu_2 = 0 \; (\textit{zapravo gledamo je li } \mu_1 = \mu_2),$$
$$\mathcal{H}_1 : \mu_1 - \mu_2 > 0 \; (\textit{zapravo gledamo je li } \mu_1 > \mu_2);$$

b)

$$\mathcal{H}_0 : \mu_1 - \mu_2 = 0 \; (\textit{zapravo gledamo je li } \mu_1 = \mu_2),$$
$$\mathcal{H}_1 : \mu_1 - \mu_2 < 0 \; (\textit{zapravo gledamo je li } \mu_1 < \mu_2).$$

### 8.1.2 Testovi o očekivanju za nezavisne slučajne uzorke s normalno distribuiranim populacijama i nepoznatim varijancama

Neka vrijedi sljedeće:

- $\mathbb{X} = (X_1, \ldots, X_{n_1})$ jednostavan slučajan uzorak iz $\mathcal{N}(\mu_1, \sigma_1^2)$,
- $\mathbb{Y} = (Y_1, \ldots, Y_{n_2})$ jednostavan slučajan uzorak iz $\mathcal{N}(\mu_2, \sigma_2^2)$,
- $\mathbb{X}$ i $\mathbb{Y}$ nezavisni,
- $\sigma_1^2 > 0$ i $\sigma_2^2 > 0$ nepoznati,
- $\mu_1 \in \mathbb{R}$ i $\mu_2 \in \mathbb{R}$ nepoznati parametri.

Test-statistika definirana je kao:

$$T(\mathbb{X}, \mathbb{Y}) = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\sqrt{\frac{S_{n_1}^2}{n_1} + \frac{S_{n_2}^2}{n_2}}}$$

gdje su $\overline{X}_{n_1} = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i$, $\overline{Y}_{n_2} = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_i$, $S_{n_1}^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \overline{X}_{n_1})^2$, $S_{n_2}^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (Y_j - \overline{Y}_{n_2})^2$. Primjećujemo sličnost test-statistike s prethodnim testom, no razlika je u tome što su nam varijance nepoznate te koristimo korigirane varijance uzorka $S_{n_1}^2$ i $S_{n_2}^2$. Ukoliko je $\mathcal{H}_0$ istinita, test-statistika $T$ ima Studentovu distribuciju s $\nu$ stupnjeva slobode, odnosno $T \sim \mathcal{T}_\nu$ gdje

$$\nu = \frac{\left( \frac{s_{n_1}^2}{n_1} + \frac{s_{n_2}^2}{n_2} \right)^2}{\frac{s_{n_1}^4}{n_1^2 (n_1 - 1)} + \frac{s_{n_2}^4}{n_2^2 (n_2 - 1)}}$$

dobijemo Welch–Satterthwaitovom formulom te se ovaj test često i zove Welchov $t$-test. On je varijacija na klasični $t$-test, no pouzdaniji je u slučaju nejednakih varijanci i različitih veličina uzoraka koje promatramo. Ostatak postupka provodi se analogno postupku kod $t$-testa na jednom uzorku. Za više detalja o testu pogledati poglavlje 11 u [1].

### 8.1.3 Testovi o očekivanju za velike nezavisne slučajne uzorke

Ukoliko je $n_1 > 30$ i $n_2 > 30$, onda uzorke smatramo velikima. U tom slučaju se za isptivanje jednakosti očekivanja možemo poslužiti Welchovim $t$-testom. No, ponovno kao i kod testova o očekivanju na jednom velikom uzorku, moramo biti oprezni kod donošenja zaključaka.

**Primjer 9.** *Tablica Concrete paketa BSDA sadrži podatke čvrstoće betonskih blokova proizvedenih dvama različitim metodama (old i new). Možemo li na razini značajnosti 0.05 tvrditi da nova metoda(new) očekivano proizvodi čvršće blokove?*

Hipoteze koje testiramo su:

$$\mathcal{H}_0 : \mu_1 = \mu_2,$$
$$\mathcal{H}_1 : \mu_1 > \mu_2.$$

gdje su

- $X_{n_1}$ - čvrstoća blokova proizvedenih novom metodom,
- $X_{n_2}$ - čvrstoća blokova proizvedenih starom metodom,
- $\mu_1$ - očekivana čvrstoća blokova proizvedenih novom metodom,
- $\mu_2$ - očekivana čvrstoća blokova proizvedenih starom metodom.

Za potrebe izračuna realizacije test-statistike računamo:

- $\overline{x}_{n_1} = \dfrac{1}{n_1}\sum_{i=1}^{n_2} x_i = \dfrac{1}{10}\sum_{i=1}^{10}(152 + \ldots + 160) = 147.2,$
- $\overline{x}_{n_2} = \dfrac{1}{n_1}\sum_{i=1}^{n_1} x_i = \dfrac{1}{10}\sum_{i=1}^{10}(132 + \ldots + 137) = 136.1,$
- $s^2_{n_1} = \dfrac{1}{n_2 - 1}\sum_{i=1}^{n_2}(x_i - \overline{x}_{n_2})^2 = \dfrac{1}{10-1}\sum_{i=1}^{10}(x_i - 147.2)^2 = 81.02,$
- $s^2_{n_2} = \dfrac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_i - \overline{x}_{n_1})^2 = \dfrac{1}{10-1}\sum_{i=1}^{10}(x_i - 136.1)^2 = 72.77.$

Realizacija test-statistike sada je jednaka:

$$\hat{t}(\mathbf{x},\mathbf{y}) = \frac{\overline{x}_{n_1} - \overline{x}_{n_2}}{\sqrt{\frac{s^2_{n_1}}{n_1} + \frac{s^2_{n_2}}{n_2}}} = \frac{147.2 - 136.1}{\sqrt{\frac{81.022}{10} + \frac{72.77}{10}}} = 2.830.$$

Za kritično područje potrebna nam je Studentova distribucija s $\nu$ stupnjeva slobode gdje

$$\nu = \frac{\left(\frac{s^2_{n_1}}{n_1} + \frac{s^2_{n_2}}{n_2}\right)^2}{\frac{s^4_{n_1}}{n_1^2(n_1-1)} + \frac{s^4_{n_2}}{n_2^2(n_2-1)}} = \frac{\left(\frac{81.02}{10} + \frac{72.77}{10}\right)^2}{\frac{81.02^2}{10^2\cdot 9} + \frac{72.77^2}{10^2\cdot 9}} = 17.948.$$

Kritično područje jednako je $C = [t_{\nu,\alpha}, \infty\rangle$ pri čemu je $t_{\nu,\alpha}$ $(1-\alpha)$-kvantil $\mathcal{T}_\nu$ distribucije. Dakle, naše kritično područje jednako je $C = [t_{18,\,0.05}, \infty\rangle = [1.734, \infty\rangle$. Vidimo kako se test-statistika realizira unutar kritičnog područja pa odbacujemo nulltu hipotezu i tvrdimo da nova metoda proizvodi čvršće blokove. Do istog zaključka dolazimo i preko dobivene $p$-vrijednosti koja iznosi $0.01112 < 0.05$.

### 8.1.4 Test o proporciji za nezavisne slučajne uzroke

Pretpostavke koje imamo ovdje su sljedeće:

- $\mathbb{X} = (X_1, \ldots, X_n)$ jednostavan slučajan uzorak iz Bernoullijeve distribucije

$$\begin{pmatrix} 0 & 1 \\ 1-\theta_1 & \theta_1 \end{pmatrix},$$

- $\mathbb{Y} = (Y_1, \ldots, Y_n)$ jednostavan slučajan uzorak iz Bernoullijeve distribucije

$$\begin{pmatrix} 0 & 1 \\ 1-\theta_2 & \theta_2 \end{pmatrix},$$

,

- $\mathbb{X}$ i $\mathbb{Y}$ nezavisni,
- $\theta_1, \theta_2 \in \langle 0, 1\rangle$.

Zanima nas je li vjerojatnost događaja od interesa jednaka u prvoj i drugoj populaciji ili ne, odnosno hipoteze koje promatramo bit će:

$$\mathcal{H}_0 : \theta_1 = \theta_2,$$
$$\mathcal{H}_1 : \theta_1 \neq \theta_2.$$

Test-statistika za ovaj problem je

$$T(\mathbb{X},\mathbb{Y}) = \frac{\left(\overline{X}_{n_1} - \overline{Y}_{n_2}\right)^2}{\frac{n_1\overline{X}_{n_1} + n_2\overline{Y}_{n_2}}{n_1+n_2}\left(1 - \frac{n_1\overline{X}_{n_1} + n_2\overline{Y}_{n_2}}{n_1+n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

U uvjetima istinitosti hipoteze $\mathcal{H}_0$, $\sqrt{T(\mathbb{X},\mathbb{Y})}$ ima asimptotski $\mathcal{N}(0,1)$ distribuciju. Ukoliko je $K \sim \mathcal{N}(0,1)$, onda je $K^2 \sim \chi^2(1)$.
Koristeći prethodnu tvrdnju dolazimo do zaključka da u uvjetima istinitosti $\mathcal{H}_0$ mora vrijediti $T(\mathbb{X},\mathbb{Y}) \sim \chi^2(1)$. Test se nadalje može provoditi kao $z$-test ili kao $\chi^2$-test. Ne smijemo zaboraviti da je test asimptotski te da stoga trebamo velike uzorke kako bi rezultati bili valjani.

**Primjer 10.** *U bazi podataka Students Performance Dataset*[1] *nalaze se podaci o* 2392 *srednjoškolca. Je li na razini značajnosti* 0.05 *proporcija prosjeka ocjena većeg od* 3 *različita među djevojčicama i dječacima?*

U našem uzorku nalazi se 1170 dječaka i 1222 djevojčice. Među njima 161 djevojčica te 160 dječaka imaju prosjek veći od 3. Neka je $\theta_1$ udio dj
Definiramo:

- $\theta_1$ - udio djevojčica kojima je prosjek ocjena veći od 3 među djevojčicama,
- $\theta_2$ - udio dječaka kojima je prosjek ocjena veći od 3 među dječacima.

Hipoteze su iduće:

$$\begin{aligned} \mathcal{H}_0 &: \theta_1 = \theta_2, \\ \mathcal{H}_1 &: \theta_1 \neq \theta_2. \end{aligned}$$

Označimo li s 1 događaj kada je prosjek veći od 3, distribuciju djevojčica možemo zapisati kao:

$$X_i \sim \begin{pmatrix} 0 & 1 \\ 1-\theta_1 & \theta_1 \end{pmatrix},$$

a dječaka

$$Y_i \sim \begin{pmatrix} 0 & 1 \\ 1-\theta_1 & \theta_2 \end{pmatrix}.$$

Za izračun realizacije test-statistike koristimo iduće:

- $(\overline{x}_{n_1} - \overline{y}_{n_2})^2 = \left(\dfrac{161}{1222} - \dfrac{160}{1170}\right)^2 = 2.500909 \cdot 10^{-5}$,
- $\dfrac{n_1\overline{x}_{n_1} + n_2\overline{y}_{n_2}}{n_1+n_2} = \dfrac{161+160}{1222+1170} = 0.1341973$,
- $\dfrac{1}{n_1} + \dfrac{1}{n_2} = \dfrac{1}{1222} + \dfrac{1}{1170} = 0.001673031$.

Sada imamo:

$$\hat{t}(\mathbf{x},\mathbf{y}) = \frac{2.500909 \cdot 10^{-5}}{0.1341973 \cdot (1 - 0.1341973) \cdot 0.001673031} = 0.1286563.$$

[1] Baza podataka pronađena je na *https://www.kaggle.com/*.

Kritično područje oblika je $C = [h_{m-1,\alpha}, \infty)$ pri čemu je $h_{m-1,\alpha} = Q_{\chi^2_{m-1}}(1-\alpha) = Q_{\chi^2_1}(0.95) = 3.84$. Kako je naša test-statistika 0.1286563, a kritično područje je $C = [3.84, \infty)$, vidimo kako se test-statistika ne realizira unutar kritičnog područja pa ne možemo odbaciti nul hipotezu. Dobivena $p$-vrijednost jednaka je $0.6174 > 0.05$ i potrvđuje naš prethodni zaključak. Ne odbacujemo nul hipotezu, odnosno ne možemo tvrditi da su proporcije onih koji imaju prosjek veći od 3 različite među dječacima i djevojčicama.

### 8.1.5 Test o lokacijskom pomaku

Pretpostavke koje koristimo u ovom slučaju:

- $\mathbb{X} = (X_1, \ldots, X_{n_1})$ j.s.u. iz neprekidne distribucije $F_X$ slučajne varijable $X$,
- $\mathbb{Y} = (Y_1, \ldots, Y_{n_2})$ j.s.u. iz neprekidne distribucije $F_Y$ slučajne varijable $Y$,
- $\mathbb{X}$ i $\mathbb{Y}$ nezavisni,
- $F_X$ i $F_Y$ imaju isti oblik, ali mogu se razlikovati za lokacijski pomak.

Lokacijski pomak označimo s $\delta \in \mathbb{R}$. Zadnju pretpostavku možemo zapisati kao

$$P(X \leq x) = F_X(x) = F_Y(x-\delta) = P(Y \leq x-\delta) = P(Y+\delta \leq x), \ \forall x \in \mathbb{R}$$

iz čega vidimo da je $X$ jednako distribuiran kao $Y+\delta$, a to vrijedi ako i samo ako je $X-\delta$ jednako distribuirano kao $Y$. Stoga se testiranje hipoteze o distribucijama svede na testiranje broja.

- dvostrani test

$$\begin{aligned}
&\mathcal{H}_0 : \delta = 0, \\
&\mathcal{H}_1 : \exists\, \delta \in \mathbb{R} \text{ } \textit{takav da } F_X(x) = F_Y(x-\delta), \ \forall x \in \mathbb{R},
\end{aligned}$$

- jednostrani testovi

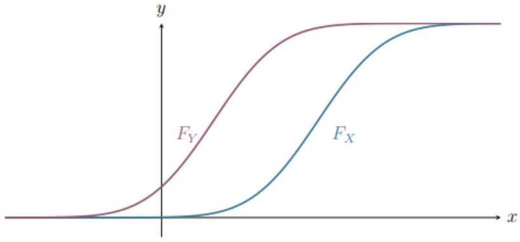
  a)

$$\begin{aligned}
&\mathcal{H}_0 : \delta = 0, \\
&\mathcal{H}_1 : \exists\, \delta > 0 \text{ } \textit{takav da } F_X(x) = F_Y(x-\delta), \ \forall x \in \mathbb{R},
\end{aligned}$$

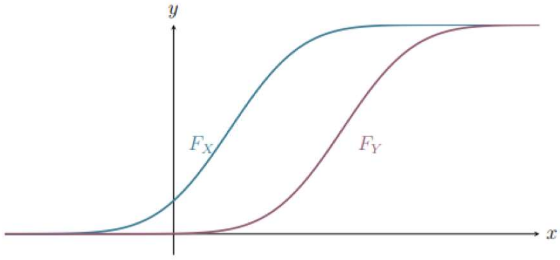
  b)

$$\begin{aligned}
&\mathcal{H}_0 : \delta = 0, \\
&\mathcal{H}_1 : \exists\, \delta < 0 \text{ } \textit{takav da } F_X(x) = F_Y(x-\delta), \ \forall x \in \mathbb{R}.
\end{aligned}$$

Slučaj pod a) zovemo pozitivan lokacijski pomak, dok je slučaj b) negativan lokacijski pomak. Iz $X \overset{d}{=} Y + \delta$ slijedi $E[X] = E[Y] + \delta \iff \mu_1 = \mu_2 + \delta \iff \mu_1 - \mu_2 = \delta$. Donošenje zaključaka koji idu u korist alternativnoj hipotezi testa implicira isti zaključak i kod $t$-testa. Ovaj test zovemo Mann-Whitney-Wilcoxonov test i on je alternativa klasičnog $t$-testa. Prednosti su te što je egzaktan (pa se kao takav može se koristiti i za male uzorke) te nema nikakve pretpostavke na distribuciju osim na njezin oblik. Za više detalja o samom testu pogledati poglavlje Mann-Whitney-Wilcoxonov test u [4].




Slika 8.1: Funkcija distribucije za slučaj pod a)




Slika 8.2: Funkcija distribucije za slučaj pod b)

**Primjer 11.** *U tablici iris iz paketa datasets u R-u nalaze se podaci o širini i duljini čašice i latice triju vrsta irisa: setosa, virginica i versicolor. Zanima nas možemo li na razini značajnosti* 0.05 *tvrditi da je širina latice irisa setosa manja od širine latice irisa virginica?*

Za svaku vrstu irisa imamo 50 podataka o širinama latice. Neka je s $F_X$ označena distribucija širine latice irisa *setosa*, a s $F_Y$ distibucija širine latice irisa *virginica*. Hipoteze koje testiramo:

$$\mathcal{H}_0 : F_X(x) = F_Y(x), \text{ za sve } x \in \mathbb{R},$$
$$\mathcal{H}_1 : \text{za } \delta < 0,\ F_X(x) = F_Y(x - \delta), \text{ za sve } x \in \mathbb{R}.$$

Možemo to zapisati i kao:

$$\mathcal{H}_0 : X \overset{d}{=} Y,$$
$$\mathcal{H}_1 : \text{za } \delta < 0, X \overset{d}{=} Y + \delta.$$

Za $p$-vrijednost dobivamo $p = 2.2 \cdot 10^{-16} < 0.05$ pa na razini značajnosti 0.05 odbacujemo hipotezu $\mathcal{H}_0$ i tvrdimo da je širina latice irisa *septosa* manja od širine latice irisa *virginica*.

### 8.1.6 Testovi o distribuciji za nezavisne slučajne uzorke

Razlikujemo dva testa koja koristimo za testiranje distribucija nezavisnih slučajnih uzoraka, a to su:

- $\chi^2$-test o homogenosti
  Ovaj test nam omogućuje testiranje jednakosti diskretnog obilježja u $l \geq 2$ populacija. Iz svake od populacija dolazi po jedan uzorak:

- populacija 1 - $(X_1^{(1)}, \dots, X_{n_1}^{(1)})$ j.s.u. iz $\begin{pmatrix} y_1, \dots, y_n \\ p_1^{(1)}, \dots, p_n^{(1)} \end{pmatrix}$,
- $\vdots$
- populacija $l$ - $(X_1^{(l)}, \dots, X_{n_l}^{(l)})$ j.s.u. iz $\begin{pmatrix} y_1, \dots, y_n \\ p_1^{(l)}, \dots, p_n^{(l)} \end{pmatrix}$,

i svi uzorci su međusobno nezavisni. U ovom testu potrebna nam je sljedeća tablica:

| | $y_1$ | $y_2$ | $\dots$ | $y_n$ | |
|---|---|---|---|---|---|
| 1 | $f_{11}$ | $f_{12}$ | $\dots$ | $f_{1n}$ | $n_1$ |
| 2 | $f_{21}$ | $f_{22}$ | $\dots$ | $f_{2n}$ | $n_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $l$ | $f_{l1}$ | $f_{l2}$ | $\dots$ | $f_{ln}$ | $n_l$ |
| | $f_1$ | $f_2$ | $\dots$ | $f_n$ | $N$ |

Tablica 8.1: Tablica frekvencija

gdje su

- $f_{ij}$ - frekvencija $y_j$ u $i$-toj populaciji,
- $f_j = \sum_{i=1}^{n} f_{ij}$ - frekvencija $y_j$ u realizacijama svih uzoraka zajedno,
- $N = \sum_{i=1}^{m} n_i$ - veličina svih uzoraka zajedno.

Hipoteze koje mi želimo testirati su:

$$\mathcal{H}_0 : p_j^{(1)} = p_j^{(2)} = \dots = p_j^{(l)}, \; j = 1, \dots, n,$$
$$\mathcal{H}_1 : ne\ \mathcal{H}_0.$$

Ukoliko je $\mathcal{H}_0$ istinita, mogli bi napraviti procjenu: $\widehat{p}_j = \frac{f_j}{N}$, $j = 1, \dots, n$, a očekivana frekvencija $y_j$ u $i$-toj populaciji je onda $\widehat{p}_j n_i$, $i = 1, \dots, l$, $j = 1, \dots, n$. Test-statistika definirana je kao

$$D_{l,n} = \sum_{i=1}^{l} \sum_{j=1}^{n} \frac{(f_{ij} - \widehat{p}_j n_i)^2}{\widehat{p}_j n_i}$$

pri čemu treba obratiti pozornost na to da su $f_{ij}$ opažene frekvencije, a $\widehat{p}_j n_i$ očekivane frekvencije iz čega vidimo sličnost s $\chi^2$-testom za jedan uzorak. Kritično područje oblika je $C = [h_{(n-1)(l-1),\alpha,} \infty)$, gdje $h_{(n-1)(l-1),\alpha} = Q_{\chi^2_{(n-1)(l-1)}}(1-\alpha)$ predstavlja $(1-\alpha)$-kvantil $\chi^2$ distribucije s $(n-1)(l-1)$ stupnjeva slobode. Ukoliko za test-statistiku dobijemo veće vrijednosti, vidimo da ona ide u prilog alternativnoj hipotezi $\mathcal{H}_1$. Trebamo imati na umu da očekivane frekvencije trebaju biti barem 5 kako bi mogli provoditi ovaj test. Za više detalja o ovom testu pogledati poglavlje *The Chi-Square Test of Homogenity* u [5].

- Kolomogorov-Smirnovljev test
  Pretpostavke koje koristimo u ovom slučaju:
  - $\mathbb{X} = (X_1, \ldots, X_{n_1})$ j.s.u. iz neprekidne distribucije $F_X$ slučajne varijable $X$,
  - $\mathbb{Y} = (Y_1, \ldots, Y_{n_2})$ j.s.u. iz neprekidne distribucije $F_Y$ slučajne varijable $Y$,
  - $\mathbb{X}$ i $\mathbb{Y}$ nezavisni,

  a hipoteze koje provjeravamo su:

$$\mathcal{H}_0 : F_X(x) = F_Y(x),\ \forall x \in \mathbb{R},$$
$$\mathcal{H}_1 : \exists\, x \in \mathbb{R} \text{ takav da } F_X(x) \neq F_Y(x).$$

  Test-statistika temelji se na empirijskim funkcijama distribucije i zadana je s

$$D_{n_1,n_2} = \sup_{x\in\mathbb{R}} |\widehat{F}_X(x) - \widehat{F}_Y(x)|$$

  pri čemu su:

  - $\widehat{F}_X(x) = \dfrac{1}{n_1}\sum_{i=1}^{n_1} I_{\{X_i \leq x\}}$,
  - $\widehat{F}_Y(x) = \dfrac{1}{n_2}\sum_{i=1}^{n_2} I_{\{Y_i \leq x\}}$.

  Prednosti ovog testa su te što je egzaktan te je primjenjiv za proizvoljne distribucije. Za više detalja o testu pogledati poglavlje X.2. u [4].

**Primjer 12.** *Pretpostavimo da je provedeno istraživanje o sportskoj aktivnosti studenata na četiri sveučilišne sastavnice Sveučilišta u Osijeku - Ekonomski fakultet, Fakultet primjenjene matematike i informatike, Medicinski fakultet i Filozofski fakultet. Neka su rezultati dani tablicom 8.2:*

| studenti | sportaši | nesportaši | $\sum$ |
|---|---|---|---|
| matematike | 105 | 245 | 350 |
| ekonomije | 36 | 164 | 200 |
| filozofskog | 231 | 69 | 300 |
| medicine | 65 | 135 | 200 |
| $\sum$ | 437 | 613 | 1050 |

Tablica 8.2: Raspodjela studenata po fakultetima i sportskoj aktivnosti

*pri čemu su sportaši oni koji se bave sportskom aktivnosti barem jednom tjedno, a nesportaši oni koji se ne bave sportskom aktivnosti uopće. Možemo li na razini značajnosti* 0.05 *tvrditi da je sportska aktivnost različito zastupljena među studentima na nekom od promatranih fakulteta?*

Hipoteze koje testiramo su:

$\mathcal{H}_0$: sportska aktivnost jednako je zastupljena među studentima svih promatranih fakulteta,

$\mathcal{H}_1$: sportska aktivnost nije jednako zastupljena među studentima svih promatranih fakulteta.

Najprije ćemo izračunati $\widehat{p}_j n_i$ za studente matematike. Dakle, ako promatramo prvo sportaše imamo:

$$\widehat{p}_j n_i = \widehat{p}_1 n_1 = \frac{f_1}{N} \cdot n_1 = \frac{437}{1050} \cdot 350 = 145.67.$$

Kako je naša test-statistika oblika

$$D_{4,2} = \sum_{i=4}^{l} \sum_{j=2}^{n} \frac{(f_{ij} - \widehat{p}_j n_i)^2}{\widehat{p}_j n_i}$$

potrebno nam je za matematičare sportaše još izračunati:

$$\frac{(f_{11} - \widehat{p}_1 n_1)^2}{\widehat{p}_1 n_1} = \frac{(105 - 145.67)^2}{145.67} = 11.35.$$

Sada imamo prvog člana sume. Ostalo računamo analogno i dobivamo:

$$D_{4,2} = 11.35 + 8.09 + 26.81 + 19.11 + 91.27 + 64.69 + 82.97 + 2.85 = 307.14,$$

a kritično područje je oblika $C = [h_{(4-1)(2-1),\alpha}, \infty) = [7.81, \infty)$. Kako se test-statistika realizira unutar kritičnog područja, zaključak je da odbacujemo nultu hipotezu. To nam potvrđuje i $p$-vrijednost jednaka $p = 2.2 \cdot 10^{-16} < 0.05$ pa na razini značajnosti 0.05 odbacujemo $\mathcal{H}_0$ i prihvaćamo $\mathcal{H}_1$, odnosno možemo tvrditi da je sport više zastupljen među studentima na nekom od promatranih fakulteta.

## 8.2 Testovi za vezane slučajne uzorke

### 8.2.1 Testovi o očekivanju za vezane slučajne uzorke

Promatramo slijedeće slučajne uzorke:

- $\mathbb{X}^1 = (X_1^{(1)}, \ldots, X_n^{(1)})$,
- $\mathbb{X}^2 = (X_1^{(2)}, \ldots, X_n^{(2)})$,
- $X_i^{(1)}$ i $X_i^{(2)}, i = 1, \ldots, n$ nezavisni i jednako distribuirani

takve da $((X_1^1, X_1^{(2)}), \ldots, (X_n^1, X_n^{(2)}))$ dvodimenzionalan j.s.u. te uzorci ne moraju biti nezavisni.
Testiramo hipotezu:

$$\mathcal{H}_0 : \mu_1 = \mu_2,$$

gdje je $\mu_1 = EX_i^{(1)}$, a $\mu_2 = EX_i^{(2)}$. Promatranjem slučajnog vektora razlika $D_i = X_i^{(1)} - X_i^{(2)}$ za $i = 1, \ldots, n$ dobivamo da je $(D_1, \ldots, D_n)$ j.s.u. te vrijedi $\mu_D = \mu_1 - \mu_2$, pa danu hipotezu možemo zapisati kao:

$$\mathcal{H}_0 : \mu_D = 0.$$

Ovime problem svodimo na test o očekivanju na jednom uzorku te možemo koristiti slijedeće testove ovisno o tome je li uzorak velik ili su razlike normalno distribuirane:

- $z$-test ako je varijanca distribucije razlika poznata,
- $t$-test ako je varijanca distribucije razlika nepoznata.

Važno je napomenuti da se test ne može primijeniti na uzorcima koji su različitih veličina. Također, ako su oba uzorka iz normalne distribucije, to nužno ne mora značiti da je i njihova razlika normalno distribuirana.

### 8.2.2 Test o proporciji za vezane slučajne uzorke

Imamo slijedeće pretpostavke:

- $\mathbb{X}^{(1)} = (X_1^{(1)}, \ldots, X_n^{(1)})$ jednostavan slučajan uzorak iz Bernoullijeve distribucije
$$\begin{pmatrix} 0 & 1 \\ 1-\theta_1 & \theta_1 \end{pmatrix}.$$
- $\mathbb{X}^{(2)} = (X_1^{(2)}, \ldots, X_n^{(2)})$ jednostavan slučajan uzorak iz Bernoullijeve distribucije
$$\begin{pmatrix} 0 & 1 \\ 1-\theta_2 & \theta_2 \end{pmatrix}.$$
- $\left((X_1^{(1)}, X_1^{(2)}), \ldots, (X_n^{(1)}, X_n^{(2)})\right)$ dvodimenzionalan jednostavan slučajan uzorak.

Za testiranje hipoteze koristimo test za razliku između proporcija u vezanim uzorcima:

$$\mathcal{H}_0 : \theta_1 = \theta_2.$$

Definiramo tablicu frekvencija kao:

| $X^{(1)} \backslash X^{(2)}$ | 0 | 1 | $\sum$ |
|---|---|---|---|
| 0 | $a$ | $b$ | $a+b$ |
| 1 | $c$ | $d$ | $c+d$ |
| $\sum$ | $a+c$ | $b+d$ | $n$ |

Tablica 8.3: Tablica frekvencija

Za zaključivanje su bitni parovi $(0,1)$ i $(1,0)$, odnosno $a$ i $d$ nisu bitni za zaključivanje o tome da li se razlikuju $\theta_1$ i $\theta_2$. U uvjetima istinitosti $\mathcal{H}_0$ broj pozitivnih razlika (parovi $(1,0)$) među ne-nul razlikama imaju $\mathcal{B}(b+c, \frac{1}{2})$ distribuciju pa se test može provesti kao test o proporciji na jednom uzorku (binomni test). Uspjesi se računaju kao $c$ od $b+c$ ponavljanja te se uspoređuju sa parametrom $\frac{1}{2}$.

**Primjer 13.** *Pretpostavimo da je provedeno istraživanje na 60 nogometaša o tome jesu li doživjeli ozbiljniju povredu koju su morali liječiti te im je isto pitanje postavljeno kada su trenirali 10 godina. Njih 15 je imalo povredu sa 5 godina, 11 sa 10 godina, 20 oba puta i 14 nije uopće bilo povrijeđeno. Možemo li na razini značajnosti 0.05 tvrditi da je pojavnost ozlijede veća što se dulje trenira?*

Ovdje se radi o usporedbi proporcija u vezanim uzorcima. Označimo li:

- $\mathbb{X}^{(1)}$ - ozlijeđeni nakon 5 godina treniranja,
- $\theta_1$ - proporcija ozlijeđenih nakon 5 godina treniranja,
- $\mathbb{X}^{(2)}$ - ozlijeđeni nakon 10 godina treniranja,
- $\theta_2$ - proporcija ozlijeđenih nakon 10 godina treniranja,

hipoteze koje imamo su:

$$\mathcal{H}_0 : \theta_1 = \theta_2,$$
$$\mathcal{H}_1 : \theta_1 < \theta_2.$$

Ako s 0 označimo događaj da igrač nije bio ozlijeđen promatrane godine, a s 1 da je bio ozlijeđen promatrane godine, tablica frekvencija oblika je:

| $X^{(1)} \backslash X^{(2)}$ | 0 | 1 | $\sum$ |
|---|---|---|---|
| 0 | 14 | 11 | 25 |
| 1 | 15 | 20 | 35 |
| $\sum$ | 29 | 31 | 60 |

Tablica 8.4: Tablica frekvencija postavljenog problema

Kako smo naveli u teoretskom djelu testa, analizu možemo provesti s testom o proporciji na jednom uzorku (binomni test). Uspjesi se računaju kao $c$ od $b+c$ ponavljanja te se uspoređuju sa parametrom $\frac{1}{2}$, odnosno u našem slučaju uspjesi su 15 od $15+11=26$ ponavljanja. Hipoteza koju testiramo je:

$$\mathcal{H}_0 : \theta_1 = \theta_2 \Leftrightarrow \mathcal{H}_0 : \hat{\theta} = 0.5,$$
$$\mathcal{H}_1 : \theta_1 < \theta_2 \Leftrightarrow \mathcal{H}_0 : \hat{\theta} < 0.5.$$

Vrijednost test-statistike i kritičnog područja izračunali bi analogno kao u Primjeru 4. Za $p$-vrijednost dobivamo $0.8365 > 0.05$ pa ne odbacujemo $\mathcal{H}_0$, odnosno ne možemo tvrditi da je pojavnost ozljede veća što se dulje trenira.

### 8.2.3 $\chi^2$ test o nezavisnosti

Promatramo slijedeće:

- $(X_i, Y_i)$ za $i = 1. \ldots, N$ nezavisni slučajni vektori,
- $(X, Y)$ je slučajan vektor pri čemu su nezavisni slučajni vektori $(X_i, Y_i)$ za svaki $i = 1, \ldots, N$ jednako distribuirani kao slučajan vektor,
- $X$ i $Y$ varijable, ne nužno nezavisne.

Za dani $j \in \{1, \ldots, n\}$ uvjetnu vjerojatnost definiramo sa:

$$p_{x_i|y_j} = P(X = x_i|Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} = \frac{p_{ij}}{p_{y_j}}, \; j = 1, \ldots, m,$$

a za dani $i \in \{1, \ldots, m\}$

$$p_{y_j|x_i} = P(Y = y_j|X = x_i) = \frac{P(X = x_i, Y = y_j)}{P(X = x_i)} = \frac{p_{ij}}{p_{x_i}}, \; i = 1, \ldots, n$$

gdje su

- $p_{x_i} = P(X = x_i) = \sum_{j=1}^{n} p_{ij}, \; i = 1, \ldots, m$ ,
- $p_{y_j} = P(Y = y_j) = \sum_{i=1}^{m} p_{ij}, \; j = 1, \ldots, n.$

Ovime smo definirali slučajne varijable $X|_{Y=y_j}$ i $Y|_{X=x_i}$. Formiramo zajedničku tablicu frekvencija:

| $X\backslash Y$ | $y_1$ | $y_2$ | $\ldots$ | $y_n$ | |
|---|---|---|---|---|---|
| $x_1$ | $f_{11}$ | $f_{12}$ | $\ldots$ | $f_{1n}$ | $f_{x_1}$ |
| $x_2$ | $f_{21}$ | $f_{22}$ | $\ldots$ | $f_{2n}$ | $f_{x_2}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $x_m$ | $f_{m1}$ | $f_{m2}$ | $\ldots$ | $f_{mn}$ | $f_{x_m}$ |
| | $f_{y_1}$ | $f_{y_2}$ | $\ldots$ | $f_{y_n}$ | $N$ |

Tablica 8.5: Tablica frekvencija

gdje su:

- $f_{ij}$ - frekvencija parova $(x_i, y_j)$ u uzorku,
- $f_{x_i}(f_{y_j})$ - frekvencija $x_i(y_j)$.

Hipoteze koje želimo testirati su:

$$\begin{aligned}
&\mathcal{H}_0 : p_{ij} = p_{x_i} \cdot p_{y_j} \text{ za sve } i \in \{1, \dots, m\},\ j \in \{1, \dots, n\},\\
&\mathcal{H}_1 : \text{postoje } i \in \{1, \dots, m\},\ j \in \{1, \dots, n\} \text{ takvi da je } p_{ij} \neq p_{x_i} \cdot p_{y_j}.
\end{aligned}$$

Zašto želimo baš ove hipoteze testirati? Zapravo želimo ispitati nezavisnost slučajnih varijabli. Nul hipoteza nam tvrdi kako su sve zajedničke vjerojatnosti $p_{ij}$ jednake umnošku marginalnih vjerojatnosti (odnosno $p_{x_i}$ i $p_{y_j}$). To je jednako pretpostavci kako su dvije varijable potpuno nezavisne, točnije ako znamo realizaciju jedne varijable to ne mijenja vjerojatnost realizacije za drugu varijablu. Ako odbacimo nul hipotezu, dolazimo do zaključka kako postoji povezanost između varijabli koje ispitujemo. Procjenu distribucije slučajnog vektora $(X, Y)$ dobivamo zajedničkom tablicom relativnih frekvencija gdje definiramo :

- $\hat{p}_{ij} = \frac{f_{ij}}{N}$ - relativna frekvencija parova $(x_i, y_j)$ u uzorku,
- $\hat{p}_{x_i} = \frac{f_{x_i}}{N}$ - procjena za $p_{x_i}$,
- $\hat{p}_{y_j} = \frac{f_{y_j}}{N}$ - procjena za $p_{y_j}$.

Definiramo test-statistiku kao:

$$D_{m,n} = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(f_{ij} - N\hat{p}_{x_i}\hat{p}_{y_j})^2}{N\hat{p}_{x_i}\hat{p}_{y_j}}.$$

U uvjetima istinitosti $\mathcal{H}_0$, $N\hat{p}_{x_i}\hat{p}_{y_j}$ je očekivana frekvencija $(x_i, y_j)$ pa test-statistika ima oblik kao kod $\chi^2$ testa o homogenosti i $\chi^2$ na jednom uzorku. Također, u uvjetima istinitosti $\mathcal{H}_0$ statistika $D_{m,n}$ ima $\chi^2_{(n-1)(m-1)}$ distribuciju (pogledati poglavlje *The Chi-Square Test of Independence* u [5]). Kako veće vrijednosti realizacije statistike $D_{m,n}$ idu u prilog $\mathcal{H}_1$ kritično područje je $C = [h_\alpha, \infty)$ pri čemu je $h_\alpha = Q_{\chi^2_{(n-1)(m-1)}}(1 - \alpha)$, dok je $p$-vrijednost dana s $p = P(\chi^2_{(n-1)(m-1)} \geq \hat{d}_{m,n})$ za realizaciju test-statistike $\hat{d}_{m,n}$. Očekivane frekvencije trebale bi biti barem 5, a test je asimptotski kao i ranije spomenuti test.

# Literatura


[1] L. J. Bain, M. Engelhardt, *Introduction to probability and mathematical statistics, Second edition*, Duxbury Press, Pacific Grove, 1992.

[2] M. Benšić, N. Šuvak, *Primijenjena statistika*, Odjel za matematiku, Osijek, 2013.

[3] M. Benšić, N. Šuvak, *Uvod u vjerojatnost i statistiku*, Odjel za matematiku, Osijek, 2014.

[4] Ž. Pauše, *Uvod u matematičku statistiku*, Školska knjiga, Zagreb, 1993.

[5] J. A. Rice, *Mathematical Statistics and Data Analysis, Third edition*, University of California, Berkeley, 2006.

[6] R. Dudley, *The Shapiro–Wilk test for normality*, Massachusetts Institute of Technology, Cambridge, 2012.

[7] *Statistics for application*, dostupno na
`https://ocw.mit.edu/courses/18-443-statistics-for-applications-fall-2006/`

# Sažetak

Tema ovog rada je testiranje statističkih hipoteza. Prvo smo ponovili neke osnovne pojmove koji su nam bili potrebni za daljnje razumijevanje rada. Zatim smo definirali bitne pojmove vezane uz statističke hipoteze kao što su statistički model, kritično područje i razina značajnosti. Nakon toga opisivali smo razne testove koje koristimo u testiranju, počevši prvo s onima nad jednim uzorkom (testovi o očekivanju, test o proporciji, test o kvantilima i test o distribuciji). Potom smo opisali one testove koji se temelje na dva uzorka, prvo za nezavisne slučajne uzorke, a potom one za vezane slučajne uzorke. Uz korištenje R-a napravili smo primjere koristeći konkretne baze podataka koje se nalaze unutar programa.

## Ključne riječi

jednostavni slučajni uzorak, statistička hipoteza, kritično područje, $p$-vrijednost, pogreške prve i druge vrste, normalna distribucija, varijanca, očekivanje,test-statistika, razina značajnosti, Bernoullijeva distribucija, vjerojatnost, funkcija distribucije

# Testing statistical hypotheses


## Summary

The topic of this thesis is hypothesis testing. Firstly we reviewed some basic concepts that are necessary for further understanding of the thesis. Then, we defined important concepts related to statistical hypotheses such as statistical model, critical region and significance level. After that, we described various tests used in hypothesis testing, starting with those involving a single sample ($t$-test, binomial test, quantile test, Kolmogorov–Smirnov test and $\chi^2$-test). Following that, we explained tests that involve two samples, beginning with independent random samples and then those involving dependent random samples. We created examples using specific databases that are involved in R.



## Keywords

simple random sample, statistical hypothesis, critical region, $p$-value, type I and type II errors, normal distribution, variance, mean,test-statistic, significance level, Bernoulli distribution, probability, cumulative distribution function

# Životopis

Rođena sam 11. rujna 1999. godine u Sisku. Završila sam Osnovnu školu Ivana Kukuljevića u Sisku te opću gimnaziju u Srednjoj školi Petrinja. Prvu godinu Prijediplomskog studija matematike, na tadašnjem Odjelu za matematiku, a sadašnjem Fakultetu primijenjene matematike i informatike u Osijeku upisala sam 2018. godine. Pas Aspi i ja trenutno boravimo u Osijeku. U slobodno vrijeme planinarim i rekreativno penjem na umjetnoj stijeni. Obitelj mi je velika podrška u svemu. Ljeto provodim u svojoj oazi mira koja se zove Karin.