Vjerojatnost i statistika

Statističko zaključivanje - jedna varijabla

---

9. siječnja 2024.

Procjena očekivanja, varijance i standardne devijacije

Procjena očekivanja, varijance i standardne devijacije

- **varijabla** - stupac u bazi podataka
- **slučajna varijabla** - model za varijablu
- **procjena očekivanja** slučajne varijable je aritmetička sredina podataka $(x_1, x_2, \ldots, x_n)$

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

- **procjena varijance** slučajne varijable je korigirana varijanca podataka $(x_1, x_2, \ldots, x_n)$

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

- **procjena standardne devijacije** slučajne varijable je $\sqrt{s_n^2}$

Zadatak - automobili.sta

a) Napravite histogram za varijablu *potrosnja*. Ima li smisla modelirati potrošnju goriva ovog modela automobila koristeći normalnu slučajnu varijablu?

b) Procijenite očekivanje i varijancu slučajne varijable kojom modeliramo potrošnju automobila.

c) Procijenite vjerojatnost da je potrošnja goriva manja od 5.5.

Jednostavni slučajni uzorak i procjenitelj

Definicija 1. Jednostavni slučajni uzorak (j.s.u.) je slučajni vektor $(X_1, X_2, \ldots, X_n)$ kojemu su komponente nezavisne i jednako distribuirane.

- izmjerene podatke $x_1, x_2, \ldots, x_n$ smatramo realizacijama komponenti jednostavnog slučajnog uzorka

Jednostavni slučajni uzorak i procjenitelj

Definicija 2. Procjenitelj je slučajna varijabla dobivena na temelju j.s.u. kojom modeliramo procjenu.

- npr. $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$ je procjenitelj za očekivanje slučajne varijable X kojom se modelira neko obilježje
- procjena je jedna realizacija procjenitelja

Intervalna procjena

Definicija 3. Neka je $\gamma \in (0,1)$. **Interval pouzdanosti** γ za procjenu neke numeričke karakteristike slučajne varijable je interval koji ima slučajne varijable kao granice i određen je zahtjevom da se stvarna vrijednost veličine nalazi u slučajnom intervalu s vjerojatnošću barem γ.

- jedna realizacija intervala pouzdanosti γ koju odredimo na temelju uzorka je običan interval realnih brojeva
- u barem $100\gamma\%$ slučajeva izračunati interval realnih brojeva sadržavat će stvarnu vrijednost veličine koju procjenjujemo

Intervalna procjena očekivanja za velike uzorke

- $\overline{X_n}$ za velike n ima približno $\mathcal{N}(\mu, \sigma^2/n)$ distribuciju
- ako $\overline{X_n}$ **standardiziramo**

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

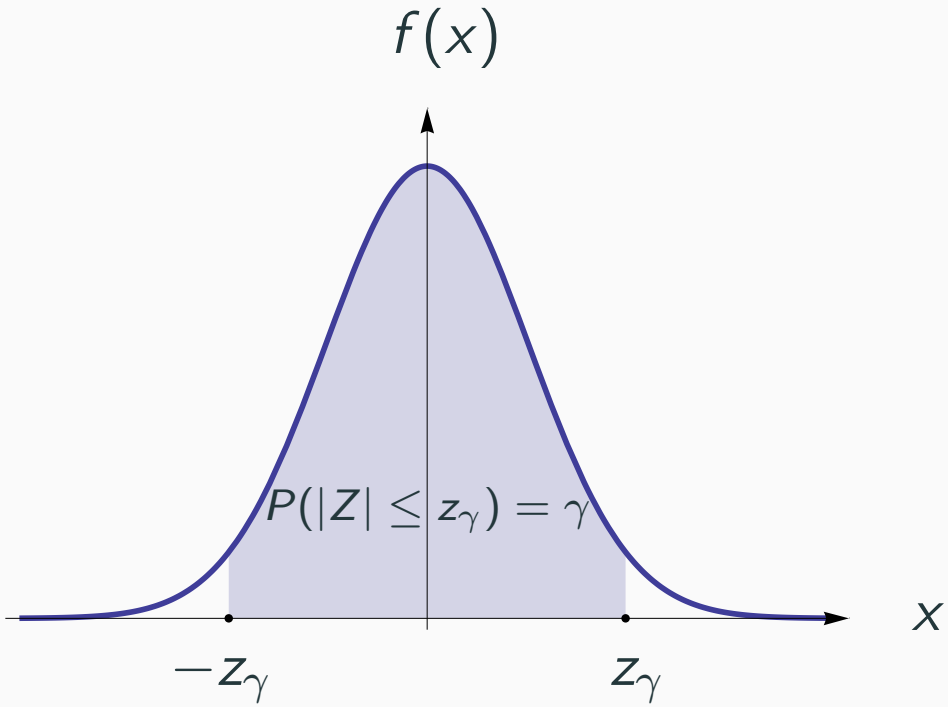
  dobivena slučajna varijabla ima približno $\mathcal{N}(0,1)$ distribuciju
- neka je $Z \sim \mathcal{N}(0,1)$ i z_γ broj za koji vrijedi

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

- uočimo da γ predstavlja površinu ispod grafa funkcije gustoće standardne normalne distribucije nad intervalom $[-z_\gamma, z_\gamma]$

$$P(|Z| \leq z_\gamma) = \frac{1}{\sqrt{2\pi}} \int_{-z_\gamma}^{z_\gamma} e^{-x^2/2}\, dx = \gamma$$

Intervalna procjena očekivanja za velike uzorke



Intervalna procjena očekivanja za velike uzorke

- ako uvrstimo $Z' = \frac{\overline{X}_n - \mu}{\sigma}\sqrt{n}$ umjesto Z, slijedi

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

- dakle, vrijedi

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

Intervalna procjena očekivanja za velike uzorke

Ako je $(x_1, \ldots, x_n)$ realizacija j.s.u. iz slučajne varijable X i $\gamma \in (0,1)$, onda će u približno $100\gamma\%$ slučajeva interval

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

sadržavati stvarnu (nepoznatu) vrijednost očekivanja μ slučajne varijable X.

$\overline{x}_n$ - aritmetička sredina uzorka

s_n - standardna devijacija uzorka

z_γ - broj za koji vrijedi $P(|Z| \leq z_\gamma) = \gamma$

Z - standardna normalna slučajna varijabla

Primjer - automobili.sta

Intervalom pouzdanosti 95% procijenite očekivanu potrošnju goriva danog modela automobila ako je

$$n = 300, \quad \overline{x}_{300} = 5.12, \quad s_{300} = 0.97.$$

- z_γ za $\gamma = 0.95$ određujemo pomoću kalkulatora vjerojatnosti u Statistici

$$z_\gamma = 1.959964 \approx 1.96$$

- uvrštavanjem slijedi

$$\overline{x}_{300} - z_\gamma \frac{s_{300}}{\sqrt{300}} \approx 5.01023$$

$$\overline{x}_{300} + z_\gamma \frac{s_{300}}{\sqrt{300}} \approx 5.22977$$

Primjer - automobili.sta

- realizacija intervala pouzdanosti 95% za očekivanje slučajne varijable kojom je modelirana potrošnja je $[5.01023, 5.22977]$
- u Statistici:

> Statistics $\rightarrow$ Basic Statistics $\rightarrow$ Descriptive Statistics $\rightarrow$ Variables $\rightarrow$ Advanced $\rightarrow$ označiti "Conf. limits for means interval" i odabrati vrijednost 95% $\rightarrow$ Summary

Zadatak - poduzetnici.sta

Intervalom pouzdanosti 95% i 97% procijenite očekivanu dob poduzetnika. U kakvom su odnosu dobiveni intervali pouzdanosti?

Rješenje:

- $I_{0.95} = [41.35088, 43.85912]$
- $I_{0.97} = [41.21490, 43.99510]$

Intervalna procjena vjerojatnosti

Intervalna procjena vjerojatnosti događaja za velike uzorke

- ukoliko varijabla poprima samo dvije vrijednosti (npr. 0 i 1), možemo ju modelirati Bernoullijevom slučajnom varijablom

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

- s obzirom da je $EX = p$, problem procjene vjerojatnosti p svodi se na problem procjene očekivanja
- za procjenu vjerojatnosti uspjeha p, na osnovu n nezavisnih ponavljanja pokusa, koristimo relativnu frekvenciju uspjeha u uzorku

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

pri čemu je f_1 frekvencija uspjeha u uzorku

Intervalna procjena vjerojatnosti događaja za velike uzorke

- $Z' = \frac{\hat{p}-p}{\sqrt{p(1-p)}}\sqrt{n}$ ima približno $\mathcal{N}(0,1)$ distribuciju
- neka je z_γ broj za koji vrijedi

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

gdje je $Z \sim \mathcal{N}(0,1)$

- pokazuje se da vrijedi

$$P\left(p \in \left[\hat{p} - z_\gamma\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_\gamma\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]\right) \approx \gamma$$

Intervalna procjena vjerojatnosti događaja za velike uzorke

Ako je $\hat{p}$ relativna frekvencija jedinica u n-dimenzionalnom uzorku iz Bernoullijeve distribucije i $\gamma \in (0,1)$, onda će u približno $100\gamma\%$ slučajeva interval

$$\left[\hat{p} - z_\gamma \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_\gamma \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right],$$

sadržavati pravu (nepoznatu) vrijednost vjerojatnosti p.

$\hat{p}$ - relativna frekvencija jedinice (uspjeha) u uzorku

z_γ - broj za koji vrijedi $P(|Z| \leq z_\gamma) = \gamma$

Z - standardna normalna slučajna varijabla

Intervalna procjena vjerojatnosti događaja za velike uzorke

Broj elemenata u uzorku n je dovoljno velik za primjenu ovakvog zaključivanja ako interval

$$\left[\hat{p} - 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

ne sadrži ni 0 ni 1.

Zadatak - tvornica

Tvornica hrane želi provesti istraživanje tržišta intervjuirajući 1000 potrošača kako bi odredila koju marku pahuljica za doručak preferiraju. Prikupljeni podaci pokazali su da 313 potrošača odabire upravo marku tvornice koja je provela istraživanje. Intervalom pouzdanosti 95% procijenite vjerojatnost da slučajno odabrani potrošač preferira pahuljice tvornice koja je provela istraživanje.

Rješenje: $[0.284, 0.342]$

Zadatak - tvornica

- u Statistici:

  Statistics $\rightarrow$ Power Analysis$\rightarrow$ Interval Estimation $\rightarrow$ One Proportion, Z, Chi-Square Test $\rightarrow$ unijeti vrijednost relativne frekvencije uspjeha, dimenziju uzorka te nivo pouzdanosti $\rightarrow$ Compute

Zadatak - vrtic.sta

U nekom poduzeću zaposleno je više od 3000 ljudi. Uprava poduzeća želi ponuditi pomoć svojim zaposlenicima oko organizacije čuvanja djece. Predložene su dvije opcije - otvaranje vrtića u sklopu poduzeća ili plaćanje dijela troškova čuvanja djece koje bi roditelji organizirali sami. Da bi se utvrdilo koja je od ovih dvaju mjera popularnija među zaposlenicima, odabran je uzorak od 60 roditelja s malom djecom koji su se izjasnili o tome koju opciju preferiraju. Njihovi odgovori su označeni na sljedeći način:

- 0 - radije bih novčanu pomoć za samostalno organiziranje čuvanja djece
- 1 - radije bih da se otvori vrtić u sklopu poduzeća.

Intervalom pouzdanosti 95% procijenite proporciju zaposlenika koji preferiraju otvaranje vrtića u okviru poduzeća.

Rješenje: $[0.51, 0.76]$