Statistika – skripta

Varijable:

| **type** | **examples** |
| --- | --- |
| *logical* | TRUE, FALSE ili T, F |
| *integer* | 2L, 5L, 123456789L |
| *double* | 4, 6, 3.14, 2e5 |
| *complex* | 5 + 2i, 7 + 1i |
| *character* | "A", "B", "Pero", "ABCDEFGHijklmnoPQRSTUVwyz" |
| *raw* | as.raw(2), charToRaw("1") |

Za kolokvij je samo bitno znati za varijable:

Numeričke – brojevi

Logičke –„true“ i „false“ varijable

Tekstualne / faktorske / string– riječi, svaka varijabla se može pretvoriti u riječ uz dodavanje ″″

Deskriptivna statistika

Područje statistike u kojem se opisuju konkretni rezultati dobiveni ispitivanjem ili mjerenjem. Njena zadaća je da opiše podatke, i to na način da ih i sredi i sažme kako bi bili što pregledniji. Za opisivanje podataka u deskriptivnoj statistici koristimo mjere centralne vrijednosti, mjere varijabilnosti, skale mjerenja, grafičke prikaze,...

Mjere centralne tendencije:

**Aritmetička sredina** – izračunava se kao suma svih rezultata podijeljeno s ukupnim brojem rezultata

**Medijan** – vrijednost koja se u nizu rezultata, poredanih po veličini, nalazi točno u sredini – koristi se kad je distribucija izrazito asimetrična ili u nizu rezultata imamo one koji izrazito odudaranju od drugih rezultata

**Mod** – dominantna vrijednost, ona vrijednost koja se u mjerenju najviše pojavljuje/dominira

Mjere varijabilnosti:

* služe da nam pokažu koliko se rezultati grupiraju oko srednje vrijednosti

**Raspon** – razlika između maksimuma i minimuma, daje potpuno širenje podataka

**Ekstremne vrijednosti** – maksimum i minimum

**Kvartili / poluinterkvartilno raspršenje** – 4 kvartila i u svakom je 25% rezultata, pokazuje gdje se nalazi „srednja polovica“ podataka, nadopunjuje medijan

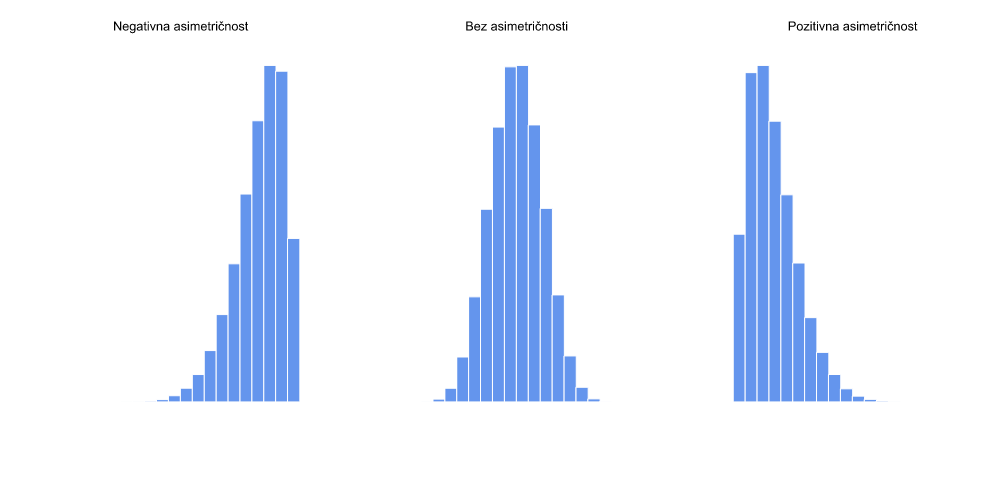
**Srednje apsolutno odstupanje** – koliko su opažanja "u prosjeku" daleko od srednje vrijednosti.

**Varijanca** – prosječno kvadratno odstupanje od srednje vrijednosti

**Standardna devijacija** – kvadratni korijen varijance, u situacijama kada je mod mjera središnje tendencije, ovo je zadana vrijednost

**Apsolutno odstupanje od medijana** – tipično (tj. srednje) odstupanje od srednje vrijednosti, u ispravljenom obliku to je robustan način za procjenu standarda odstupanja

Raspodjele:



1. Ako postoji više izuzetno velikih vrijednosti od izuzetno malih (desna ploča) kažemo da su podaci pozitivno iskrivljeni.
2. Ako podaci imaju puno ekstremno malih vrijednosti i ne toliko izuzetno velikih vrijednosti (lijeva ploča), tada kažemo da su podaci negativno iskrivljeni
3. Ako su podatci simetrično raspodijeljeni tj. imaju i tendenciju grupiranja i tendenciju raspršivanja oko srednje vrijednosti, tada kažemo da imamo normalnu raspodjelu podataka

Deskriptivna statistika na varijabli:

Numerička varijabla

*# Pregled numeričke varijable*

summary( object = afl.margins ) *# Deskriptivna stat*

#> Min. 1st Qu. Median Mean 3rd Qu. Max.

#> 0.00 12.75 30.50 35.30 50.50 116.00

minimum prvi medijan srednja treći maksimum

kvadrant vrijednost kvadrant

Daje nam prilično dobru zbirku opisne statistike povezane sa središnjom tendencijom i širenjem podataka. - min + max = raspon; prvi i treći kvartil – interkvartilno raspršenje

Logička varijabla

*# Pregled logičke varijable*

ekstremi <- afl.margins > 50 *# Stvori log varijablu*

head(ekstremi,5) *# Pogledaj podatke*

#> [1] TRUE FALSE TRUE FALSE FALSE

summary(ekstremi) *# Deskriptivna stat*

#> Mode FALSE TRUE

#> logical 132 44

\*True ili fals je rezultat uvjeta „>50“

U ovom kontekstu, funkcija „summary ()“ daje nam broj broja istinitih vrijednosti, broj pogrešnih vrijednosti i broj vrijednosti koje nedostaju

Faktorska varijabla

*# Pregled tekstualne varijable*

txt <- as.character( afl.finalists ) *# Stvori txt var*

summary( object = txt ) *# Deskriptivna stat*

#> Length Class Mode

#> 400 character character

\*Lista 400 tekstualnih objekata/stringova

Ovo je jedna od onih situacija u kojima je korisno izjaviti svoju nominalnu varijablu ljestvice kao faktor, a ne kao vektor znakova. Budući da smo „afl.finaliste“ definirali kao faktor, R zna da bi to trebao tretirati kao nominalnu varijablu ljestvice, i zato daje puno više detaljan (i koristan) sažetak nego što bi ga imao da je ostavljen kao vektor znakova.

Deskriptivna statistika na podatkovnom skupu (data frame):

#> -- Name -- -- Class -- -- Size --

#> effort data.frame 10 x 2 -> 10 redaka, 2 stupca

#> $hours numeric 10

#> $grade numeric 10

*# Deksriptivna statistika na podatkovnom okviru*

summary(clin.trial) *# Desktiptivna stat*

#> drug therapy mood.gain

#> placebo :6 no.therapy:9 Min. :0.1000

#> anxifree:6 CBT :9 1st Qu.:0.4250

#> joyzepam:6 Median :0.8500

#> Mean :0.8833

#> 3rd Qu.:1.3000

#> Max. :1.8000

Tri su lijeka: placebo, nešto što se naziva "anxifree" i nešto što se zove "joyzepam"; bilo je po 6 ljudi koji su primali svaki lijek. Bilo je 9 osoba liječenih kognitivnim ponašanjem terapiji (CBT) i 9 osoba koje nisu imale psihološkog liječenja. I možemo vidjeti iz gledanja sažetak varijable mood.gain koja je kod većine ljudi pokazala porast raspoloženja (znači ".88)

Alternativna funkcija

*# Deksriptivna statistika na podatkovnom okviru*

describe(clin.trial) *# Desktiptivna stat/ druga funkcija*

#> vars n mean sd median trimmed mad min max range skew kurtosis se

#> drug\* 1 18 2.00 0.84 2.00 2.00 1.48 1.0 3.0 2.0 0.00 -1.66 0.20

#> therapy\* 2 18 1.50 0.51 1.50 1.50 0.74 1.0 2.0 1.0 0.00 -2.11 0.12

#> mood.gain 3 18 0.88 0.53 0.85 0.88 0.67 0.1 1.8 1.7 0.13 -1.44 0.13

* vars – samo indeks: 1 za prvu varijablu, 2 za drugu varijablu itd.
* n – veličina uzorka: točnije, to je broj vrijednosti koje nedostaju.
* mean – srednja vrijednost uzorka – srednje apsolutno odstupanje
* sd – standardno odstupanje
* median – medijan
* trimmed – podrezana srednja vrijednost. Prema zadanim postavkama to je srednja vrijednost od 10%
* mad – medijan apsolutnog odstupanja
* min – minimalna vrijednost.
* max – maksimalna vrijednost.
* range – raspon obuhvaćen podacima
* skew – asimetrija
* kurtosis – zaobljenost
* se – standardna pogreška srednje vrijednosti

Grupirani pregled

*# Pregledaj grupirano prema terapiji*

by(data = clin.trial, *# Izvor podataka*

INDICES = clin.trial$therapy, *# Odredi grupiranje*

FUN = summary) *# Odredi funkciju*

#> clin.trial$therapy: no.therapy

#> drug therapy mood.gain

#> placebo :3 no.therapy:9 Min. :0.1000

#> anxifree:3 CBT :0 1st Qu.:0.3000

#> joyzepam:3 Median :0.5000

#> Mean :0.7222

#> 3rd Qu.:1.3000

#> Max. :1.7000

#> ------------------------------------------------------------

#> clin.trial$therapy: CBT

#> drug therapy mood.gain

#> placebo :3 no.therapy:0 Min. :0.300

#> anxifree:3 CBT :9 1st Qu.:0.800

#> joyzepam:3 Median :1.100

#> Mean :1.044

#> 3rd Qu.:1.300

#> Max. :1.800

Ovo su samo izvučeni podatci iz testiranja.

Na ovom pregledu možemo vidjeti koja je terapija bolje prošla. Prvi dio je uzorak s „no.therapy“ i vidimo određene rezultate. Drugi dio tj. „CBT“ ima bolje rezultate u odnosu na prvi dio.

*# Pregledaj grupirano prema razlici u raspoloženju*

aggregate(formula = mood.gain ~ drug + therapy, *# Prikaz*

data = clin.trial, *# Podatci*

FUN = mean) *# AS*

#> drug therapy mood.gain

#> 1 placebo no.therapy 0.300000

#> 2 anxifree no.therapy 0.400000

#> 3 joyzepam no.therapy 1.466667

#> 4 placebo CBT 0.600000

#> 5 anxifree CBT 1.033333

#> 6 joyzepam CBT 1.500000

Ovo su grupirani podatci gdje se vidi da je „mood.gain“ veći kod testiranja s „CBD“ terapijom.

Grupirane su kombinacije lijekova i njihovih terapija, tj. 3 lijeka i 2 terapije. Najveći mood.gain je postignut u slučaju korištenja lijeka joyzepam uz terapiju CBT.

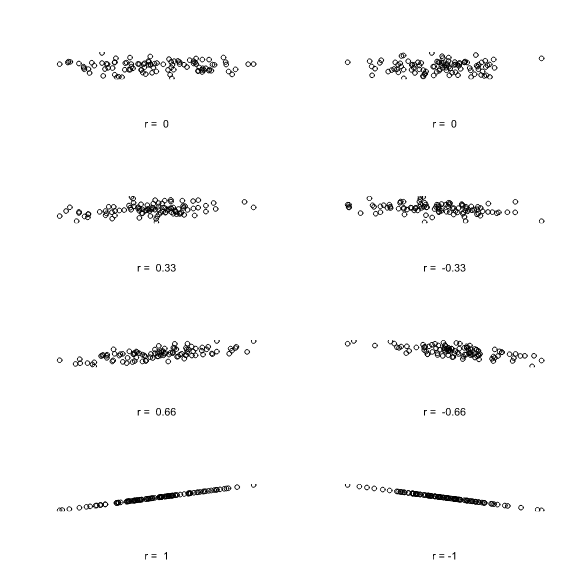
Standardne vrijednosti

**Z-vrijednost / standardni rezultat** – daje predodžbu koliko je udaljena srednja vrijednost podataka. No, tehnički je mjera koliko standardnih odstupanja ispod ili iznad populacije znače neobrađeni rezultat. Za z-vrijednost treba znati srednje apsolutno odstupanje i standardnu devijaciju populacije

Korelacija

**Kovarijanca** - generalizacija pojma odstupanje između dvije varijable X i Y; to je matematički jednostavan način opisivanja odnosa između dvije varijable

**Personov korelacijski koeficijent; standardizacija kovarijance** – koeficijent od -1 do 1 koji određuje jačinu korelacije dvije varijable



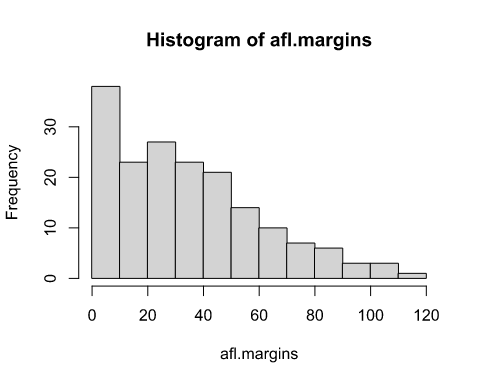
1. r =0 nema korelacije
2. r = 0.33 pozitivna korelacija ali slaba / r = -0.33 negativna korelacija, slaba
3. r = 0.66 pozitivna korelacija umjerena do jaka / r = -0.66 negativna korelacija, umjerena do jaka
4. r = 1 izrazito jaka pozitivna korelacija / r = -1 izrazito jaka negativna korelacija

Pozitivni broj označava pozitivnu korelaciju / minus (negativni broj) negativnu

**Spermanova korelacija** – bazira se na tome da se izmjeri dosljednost povezanosti između poredanih varijabli, a oblik povezanosti nije bitan. Slučajevi u kojima se koristi Spearmanov koficijent su npr. kada među varijablama ne postoji linearna povezanost, a nije moguće primijeniti odgovarajuću transformaciju kojom bi se povezanost prevela u linearnu.

Vizualizacija podataka

**Histogram** – grafički prikaz učestalosti pojave rezultata između dviju granica koje definiraju skupinu. Služe za stjecanje ukupnog dojma o podatcima i najčešće se koriste kad imamo frekvenciju nekog podatka

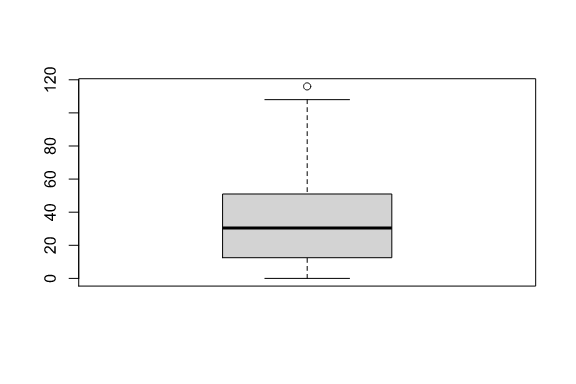


\*Koliko bodova razlike najčešće bude po utakmici – frekvencija / učestalost nekog podatka

\*Najmanje učestala razlika je 110 do 120 bodova, a najveća od 0 do 10.

**Boxplot** – standardizirani način prikazivanja raspodjele podataka na temelju sažetka od pet brojeva („minimum“, prvi kvartil (Q1), medijan, treći kvartil (Q3) i „maksimum“). Može reći o izvanrednim situacijama i koje su njihove vrijednosti.

odstupanje



medijan

minimum

1.kvartil

3.kvartil

maksimum

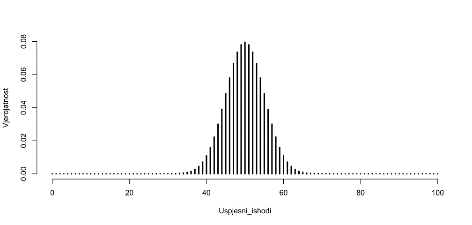
**Dijagram rasipanja** – pokazuje jačinu korelacije tj. koliko jedna varijabla utječe na drugu, što

su podatci gušće raspodijeljeni korelacija je snažnija

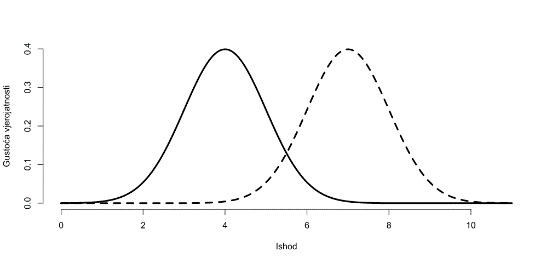
Inferencijalna statistika

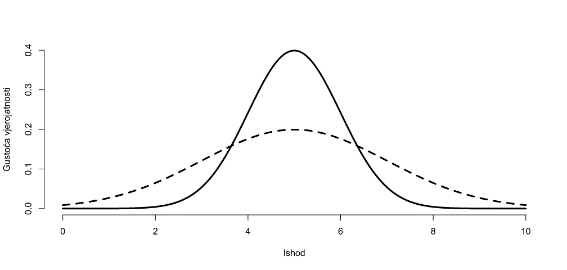
**Inferencijalna statistika** – iz pojedinačnog slučaja (s uzorka) želimo zaključiti na neku pojavu koja je možda opći zakon

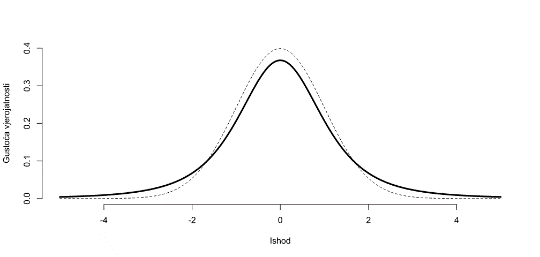
|  |  |
| --- | --- |
| Deskriptivna statistika | Inferencijalna statistika |
| 1. Numerički opis podataka | 1. Osnova je teorija vjerojatnosti |
| 1. Vizualizacije | 1. Razrađeni modeli za analizu strukture i odnosa među varijablama |
| 1. Nema mnogo teoretske pozadine | 1. Moćniji analitički alat |

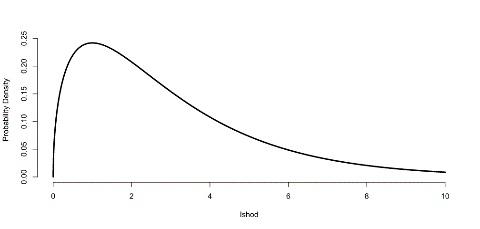
**Binomna distribucija** – broj ishoda u skupinama, Svaki stupac prikazuje vjerojatnost jednog ishoda (i.e., jedna moguća vrijednost (X)). Pošto je riječ o distribuciji vjerojatnosti, svaka od pojedinačnih vrijednosti mora biti broj od 0 do 1 dok zbroj svih stupaca mora biti jednak 1.

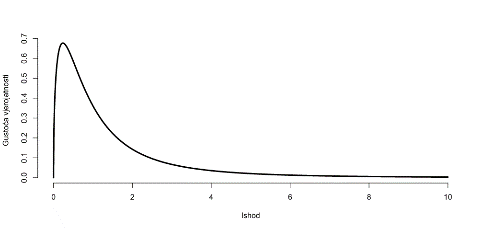
**Standardna distribucija** – vrsta kontinuirane raspodjele vjerojatnosti za stvarnu vrijednost slučajne varijable. Služi za predstavljanje stvarnih vrijednosti slučajnih varijabli čija raspodjela nije poznata. Vrijedi zbog središnjeg graničnog teorema koji nalaže da pod nekim uvjetima, prosjek puno uzoraka (promatranja) slučajne varijable s konačnom sredinom i varijancom koja je po sebi slučajna varijabla - čija raspodjela konvergira u normalnu raspodjelu kako se povećava broj uzoraka

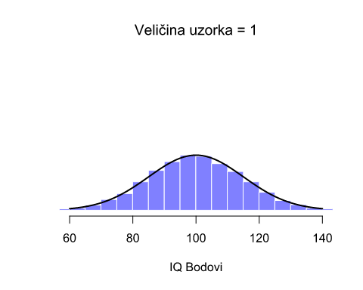
Standardna distribucija s 2 različita prosjeka

Standardna distribucija s istim prosjekom, ali 2 različite standardne devijacije

* **T-distribucija** – slična po izgledu standardnoj, ali ima „teže“ repove pa je teže označiti graničnu vrijednost. T-distribucija => distribucija t-izraza koja se dobiva kada se računski određuje standardna pogreška aritmetičke sredine, odnosno standardna pogreška razlike između aritmetičkih sredina a pri tome se u računu koristi standardna devijacija uzorka.

**Chi-sq distribucija** – ishodi moraju biti veći od 0 i prilično je zakrivljena, koristi se kod raspodjele vjerojatnosti u inferencijalnoj statistici, posebno u ispitivanju hipoteza i u izgradnji intervala pouzdanosti

**F-distribucija**

**Sampling distribucija** – raspodjela uzoraka ili raspodjela konačnog uzorka je raspodjela vjerojatnosti zadane statistike na temelju slučajnog uzorka. Raspodjela uzoraka važna je u statistici jer pruža pojednostavljenje na putu do statističkog zaključivanja. Točnije, dopušta da se analitička razmatranja temelje na raspodjeli vjerojatnosti statistike, a ne na zajedničkoj raspodjeli vjerojatnosti svih pojedinačnih vrijednosti uzorka.

Uzorak vs. populacija

**Kako povećavamo veličinu uzorka, standardna devijacija je sve manja**

* U pravilu je populacija prevelika da bismo mogli istražiti sve jedinice => koristimo uzorke
* Rezultat dobiven na uzorku zove se **statistik**
* Vrijednost populacije najčešće se naziva **parametar**
* Uzorak je manji broj definiranih statističkih jedinica ili elemenata koji čine veću cjelinu => populaciju
* Populacija (univerzum ili osnovni skup) razumijevamo sve članove neke skupine (statističke jedinice) s određenom karakteristikom koju mjerimo
* Postupak odabiranja uzorka ili preciznije postupak kojim se statističke jedinice odabiru u uzorak nazivamo **uzorkovanje**
* Statistika počiva na **slučajnim uzorcima** ali se često koriste i **neslučajni uzorci**
* Uzorak bi trebao biti takav da reprezentira populaciju u svim važnim značajkama za predmet mjerenja
* Zavisni uzorak => jedna skupina na kojoj je dva ili više puta provedeno mjerenje.
* Nezavisni uzorak => različite skupine na kojima je provođeno mjerenje (a)

Vrste uzoraka:

* Slučajni uzorci
  + Jednostavni slučajni uzorak
  + Stratificirani uzorak
  + Klaster uzorak (jednostupanjski, dvostupanjski, trostupanjski) – višeetapni ...
  + Sustavni (sistematski) slučajni uzorak
* Neslučajni uzorci
  + Prigodni uzorak
  + Namjerni ili svrhoviti uzorak
  + Uzorak eksperata (nekada ga se poistovjećuje s namjernim)
  + Kvotni uzorak (proporcionalni kvota uzorak i neproporcionalni kvota uzorak)

**Zakon velikih brojeva** – Kako se veličina uzorka povećava, sve je vjerojatnije da će vrijednost uzorka biti vrlo blizu vrijednosti pravog stanja u populaciji

**Teorem centralne tendencije –** s povećanjem uzorka, podatci teže normalnoj raspodjeli

**Interval pouzdanosti** –raspon vrijednosti koji vjerojatno uključuje vrijednost populacije s određenim stupnjem pouzdanosti. Često se izražava kao% pri čemu stanovništvo znači da leži između gornjeg i donjeg intervala. Intervali pouzdanosti mjere stupanj nesigurnosti ili sigurnosti u metodi uzorkovanja. Najčešća je razina pouzdanosti 95% ili 99%.

Testiranje hipoteza

**Istraživačka hipoteza** – iznošenje značajne, provjerljive znanstvene tvrdnje

* Npr. „Slušanje glazbe smanjuje vašu sposobnost obraćanja pažnje na druge stvari.“

**Statistička hipoteza** – mora biti matematički precizna i odgovarati određenim tvrdnjama o karakteristikama mehanizma za generiranje podataka

* Npr. „za vrijeme karantene, nasilje u obitelji se povećalo za 20% u odnosu na vrijeme prije karantene“

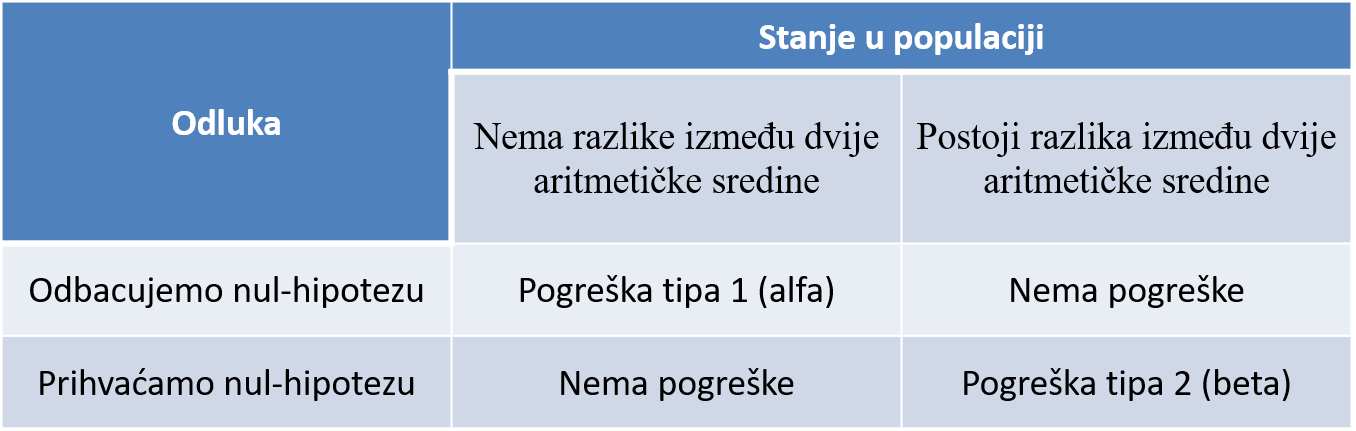
**Nul-hipoteza** – početna pretpostavka koju izlažemo testu i koju pokušavamo opovrgnuti. Postavljenu hipotezu najčešće testiramo uz rizik od 5% ili 1% (p<0,05 ili p<0,01). Niža razina rizika uzima se u ovisnosti o predmetu istraživanja. Nulta hipoteza je negativno formulirana i obično tvrdi suprotno od hipoteze istraživanja, tj. da nema utjecaja ili razlike u populacijama iz kojih je uzet uzorak ili da su parametri dviju populacija jednaki.

* Npr. „Nema razlike u visini između Dalmatinaca i Zagoraca.“

**Alternativna hipoteza** – Suprotno nultoj hipotezi, alternativne hipoteze tvrde da postoji učinak ili razlika u populacijama iz kojih je izvađen uzorak.

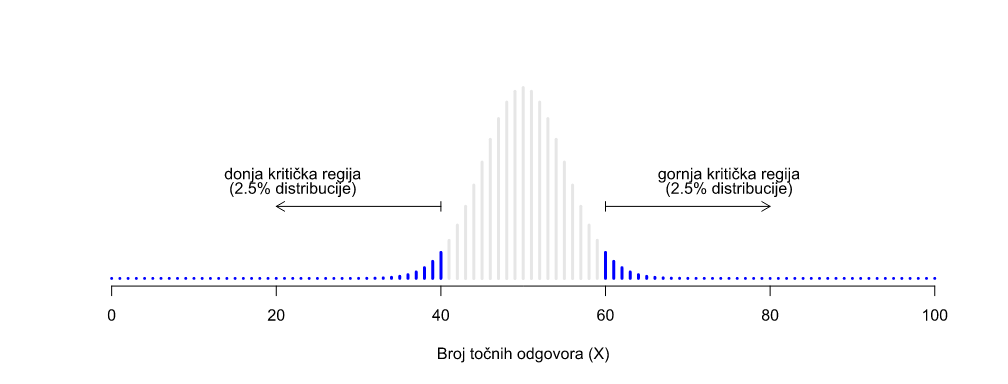
* Npr. „Postoji razlika u visini između Dalmatinaca i Zagoraca.“

Vrste pogrešaka kod testiranja hipoteza:



Ako zadržimo nultu hipotezu a ona nije ispravna

Ako odbacimo nultu hipotezu koja je točna

**Kritičke granice**

Kritičko područje testa odgovara onim vrijednostima koje bi nas dovele do odbacivanja nulte hipoteze (zbog čega se kritična regija također naziva i regijom odbijanja).

1. X bi trebao biti vrlo velik ili vrlo malen da bi odbacio nultu hipotezu.
2. Ako je nulta hipoteza istinita, distribucija uzorka X je Binomial p=0,5, Nq.
3. Ako je α 0.05, kritično područje mora pokriti 5% ove distribucije uzorkovanja.

Ispostavilo se da te tri stvari jedinstveno rješavaju problem: kritično područje sastoji se od najekstremnijih vrijednosti, poznatih kao repovi distribucije.

**P-vrijednosti –** kriteriji za odluku, u statističkom ispitivanju p-vrijednost je vjerojatnost dobivanja rezultata ispitivanja barem onoliko ekstremnih koliko su stvarno promatrani rezultati, pod pretpostavkom da je nulta hipoteza točna.

**Fisher** – izračunava vjerojatnosti da će nulta hipoteza biti odbačena

**Neyman** – izračunava kritičke vrijednosti

Efekt veličine i snaga testa

**Efekt veličine** – statistički koncept koji mjeri snagu odnosa između dvije varijable na numeričkoj skali

**Snaga testa** – vjerojatnost da će test značajnosti otkriti odstupanje od nulte hipoteze, ukoliko takvo odstupanje postoji – vjerojatnost izbjegavanja pogreške tipa II