II Uvod u statistiku

Statistika se bavi

(i) prikupljanjem podataka ("dizajn pokusa")

(ii) prikazom podataka ("deskriptivna statistika")

(iii) zaključivanjem na osnovu podataka ("statističko zaključivanje")

(1) Populacija i uzorak

Populacija – skupina jedinki (ili stvari) o kojima želimo nešto znati

za danu populaciju, tipično nas zanima jedna ili više varijabli, npr.

- skup svih punoljetnih ljudi u RH, a varijabla je stranka koju osoba podržava

- sve baterije ~~za~~ automobile proizvedene u nekoj tvornici, a varijabla je trajanje baterije.
- svi studenti na PMF-MO, a varijable su visina i težina.

Općenito, varijable mogu biti

(i) Kategorijalne (ili kvalitativne)

np. spol (M/Ž), stranka, pušač (DA/NE)

(ne možemo ih prirodno predstaviti na nekoj numeričkoj skali)

iii) Numeričke (ili kvantitativne)

(vrijednost im označavamo brojem)

neprekidne

- trajanje baterije
- mjesečne plaće
- visina, težina

diskretne

- broj članova obitelji
- broj položenih kolegija u semestru

Uzorak - grupa jedinki iz populacije na kojima "mjerimo" promatranu varijablu (u kapro $n \in \mathbb{N}$)

Uzorak (mjerenja) - opažena mjerenja, oznaka će biti

"mali" x-evi $\rightarrow$ $\boxed{x_1, \dots, x_n \in \mathbb{R} \ (\text{ili } \mathbb{R}^d)}$

(x_i - mjerenje za i-tu osobu)

Intuitivno, <u>cilj</u> stat. zaključivanja je donijeti zaključke o karakteristikama promatrane varijable u <u>cijeloj</u> populaciji <u>samo</u> na temelju uzorka $x_1, \ldots, x_n$.

<u>Slučajnost</u>: Ako je

$X :=$ vrijednost varijable za <u>slučajno</u> odabranu osobu u populaciji,

te $F := F_X$, <u>pretpostavljamo</u> da su podaci $x_1, \ldots, x_n$ konkretne <u>realizacije</u> $\boxed{\text{njd}}$ <u>slučajnih varijabli</u> $X_1, \ldots, X_n$ t.d. $\boxed{F_{X_i} = F}$, $\forall i$.

Kažemo da je $X_1, \dots, X_n$ **slučajan uzorak** za X (ili iz F).

$\boxed{\text{Formalno, } \underline{\text{cilj}} \text{ s tat. zaključivanja je iz } X_1, \dots, X_n \text{ zaključiti nešto o } \underline{F}.}$

Napl Definicija slučajnog uzorka odgovara ideji da želimo $X_1, \dots, X_n$ koji su "**reprezentativni**" za cijelu populaciju.

↳ Loši primjeri

- Zanima nas % ljudi u RH koji navijaju za Dinamo, a uzorak uzmemo u Splitu.
- Zanima nas prosječan broj članova obitelji u Zagrebu, a uzorak uzmemo tako da nasumično ispitujemo ljude na glavnom trgu. ⟶ $\boxed{\text{Problem?}}$

2 Deskriptivna statistika

→ primjeri u R-u

Histogram

Podatke $x_1, \dots, x_n$ grupiramo u $\boxed{k}$ razreda

$$I_1 = [a_0, a_1), \; I_2 = [a_1, a_2), \dots, \; I_k = [a_{k-1}, a_k)$$

$f_1 :=$ # x_i-eva koji se upali u I_1 | f_2 | f_k

$\nu_1 :=$ % — " — $= f_1/n$ | ν_2 | ν_k

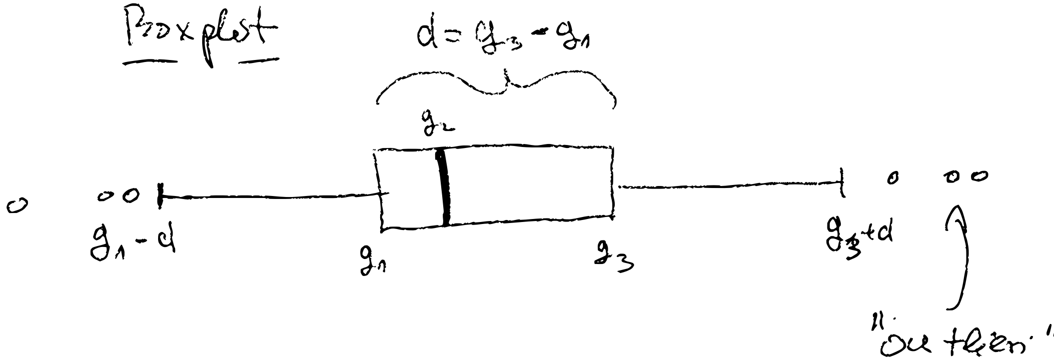
→ iznad razreda I_j crtamo pravokutnik visine

$$\boxed{\frac{\nu_j}{a_j - a_{j-1}}}$$

vidi Nap. prije Pr. 6.2

→ ukupna površina $\boxed{\text{jednaka } 1}$, te je histogram zapravo procjenitelj za $\boxed{\text{gustoću}}$ od X.

Boxplot



g_1, g_2, g_3 t.d.

- 50% podataka veće, a 50% manje od g_2

  (tzv. medijan uzorka)

- 75% —"— , a 25% —"— g_1
- 25% —"— , a 75% —"— g_3

npr. 50% podataka se nalazi u $[g_1, g_3]$

3. Testiranje statističkih hipoteza

[spada pod stat. zaključivanje]

Pretpostavimo se, za slučajnu varijablu X s nepoznatom f-jom distribucije F, iz uzorka $X_1, \dots, X_n$ želimo nešto zaključiti o F.

Tipično pretpostavljamo da F ovisi o nepoznatom parametru Θ pa se problem svodi na njegovu procjenu.

$$\boxed{\begin{array}{l} \text{npr.} \cdot X \sim \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix}, \Theta = p \\ \quad \cdot X \sim N(\mu, \sigma^2), \Theta = (\mu, \sigma^2) \end{array}}$$

Elementi statističkog testa

(i) nul-hipoteza (H_0) i alternativna hipoteza (H_1)

Želimo utvrditi imamo li na temelju podataka dovoljno dokaza da bismo odbacili H_0 i prihvatili H_1.

npr. Standardni lijek pomaže u 60% slučajeva, a želimo provjeriti je li novi lijek efikasniji.

→ model je $X \sim \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix}$ ← poboljšanje od novog lijeka

$H_0: p = 0.6$ (tj. nema poboljšanja)

$H_1: p > 0.6$ (ima poboljšanja)

↳ tipično, H_1 predstavlja ono što želimo provjeriti.

(ii) Druge vrste pogreške

odbacujemo H_0, a ona je istinita ("pogreška prve vrste")

ne odbacujemo H_0, a ona nije istinita ("pogreška druge vrste")

Za mali α (tipično $\alpha = 0.05, 0.01$), zahtijevamo da vrijedi

$$\boxed{\underbrace{P(\text{odbacujemo } H_0 \mid H_0)}_{\text{vjerojatnost pogreške 1. vrste}} = \alpha}$$

→ tzv. razina značajnosti testa

[tipično je greška 1. vrste "skuplje", npr. odobravamo novi lijek iako nije bolji od starog; osudujemo nevinu osobu.]

(iii) Testna statistika: $\boxed{T = f(X_1, \dots, X_n)}$

Kritično područje: skup $\boxed{C_\alpha \subseteq \mathbb{R}}$ t.d.

$$\mathbb{P}(\underbrace{T \in C_\alpha}_{\text{tada ćemo odbaciti } H_0} \mid H_0) = \alpha \quad \left[= \begin{matrix}\text{vjerojatnost} \\ \text{greške} \\ \text{prve vrste}\end{matrix} \right]$$

F-ju f i skup C_α biramo tako da znamo odrediti razdiobu od T pod uvjetom da vrijedi H_0, te tako da je

$$\mathbb{P}(T \in C_\alpha \mid H_1)$$

što veća.

U našem primjeru, tipično se uzima

$$T = \frac{\sqrt{n}\,(\overline{X}_n - 0.6)}{\sqrt{0.6 \cdot 0.4}}$$

gdje je $\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ $\left(X_i = \mathbb{1}_{\{i\text{-ta osoba doživjela poboljšanje}\}} \right)$

Ako vrijedi H_0, tj. $X_1, \dots, X_n$ su njd t.d.

$$X_i \sim \begin{pmatrix} 0 & 1 \\ 0.4 & \boxed{0.6} \end{pmatrix},$$

po CGT-u znamo da je za velike n,

$$(*) \quad \boxed{T = \frac{S_n - n \cdot 0.6}{\sqrt{n \cdot 0.6 \cdot 0.4}} \overset{d}{\approx} N(0,1)} \quad \left(S_n = \sum_{i=1}^{n} X_i\right)$$

Uočimo,

$\overline{X}_n$ = % ljudi s poboljšanjem nakon korištenja novog lijeka.

Ako je lijek zaista bolji, tj. ako $p > 0.6$, očekujemo da $\overline{X}_n > 0.6$, tj. $T > 0$.

$$\boxed{C_\alpha = [z_\alpha, +\infty)}, \text{ gdje je}$$

$z_\alpha \in \mathbb{R}$ t.d. $\boxed{\mathbb{P}(Z \geq z_\alpha) = \alpha \text{ za } Z \sim N(0,1).}$

npr. $z_{0.05} = 1.65$

Uočimo, zbog $(*)$ zaista imamo

$$\mathbb{P}(T \in C_\alpha \mid H_0) \approx \mathbb{P}(Z \geq z_\alpha) = \alpha.$$

Nakon što smo odredili sve elemente, test dakle provodimo na sljedeći način:

- Iz podataka $x_1, \dots, x_n$ odredimo vrijednost testne statistike $\boxed{t = f(x_1, \dots, x_n)}$

ako $t \in C_\alpha$

odbacujemo H_0
u korist H_1
(na razini značajnosti α)

ako $t \notin C_\alpha$

ne ——— 〃 ———
——— 〃 ———
——— 〃 ———

U našem primjeru, ako je $\underline{\underline{n = 1000}}$ i $\underline{\underline{\alpha = 0.05}}$

($z_{0.05} = 1.65$),

- u slučaju $\bar{x}_n = 0.65$, imali bismo

$$t = \sqrt{1000} \cdot \frac{(0.65 - 0.6)}{\sqrt{0.6 \cdot 0.4}} \approx \boxed{3.23 > 1.65}$$

$\Rightarrow$ <u>odbacili</u> bi H_0 ("nema promjene")
na raz. znač. 0.05.

- u slučaju $\boxed{\bar{x}_n = 0.61}$,

$$t \approx 0.64 \le 1.65$$

$\Rightarrow$ na <u>ovoj</u> razin. znač. $\underline{\underline{\text{ne}}}$ bi
odbacili H_0

(tj., iako $\bar{x}_n > 0.6$, razlika nije velika te je možda biti samo plod slučajnosti u odabiru uzorka)

Općenito, ako je $X \sim \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix}$ za $p \in \langle 0,1 \rangle$, $X_1, \ldots, X_n$ uzorak za X, a želimo testirati

$$H_0 : p = p_0 \longleftarrow \text{neka fiksna vrijednost}$$

u odnosu na

(a) $H_1 : p > p_0$, ili

(b) $H_1 : p < p_0$, ili

(c) $H_1 : p \neq p_0$,

koristimo

$$T := \sqrt{n}\,\frac{\bar{X}_n - p_0}{\sqrt{p_0(1-p_0)}} \overset{d}{\underset{\text{uz } H_0}{\approx}} N(0,1),$$

te

(a) $C_\alpha = [z_\alpha, +\infty\rangle$, (b) $C_\alpha = \langle -\infty, -z_\alpha]$,

(c) $C_\alpha = \langle -\infty, -z_{\alpha/2}] \cup [z_{\alpha/2}, +\infty\rangle$

$z_{\alpha/2} > z_\alpha$

$\hookrightarrow$ u sva tri slučaja imamo $\boxed{\mathbb{P}(T \in C_\alpha \mid H_0) \approx \alpha}$.

Nap.] H_1 određujemo prije nego što smo vidjeli

podatke → vidi Zad. 8.os. i Nap. 9.10.

na vježbama

Nap.] Ako zaista vrijedi H_0 te puno puta

ponovimo test, u prosjeku ćemo u približno

$\boxed{100 \cdot \alpha \text{ \% slučajeva odbaciti } H_0}$

(iako je točna)

[intuicija o razini značajnosti α]

[ostatak na vježbama!]