Teorija uzoraka

Metoda uzorka

- Statističko zaključivanje – na osnovu uzorka donosi zaključke o cijeloj populaciji.
- Takvi zaključci se temelje na podacima samo dijela jedinica, te stoga u sebi sadrže grešku nastalu zbog primjene uzorka (sampling error).
- Poželjno je stoga istraživanjem obuhvatiti sve jedinice statističkog skupa.
- Nedostaci takvog pristupa:
  - osnovni skup može biti beskonačan (nemoguće je prikupiti sve podatke)
  - cijena
  - dugo vremensko trajanje
- Dva osnovna zadatka metode uzoraka:
  - procjena nepoznatih parametara
  - ispitivanje pretpostavki o osobinama populacije. Pretpostavke se zovu hipoteze.

Procjena parametara

- Na primjer, zanima nas prosječna visina stanovnika Crne Gore. Uzmemo uzorak od n ljudi, odredimo srednju vrijednost, te nam je ona procjena za nepoznati parametar – srednju vrijednost svih stanovnika.
  - Koliko je ta procjena vjerodostojna?
  - Koliko dobijena srednja vrijednost odstupa od traženog parametra?
- Oznake za srednju vrijednost i varijansu uzorka, odnosno populacije

Uzorak	Populacija
$\bar{x}$	μ ili $E(X)$
s^2	σ^2 ili $V(X)$

Odabir uzorka

- Teorija koja daje odgovor na prethodna pitanja zasniva se na pretpostavci slučajnog biranja uzorka. Matematičkim jezikom to znači: – odabir pojedinih elemenata u uzorak mora biti međusobno nezavisan.
- Postoji više modela izbora slučajnog uzorka: jednostavni slučajni uzorak, stratifikovani uzorak ...
- Kod jednostavnog slučajnog uzorka njegov izbor se sprovodi na način da svaki član ima jednaku vjerojatnoću izbora. Pri tom se koristimo nekom objektivnom metodom odabiranja elemenata u uzorak (tablica slučajnih brojeva). Ovaj model primjenjuje se kada osnovni skup sadrži malu promjenljivost svojstva koje se ispituje – *homogen skup*.

Odabir uzorka 2

- Stratifikovani uzorak se koristi kad imamo veći stepen varijabilnosti podataka, odnosno kad želimo osigurati da su nam obuhvaćeni podaci iz raznih kategorija.
- U prethodnom primjeru sa visinama želimo da osiguramo da su u uzorku zastupljene sve regije Crne Gore.
- To možemo postići tako da osnovni skup podijelimo na nekoliko disjunktnih dijelova, te u svakom izaberemo odgovarajući broj podataka.
  - Pri tome je važno da podskupovi nastali gornjom podjelom imaju manji stepen varijabilnosti od same populacije

Uzorci iz normalne raspodjele

- Neka je na osnovnom skupu definisana varijabla X normalne raspodjele $N(\mu,\sigma^2)$. Uzmemo uzorak veličine n, te mu odredimo aritmetičku sredinu

$$x_{avg} = (x_1 + x_2 + ... + x_n) / n.$$

- Prvom sabirku možemo pridružiti slučajnu varijablu X_1 – ona predstavlja vrijednost prvog elementa uzorka, odnosno x_1. Analogno definišemo varijable X_2, X_3, ..., X_n. Prema pretpostavci slučajnosti uzorka, ove varijable su nezavisne, i svaka od njih ima normalnu raspodjelu jednaku polaznoj. Stoga je varijabla

$$X_{avg} = (X_1 + X_2 + ... + X_n) / n$$

linearna kombinacija normalnih varijabli, te je stoga i sama takva. Zanima nas koji su parametri raspodjele X_{avg}.

Uzorci iz normalne raspodjele 2

- Kako matematičko očekivanje ima svojstvo linearnosti, to je

$$E\left(\bar{X}\right) = E\left(\frac{1}{n}\sum_i X_i\right) = \frac{1}{n}\sum_i E\left(\bar{X}_i\right) = \frac{1}{n} n\mu = \mu$$

odnosno aritmetičke sredine uzorka se rasipavaju oko očekivanja osnovnog skupa.

- Primijetimo da ovdje nijesmo koristili zakon raspodjele varijable X, pa tvrđenje vrijedi i za uzorke iz ostalih distribucija.
- Izračunajmo i varijansu varijable X_{avg}.

Uzorci iz normalne raspodjele 3

- Za nezavisne varijable X i Y važi

$$V(\alpha X + \beta Y) = \alpha^2 V(X) + \beta^2 V(Y)$$

- Za X_{avg} imamo

$$V(\bar{X}) = V\left(\frac{1}{n}\sum_i X_i\right) = \frac{1}{n^2}\sum_i V\left(\bar{X}_i\right) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

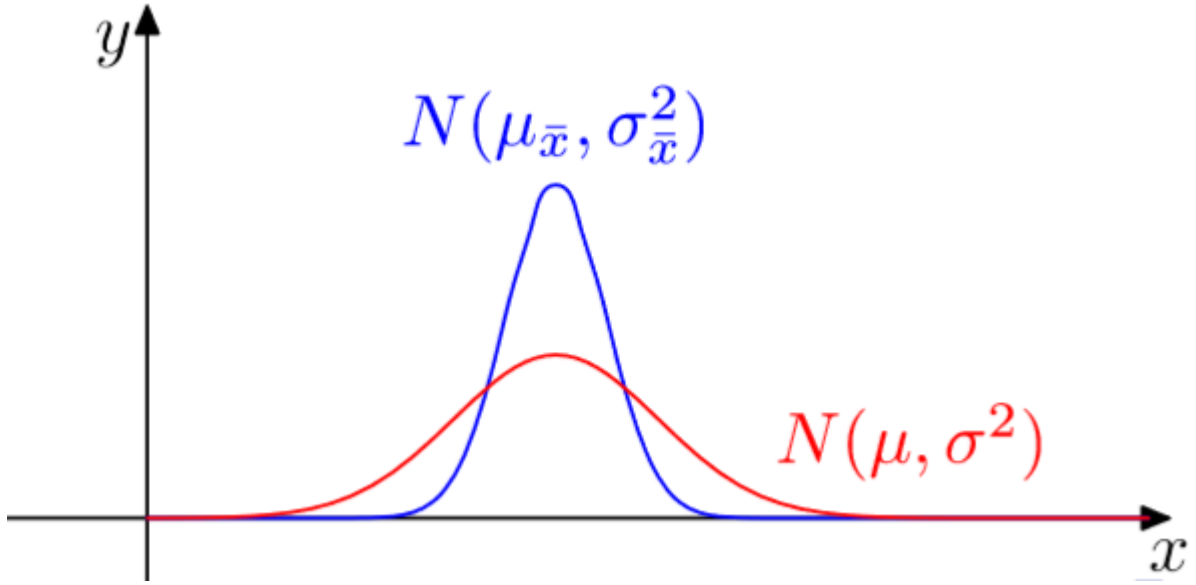
Uzorci iz normalne raspodjele 4

- Zaključimo: ako je X normalno distribuirana varijabla, tada aritmetička sredina uzorka ima raspodjelu

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- Iz slike vidimo da je normalna raspodjela za X_{avg} uža od raspodjele za X. Aritmetičke sredine uzoraka

rasipavaju se oko očekivanja μ u užem području negoli same vrijednosti varijable X. To područje će biti uže što je uzorak veći, jer je u tom slučaju standardna devijacija manja.



Primjer 1

- Zadata je normalna varijabla X ~ N(50,4). U kojim granicama će se kretati aritmetičke sredine slučajnih uzoraka veličine n = 16 elemenata?

Rješenje: Standardna devijacija aritmetičkih sredina je $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2}{4} = 0.5$

Budući da su aritmetičke sredine normalno raspoređene oko očekivanja μ = 50, to će njih oko 95% ležati u intervalu

$$\mu \pm 2\sigma_{\bar{x}}$$

odnosno

$$49 \leqslant \bar{x} \leqslant 51$$

Primjer 2

- Visina čovjeka ima približno normalnu raspodjelu N(167.5, 8^2cm).
  - Ako je slučajno odabran jedan čovjek, koja je vjerovatnoća da je njegova visina izmedu 152.5 i 185cm?
  - Ako je slučajno izabran uzorak od 64-ro ljudi, koja je vjerovatnoća da je njihova prosječna visina izmedu 165 i 170cm?

a)

$$z_1 = \frac{152,5 - 167,5}{8} = -1.88, \quad z_2 = \frac{185 - 167,5}{8} = 2.19$$

Stoga je

$$P\{152,5 \leqslant X \leqslant 167,5\} = P\{-1.88 \leqslant z \leqslant 2.19\}$$
$$= \Phi(2.19) - \Phi(-1.88) = 0.9556.$$

Primjer 2

b)

$$\mu_{\bar{x}} = \mu = 167.5, \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{8}{\sqrt{64}} = 1\,.$$

$$z_1 = \frac{165 - 167,5}{1} = -2.5, \quad z_2 = \frac{170 - 167,5}{1} = 2.5$$

$$P\{165 \leqslant \bar{X} \leqslant 170\} = \Phi(2.5) - \Phi(-2.5) = 0.9876.$$

Centralna granična teorema

- Šta ako polazna varijabla nema normalnu raspodjelu? Tada ni aritmetičke sredine ne moraju biti normalno raspoređene. Međutim, za njih i dalje vrijede izvedene formule za očekivanje i varijansu, jer su te formule izvedene bez pretpostavke o raspodjeli polazne varijable X.

$$E(\bar{X}) = \mu, \quad V(\bar{X}) = \frac{\sigma^2}{n}$$

- Centralna granična teorema: Raspodjela aritmetičkih sredina uzoraka teži ka normalnoj raspodjeli $N\left(\mu, \frac{\sigma^2}{n}\right)$ kad veličina uzorka n teži u beskonačnost.

- Običnim riječima (i matematički manje precizno) možemo reći da se aritmetičke sredine približno pokoravaju zakonu raspodjele, bez obzira na zakon raspodjele osnovnog skupa.

$$N\left(\mu, \frac{\sigma^2}{n}\right)$$

Centralna granična teorema 2

- Vidimo da za dovoljno veliko n imamo sličan rezultat kao i za varijable iz normalne raspodjele. Pri tome će u praksi dovoljno veliko značiti
  $n > 30$, te ćemo u tom slučaju smatrati da je aproksimacija normalnom raspodjelom dovoljno tačna za praktična izračunavanja.

Primjer 3

- Fabrika cigareta tvrdi da raspodjela katrana u njenim cigaretama ima aritmetičku sredinu μ = 3.9 mg po cigareti, te standradnu devijaciju σ = 1 mg. Pretpostavimo da je inspekcija uzela uzorak od 80 cigareta, te mjerila vrijednost katrana u njima. Uz pretpostavku da su tvrdnje proizvođača istinite, nađite vjerovatnoću da je srednja vrijednost katrana u uzorku veća od 4.15 mg.

$$\mu_{\bar{x}} = \mu = 3.9, \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{80}} = 0.11$$

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{4.15 - 3.9}{0.11} = 2.27$$

$$P\{\bar{X} \geqslant 4.15\} = P\{z \geqslant 2.27\} = 1 - \Phi(2.27) = 1 - 0.9884 = 0.0116 = 1.16\%$$

Procjena srednje vrijednosti

- Vidjeli smo da za dovoljno veliko n aritmetička sredina ima približno normalnu raspodjelu $N\left(\mu, \frac{\sigma^2}{n}\right)$
- Na osnovu toga možemo izračunati vjerovatnoću $P\{\bar{x} \in \mu \pm k\sigma_{\bar{x}}\}$

odnosno vjerovatnoću da je

$$|\bar{x} - \mu| \leqslant k\sigma_{\bar{x}}$$

za neko k > 0. Međutim, najčešće ne znamo μ, (a ni σ), nego ih želimo odrediti na osnovu uzorka.

Drugim riječima, uzmemo uzorak, nađemo njegovu aritmetičku sredinu, te nam on predstavlja procjenu za nepoznati parametar μ.

Koliko je ta procjena dobra?

- Odgovor nam omogućuje centralna granična teorema

Intervali povjerenja

- Poznata nam je vjerovatnoća $|\bar{x} - \mu| \leqslant k\sigma_{\bar{x}}$

$$\bar{x} - k\sigma_{\bar{x}} \leqslant \mu \leqslant \bar{x} + k\sigma_{\bar{x}}$$

$$\mu \in \langle \bar{x} - k\sigma_{\bar{x}}, \bar{x} + k\sigma_{\bar{x}} \rangle$$

- Za k = 2, vjerovatnoća da važi posljednja formula je ~95%, odnosno sa tom vjerovatnoćom možemo da tvrdimo da se μ nalazi u intervalu

$$\langle \bar{x} - 2\sigma_{\bar{x}}, \bar{x} + 2\sigma_{\bar{x}} \rangle$$

- Interval ovakvog oblika naziva se interval povjerenja
- Kako postići veću pouzdanost, na primjer 99%?

Intervali povjerenja 2

- U opštem slučaju želimo da odredimo (1-α)% interval povjerenja za μ
- Tada je α obično mali broj 1%-5%
- Interval koji tražimo je oblika

$$\bar{x} \pm d$$

$$P\{\mu \in \langle \bar{x} - d, \bar{x} + d \rangle\} = 1 - \alpha$$

- Širina intervala zavisi od α

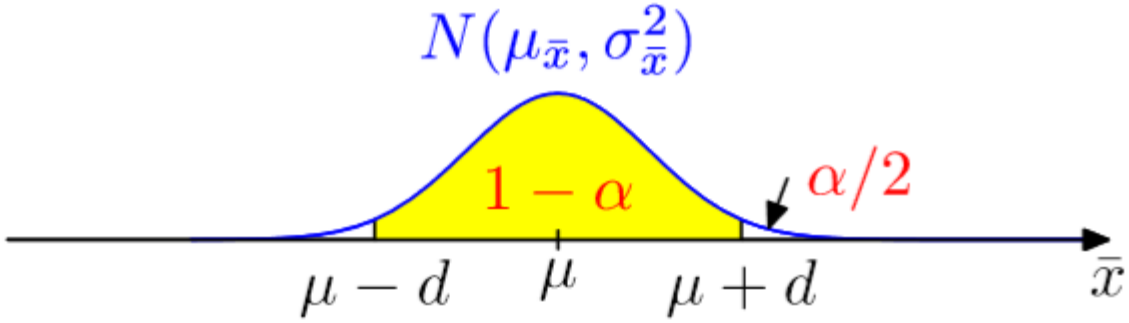
Intervali povjerenja 2

- Kako je za velike uzorke

$$P\{\mu \in \langle \bar{x} - d, \bar{x} + d \rangle\} = P\{\bar{x} \in \langle \mu - d, \mu + d \rangle\}$$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

- Prethodna vjerovatnoća jednaka je označenoj površini na slici



Intervali povjerenja 3

- Tražena širina intervala d računa se

$$d = z_{\alpha/2} \cdot \sigma_{\bar{x}}$$

$$\Phi(z_{\alpha/2}) = 1 - \frac{\alpha}{2}$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$

- Što je uzorak veći, to je procjena bolja (manja širina intervala)
- Zamjenom standardne devijacije σ s procjenom s dobijenom iz velikog uzorka, pravi se neznatna greška, koja se u praksi može tolerisati. Zamjenu je bilo nužno sprovesti, budući da u praksi obično ne znamo σ (kao ni μ)

Primjer 4

- Procijeniti prosječnu zaradu na uzorku od 100 ljudi, pri čemu je x_{avg} = 4200, s = 1300. Nađite 95%-tni interval pouzdanosti za μ.

$$\Phi(z_{0.025}) = 0.975$$

$$z_{0.025} = 1.96$$

$$d = z_{0.025}\,\sigma_{\bar{x}} = 1.96 \cdot \frac{1300}{10} = 254.8$$

$$\bar{x} \pm d = \langle 3945.2, 4454.8 \rangle$$

Primjer 5

- Nađite 99%-tni interval povjerenja za μ, ako su parametri uzorka n = 50, x_{avg} = 1.315, s = 0.366

$$\Phi(z_{0.005}) = 0.995$$

$$z_{0.005} = 2.58$$

$$1.315 \pm 2.58\frac{0.366}{\sqrt{50}} = \langle 1.181, 1.449 \rangle$$