Primijenjena i inženjerska matematika

Statistika - predavanje 4

Statističko zaključivanje - jedna varijabla

3.11.2023.

Varijable baze podataka i slučajne varijable

- varijabla baze podataka – sadrži **vrijednosti obilježja** promatranog na jedinkama neke populacije obuhvaćenim nekim istraživanjem (stupac baze podataka)
- slučajna varijabla – model za **obilježje cijele populacije**
- jedna varijabla baze podataka sadrži podatke $x_1, \ldots, x_n$ – **uzorak**
- **svaki podatak iz varijable baze podataka je realizacija jedne slučajne varijable**

Varijable baze podataka i slučajne varijable

varijabla baze podataka	slučajne varijable
x_1	$\leftarrow X_1$
x_2	$\leftarrow X_2$
$\vdots$	$\vdots$
x_n	$\leftarrow X_n$

- niz slučajnih varijabli $X_1, \dots X_n$ nazivamo **slučajni uzorak**

Varijable baze podataka i slučajne varijable

- podaci varijable baze podataka $x_1, \ldots, x_n$ dolaze iz iste populacije
- zato pretpostavljamo da su $X_1, \ldots X_n$ sve **jednako distribuirane** i to kao neka slučajna varijabla X:

$$P(X_1 \in C) = P(X_2 \in C) = \cdots = P(X_n \in C) = P(X \in C)$$

  za sve događaje C
- X modelira obilježje u cijeloj populaciji

Varijable baze podataka i slučajne varijable

- postupak prikupljanja podataka mora biti takav da su mjerenja međusobno nezavisna – $X_1, \ldots X_n$ nezavisne slučajne varijable
- **jednostavan slučajni uzorak** – $X_1, \ldots, X_n$ nezavisne slučajne varijable i jednako distribuirane kao X (kraće ćemo govoriti samo slučajni uzorak)
- u tom smislu **varijabla baze podataka (uzorak) dolazi iz slučajne varijable** X

Statističko zaključivanje

- distribucija slučajne varijable X je nepoznata – ne možemo računati $P(X \in C)$
- **statističko zaključivanje** – na osnovu podataka želimo zaključivati o distribuciji slučajne varijable X

$$x_1, \ldots, x_n \quad \dashrightarrow \quad \text{distribucija od X}$$

- na osnovu podataka možemo **procijeniti** distribuciju slučajne varijable X, a onda procijeniti i vjerojatnosti oblika $P(X \in C)$
- osim distribucije mogu nas zanimati i neke numeričke karakteristike slučajne varijable X (npr. očekivanje i varijanca) koje su također nepoznate

Procjena distribucije, očekivanja i varijance

Procjena distribucije

- procijeniti distribuciju znači procijeniti vjerojatnosti oblika $P(X \in [a, b])$ za $a, b \in \mathbb{R}$
- empirijskom distribucijom možemo procijeniti $P(X \in [a, b])$
- veći broj podataka – opravdanije korištenje empirijske distribucije za računanje vjerojatnosti
- **empirijska distribucija** je **procjena** za nepoznatu distribuciju slučajne varijable X
- nedostatak: empirijska distribucija uvijek je diskretna
- nekad distribuciju (npr. normalnu) možemo naslutiti iz histograma podataka

Procjena distribucije, očekivanja i varijance

Procjena očekivanja

- za procjenu očekivanja koristimo aritmetičku sredinu uzorka $x_1, x_2, \ldots, x_n$

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Procjena varijance

- za procjenu varijance koristimo korigiranu varijancu uzorka $x_1, x_2, \ldots, x_n$

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$$

primjer 5.1, primjer 5.2

Procjenitelj

- uzorak $x_1, \ldots, x_n$ je realizacija slučajnog uzorka $X_1, \ldots, X_n$
- procjenu neke numeričke karakteristike slučajne varijable X računamo na osnovu uzorka $x_1, \ldots, x_n$ — tako je i procjena realizacija slučajne varijable koju nazivamo **procjenitelj**
- na konkretnoj realiziciji procjenitelj poprima jednu vrijednost
- procjenitelj za očekivanje je slučajna varijabla

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- procjenitelj za varijancu je slučajna varijabla

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

- primjer 5.3

Pouzdani intervali

- nepoznatu numeričku karakteristiku procjenjujemo brojem – ne daje informaciju o pouzdanosti takve procjene (ne uzima u obzir raspršenost)
- Primjer: neka su zadana tri uzroka
  - $0, 0, -1, 1$
  - $-200, 100, 200, -100$
  - $0, 0, -1, 1, 0, -1, 1$

  Procjena očekivanja je 0 u sva tri slučaja, no u kojem slučaju je *pouzdanija*?

Pouzdani intervali

- procjenu možemo napraviti i intervalom – procjenjujemo interval koji sadrži nepoznatu numeričku karakteristiku
- odaberemo $\gamma \in (0, 1)$, npr. $\gamma = 0.95$ ili $\gamma = 0.99$ – **pouzdanost**
- γ-**pouzdani interval** (ili interval pouzdanosti γ) je slučajan interval (granice su mu slučajne varijable) takav da se stvarna vrijednost numeričke karakteristike koju procjenjujemo nalazi u takvom intervalu s vjerojatnošću barem γ
- jedna **realizacija pouzdanog intervala** određena na osnovu prikupljenog uzorka je običan interval realnih brojeva

Pouzdani interval

- na osnovu podataka i formula za određivanje granica intervala pouzdanosti γ izračunavamo običan interval s realnim brojevima kao granicama
- ne možemo reći da baš taj interval sadrži stvarnu vrijednost numeričke karakteristike
- možemo samo reći da ako bi računanje ponavljali na uzorcima iz iste populacije, tada bi izračunati interval realnih brojeva u približno $100\gamma\,\%$ slučajeva sadržavao stvarnu vrijednost karakteristike koju procjenjujemo

Procjena očekivanja pouzdanim intervalom normalno distribuirana populacija, poznata varijanca

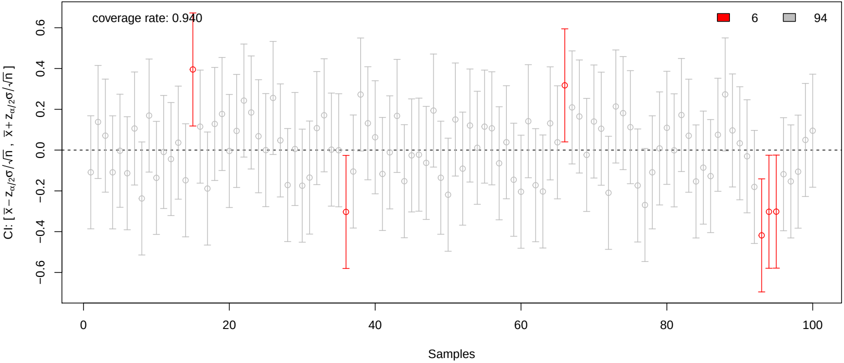

Figure 1: Realizacije pouzdanih intervala za očekivanje izračunatih na 100 uzoraka iz $\mathcal{N}(0, 1)$ distribucije – približno 95% njih sadrži stvarnu vrijednost parametra (0)

Procjena očekivanja pouzdanim intervalom normalno distribuirana populacija, poznata varijanca

- neka je $X_1, \dots, X_n$ slučajan uzorak iz slučajne varijable X i pretpostavimo da znamo da $X \sim \mathcal{N}(\mu, \sigma^2)$ pri čemu je očekivanje μ nepoznato, a σ^2 poznato
- $\overline{X}_n$ - aritmetička sredina uzorka veličine n

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- vrijedi

$$Z = \frac{\overline{X}_n - \mu}{\sigma} \sqrt{n} \sim \mathcal{N}(0, 1)$$

- odredimo broj z_γ za koji vrijedi

$$P(|Z| \leq z_\gamma) = P(-z_\gamma \leq Z \leq z_\gamma) = \gamma$$

Procjena očekivanja pouzdanim intervalom normalno distribuirana populacija, poznata varijanca

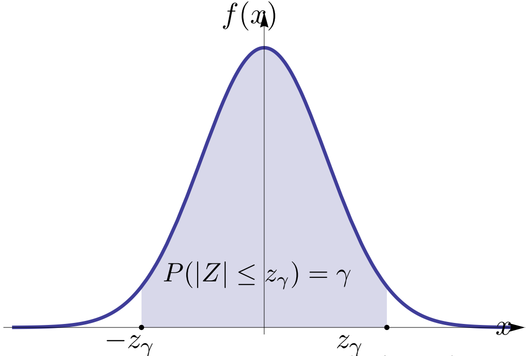

Figure 2: Vjerojatnost $P(|Z| \leq z_\gamma)$.

Procjena očekivanja pouzdanim intervalom normalno distribuirana populacija, poznata varijanca

- z_γ je $\frac{1-\gamma}{2} + \gamma = \frac{1+\gamma}{2}$ kvantil slučajne varijable Z (možemo ga izračunati u Probability calculatoru)
- sada imamo:

$$\begin{aligned}\gamma &= P\left(|Z| \leq z_\gamma\right)\\ &= P\left(-z_\gamma \leq \frac{\overline{X}_n - \mu}{\sigma}\sqrt{n} \leq z_\gamma\right)\\ &= P\left(\overline{X}_n - z_\gamma\,\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}_n + z_\gamma\,\frac{\sigma}{\sqrt{n}}\right)\end{aligned}$$

odnosno

$$P\left(\mu \in \left[\overline{X}_n - z_\gamma\frac{\sigma}{\sqrt{n}},\ \overline{X}_n + z_\gamma\frac{\sigma}{\sqrt{n}}\right]\right) = \gamma$$

Procjena očekivanja pouzdanim intervalom normalno distribuirana populacija, poznata varijanca

- pouzdani interval (z-**interval**)

$$\left[\overline{X}_n - z_\gamma \frac{\sigma}{\sqrt{n}},\ \overline{X}_n + z_\gamma \frac{\sigma}{\sqrt{n}}\right]$$

- to je slučajni interval koji sadrži očekivanje μ s vjerojatnošću γ

Procjena očekivanja pouzdanim intervalom normalno distribuirana populacija, poznata varijanca

- na osnovu uzorka $x_1, \ldots, x_n$ iz slučajne varijable X (realizaciju) pouzdanog intervala računamo po formuli

$$\left[\overline{x}_n - z_\gamma \frac{\sigma}{\sqrt{n}},\ \overline{x}_n + z_\gamma \frac{\sigma}{\sqrt{n}}\right]$$

$\overline{x}_n$ - aritmetička sredina uzorka
σ - standardna devijacija slučajne varijable X
z_γ - broj za koji vrijedi da je $P(|Z| \leq z_\gamma) = \gamma$
Z - standardna normalna slučajna varijabla

- u približno $100\gamma\,\%$ slučajeva tako izračunati interval će sadržavati stvarnu vrijednost očekivanja μ slučajne varijable X
- primjer 5.4

Procjena očekivanja pouzdanim intervalom normalno distribuirana populacija, nepoznata varijanca

- u praktičnim problemima σ gotovo nikad nije poznat
- neka je $X_1, \ldots, X_n$ slučajan uzorak iz slučajne varijable X i pretpostavimo da znamo da $X \sim \mathcal{N}(\mu, \sigma^2)$ pri čemu su očekivanje μ i varijanca σ^2 nepoznati
- procjenitelj za standardnu devijaciju

$$S_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}$$

- vrijedi

$$T = \frac{\overline{X}_n - \mu}{S_n}\sqrt{n} \sim \mathcal{T}_{n-1}$$

gdje je $\mathcal{T}_{n-1}$ Studentova distribucija (t-distribucija) s $n-1$ stupnjeva slobode

Procjena očekivanja pouzdanim intervalom normalno distribuirana populacija, nepoznata varijanca

- odredimo broj $t_{\gamma,n-1}$ za koji vrijedi

$$P(|T| \leq t_{\gamma,n-1}) = P(-t_{\gamma,n-1} \leq T \leq t_{\gamma,n-1}) = \gamma$$

- $t_{\gamma,n-1}$ je $\frac{1-\gamma}{2} + \gamma = \frac{1+\gamma}{2}$ kvantil slučajne varijable T (možemo ga izračunati u Probability calculatoru)
- slijedi

$$P\left(\mu \in \left[\overline{X}_n - t_{\gamma,n-1}\frac{S_n}{\sqrt{n}},\ \overline{X}_n + t_{\gamma,n-1}\frac{S_n}{\sqrt{n}}\right]\right) = \gamma$$

Procjena očekivanja pouzdanim intervalom normalno distribuirana populacija, nepoznata varijanca

- pouzdani interval (t-**interval**):

$$\left[\overline{X}_n - t_{\gamma,n-1}\frac{S_n}{\sqrt{n}},\ \overline{X}_n + t_{\gamma,n-1}\frac{S_n}{\sqrt{n}}\right]$$

- to je slučajni interval koji sadrži očekivanje μ s vjerojatnošću γ

Procjena očekivanja pouzdanim intervalom normalno distribuirana populacija, nepoznata varijanca

- na osnovu uzorka $x_1, \ldots, x_n$ iz slučajne varijable X (realizaciju) pouzdanog intervala računamo po formuli

$$\left[\overline{x}_n - t_{\gamma,n-1}\frac{s_n}{\sqrt{n}},\ \overline{x}_n + t_{\gamma,n-1}\frac{s_n}{\sqrt{n}}\right]$$

$\overline{x}_n$ - aritmetička sredina uzorka
s_n - standardna devijacija uzorka
$t_{\gamma,n-1}$ - broj za koji vrijedi da je $P(|T| \leq t_{\gamma,n-1}) = \gamma$
T - slučajna varijabla sa Studentovom distribucijom
s $n-1$ stupnjeva slobode

- u približno $100\gamma\,\%$ slučajeva tako izračunati interval će sadržavati stvarnu vrijednost očekivanja μ slučajne varijable X
- zadatak

Procjena očekivanja pouzdanim intervalom veliki uzorci

- neka je $X_1, \ldots, X_n$ slučajan uzorak iz slučajne varijable X s očekivanjem μ i varijancom σ^2 nepoznatima
- i za velike uzorke možemo korisiti t-interval kao pouzdani interval za očekivanje
- primjer 5.5

Procjena vjerojatnosti događaja pouzdanim intervalom veliki uzorci

- neka je $X_1, \ldots, X_n$ slučajan uzorak iz Bernoullijeve slučajne varijable

$$X \sim \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}, \quad p \in (0,1)$$

  pri čemu je p nepoznat
- procjenitelj za p: relativna frekvencija "uspjeha" (jedinice) u uzorku:

$$\hat{p} = \frac{f_1}{n},$$

  gdje je f_1 frekvencija jedinice u uzorku
- uočimo: $\hat{p}$ je aritmetička sredina uzorka

Procjena vjerojatnosti događaja pouzdanim intervalom veliki uzorci

- a osnovu uzorka $x_1, \ldots, x_n$ iz slučajne varijable X (realizaciju) pouzdanog intervala računamo po formuli

$$\left[\hat{p} - z_\gamma \sqrt{\frac{\hat{p}\hat{q}}{n}},\ \hat{p} + z_\gamma \sqrt{\frac{\hat{p}\hat{q}}{n}}\right],$$

$\hat{p}$ - relativna frekvencija jedinice (uspjeha) u uzorku
$\hat{q}$ - relativna frekvencija nula (neuspjeha) u uzorku, $\hat{q} = 1 - \hat{p}$
z_γ - broj za koji vrijedi $P(|Z| \leq z_\gamma) = \gamma$
Z - standardna normalna slučajna varijabla

- u približno $100\gamma\,\%$ slučajeva tako izračunati interval će sadržavati stvarnu vrijednost vjerojatnosti p
- primjer 5.7

Testiranje statističkih hipoteza

- hipoteze koje želimo testirati treba izraziti kao hipoteze o distribuciji slučajne varijable
- **statistička hipoteza** – slutnja koja se odnosi na distribuciju slučajnog uzorka
- statističku hipotezu standardno označavamo s H ili $\mathcal{H}$
- **statistički test** – pravilo na osnovu kojeg donosimo odluku o *odbacivanju* ili *ne odbacivanju* hipoteze
- kod testiranja postoje dvije hipoteze: **nul-hipoteza** – H_0 i **alternativna hipoteza** – H_1
- alternativna hipoteza je ona koju prihvaćamo u slučaju odbacivanja nul-hipoteze (suprotno od H_0)

Testiranje statističkih hipoteza

- Hipoteze se uvijek postavljaju tako da se prije provođenja testa smatra da vrijedi nul-hipoteza H_0. Ako ne odbacimo H_0, ništa se neće dogoditi. Osim toga, hipoteze biramo tako da alternativna hipoteza odgovara tvrdnji koju želimo dokazati.
- To možemo usporediti sa suđenjem: Nitko nije kriv dok mu se ne dokaže krivnja.
  U tom slučaju hipoteze su

$$H_0 : \text{optuženi nije kriv}$$
$$H_1 : \text{optuženi je kriv}$$

Testiranje statističkih hipoteza

- Uvijek govorimo o ***odbacivanju*** (postoji dovoljno dokaza za krivnju) ili ***ne odbacivanju*** (ne postoji dovoljno dokaza za krivnju) nul-hipoteze.
- Nije dobro reći *prihvaćamo* nul-hipotezu - to bi značilo da je ona točna samo zato jer je nismo uspjeli opovrgnuti (nije dokazano da je osoba nevina, već samo da ne postoji dovoljno dokaza da se proglasi krivom)
- Kod odbacivanja, često kažemo *odbacujemo hipotezu* H_0 *u korist hipoteze* H_1

Pogreške statističkog testa

- sve odluke temeljene na uzorcima iz populacije nisu 100% pouzdane pa ni zaključak testa ne mora biti 100% pouzdan
- odluka koja je donesena statističkim testom može biti ili pogrešna ili ispravna

istina	zaključak testa	
	ne odbaciti H_0	odbaciti H_0
H_0	dobra odluka	pogreška I. tipa
H_1	pogreška II. tipa	dobra odluka

pogreška I. tipa: odbaciti H_0 ako je ona istinita

pogreška II. tipa: ne odbaciti H_0 ako je H_1 istinita.

Pogreške statističkog testa

- za primjer:
  pogreška I. tipa - osuditi nevinu osobu
  pogreška II. tipa - osloboditi osobu koja je stvarno kriva
- statistički test se dizajnira tako da dopušta izbor maksimalne vjerojatnosti pogreške prvog tipa (npr. 0.01, 0.05 ili 0.1)
- odabrana maksimalna vjerojatnost pogreške prvog tipa zove se **razina značajnosti testa** (ili nivo signifikantnosti), oznaka α

Zaključak testa može biti:

- *odbacujemo H_0 na razini značajnosti α i prihvaćamo hipotezu H_1*
- *ne odbacujemo H_0 na razini značajnosti α (nema dovoljno dokaza da H_0 ne vrijedi)*

Rezultat statističkog testa

- kao rezultat testiranja dobit ćemo p-**vrijednost**
- p-**vrijednost** – vjerojatnost da, pod pretpostavkom da H_0 vrijedi, dobijemo uzorak jednak ili ekstremniji od onoga koji smo dobili (jednako ili manje vjerojatan)
- ako je ta vjerojatnost mala (manja od razine značajnosti), onda H_0 možemo odbaciti, u suprotnom nemamo razloga za odbacivanje
- vjerojatnost ćemo računati na temelju distribucije neke funkcije uzorka – **test statistike**

Pomoću p-vrijednosti možemo zaključiti sljedeće:

- ♣ ako je $p \leq \alpha$ odbacujemo H_0 na razini značajnosti α
- ♣ ako je $p > \alpha$ ne odbacujemo H_0 na razini značajnosti α

Testiranje hipoteza o očekivanju

- testiramo je li očekivanje uzorka μ jednako nekoj unaprijed zadanoj vrijednosti μ_0
- nul-hipoteza:

$$H_0 : \quad \mu = \mu_0$$

- alternativne hipoteze mogu biti:
  - $H_1 : \quad \mu > \mu_0$
  - $H_1 : \quad \mu < \mu_0$
  - $H_1 : \quad \mu \neq \mu_0$
- odabir testa ovisi o pretpostavkama koje uzorak zadovoljava

Normalno distribuirana populacija, poznata varijanca (z-test)

- neka je $X_1, \ldots, X_n$ slučajan uzorak iz slučajne varijable X i pretpostavimo da znamo da $X \sim \mathcal{N}(\mu, \sigma^2)$ pri čemu je očekivanje μ nepoznato, a σ^2 poznato
- test-statistika:

$$Z = \frac{\overline{X}_n - \mu_0}{\sigma} \sqrt{n}$$

- ako H_0 vrijedi, $Z \sim \mathcal{N}(0, 1)$
- na temelju podataka izračunamo vrijednost za Z – označit ćemo je $\hat{z}$
- ako je H_0 istinita, očekujemo da je $\hat{z}$ blizu 0, npr.
    - 95% realizacija od Z je u intervalu $(-\infty, 1.6448]$
    - 95% realizacija od Z je u intervalu $[-1.6448, \infty)$
    - 95% realizacija od Z je u intervalu $[-1.96, 1.96]$

Normalno distribuirana populacija, poznata varijanca (z-test)

- p-vrijednost – vjerojatnost da test-statistika bude jednaka ili ekstremnija od $\hat{z}$:
  - $p = P(Z \geq \hat{z})$ ako je alternativna hipoteza oblika $H_1 : \mu > \mu_0$
  - $p = P(Z \leq \hat{z})$ ako je alternativna hipoteza oblika $H_1 : \mu < \mu_0$
  - $p = P(Z \in \langle -\infty, -|\hat{z}|] \cup [|\hat{z}|, \infty\rangle)$ ako je alternativna hipoteza oblika $H_1 : \mu \neq \mu_0$
- uočimo:

$$\begin{aligned} P(Z \in \langle -\infty, -|\hat{z}|] \cup [|\hat{z}|, \infty\rangle) &= P(Z \geq |\hat{z}|) + P(Z \leq -|\hat{z}|) \\ &= 2P(Z \geq |\hat{z}|) = 2P(Z \leq -|\hat{z}|) \end{aligned}$$

  p-vrijednost dvostranog testa je dvostruka p-vrijednost jednostranog testa
- tako izračunatu p-vrijednost uspoređujemo s razinom značajnosti α:
  - ako je $p \leq \alpha$ odbacujemo nul-hipotezu na razini značajnosti α
  - ako je $p > \alpha$ ne odbacujemo nul-hipotezu na razini značajnosti α
- zadatak

Normalno distribuirana populacija, nepoznata varijanca (t-test)

- u praktičnim problemima σ gotovo nikad nije poznat
- neka je $X_1, \ldots, X_n$ slučajan uzorak iz slučajne varijable X i pretpostavimo da znamo da $X \sim \mathcal{N}(\mu, \sigma^2)$ pri čemu su očekivanje μ i varijanca σ^2 nepoznati
- test-statistika:

$$T = \frac{\overline{X}_n - \mu_0}{S_n}\sqrt{n} \sim \mathcal{T}_{n-1}$$

- ako H_0 vrijedi, $T \sim \mathcal{T}_{n-1}$ gdje je $\mathcal{T}_{n-1}$ Studentova distribucija (t-distribucija) s $n-1$ stupnjeva slobode
- na temelju podataka izračunamo vrijednost za T – označit ćemo je $\hat{t}$
- ako je H_0 istinita, očekujemo da je $\hat{t}$ blizu 0, npr. za $n = 100$
  - 95% realizacija od T je u intervalu $(-\infty, 1.6604]$
  - 95% realizacija od T je u intervalu $[-1.6604, \infty)$
  - 95% realizacija od T je u intervalu $[-1.9842, 1.9842]$

Normalno distribuirana populacija, nepoznata varijanca (t-test)

- p-vrijednost – vjerojatnost da test-statistika bude jednaka ili ekstremnija od $\hat{t}$:
    - $p = P(T \geq \hat{t})$ ako je alternativna hipoteza oblika $H_1 : \mu > \mu_0$
    - $p = P(T \leq \hat{t})$ ako je alternativna hipoteza oblika $H_1 : \mu < \mu_0$
    - $p = P(T \in \langle -\infty, -|\hat{t}|] \cup [|\hat{t}|, \infty\rangle) = 2P(T \geq |\hat{t}|)$ ako je alternativna hipoteza oblika $H_1 : \mu \neq \mu_0$
- tako izračunatu p-vrijednost uspoređujemo s razinom značajnosti α:
    - ako je $p \leq \alpha$ odbacujemo nul-hipotezu na razini značajnosti α
    - ako je $p > \alpha$ ne odbacujemo nul-hipotezu na razini značajnosti α
- zadatak

Veliki uzorci

- neka je $X_1, \ldots, X_n$ slučajan uzorak iz slučajne varijable X s očekivanjem μ i varijancom σ^2 nepoznatima
- i za velike uzorke možemo koristiti t-test

$\leftrightsquigarrow$

Testiranje hipoteza o vjerojatnosti događaja veliki uzorci

- neka je $X_1, \ldots, X_n$ slučajan uzorak iz Bernoullijeve slučajne varijable
$$X \sim \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}, \quad p \in (0,1)$$
pri čemu je p nepoznat
- nul-hipoteza:
$$H_0 : \quad p = p_0$$
- test-statistika:
$$Z' = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$
gdje je n dimenzija uzorka, a $\hat{p}$ relativna frekvencija "uspjeha"
- ako H_0 vrijedi, Z' ima približno $\mathcal{N}(0,1)$ distribuciju
- na temelju podataka izračunamo vrijednost za Z' – označit ćemo je $\hat{z}$
- ako je H_0 istinita, očekujemo da je $\hat{z}$ blizu 0

Testiranje hipoteza o vjerojatnosti događaja veliki uzorci

- p-vrijednost – vjerojatnost da test-statistika bude jednaka ili ekstremnija od $\hat{z}$:
    - p-vrijednost $= P(Z \geq \hat{z})$ ako je alternativna hipoteza oblika $H_1 : p > p_0$
    - p-vrijednost $= P(Z \leq \hat{z})$ ako je alternativna hipoteza oblika $H_1 : p < p_0$
    - p-vrijednost $= P(Z \in \langle -\infty, -|\hat{z}|] \cup [|\hat{z}|, \infty\rangle) = 2P(Z \geq |\hat{z}|)$ ako je alternativna hipoteza oblika $H_1 : p \neq p_0$
- tako izračunatu p-vrijednost uspoređujemo s razinom značajnosti α:
    - ako je p-vrijednost $\leq \alpha$ odbacujemo nul-hipotezu na razini značajnosti α
    - ako je p-vrijednost $> \alpha$ ne odbacujemo nul-hipotezu na razini značajnosti α
- primjer 5.10

Testiranje hipoteza o distribuciji diskretne distribucije

- kao procjenu za stvarnu (nepoznatu) distribuciju slučajne varijable koristimo empirijsku distribuciju određenu na temelju podataka
- želimo testirati ima li slučajna varijabla neku pretpostavljenu **teorijsku distribuciju**
- testiranje hipoteze da podaci dolaze iz pretpostavljene teorijske distribucije može se provesti tzv. χ^2-testom
- neka je teorijska distribucija

$$\begin{pmatrix} x_1 & x_2 & \dots & x_k \\ p_1 & p_2 & \dots & p_k \end{pmatrix}$$

- test se provodi usporedbom opaženih i teorijskih frekvencija
- opažene (eksperimentalne) frekvencije

x_1	x_2	$\dots$	x_k
f_1	f_2	$\dots$	f_k

$\sum_{i=1}^{k} f_i = n$

Testiranje hipoteza o distribuciji diskretne distribucije

- teorijske (očekivane) frekvencije

x_1	x_2	$\dots$	x_k
np_1	np_2	$\dots$	np_k

- hipoteze χ^2-testa:

H_0 : distribucija iz koje dolaze podaci jednaka je teorijskoj

H_1 : distribucija iz koje dolaze podaci razlikuje se od teorijske

- primjer 5.11

Testiranje hipoteza o distribuciji normalna distribucija

- za prvi uvid u moguća odstupanja od normalne distribucije možemo koristiti razne mjere deskriptivne statistike i grafičke prikaze (npr. histograme relativnih frekvencija), no to nije dovoljno za donošenje zaključka o normalnoj distribuiranosti varijable
- Shapiro-Wilk W test - statistički test koji se mogu koristiti za testiranje sljedećih hipoteza:

$$H_0 : \text{varijabla ima normalnu distribuciju}$$
$$H_1 : \text{varijabla nema normalnu distribuciju}$$

- primjer 5.12