Matematika i statistika

Doktorski studij Strojarskog fakulteta

Sveučilište u Slavonskom Brodu

Predavanje 2

Predavač:
izv.prof.dr.sc. Nenad Šuvak, Odjel za matematiku, Sveučilište u Osijeku

Slučajna varijabla - motivacija

- varijable modeliramo kao **slučajne varijable** - karakterizirane skupom vrijednosti kojima se mogu realizirati, no u trenutku mjerenja njihove vrijednosti ne možemo sagledati sve uvjete pod kojima će se realizirati neka od tih vrijednosti
- osnova za proučavanja slučajnih varijabli - **skup svih mogućih realizacija** ili **slika** slučajne varijable, oznaka $\mathcal{R}(X)$
- promotrimo neki neprazan skup i označimo ga Ω (za Ω možemo uzeti skup svih mogućih vrijednosti slučajne varijable, tj. $\mathcal{R}(X)$)

SLUČAJNA VARIJABLA - VJEROJATNOST

- **događaj** - podskup nepraznog skupa Ω
- mogućnost realizacije događaja interpretiramo kao "stupanj vjerovanja u realizaciju tog događaja"
- **vjerojatnost** - mjera koja modelira stupanj vjerovanja da će se realizirati događaj $A \subset \Omega$
- od određenih podskupova skupa Ω formiramo familiju događaja $\mathfrak{F}$ (mora sadržavati Ω, biti zatvorena na komplement i uniju)
- vjerojatnost (oznaka P) je funkcija koja svakom događaju $A \in \mathfrak{F}$ pridružuje realan broj iz intervala $[0, 1]$ tako da vrijede sljedeći zahtjevi:
  - $P(\Omega) = 1$
  - ako su A_1 i A_2 događaji iz $\mathfrak{F}$ koji nemaju zajedničkih elemenata, tj. $A_1, A_2 \in \mathfrak{F}$ i $A_1 \cap A_2 = \emptyset$, tada je

$$P(A_1 \cup A_2) = P(A_1) + P(A_2)$$

Slučajna varijabla - zadavanje

- $\mathcal{R}(X)$ - skup svih mogućih realizacija slučajne varijable X
- $\mathfrak{F}$ - familija događaja, tj. familija podskupova od $\mathcal{R}(X)$ na kojoj je definirana vjerojatnost P
- događaj $A \in \mathfrak{F}$ će se dogoditi (realizirati) ako se slučajna varijabla X realizira vrijednošću iz skupa A
- slučajna varijabla X je zadana ako je zadan skup $\mathcal{R}(X)$ i vjerojatnost P na familiji događaja $\mathfrak{F}$ - kažemo da smo zadali **razdiobu (distribuciju) slučajne varijable** X
- za određivanje vjerojatnosti koristimo jedan od dvaju standardnih pristupa - **klasični** i **statistički**

Slučajna varijabla - primjer

- neka slučajna varijabla X opisuje rezultat bacanja pravilne igraće kockice
- svi mogući ishodi jednog bacanja čine skup Ω:

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

- kako je kockica pravilna, znamo da je

$$P(X=1) = \ldots = P(X=6) = \frac{1}{6}$$

- distribuciju slučajne varijable X pregledno zapisujemo tablicom

$$X \sim \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{pmatrix}$$

- ovdje smo zapravo iskorsitili klasični pristup određivanju vjerojatnosti

VJEROJATNOST - KLASIČNI PRISTUP

- neka vrijedi sljedeće:
  (1) skup $\Omega \neq \emptyset$ ima konačno mnogo elemenata ($\Omega = \{\omega_1, \ldots, \omega_n\}, n \in \mathbb{N}$)
  (2) svi jednočlani podskupovi od Ω su jednako vjerojatni, tj.

$$P(\{\omega_i\}) = P\left(\{\omega_j\}\right), \qquad \text{za sve } i, j \in \{1, \ldots, n\}$$

- tada vjerojatnost događaja A definiramo na sljedeći način:

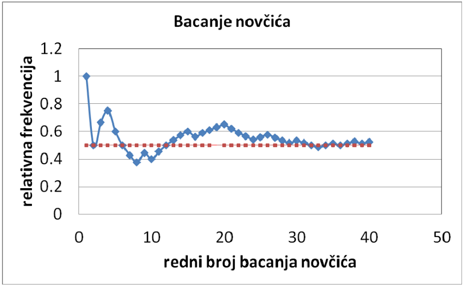
$$P(A) = \frac{\text{broj elemenata od } A}{\text{broj elemenata od } \Omega} = \frac{k(A)}{k(\Omega)},$$

  gdje je $k(\cdot)$ oznaka za broj elemenata skupa
- analogno za $\Omega = \mathcal{R}(X)$

Vjerojatnost - statistički pristup

- temelji se na nezavisnom ponavljanju pokusa - činjenica da se neki događaj realizirao pri izvođenju pokusa nikako ne utječe na realizaciju bilo kojeg ishoda narednih izvođenja tog pokusa
- npr. bacanje igraće kockice dva puta čini dva nezavisna pokusa, ali izvlačenje drugog broja u igri loto pokus je koji nije nezavisan od izvlačenja prvog broja u toj igri
- ako je pokus takav da ga možemo nezavisno ponavljati mnogo puta, relativna frekvencija pojavljivanja događaja A će se s povećanjem broja ponavljanja pokusa stabilizirati oko nekog broja koji predstavlja statistički definiranu vjerojatnost realizacije tog događaja

Vjerojatnost - statistički pristup



Grafički prikaz relativnih frekvencija pojavljivanja pisma za 40 bacanja pravilnog novčića

SLUČAJNA VARIJABLA - TIPOVI

- prije smo uočili bitnu razliku u opisu diskretnih i neprekidnih numeričkih varijabli
- iste razlike vidljive su u načinu zadavanja slučajnih varijabli kojima modeliramo diskretne i neprekidne numeričke varijable
- razlikovat ćemo dva tipa slučajnih varijabli:
  - diskretne
  - neprekidne

Diskretna slučajna varijabla - zadavanje

- **ako je $\mathcal{R}(X)$ konačan ili prebrojiv skup kažemo da je slučajna varijabla X diskretna**
- konačan $\mathcal{R}(X)$ - $\{x_1, x_2, x_3, \ldots, x_n\}$
- prebrojiv $\mathcal{R}(X)$ - $\{x_1, x_2, x_3, \ldots\}$
- za svaku pojedinu realizaciju x_i definiramo vjerojatnost $p_i = P\{X = x_i\}$
- DSV X je u potpunosti zadana skupom $\mathcal{R}(X)$ i pripadnim vjerojatnostima $p_i = P\{X = x_i\}$ za koje vrijede sljedeća svojstva:
  1. $p_i \geq 0$ za sve $x_i \in \mathcal{R}(X)$
  2. $\sum_{\text{svim } x_i \in \mathcal{R}(X)} p_i = 1$
- računanje vjerojatnosti da diskretna slučajna varijabla X primi vrijednosti iz nekog skupa $A \subseteq \mathcal{R}(X)$:

$$P(X \in A) = \sum_{x_i \in A} p_i$$

Diskretna slučajna varijabla - tablica distribucije

- distribucija DSV pregledno se zadaje **tablicom distribucije**:

$$X \sim \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ p_1 & p_2 & \dots & p_n \end{pmatrix} \quad \text{tj.} \quad X \sim \begin{pmatrix} x_1 & x_2 & x_3 & \dots \\ p_1 & p_2 & p_3 & \dots \end{pmatrix}$$

- prva tablica odnosi se na DSV s konačnim skupom $\mathcal{R}(X)$, a druga na DSV s prebrojivim skupom $\mathcal{R}(X)$
- slučajna varijabla jednoznačno je zadana svojom distribucijom - često umjesto o slučajnim varijablama govorimo o distribucijama

DISKRETNE DISTRIBUCIJE - BERNOULLIJEVA I BINOMNA DISTRIBUCIJA

- Bernoullijeva distribucija - modelira (opisuje) rezultat pokusa koji ima samo dva moguća ishoda, uspjeh (oznaka 1) i neuspjeh (oznaka 0):

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

- parametar $p \in (0,1)$ je vjerojatnost da se X realizira vrijednošću 1
- binomna $\mathcal{B}(n,p)$ distribucija - služi za modeliranje broja realiziranih uspjeha (oznaka 1) pri n nezavisnih ponavljanja pokusa sa samo dva moguća ishoda, pri čemu se uspjeh u jednom izvođenju pokusa realizira s vjerojatnošću $p \in (0,1)$

Diskretne distribucije - primjeri

Primjer

Ako imamo jako preciznu vagu i mjerimo masu šećera koji je pakiran u vrećice deklarirane mase 1 *kg, hoćemo li dobiti točno* 1 *kg? Ako uzmemo drugo pakiranje istog tipa, koliko vam se čini izvjesno da će masa biti ista kao u prethodno vaganom pakiranju? Očekujete li velika odstupanja? Ako masu šećera u toj seriji pakiranja modeliramo slučajnom varijablom* X*, koji biste skup uzeli za skup* $\mathcal{R}(X)$*?*

Primjer

Poznato je da je u velikom skladištu trgovine informatičkom opremom vjerojatnost pojavljivanja prijenosnog računala s greškom nastalom u proizvodnji jednaka 0.02*. Pretpostavimo da iz tog skladišta nasumično biramo* 10 *prijenosnih računala. Poslodavca zanima je li manje od pet prijenosnih računala koje smo izabrali s greškom. Koju distribuciju i s kojim parametrima smatrate prikladnom za opis ovog problema? (R - primjer 1)*

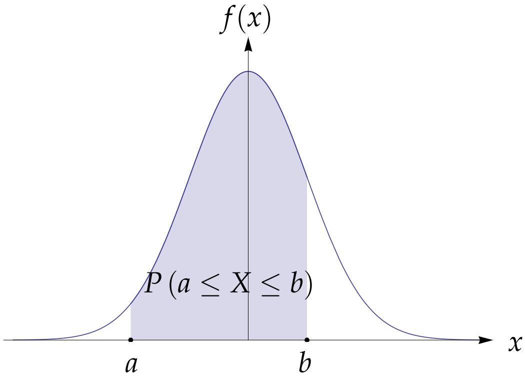
Neprekidne distribucije - motivacija i definicija

- model za **neprekidnu numeričku varijablu** je slučajna varijabla za koju je $\mathcal{R}(X)$ interval realnih brojeva ili je $\mathcal{R}(X) = \mathbb{R}$
- za slučajnu varijablu X kažemo da je **neprekidna** ako postoji funkcija $f : \mathbb{R} \to [0, \infty\rangle$ takva da je vjerojatnost

$$P\left(a \leq X \leq b\right) = P\left(a < X \leq b\right)$$

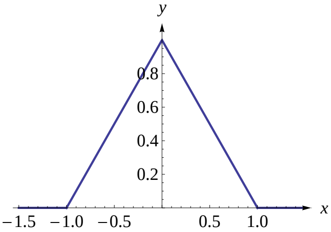
  jednaka površini između x-osi i grafa funkcije f nad intervalom $[a, b]$ (slika na sljedećem slide-u)
- takvu funkciju f zovemo **funkcija gustoće** NSV X
- NSV je zadana ako je poznata njena funkcija gustoće - tada kažemo da poznajemo **razdiobu** ili **distribuciju** NSV

Neprekidne distribucije - računanje vjerojatnosti



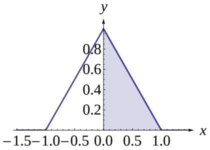
Vjerojatnost kao površina između x-osi i grafa funkcije f nad intervalom $[a, b]$

Neprekidne distribucije - računanje vjerojatnosti

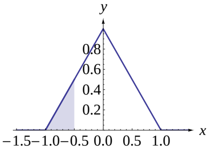


Graf funkcije gustoće jedne neprekidne distribucije

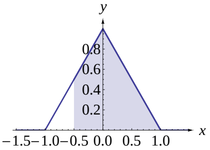
Neprekidne distribucije - računanje vjerojatnosti



(a) $P\left(X \in (0,1)\right) = \frac{1}{2}$



(b) $P\left(X \in (-1, -\frac{1}{2})\right) = \frac{1}{8}$



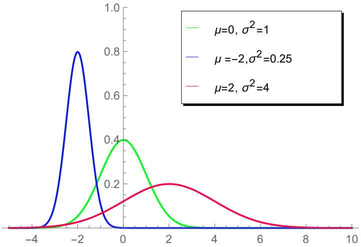
(c) $P\left(X \in (-\frac{1}{2}, 1)\right) = \frac{7}{8}$

Računanjem površine ispod grafa funkcije gustoće slučajne varijable X određujemo vjerojatnost da se X realizira realnim brojem iz zadanog intervala

Normalna distribucija

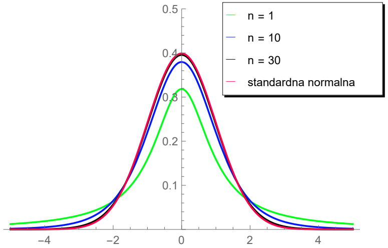
- normalna $X \sim \mathcal{N}(\mu, \sigma^2)$ distribucija je neprekidna distribucija za koju je
$$\mathcal{R}(X) = \mathbb{R},$$
a funkcija gustoće definirana je izrazom
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R},$$
$\mu \in \mathbb{R}$ i $\sigma > 0$ zovemo parametrima te distribucije
- $\mu = 0, \sigma^2 = 1$ - **standardna normalna distribucija**
- graf funkcije gustoće normalne distribucije ovisi o izboru parametara μ i σ^2 (slika na sljedećem slide-u)

Normalna distribucija



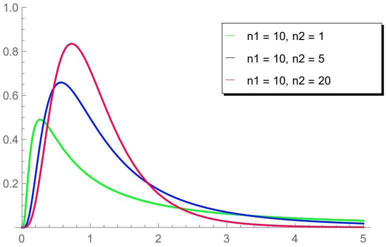
Graf funkcije gustoće normalne distribucije za različite μ and σ^2

Studentova distribucija



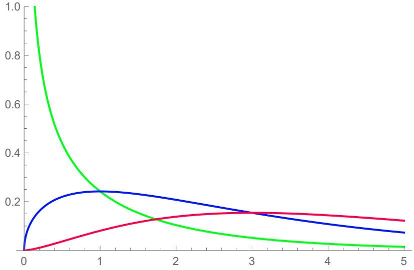
Graf funkcije gustoće Studentove distribucije za različite vrijednosti parametra

FISHEROVA ILI F DISTRIBUCIJA



Graf funkcije gustoće F distribucije za različite vrijednosti parametara

χ^2 DISTRIBUCIJA



Graf funkcije gustoće studentove distribucije za različite vrijednosti parametra ($n = 1, 3, 5$)

MJERE SREDINE - MEDIJAN

- **medijan** slučajne varijable X je realan broj m za kojega vrijedi da je

$$P(X \geq m) \geq \frac{1}{2} \quad \text{i} \quad P(X \leq m) \geq \frac{1}{2}$$

- medijan je također jedna mjera sredine
- ne mora biti jedinstven

MJERE SREDINE I RASPRŠENOSTI DSV

- **diskretna slučajna varijabla X** - ako postoji, **očekivanje** je realan broj

$$\mu = EX = \sum_{\text{svim } x_i \in \mathcal{R}(X)} x_i p_i,$$

a **varijanca** realan broj

$$\sigma^2 = Var\, X = \sum_{\text{svim } x_i \in \mathcal{R}(X)} (x_i - \mu)^2 p_i$$

- očekivanje - mjera sredine
- varijanca i standardna devijacija - mjere raspršenja vrijednosti slučajne varijable oko očekivanja

Mjere sredine i raspršenosti NSV

- **neprekidna slučajna varijabla X** - ograničit ćemo se na normalnu slučajnu varijablu za koju je

$$EX = \mu, \quad Var\,X = \sigma^2$$

- **standardna devijacija** σ - kvadratni korijen iz varijance

Čebiševljeva nejednakost

- neka je X slučajna varijabla s očekivanjem μ i standardnom devijacijom σ
- tada za svaki prirodan broj k vrijedi:

$$P\{|X-\mu| \geq k\sigma\} \leq \frac{1}{k^2}, \text{ tj.}$$

$$P\{|X-\mu| < k\sigma\} \geq 1 - \frac{1}{k^2}$$

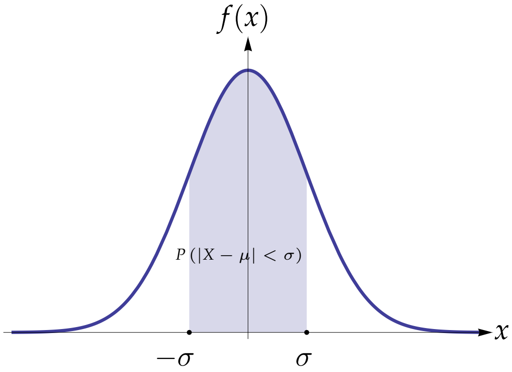
Čebiševljeva nejednakost - interpretacija

sljedeće tvrdnje vrijede za sve slučajne varijable koje imaju varijancu:

- vjerojatnost da se slučajna varijabla realizira vrijednostima koje su od očekivanja μ udaljene više ili jednako $k\sigma$ manja je ili jednaka $1/k^2$
- vjerojatnost da se slučajna varijabla realizira vrijednostima koje su od očekivanja μ udaljene manje od $k\sigma$ veća je od $1 - 1/k^2$
- uvrštavajući $k = 3$ vidimo da realizacija slučajne varijable upada u interval $(\mu - 3\sigma, \mu + 3\sigma)$ s vjerojatnošću većom od 0.88 (≈ 0.9)
- to praktično znači da se više od 88% realizacija slučajne varijable X nalazi u intervalu $(\mu - 3\sigma, \mu + 3\sigma)$

Čebiševljeva nejednakost - geometrijska interpretacija

$P(|X-\mu|<k\sigma)$ za $k=1$ za standardnu normalnu slučajnu varijablu X

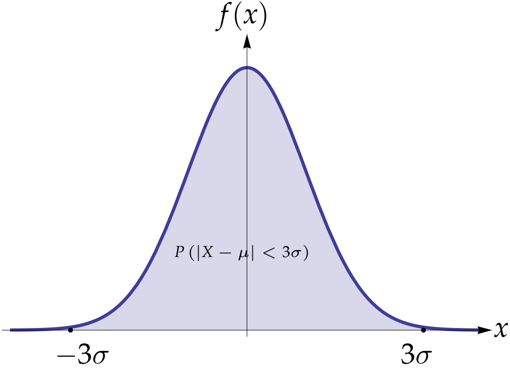


prema Čebiševljevoj nejednakosti je

$$P(|X-\mu|<\sigma)=P(X\in(\mu-\sigma,\mu+\sigma))\geq 0$$

Čebiševljeva nejednakost - geometrijska interpretacija

$P\left(|X-\mu|<k\sigma\right)$ za $k=3$ za standardnu normalnu slučajnu varijablu X



prema Čebiševljevoj nejednakosti je

$$P\left(|X-\mu|<3\sigma\right)=P\left(X\in\left(\mu-3\sigma,\mu+3\sigma\right)\right)\geq\frac{8}{9}$$

Normalna distribucija - primjer izračunavanja vjerojatnosti

Primjer

Slučajna varijabla Z ima standardnu normalnu distribuciju, $Z \sim \mathcal{N}(0,1)$. Pomoću R-a odredimo sljedeće vjerojatnosti (R - primjer 2):

1. $P\left(\mu - \sigma < Z < \mu + \sigma\right) = 0.68$
2. $P\left(\mu - 2\sigma < Z < \mu + 2\sigma\right) = 0.95$
3. $P\left(\mu - 3\sigma < Z < \mu + 3\sigma\right) = 0.9972$
4. $P\left(-0.5 \leq Z \leq 1.1\right) = 0.555796$
5. $P\left(-0.38 \leq Z \leq 1.72\right) = 0.605311$
6. $P\left(Z \geq 1.6\right) = 0.054799$
7. $P\left(Z \leq -1.8\right) = 0.035930$

Problem procjene

- **obilježja** obuhvaćena istraživanjem nazivamo **varijablama** - na razini populacije modeliramo ih **slučajnim varijablama**
- **vrijednosti varijable izmjerene na uzorku** (stupac baze podataka) - **nezavisne realizacije slučajne varijable**
- slučajna varijabla - u potpunosti zadana svojom **distribucijom**
- distribucija omogućuje izračunavanje **vjerojatnosti** vezanih uz realizacije slučajne varijable i njezinih **numeričkih karakteristika** (očekivanje, varijanca, standardna devijacija i sl.)
- **nepoznata distribucija** slučajne varijable - problem!

Problem procjene

- X - slučajna varijabla kojom modeliramo neko obilježje
- x - podatak, tj. vrijednost promatranog obilježja za jedinku iz uzorka, tj. jedna realizacija slučajne varijable X
- na uzorku koji se sastoji od n jedinki izmjerimo/zabilježimo n podataka $x_1, \ldots, x_n$ - svaki podatak x_i jedna je realizacija slučajne varijable X_i koja ima istu distribuciju kao slučajna varijabla X
- postupak prikupljanja podataka na uzorku mora biti takav da su mjerenja međusobno nezavisna
- podaci $(x_1, \ldots, x_n)$ - realizacija n-torke slučajnih varijabli $(X_1, \ldots, X_n)$ koje su međusobno nezavisne i imaju istu distribuciju kao slučajna varijabla X
- $(X_1, \ldots, X_n)$ - **jednostavni slučajni uzorak** (JSU)
- $(x_1, \ldots, x_n)$ - realizacija JSU, zovemo ju samo **uzorak**

PROCJENA DISTRIBUCIJE DISKRETNE SLUČAJNE VARIJABLE

- **empirijska distribucija slučajne varijable** - procjena distribucije DSV
- npr. za DSV X sa slikom $\mathcal{R}(X) = \{y_1, \ldots, y_m\}$ i nepoznatim vjerojatnostima $p_i = P(X = y_i)$ empirijska distribucija je dana tablicom

$$\begin{pmatrix} y_1 & y_2 & \ldots & y_m \\ f_1/n & f_2/n & \ldots & f_m/n \end{pmatrix},$$

  gdje je n dimenzija uzorka, a f_i frekvencija pojavljivanja podatka/vrijednosti y_i u uzorku $(x_1, \ldots, x_n)$
- veći broj podataka - opravdanje korištenje empirijske distribucije, u skladu sa statističkim pristupom modeliranju vjerojatnosti

Procjena distribucije općenito

- empirijska distribucija slučajne varijable temeljena na uzorku (podacima) uvijek je diskretna
- temeljem poznavanja svojstava varijable koju proučavamo često možemo unaprijed odrediti tip neprekidne distribucije koju je opravdano koristiti za njezino modeliranje

Procjena očekivanja

- problem procjene očekivanja slučajne varijable X na temelju realizacije, tj. podataka, $(x_1, \ldots, x_n)$ jednostavnog slučajnog uzorka $(X_1, \ldots, X_n)$
  - procjenitelj - aritmetička sredina JSU $(X_1, \ldots, X_n)$
$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$
  - procjena - aritmetička sredina uzorka (podataka) $(x_1, \ldots, x_n)$
$$\overline{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$$
  - procjenitelj $\overline{X}_n$ je slučajna varijabla s nekom distribucijom
  - procjena $\overline{x}_n$ je realan broj - realizacija procjenitelja $\overline{X}_n$

Procjena varijance

- problem procjene varijance slučajne varijable X na temelju realizacije, tj. podataka, $(x_1, \ldots, x_n)$ jednostavnog slučajnog uzorka $(X_1, \ldots, X_n)$
  - procjenitelj - korigirana varijanca JSU $(X_1, \ldots, X_n)$

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

  - procjena - korigirana varijanca uzorka (podataka) $(x_1, \ldots, x_n)$

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

  - procjenitelj S_n^2 je slučajna varijabla s nekom distribucijom
  - procjena s_n^2 je realan broj - realizacija procjenitelja S_n^2

INTERVALNA PROCJENA

- pri procjeni očekivanja i varijance jednim brojem treba uzeti u obzir slučajni karakter procjene - radi se o realizaciji slučajnih varijabli $\overline{X}_n$ i S_n^2
- **intervalna procjena** - procjena parametra **intervalom unaprijed izabrane pouzdanosti** γ u kojemu se stvarna, nama nepoznata vrijednost parametra, nalazi s vjerojatnošću $100\gamma\%$
- **interval izabrane pouzdanosti** γ - interval koji ima slučajne varijable kao granice i određen je temeljem zahtjeva da se stvarna vrijednost parametra kojeg procjenjujemo nalazi u takvom, slučajnom, intervalu s vjerojatnošću barem γ
- npr. 95%-tni pouzdani interval - interval poudanosti $\gamma = 0.95$; s vjerojatnošću 0.95 sadrži stvarnu vrijednost nepoznatog parametra
- izračunavanje granica intervala pouzdanosti γ - na temelju podataka $(x_1, \ldots, x_n)$ izračunavamo običan interval s realnim brojevima kao granicama, koji u $100\gamma\%$ slučajeva sadrži stvarnu vrijednost karakteristike koju procjenjujemo (simulacija u R-u)

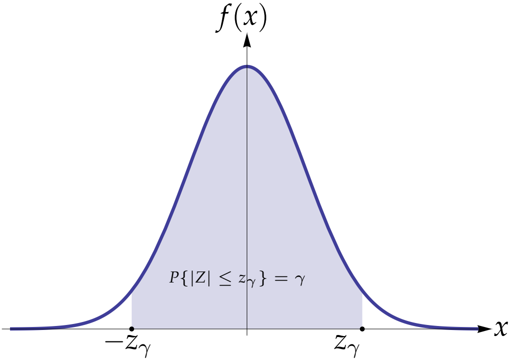
Intervalna procjena očekivanja

- $\overline{X}_n$ - aritmetička sredina JSU $(X_1, \ldots, X_n)$ velike dimenzije n iz distribucije slučajne varijable X s nepoznatim očekivanjem μ i poznatom varijancom σ^2
- aritmetička sredina $\overline{X}_n$ za veliki n ima približno normalnu distribuciju s očekivanjem μ i varijancom $\frac{\sigma^2}{n}$
- slučajna varijabla

$$Z' = \frac{\overline{X}_n - E\overline{X}_n}{\sqrt{Var(\overline{X}_n)}} = \frac{\overline{X}_n - \mu}{\sigma}\sqrt{n}$$

  ima **približno standardnu normalnu distribuciju** $\mathcal{N}(0,1)$
- standardnu normalnu distribuciju označimo $Z \sim \mathcal{N}(0,1)$
- neka je z_γ broj za koji vrijedi $P\{|Z| \leq z_\gamma\} = \gamma$
- vrijednost γ predstavljaju površinu ispod grafa funkcije gustoće standardne normalne distribucije nad intervalom $[-z_\gamma, z_\gamma]$

Intervalna procjena očekivanja



vjerojatnost $P\{|Z| \leq z_\gamma\}$

Intervalna procjena očekivanja

- uvrštavanjem izraza $Z' = \frac{\overline{X}_n - \mu}{\sigma}\sqrt{n}$ u izraz $P\{|Z| \leq z_\gamma\}$ umjesto Z slijedi:

$$P\left\{|Z'| \leq z_\gamma\right\} = P\left\{-z_\gamma \leq Z' \leq z_\gamma\right\} =$$

$$= P\left\{-z_\gamma \leq \frac{\overline{X}_n - \mu}{\sigma}\sqrt{n} \leq z_\gamma\right\} =$$

$$= P\left\{\overline{X}_n - z_\gamma \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}_n + z_\gamma \frac{\sigma}{\sqrt{n}}\right\}$$

- slijedi da je

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

- za velike uzorke procjenjujemo σ korijenom korigirane varijance uzorka, tj. brojem s_n

Intervalna procjena očekivanja

ako je $(x_1, \ldots, x_n)$ realizacija JSU iz distribucije slučajne varijable X i $\gamma \in (0,1)$, onda će u približno $100\gamma\%$ slučajeva interval izračunat 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 $(x_1, \ldots, x_n)$
σ - poznata standardna devijacija slučajne varijable X
z_γ - broj za koji vrijedi da je $P\{|Z| \leq z_\gamma\} = \gamma$
Z - standardna normalna slučajna varijabla

sadržavati stvarnu (nepoznatu) vrijednost očekivanja μ slučajne varijable X

R - primjeri 3 i 4

STATISTIČKE HIPOTEZE

- **statistička hipoteza** $\mathcal{H}$ - pretpostavka vezana uz distribuciju slučajne varijable kojom modeliramo neko obilježje populacije
- **testirati hipotezu** znači donijeti odluku o tome hoćemo li $\mathcal{H}$ odbaciti ili ne
- zbog toga često govorimo o testiranju dvije hipoteze - **nul-hipoteze** $\mathcal{H}_0$ i **alternativne hipoteze** $\mathcal{H}_1$
- $\mathcal{H}_1$ **prihvaćamo u slučaju odbacivanja** $\mathcal{H}_0$
- **nul-hipotezu** $\mathcal{H}_0$ **nikada ne možemo prihvatiti** jer se odlučivanje u statističkom testu provodi uz toleranciju malih vjerojatnosti pogrešne odluke

Statističke hipoteze

- **odluka donesena temeljem statističkog testa može biti ili pogrešna ili ispravna**
- tipovi pogrešne odluke:
  - **pogreška I. tipa: odbaciti $\mathcal{H}_0$ ako je ona istinita**
  - **pogreška II. tipa: ne odbaciti $\mathcal{H}_0$ ako je $\mathcal{H}_1$ istinita**
- statistički test se dizajnira tako da dopušta izbor maksimalne vjerojatnosti pogreške prvog tipa (npr. 0.01, 0.05 ili 0.1)
- α - odabrana maksimalna vjerojatnost pogreške prvog tipa zove se **razina značajnosti statističkog testa**

TESTIRANJE HIPOTEZE O OČEKIVANJU - T-TEST

- hipoteze:

$$\mathcal{H}_0 : \quad \mu = \mu_0$$

$$\mathcal{H}_1 : \quad \mu > \mu_0 \text{ ili } \mathcal{H}_1 : \quad \mu < \mu_0$$

- vrijednost test-statistika:

$$\hat{t} = \frac{\overline{x}_n - \mu_0}{s_n/\sqrt{n}},$$

gdje je n dimenzija uzorka, $\overline{x}_n$ aritmetička sredina uzorka, a s_n standardna devijacija uzorka

- p-vrijednost testa:
  - $p = P\{T \geq \hat{t}\}$ ako je $\mathcal{H}_1 : \mu > \mu_0$
  - $p = P\{T \leq \hat{t}\}$ ako je $\mathcal{H}_1 : \mu < \mu_0$
- slučajna varijabla T ima Studentovu distribuciju s $(n-1)$ stupnjeva slobode

TESTIRANJE HIPOTEZA O OČEKIVANJU - T-TEST

- p-vrijednost uspoređujemo s razinom značajnosti α:
  - ako je $p < \alpha$ odbacujemo $\mathcal{H}_0$ na razini značajnosti α i prihvaćamo $\mathcal{H}_1$
  - ako je $p > \alpha$ zaključujemo da nemamo dovoljno informacija koje bi poduprle odluku o odbacivanju $\mathcal{H}_0$
- ovaj test adekvatan je za male uzorke iz normalne distribucije ($n < 30$) te za velike uzorke iz nespecificirane distribucije s očekivanjem μ i varijancom σ^2
- R - primjeri 5 i 6

Testiranje hipoteza o vjerojatnosti - binomni test

- hipoteze:

$$\mathcal{H}_0: \quad p = p_0$$

$$\mathcal{H}_1: \quad p > p_0 \text{ ili } \mathcal{H}_1: \quad p < p_0$$

- iz podataka znamo broj realizacija događaja o čijoj vjerojatnosti zaključujemo - frekvencija uspjeha f_1
- pomoću f_1 procjenjujemo nepoznatu vjerojatnost tog događaja p relativnom frekvencijom (proporcijom) $\hat{p} = f_1/n$
- p-vrijednost testa
    - $p = P(B \geq f_1)$ ako je $\mathcal{H}_1: p > p_0$
    - $p = P(B \leq f_1)$ ako je $\mathcal{H}_1: p < p_0$
- $B \sim \mathcal{B}(n, p_0)$
- odluku donosimo usporedbom p-vrijednosti s razinom značajnosti α
- R - primjeri 7 i 8

TESTIRANJE HIPOTEZA O DISTRIBUCIJI - χ^2 TEST

- X slučajna varijabla s (nepoznatom) teorijskom distribucijom:

$$X \sim \begin{pmatrix} y_1 & y_2 & \dots & y_m \\ p_1 & p_2 & \dots & p_m \end{pmatrix}.$$

- testiranje hipoteze da podaci $(x_1, \dots, x_n)$ dolaze iz distribucije slučajne varijable X zadane gornjom tablicom može se provesti χ^2-testom:

$\mathcal{H}_0$: distribucija iz koje dolaze podaci jednaka je teorijskoj
$\mathcal{H}_1$: distribucija iz koje dolaze podaci razlikuje se od teorijske

TESTIRANJE HIPOTEZA O DISTRIBUCIJI - χ^2 TEST

- vrijednost test-statistike

$$\widehat{d} = \sum_{i=1}^{m} \frac{(np_i - f_i)^2}{np_i}$$

  temeljena na odstupanju stvarnih frekvencija podataka $(f_i, i = 1, \ldots, m)$ od teorijskih frekvencija $(np_i, i = 1, \ldots, m)$
- p-vrijednost - $p \approx P(D > \widehat{d})$
- ako je $\mathcal{H}_0$ istinita, za velike n slučajna varijabla D ima približno $\chi^2(m-1)$ distribuciju
- korištenje χ^2 testa - prikladno ako su sve teorijske frekvencije veće od 5
- R - primjer 9

TESTIRANJE NORMALNOSTI OBILJEŽJA

- statističke hipoteze:

$$\begin{aligned} \mathcal{H}_0: &\quad \text{varijabla } X \text{ ima normalnu distribuciju} \\ \mathcal{H}_1: &\quad \text{varijabla } X \text{ nema normalnu distribuciju,} \end{aligned}$$

- dva testa koji se mogu koristiti za testiranje hipoteze o normalnosti:
  - Lillieforsova inačica Kolmogorov-Smirnovljevog testa
  - Shapiro-Wilkov W test
- R - primjeri 10 i 11

Literatura

- Benšić, M. i Šuvak, N., *Primijenjena statistika*, Odjel za matematiku, Sveučilište J.J. Strossmayera, Osijek, 2013.
- Benšić, M. i Šuvak, N., *Uvod u vjerojatnost i statistiku*, Odjel za matematiku, Sveučilište J.J. Strossmayera, Osijek, 2014.