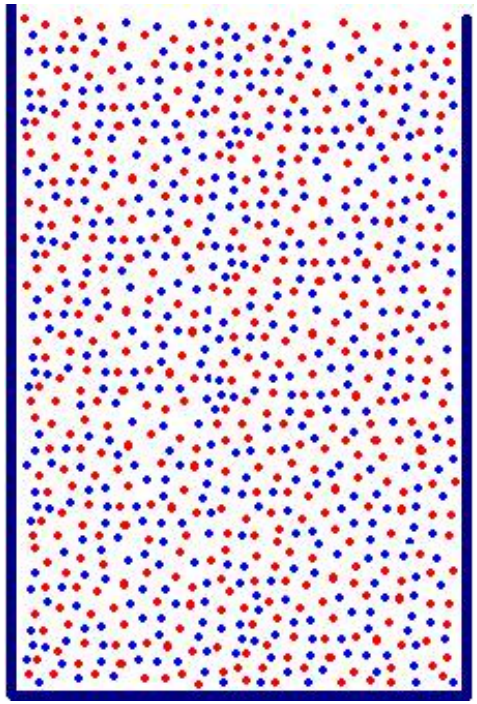
PROCJENA PARAMETARA POPULACIJE

Problem. U kutiji se nalaze crvene i plave kuglice. Koliki je udio crvenih kuglica?



Brojanje svih kuglica je prezahtjevan zadatak.
Može li se približni rezultat dobiti brže?

Rješenje. Slučajno izaberemo (izvučemo) određen broj kuglica i prebrojimo crvene i plave te izračunamo udio crvenih.

Ukoliko je od 10 izvučenih kuglica 4 crvene boje tada je udio crvenih 0,4 (40%).

Koliko je ova procjena točna?

Uočimo da se moglo dogoditi da izvučemo i 10 crvenih ili 10 bijelih kuglica.

Zbog slučajnog izbora kuglica, ukupan broj crvenih kuglica (a s time i njihov udio) je slučajan.

Ne možemo govoriti u terminima apsolutne točnosti već o vjerojatnosti da je naša procjena točna (u nekim granicama).

Da bismo mogli računati potrebne vjerojatnosti trebamo poznavati distribuciju vjerojatnosti broja crvenih kuglica. Koja je distribucija vjerojatnosti broja crvenih kuglica?

Kuglica može biti crvena ili plava (1 - crvena, 0 - plava). Izvukli smo 10 kuglica (10 puta smo izvlačili kuglicu).

Da li se radi o binomnoj distribuciji (ukupan broj pozitivnih ishoda)?

NE. Distribucija nije binomna. Ne radi se o 10 ponavljanja **ISTOG** pokusa.

Izvlačenje 10 kuglica je isto što i 10 uzastopnih izvlačenja kuglice. Ponavljamo izvlačenje!

PROBLEM. U prvom izvlačenju su sve kuglice u kutiji. U drugom izvlačenju je jedna manje (izvučena u prvom izvlačenju). U trećem izvlačenju su dvije manje u kutiji itd. Ne ponavljamo isti pokus (mjenjaju se vjerojatnosti)!

- Problem je u tome što ne vraćamo kuglicu u kutiju.
- Ukoliko bismo nakon svakog izvlačenja vratili kuglicu u kutiju, tada bi se radilo o ponavljanju istog pokusa i distribucija ukupnog broja crvenih kuglica bi bila baš binomna distribucija.

- Razliku između vraćanja i ne vraćanja kuglica u kutiju ilustrirat ćemo sljedećim primjerom.

Primjer. U kutiji se nalazi 5 crvenih i 5 plavih kuglica. Iz kutije se izvlače dvije kuglice, jedna po jedna. Izračunajte vjerojatnost da su izvučene jedna crvena i jedna plava kuglica ukoliko:

- Nakon svakog izvlačenja se kuglica vraća u kutiju.
- Nakon izvlačenja se kuglica ne vraća u kutiju.

Rješenje. Da bi se izvukle crvena i plava kuglica prvo treba izvući crvenu pa onda plavu (oznaka CP) ili prvo plavu pa onda crvenu (PC).

a) Ukoliko kuglice vraćamo tada je vjerojatnost izvlaćenja crvene (a i plave) kuglice u oba izvlaćenja jednaka 0.5 (u oba izvlaćenja se u kutiji nalazi 5 crvenih kuglica od 10 ukupno).

Tada je

$$\begin{aligned} P(CP) &= 0.5 \cdot 0.5 = 0.25 \\ P(PC) &= 0.5 \cdot 0.5 = 0.25. \end{aligned}$$

Sada je vjerojatnost da u dva izvlačenja izvućemo crvenu i plavu kuglicu jednaka

$$P(CP \text{ ili } PC) = P(CP) + P(PC) = 0.25 + 0.25 = 0.5$$

b) Ukoliko kuglice ne vraćamo tada je vjerojatnost da prvo izvućemo crvenu kuglicu jednaka 0.5, a da nakon toga izvućemo plavu, vjerojatnost je 5/9 (ako smo prvo izvukli crvenu, tada je u kutiji ostalo 5 plavih i 4 crvene kuglice), pa je

$$P(CP) = 0.5 \cdot \frac{5}{9}$$

Zbog istih razloga je i

$$P(PC) = 0.5 \cdot \frac{5}{9}.$$

Sada je vjerojatnost da u dva izvlačenja izvućemo crvenu i plavu kuglicu jednaka

$$P(CP \text{ ili } PC) = P(CP) + P(PC) = 0.5 \cdot \frac{5}{9} + 0.5 \cdot \frac{5}{9} = \frac{5}{9}$$

Izvlačenja s vraćanjem i bez vraćanja se razlikuju. Koristit ćemo izvlačenje s vraćanjem.

Napomena. Ukoliko je broj kuglica u kutiji velik, tada izvlačenje jedne kuglice minimalno utječe na vjerojatnost izvlačenja druge. Možemo izvlačiti bez vraćanja.

Ukoliko se u kutiji nalazi 1000 kuglica, 500 plavih i 500 crvenih, tada je vjerojatnost da ćemo izvući jednu plavu i jednu crvenu kuglicu jednaka

$$P(CP \text{ ili } PC) = 2 \cdot 0.5 \cdot \frac{500}{999} = 0.500501$$

Pitanje: Ukoliko je udio crvenih kuglica 0.4, kolika je vjerojatnost da se ukupan broj crvenih kuglica od 10 izvučenih nalazi između 3 i 5?

$$P(3 \leq X \leq 5) = P(X \leq 5) - P(X \leq 2)$$

Uz $n = 10$ i $p = 0.4$

$$P(3 \leq X \leq 5) = 0.834 - 0.167 = 0.667$$

Dakle, vjerojatnost da nam udio crvenih kuglica nakon 10 izvlačenja bude između 0.3 i 0.5 iznosi 0.667.

Uzorak

Populacija je skup jedinki (objekata) koje imaju neku zajedničku mjerljivu karakteristiku (obilježje).

- Skup o kojem želimo donijeti zaključak.

Uzorak je (bilo koji) podskup populacije.

- Na temelju poznavanja vrijednosti obilježja u uzorku želimo donijeti zaključak o cijeloj populaciji.

Uzorak

S obzirom na način izbora

- namjerni uzorak i
- vjerojatnosni uzorak.

NAMJERNI - prigodni, kvotni

NE MOŽE SE IZRAČUNATI POGREŠKA!!

Vjerojatnosni uzorak

Svaka jedinica populacije može biti izabrana u uzorak s određenom vjerojatnošću.

Svaki član populacije ima vjerojatnost izbora u uzorak.

Vjerojatnosni uzorak omogućava računanje pogrešaka procjena nastalih zbog primjene uzorka.

Vjerojatnosni uzorak

- (Jednostavni) slučajni uzorak (random sample)
- Stratificirani slučajni uzorak (stratified random sample)
- Uzorak skupina (cluster sampling)

Jednostavni slučajni uzorak

Svaki član skupa ima **JEDNAKU** vjerojatnost izbora u uzorak.

Izabiremo n elemenata iz skupa od N elemenata ($n < N$) tako da svaki element ima jednaku vjerojatnost izbora.

- Izvlačenje
- Bacanje kocke
- Tablice slučajnih brojeva...
- Sistematski izbor

Stratificirani slučajni uzorak

Prije izbora uzorka, razvrstavanje elemenata osnovnog skupa u stratume (podskupove populacije koji se međusobno ne preklapaju).

Razvrstavanjem u stratume dobijemo skupove s manjim stupnjem varijabilnosti.

Jedinke unutar istog stratuma imaju istu vjerojatnost izbora u uzorak

Uzorak skupina

Populaciju podijelimo u skupine (clusters)

Slučajno biramo skupine

Jednostavni slučajni uzorak

Svaka jedinka ima istu vjerojatnost izbora u uzorak

Vrijednost obilježja kod slučajno izabrane jedinke je slučajna varijabla.

Distribucija vjerojatnosti te slučajne varijable je jednaka distribuciji obilježja u populaciji.

Posebno, ukoliko su μ i σ^2 srednja vrijednost i varijanca populacije, tada varijabla X, definirana vrijednošću obilježja slučajno izabrane jedinke, zadovoljava

$$E(X) = \mu \quad \text{i} \quad Var(X) = \sigma^2.$$

Jednostavni slučajni uzorak

Uzorak dobijemo s n **nezavisnih** slučajnih izbora jedinki iz populacije

Vrijednosti obilježja na n jedinki u uzorku:

$$X_1, X_2, X_3, \ldots, X_n$$

X_i je slučajna varijabla

Slučajne varijable $X_1, X_2, X_3, \ldots, X_n$ su jednako distribuirane.

Slučajne varijable $X_1, X_2, X_3, \ldots, X_n$ su nezavisne.

Procjenitelj
Parametar populacije

Parametar - funkcija vrijednosti obilježja svih jedinki iz populacije

Primjeri parametara populacije:

- aritmetička sredina,
- varijanca,
- standardna devijacija,
- medijan,
- kvantili

Procjenitelj

Parametar populacije je najčešće nepoznat.

Može ga se pokušati odrediti (procijeniti) pomoću uzorka.

Ako se iz populacije s N elemenata izabere uzorak od n elemenata ($n < N$), parametar populacije se **PROCJENJUJE**, a izraz s pomoću kojeg se procjenjuje naziva se **PROCJENITELJ**.

STATISTIKA je funkcija vrijednosti iz uzorka.

PROCJENITELJ je statistika kojom procjenjujemo parametar osnovnog skupa

Primjer.

PARAMETAR - srednja vrijednost populacije koja ima N jedinki:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

PROCJENITELJ srednje vrijednosti populacije je srednja vrijednost (aritmetička sredina) uzorka (s n jedinki):

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

Parametar se procjenjuje

- brojem
- intervalom.

Intervalni procjenitelj - raspon vrijednosti za koji se očekuje da se u njemu nalazi parametar.

Primjer. Procjenjuje se prosječna visina učenika neke škole.

Parametar - aritmetička sredina visine svih učenika u školi.

Procjenitelj - aritmetička sredina visine učenika iz uzorka, npr. 152 cm.

Intervalni procjenitelj - s 95%-tnom vjerojatnošću očekujemo da je prosječna visina učenika između 146 i 157 cm.

Zaključujemo o preciznosti procjene

Procjena srednje vrijednosti

Populacija: $x_1, x_2, x_3, \ldots, x_N$

Uzorak: $X_1, X_2, X_3, \ldots, X_n$

PARAMETAR - srednja vrijednost populacije koja ima N jedinki:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

PROCJENITELJ srednje vrijednosti populacije je srednja vrijednost (aritmetička sredina) uzorka (s n jedinki):

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

Primjer. Od 36 studenata koji su pisali kolokvij, slučajno je izabrano 5 testova. Broj bodova za izabrane testove je

$$72 \quad 62 \quad 78 \quad 60 \quad 48.$$

Procijenite prosječan broj bodova na kolokviju.

Rješenje. Aritmetička sredina uzorka je:

$$\bar{X} = \frac{1}{n}\sum X_i = \frac{1}{5} \cdot (72 + 62 + 78 + 60 + 48) = 63.6.$$

Napomena. Stvarna srednja vrijednost bodova na kolokviju je 52.6.

Ponovljeni slučajni izbor je rezultirao sljedećim vrijednostima

$$60 \quad 59 \quad 58 \quad 58 \quad 68.$$

Aritmetička sredina uzorka je:

$$\bar{X} = \frac{1}{n}\sum X_i = \frac{1}{5} \cdot (60 + 59 + 58 + 58 + 68) = 60.6.$$

Svojstva procjenitelja srednje vrijednosti.

Procjenitelj je slučajna varijabla (ovisi o slučajnom izboru uzorka)

$X_1, X_2, X_3, \ldots, X_n$ - slučajne varijeble

Aritmetička sredina slučajnih varijabli je slučajna varijabla.

Aritmetička sredina uzorka je **normalno distribuirana** ako je obilježje (populacija) normalno distribuirano.

Srednja vrijednost uzorka je **PRIBLIŽNO normalno distribuirana** ako je uzorak relativno velik ($n \geq 30$) (**Centralni granični teorem**).

Očekivanje aritmetičke sredine uzorka jednako je srednjoj vrijednosti populacije:

$$\mathrm{E}(\bar{X}) = \mu.$$

Varijanca aritmetičke sredine uzorka jednaka je:

$$\mathrm{Var}(\bar{X}) = \frac{1}{n}\sigma^2.$$

Ako je očekivanje procjenitelja jednako procjenjivanom parametru tada kažemo da je procjenitelj **nepristran**.

$\bar{X}$ je nepristrani procjenitelj za μ.

Uzorak: $X_1, X_2, X_3, \ldots, X_n$

Distribucija slučajnih varijabli X_i jednaka je distribuciji obilježja, posebno

$$\mathrm{E}(X_i) = \mu \quad \text{i} \quad \mathrm{Var}(X_i) = \sigma^2.$$

$$\mathrm{E}(\bar{X}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \mathrm{E}\left(\frac{1}{n}(X_1 + X_2 + \ldots + X_n)\right) =$$

$$= \frac{1}{n}\mathrm{E}(X_1 + X_2 + \ldots + X_n) = \frac{1}{n}(\mathrm{E}(X_1) + \mathrm{E}(X_2) + \ldots + \mathrm{E}(X_n)) =$$

$$= \frac{1}{n}(\mu + \mu + \ldots + \mu) = \frac{1}{n}(n \cdot \mu) = \mu$$

$$\begin{aligned}
\operatorname{Var}(\bar{X}) &= \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \operatorname{Var}\left(\frac{1}{n}(X_1 + X_2 + \ldots + X_n)\right) = \\
&= \frac{1}{n^2}\operatorname{Var}(X_1 + X_2 + \ldots + X_n) = \\
&= \frac{1}{n^2}(\operatorname{Var}(X_1) + \operatorname{Var}(X_2) + \ldots + \operatorname{Var}(X_n)) = \\
&= \frac{1}{n^2}\left(\sigma^2 + \sigma^2 + \ldots + \sigma^2\right) = \frac{1}{n^2}\left(n \cdot \sigma^2\right) = \frac{1}{n}\sigma^2
\end{aligned}$$

Napomena. Za računanje varijance aritmetičke sredine uzorka važna pretpostavka je da su $X_1, X_2, X_3, ..., X_n$ **nezavisne** slučajne varijable!

Jedinke u uzorak moramo birati nezavisno.

Preciznija definicija jednostavnog slučajnog uzorka:

Sve moguće skupine od n jedinki imaju jednaku vjerojatnost izbora u uzorak.

Standardna pogreška

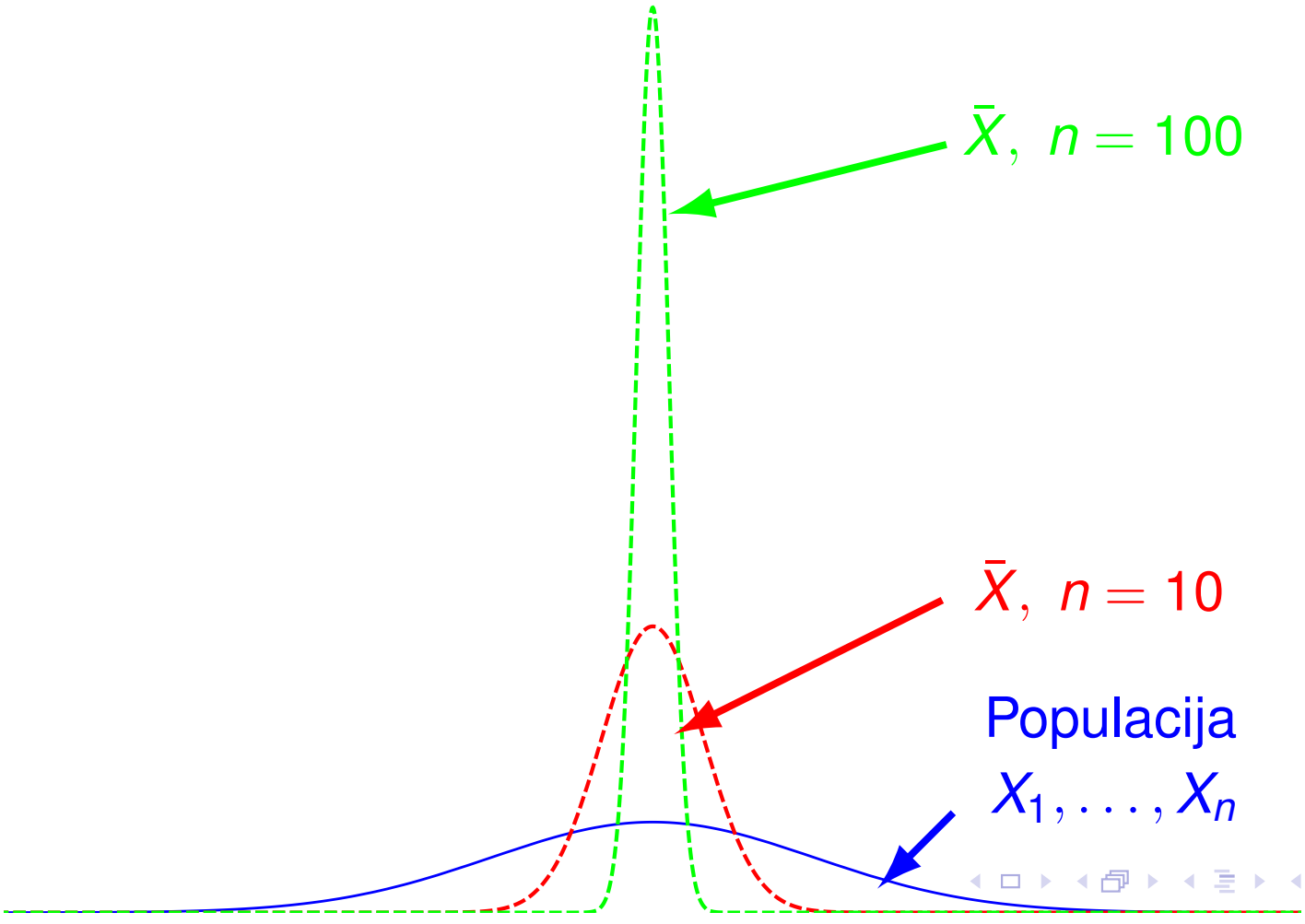
Standardna pogreška je standardna devijacija procjenitelja.

Standardna pogreška srednje vrijednosti je:

$$sem = \sqrt{\mathrm{Var}\,(\bar{X})} = \frac{\sigma}{\sqrt{n}}$$

- Povećanjem veličine uzorka (n) smanjuje se varijanca procjenitelja.
- Velika varijacija obilježja - velika varijanca procjenitelja.
- Varijanca procjenitelja ne ovisi o veličini populacije (N)
- Varijanca procjenitelja, tj. standardna pogreška je mjera točnosti procjenitelja.

Aritmetička sredina uzorka - funkcija gustoće vjerojatnosti



Procjena varijance

Varijanca populacije

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

se procjenjuje **varijancom uzorka**

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Napomena. Uočite da se dijeli s $n-1$ a ne s n.

Ovako definiran procjenitelj zadovoljava

$$\mathrm{E}(S^2) = \sigma^2.$$

S^2 je nepristrani procjenitelj varijance.

Standardnu devijaciju procjenjujemo s

$$S = \sqrt{S^2}.$$

Napomena. Ovo nije nepristrani procjenitelj.

Procjena standardne pogreške

$$\text{procjena } sem = \frac{S}{\sqrt{n}}$$

Procjena proporcije

$$p = \text{udio} = \frac{\text{broj jedinki s određenim oblikom obilježja}}{\text{ukupan broj jedinki}}$$

Izaberemo uzorak veličine n.

Neka je m broj elemenata u uzorku koji ima određeni oblik obilježja.

Tada je **procjenitelj proporcije populacije**:

$$\hat{p} = \frac{m}{n}$$

Uzorak: $X_1, X_2, \ldots, X_n$.

$$X_i = \begin{cases} 1 & \text{ukoliko se obilježje pojavilo} \\ 0 & \text{ukoliko se obilježje nije pojavilo} \end{cases}$$

X_i je Bernoullijeva slučajna varijabla.

$$\mathrm{E}(X_i) = p, \quad \mathrm{Var}(X_i) = p \cdot (1 - p)$$

$$m = \sum X_i$$

m je binomna slučajna varijabla

$$\mathrm{E}(m) = n \cdot p, \quad \mathrm{Var}(m) = n \cdot p \cdot (1 - p)$$

Sada je

$$\mathrm{E}(\hat{p}) = \mathrm{E}\left(\frac{m}{n}\right) = \frac{1}{n}\mathrm{E}(m) = \frac{1}{n} n \cdot p = p.$$

$\hat{p}$ je nepristrani procjenitelj za p.

Varijanca procjenitelja:

$$\mathrm{Var}(\hat{p}) = \mathrm{Var}\left(\frac{m}{n}\right) = \frac{1}{n^2}\mathrm{Var}(m) = \frac{1}{n^2} n \cdot p \cdot (1-p) = .\frac{p \cdot (1-p)}{n}$$

Standardna pogreška za proporciju: $\sigma_{\hat{p}} = \sqrt{\frac{p \cdot (1-p)}{n}}$

Procjena standardne pogreške za proporciju: $\sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}}$

Oznake

	Populacija	Uzorak
Veličina	N	n
Srednja vrijednost	μ	$\bar{X}$
Varijanca	σ^2	S^2
Standardna devijacija	σ	S
Proporcija	p	$\hat{p}$

Procjena medijana

Medijan populacije procjenjujemo medijanom uzorka.

Oznaka za procjenitelja medijana: $\tilde{X}$

$\tilde{X}$ je nepristrani procjenitelj.

Standardna pogreška (za veliki uzorak):

$$\sigma_{\tilde{X}} = \frac{1}{4\, n\, f^2(m)}$$

m - medijan populacije, f - funkcija gustoće za distribuciju populacije

Na žalost, distribucija populacije je najčešće nepoznata.

Efikasnost procjenitelja

Ukoliko je distribucija simetrična s obzirom na medijan, tada je medijan jednak srednjoj vrijednosti.

Tada aritmetička sredina uzorka ($\bar{X}$) i medijan uzorka ($\tilde{X}$) procjenjuju isti parametar ($\mu = m$).

Koji je procjenitelj bolje koristiti?

Onaj s manjom standardnom pogreškom.

$\bar{X}$, aritmetička sredina uzorka je procjenitelj s najmanjom standardnom pogreškom među svim procjeniteljima srednje vrijednosti.

Takav procjenitelj nazivamo **efikasni procjenitelj**.

Efikasnost procjenitelja je omjer standardne pogreške efikasnog procjenitelja i standardne pogreške procjenitelja.

Za populaciju s normalnom distribucijom, efikasnost medijana je za veliku veličinu uzorka približno

$$\frac{2}{\pi} = 0.637.$$

Intervalna procjena srednje vrijednosti

Koliko je pojedina procjena točna, pouzdana?

Standardna pogreška je mjera točnosti.

Prisjetimo se:

Ukoliko je razdioba obilježja u populaciji normalna, tada aritmetička sredina uzorka ima normalnu razdiobu uz

$$\mathrm{E}(\bar{X}) = \mu \quad \text{i} \quad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

Standardizirana varijabla

$$Y = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

ima jediničnu normalnu razdiobu - $N(0, 1)$.

Pretpostavke:

- Obilježje u populaciji je distribuirano prema normalnoj distribuciji.

  $\Longrightarrow \bar{X}$ je distribuiran prema normalnoj distribuciji

- ili: Uzorak je dovoljno velik ($n \geq 30$)

  Centralni granični teorem $\Longrightarrow \bar{X}$ je **približno** distribuiran prema normalnoj distribuciji

Kolika je vjerojatnost da Y poprimi vrijednost između -1.96 i 1.96?

Ako je F funkcija razdiobe za jediničnu normalnu razdiobu, tada je

$$P(-1.96 \leq Y \leq 1.96) = F(1.96) - F(-1.96) = 0.975 - 0.025 = \mathbf{0.95}$$

Znači, vjerojatnost da aritmetička sredina uzorka ($\bar{X}$) zadovoljava

$$\mu - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

iznosi **0.95**.

Budući da je srednja vrijednost populacije μ nepoznata, gornji uvjet je praktičnije napisati u obliku

$$\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

$\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}$ i $\bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}$ su slučajne varijable.

Vjerojatnost da interval

$$\left[\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right]$$

sadrži srednju vrijednost μ je **0.95**.

Interval

$$\left[\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right]$$

se naziva 95%-tni **interval pouzdanosti.**

Uočimo da je za određivanje intervala pouzdanosti nužno poznavati standardnu devijaciju populacije (σ).

Interval

$$\left[\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right]$$

je određen na osnovu uzorka:

- aritmetičke sredine uzorka $\bar{X}$,
- veličine uzorka n i
- standardne devijacije populacije σ

95%-tni interval pouzdanosti.

$$\left[\bar{X} - \mathbf{1.96} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + \mathbf{1.96} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

99%-tni interval pouzdanosti.

$$\left[\bar{X} - \mathbf{2.58} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + \mathbf{2.58} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

Napomena. Za jediničnu normalnu varijablu X je

$$P(-2.58 \leq X \leq 2.58) = 0.99.$$

Općenito, $100 \cdot (1 - \alpha)$%-tni interval pouzdanosti je dan s

$$\left[\bar{X} + \mathbf{z_{\alpha/2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + \mathbf{z_{1-\alpha/2}} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

gdje su $z_{\alpha/2}$ i $z_{1-\alpha/2}$ takvi da funkcija distribucije (F) za jediničnu normalnu distribuciju zadovoljava:

$$F\left(z_{\alpha/2} = \alpha/2\right) \quad \text{i} \quad F\left(z_{1-\alpha/2} = 1 - \alpha/2\right)$$

Zbog simetričnosti normalne distribucije, vrijedi:

$$z_{\alpha/2} = -z_{1-\alpha/2}$$

Primjer. Odredite 90%-tni interval pouzdanosti za srednju vrijednost.

Rješenje.

$$1-\alpha = 0.90 \quad \Longrightarrow \quad \alpha = 0.1 \quad \Longrightarrow \quad \alpha/2 = 0.05, \quad 1-\alpha/2 = 0.95.$$

$z_{\alpha/2} = z_{0.05}$ i $z_{1-\alpha/2} = z_{0.95}$ računamo pomoću R-a.

```
> qnorm(0.05,mean=0,sd=1)
-1.644854
> qnorm(0.95,mean=0,sd=1)
1.644854
```

$$z_{0.05} = -1.644854 \quad \text{i} \quad z_{0.95} = 1.644854.$$

90%-tni interval pouzdanosti:

$$\left[\bar{X} - \mathbf{1.64} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + \mathbf{1.64} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

Primjer. Na osnovu 5 slučajno izabranih vremena vožnje slaloma:

$$51.94 \quad 51.57 \quad 50.40 \quad 51.91 \quad 50.21$$

procijenite srednju vrijednost vremena vožnje te odredite 95%-tni i 99%-tni interval pouzdanosti ukoliko je standardna devijacija svih vremena jednaka 0.807079853.

Rješenje. Srednja vrijednost:

$$\bar{X} = \frac{1}{5}(51.94 + 51.57 + 50.40 + 51.91 + 50.21) = 51.206$$

95%-tni interval pouzdanosti:

$$\left[\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right] =$$

$$\left[51.206 - 1.96 \cdot \frac{0.807079853}{\sqrt{5}}, 51.206 + 1.96 \cdot \frac{0.807079853}{\sqrt{5}}\right] =$$

[50.49856332, 51.91343668]

99%-tni interval pouzdanosti:

$$\left[\bar{X} - 2.58 \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + 2.58 \cdot \frac{\sigma}{\sqrt{n}}\right] =$$

$$\left[51.206 - 2.58 \cdot \frac{0.807079853}{\sqrt{5}}, 51.206 + 2.58 \cdot \frac{0.807079853}{\sqrt{5}}\right] =$$

[50.27478233, 52.13721767]

Srednja vrijednost se s pouzdanošću od 95% nalazi u intervalu

$$[50.49856332, 51.91343668] .$$

Ne možemo reći da je vjerojatnost da se srednja vrijednost nalazi u ovom intervalu 95%.

Srednja vrijednost i granice intervala nisu slučajne varijable te ne možemo koristiti termin vjerojatnost.

Intervalna procjena srednje vrijednosti - standardna devijacija nepoznata

Ukoliko je standardna devijacija populacije (σ) nepoznata ne možemo izračunati interval pouzdanosti

$$\left[\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right]$$

Standardnu devijaciju populacije (σ) procjenjujemo pomoću uzorka s

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Međutim,

$$\left[\bar{X} - 1.96 \cdot \frac{S}{\sqrt{n}}, \bar{X} + 1.96 \cdot \frac{S}{\sqrt{n}}\right]$$

nije 95%-tni interval pouzdanosti.

Broj 1.96 smo izračunali iz normalne distribucije jer je standardizirana varijabla

$$Y = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

distribuirana prema jediničnoj normalnoj razdiobu - $N(0, 1)$.

$\bar{X}$ je distribuiran prema normalnoj razdiobi a μ i σ su konstantne.

S je slučajna varijabla te

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

nema normalnu razdiobu.

Varijabla

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

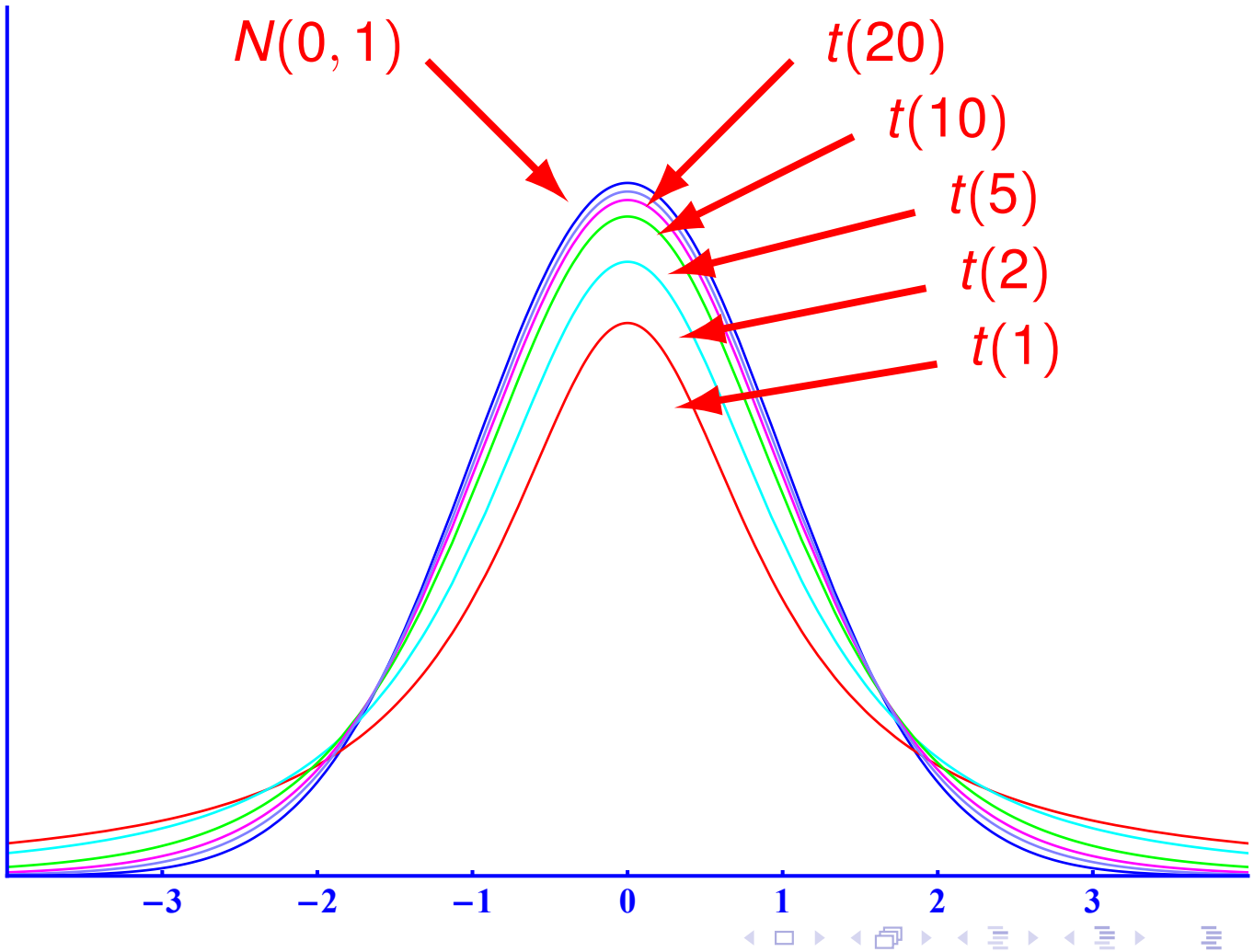
distribuirana je prema tzv. Studentovoj (t) distribuciji s $n - 1$ stupnjeva slobode (n je veličina uzorka).

Broj stupnjeva slobode je parametar Studentove distribucije.

Oznaka: $t(\nu)$

ν - broj stupnjeva slobode, $\nu = 1, 2, 3, \ldots$

Usporedba jedinične normalne razdiobe i različitih Studentovih razdioba



$100 \cdot (1-\alpha)$%-tni interval pouzdanosti:

$$\left[\bar{X} + t_{\alpha/2}(n-1) \cdot \frac{S}{\sqrt{n}}, \bar{X} + t_{1-\alpha/2}(n-1) \cdot \frac{S}{\sqrt{n}}\right]$$

$t_{\alpha/2}(n-1)$ i $t_{1-\alpha/2}(n-1)$ su brojevi takvi da slučajna varijabla T distribuirana prema Studentovoj razdiobi s $n-1$ stupnjeva slobode ($t(n-1)$) zadovoljava

$$P(T \leq t_{\alpha/2}(n-1)) = \alpha/2 \quad \text{i} \quad P(T \leq t_{1-\alpha/2}(n-1)) = 1-\alpha/2$$

odnosno

$$P(t_{\alpha/2}(n-1) \leq T \leq t_{1-\alpha/2}(n-1)) = 1-\alpha.$$

Napomena. Zbog simetričnosti Studentove razdiobe je

$$t_{\alpha/2}(n-1) - t_{1-\alpha/2}(n-1))$$

Primjer. Na osnovu 10 slučajno izabranih vremena vožnje slaloma:

52.22	51.37	51.50	51.94	52.78
53.13	51.95	51.91	51.20	51.55

procijenite 95%-tni i 99%-tni interval pouzdanosti.

Rješenje. Procjena srednje vrijednosti:

$$\bar{X} = \frac{1}{10}(52.22 + 51.37 + 51.50 + 51.94 + 52.78 + 53.13 + 51.95 + 51.9$$

Procjena varijance:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = 0.379672222$$

Procjena standardne devijacije:

$$S = \sqrt{S^2} = \sqrt{0.379672222} = 0.61617548$$

95%-tni interval pouzdanosti:

$$\left[\bar{X} + t_{0.025}(n-1) \cdot \frac{S}{\sqrt{n}}, \bar{X} + t_{1-0.025}(n-1) \cdot \frac{S}{\sqrt{n}}\right] =$$

$$\left[51.955 + t_{0.025}(9) \cdot \frac{0.616}{\sqrt{10}}, 51.955 + t_{1-0.025}(9) \cdot \frac{0.616}{\sqrt{10}}\right]$$

$t_{0.025}(9)$ i $t_{1-0.025}$ izračunamo pomoću R-a.

```
> qt(0.025,df=9)
-2.262157
> qt(0.975,df=9)
2.262157
```

$$t_{0.025}(9) = -2.262157 \quad \text{i} \quad t_{1-0.025} = 2.262157.$$

95%-tni interval pouzdanosti:

$$\left[51.955 - 2.262157 \cdot \frac{0.616}{\sqrt{10}},\ 51.955 + 2.262157 \cdot \frac{0.616}{\sqrt{10}}\right]$$

$$[51.51421462,\ 52.39578538]$$

Intervalna procjena proporcije

Procjena proporcije: $\hat{p}$

$n \cdot \hat{p}$ uspjeha u n pokusa.

Binomna distribucija.

$100(1-\alpha)\%$-interval pouzdanosti za broj uspjeha $(n \cdot \hat{p})$:

$$[b_{n,\alpha/2}, b_{n,1-\alpha/2}]$$

$$P(X \leq b_{n,p}) = p$$

Nemoguće. b-ova je konačno $(0, \ldots, n)$. Nema rješenja za svaki p.

Rješenje: uzima se prvi broj b za koji je $P(X \leq b) \geq p$

Npr, $x \sim B(30, 0.2)$

```
> qbinom(0.025,size=30,prob=0.4)
7
> pbinom(7,size=30,prob=0.4)
0.04352412
> pbinom(6,size=30,prob=0.4)
0.01718302
```

Normalna aproksimacija - Waldov interval

Praktičnije je koristiti pojednostavljeni pristup.

Procjena proporcije: $\hat{p}$

Standardna pogreška za proporciju: $\sigma_{\hat{p}} = \sqrt{\frac{p \cdot (1-p)}{n}}$

Procjena standardne pogreške za proporciju: $\hat{\sigma}_{\hat{p}} = \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}}$

Za **veliki** n je $\hat{p}$ distribuiran približno prema **normalnoj** distribuciji $N(p, \sigma_{\hat{p}})$.

Veliki n $\Rightarrow$ barem 10 uspjeha i barem 10 neuspjeha u uzorku.

Ili barem 5. Slabiji uvjet.

95%-tni interval pouzdanosti:

$$\left[\hat{p} - \mathbf{1.96} \cdot \hat{\sigma}_{\hat{p}}, \hat{p} + \mathbf{1.96} \cdot \hat{\sigma}_{\hat{p}}\right]$$

Primjer. U istraživanju o motivima posjete muzeju istraživači su bilježili i spol posjetitelja. Od 30 slučajno izabranih posjetitelja, bilo je 16 muŕaraca i 14 žena. Procijenite udio muškaraca među posjetiteljima muzeja. Procijenite i standardnu pogrešku te odredite 95%-tni i 99%-tni interval pouzdanosti.

Rješenje. Veličina uzorka: 30 posjetitelja ($n = 30$)

Broj muškaraca: 16 ($m = 16$)

Procjena proporcije:

$$\hat{p} = \frac{16}{30} = 0.533.$$

Procjena standardne pogreške:

$$\hat{\sigma}_{\hat{p}} = \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} = \sqrt{\frac{0.533 \cdot (1 - 0.533)}{30}} = 0.09.$$

$$\hat{p} = 0.533, \quad \hat{\sigma}_{\hat{p}} = 0.09$$

95%-tni interval pouzdanosti:

$$\left[\hat{p} - 1.96 \cdot \hat{\sigma}_{\hat{p}}, \hat{p} + 1.96 \cdot \hat{\sigma}_{\hat{p}}\right] =$$

$$[0.533 - 1.96 \cdot 0.09, 0.533 + 1.96 \cdot 0.09] =$$

$$[0.35, 0.71]$$

99%-tni interval pouzdanosti:

$$\left[\hat{p} - 2.58 \cdot \hat{\sigma}_{\hat{p}}, \hat{p} + 2.58 \cdot \hat{\sigma}_{\hat{p}}\right] =$$

$$[0.533 - 2.58 \cdot 0.09, 0.533 + 2.58 \cdot 0.09] =$$

$$[0.30, 0.77]$$

Ovo smo mogli riješiti i pomoću binomne distribucije.

```
> qbinom(0.025,size=30,prob=16/30)
11
> qbinom(0.975,size=30,prob=16/30)
21
```

95%-tni interval pouzdanosti za **broj** muškaraca:

$$[11, 21]$$

95%-tni interval pouzdanosti za **udio** muškaraca:

$$\left[\frac{11}{30}, \frac{21}{30}\right] = [0.3666667, 0.7]\,.$$

u odnosu na normalnu aproksimaciju

$$[0.35, 0.71]$$

Određivanje veličine uzorka za procjene

Procjena srednje vrijednosti

Za zadanu pouzdanost $1 - \alpha$ i zadani broj d treba odrediti veličinu uzorka tako da $(1 - \alpha)$%-tni interval pouzdanosti ne bude širi od $2d$.

Radi jednostavnosti ćemo uzeti da je $\alpha = 0.05$.

95%-tni interval pouzdanosti:

$$\left[\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right]$$

Širina intervala:

$$2 \cdot 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

Da bi širina intervala bila manja od $2d$ treba vrijediti

$$2 \cdot 1.96 \cdot \frac{\sigma}{\sqrt{n}} \leq 2d$$

odnosno

$$n \geq \left(1.96 \frac{\sigma}{d}\right)^2.$$

Za određivanje veličine uzorka moramo znati varijancu populacije.

To je očekivano jer veća varijacija populacije znači i veću varijabilnost procijenitelja.

Za proizvoljni α, veličina uzorka je dana s

$$n \geq \left(z_{1-\alpha/2} \frac{\sigma}{d}\right)^2.$$

Procjena proporcije

Kod procjene proporcije, $(1-\alpha)$%-tni interval pouzdanosti je dan s:

$$\left[\hat{p} - z_{1-\alpha/2} \cdot \sigma_{\hat{p}},\ \hat{p} + z_{1-\alpha/2} \cdot \sigma_{\hat{p}}\right]$$

$$\sigma_{\hat{p}} = \frac{\sigma}{\sqrt{n}}$$

Veličina uzorka bi trebala opet zadovoljavati

$$n \geq \left(z_{1-\alpha/2}\frac{\sigma}{d}\right)^2.$$

Međutim, kod proporcije je

$$\sigma = \sqrt{p \cdot (1-p)}$$

Pa je poznavanje standardne devijacije ekvivalentno poznavanju proporcije p, parametru kojeg želimo procijeniti.

Jedno rješenje je određivanje standardne devijacije pomoću grube procjene proporcije p.

Drugi način je da je kod binomne razdiobe standardna devijacija najveća za $p = 0.5$.

Tako je standardna devijacija uvijek omeđena s

$$\sigma = \sqrt{p \cdot (1-p)} \leq \sqrt{0.5 \cdot (1-0.5)} = 0.5.$$

Ako to uvrstimo u izraz za veličinu uzorka, tada možemo zahtjevati

$$n \geq \left(z_{1-\alpha/2} \frac{0.5}{d} \right)^2.$$

Ovako dobiveni n je sigurno veći od optimalnog (dobivena točnost će biti veća od tražene).

Primjer. Odredite veličinu uzorka potrebnu da širina 95%-tnog intervala pouzdanosti bude manja od 2 postotna poena u procjeni udjela glasača za kandidata X.Y. u 2. krugu predsjedničkih izbora 20xx. godine.

Rješenje.

$d = 0.02/2 = 0.01.$

Kod predsjedničkih izbora je $p \approx 0.5$ pa nećemo jako povećati uzorak ako stavimo $\sigma = 0.5$.

$$n \geq \left(z_{1-\alpha/2} \frac{\sigma}{d} \right)^2 = \left(1.96 \cdot \frac{0.5}{0.01} \right)^2 = 9604.$$