Увод у статистику

Јован Самарџић, 13/2019

Професорка: Бојана Милошевић

- дефиниције
- ознаке
- теореме
- докази
- примери

Година курса: 2020/21

Молим да ми све грешке пријавите преко мејла или друштвених мрежа.

0. Преглед познатих расподела

деф. Случајна величина је мерљиво пресликавање из Ω у $\mathbb{R}$.
↳ простор исхода

Могу бити: 1° дискретне;
2° апсолутно непрекидне.

0.1. Дискретне случајне величине

* Бернулијева случ. величина, $X \sim Ber(p)$: $\begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$

* Биномна случ. величина, $X \sim B(n,p)$: $\begin{pmatrix} 0 & 1 & \dots & n \\ P_n(0) & P_n(1) & \dots & P_n(n) \end{pmatrix}$, где $P_n(k) = P\{X=k\} = \binom{n}{k} p^k (1-p)^{n-k}$

* Геометријска случ. величина, $X \sim G(p)$: $\begin{pmatrix} 1 & 2 & \dots & k & \dots \\ p & p(1-p) & \dots & p(1-p)^{k-1} & \dots \end{pmatrix}$

* Негативна биномна случ. величина, $X \sim NB(r,p)$: $\begin{pmatrix} r & r+1 & \dots & k \\ p^r & r \cdot p^r(1-p) & \dots & \binom{k-1}{r-1} p^r (1-p)^{k-r} \end{pmatrix}$

* Пуасонова случајна величина, $X \sim \mathcal{P}(\lambda)$: одређена законом расподеле:

$$X: \begin{pmatrix} 0 & 1 & \dots & k & \dots \\ \Pi_\lambda(0) & \Pi_\lambda(1) & \dots & \Pi_\lambda(k) & \dots \end{pmatrix}, \quad \text{где} \quad \Pi_\lambda(k) := P\{X=k\} = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k \in \mathbb{N}_0,\ \lambda > 0.$$

расподела	EX	DX	параметри
Бернулијева	p	$p(1-p)$	$p \in (0,1)$
Биномна	np	$np(1-p)$	$n \in \mathbb{N}$, $p \in (0,1)$
Геометријска	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$p \in (0,1)$
Негативна биномна	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$	$r \in \mathbb{N}$, $p \in (0,1)$
Пуасонова	λ	λ	$\lambda > 0$

0.2. Апсолутно непрекидне случајне величине

Особине апс. непрекидних: 1) $F(x)=\int_{-\infty}^{x} f(u)\,du$, где $f(u)\geq 0$ и $\int_{-\infty}^{+\infty} f(u)\,du=1$ 2) $f(x)=F'(x)$

3) $EX=\int_{-\infty}^{+\infty} x f(x)\,dx$ 4) $Eg(X)=\int_{-\infty}^{+\infty} g(x) f(x)\,dx$

* Униформна расподела, $X\sim \mathcal{U}[a,b]$: $f(x)=\begin{cases} \frac{1}{b-a}, & x\in[a,b] \\ 0, & \text{иначе} \end{cases}$; $F(x)=\begin{cases} 0, & x<a \\ \frac{x-a}{b-a}, & x\in[a,b] \\ 1, & x>b \end{cases}$

* Експоненцијална расподела, $X\sim\mathcal{E}(\lambda)$: $f(x)=\begin{cases} \lambda\cdot e^{-\lambda x}, & x\geq 0 \\ 0, & x<0 \end{cases}$; $F(x)=\begin{cases} 1-e^{-\lambda x}, & x\geq 0 \\ 0, & x<0 \end{cases}$

* Гама расподела, $X\sim\gamma(\alpha,\beta)$: $f(x)=\begin{cases} \frac{x^{\alpha-1}\beta^{\alpha}}{\Gamma(\alpha)}\cdot e^{-\beta x}, & x\geq 0 \\ 0, & x<0 \end{cases}$, $\alpha,\beta>0$.

деф. Гама функција је $\Gamma(x)=\int_{0}^{+\infty} t^{x-1}e^{-x}\,dt$.

Својства: 1) $\Gamma(x)=(x-1)\cdot\Gamma(x-1)$

2) Ако $x\in\mathbb{Z}$, тада $\Gamma(x)=(x-1)!$

* Нормална расподела, $X\sim\mathcal{N}(m,\sigma^2)$: $f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\cdot e^{-\frac{(x-m)^2}{2\sigma^2}}$, $m\in\mathbb{R}$, $\sigma^2>0$.

Стандардизација: $Y=\frac{X-m}{\sigma}$ има $\mathcal{N}(0,1)$ расподелу.

* χ^2_n расподела са n степени слободе: $f(x)=\frac{x^{\frac{n}{2}-1}}{2^{\frac{n}{2}}\cdot\Gamma(\frac{n}{2})}\cdot e^{-\frac{x}{2}}$, $x>0$.

Алтернативна дефиниција: $X=X_1^2+\ldots+X_n^2$, где су $X_1,\ldots,X_n$ независне са $\mathcal{N}(0,1)$ расподелом.

* Студентова расподела са n степени слободе, $X\sim t_n$: $f(x)=\frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\cdot\Gamma(\frac{n}{2})}\cdot\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$, $n\in\mathbb{R}^+$, $x>0$.

Алтернативна дефиниција: $X=\frac{Z}{\sqrt{\frac{Y}{n}}}$, где $Z\sim\mathcal{N}(0,1)$ и $Y\sim\chi^2_n$.

* Фишерова расподела, F_{n_1,n_2}: $X=\frac{Y_1/n_1}{Y_2/n_2}$, где су $Y_1\sim\chi^2_{n_1}$, $Y_2\sim\chi^2_{n_2}$ независне.

расподела	EX	DX	параметри
униформна	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$a<b\in R$
експоненцијална	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\lambda>0$
гама	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$	$\alpha,\beta>0$
нормална	m	σ^2	$m\in R$, $\sigma^2: \sigma\in(0,\infty)$
χ^2_n	n	$2n$	
Студентова	0	$\frac{n}{n-2}$ $(n>2)$	$n\in R^+$
Фишерова	$\frac{n_2}{n_2-2}$ $(n_2>2)$	$\frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}$ $(n_2>4)$	

Напомене: 1) Збир n независних сл. вел. са $\mathcal{E}(\beta)$ има $\gamma(n,\beta)$ расподелу. Важи и обрнуто.

2) $\gamma(1,\beta) \Leftrightarrow \mathcal{E}(\beta)$;

3) $\gamma(\frac{n}{2},\frac{1}{2}) \Leftrightarrow \chi^2_n$.

Извођења за EX и DX:

1) $X\sim \mathcal{U}[a,b]$: $EX=\frac{a+b}{2}$, $DX=\frac{(b-a)^2}{12}$

$$* EX=\int_{-\infty}^{+\infty} x\cdot f(x)\,dx = \int_a^b x\cdot\frac{1}{b-a}\,dx = \frac{1}{b-a}\int_a^b x\,dx = \frac{1}{b-a}\left(\frac{x^2}{2}\right)\Big|_a^b = \frac{(b-a)(b+a)}{(b-a)\cdot 2} = \frac{a+b}{2}.$$

$$* DX=\int_{-\infty}^{+\infty}(x-EX)^2 f(x)\,dx = \int_a^b\left(x-\frac{a+b}{2}\right)^2\cdot\frac{1}{b-a}\,dx = \frac{1}{b-a}\int_a^b\left(x^2-(a+b)x+\frac{(a+b)^2}{4}\right)dx$$

$$=\frac{1}{b-a}\left(\frac{x^3}{3}-(a+b)\frac{x^2}{2}+\frac{(a+b)^2}{4}x\right)\Big|_a^b = \frac{1}{b-a}\left(\frac{b^3-a^3}{3}-(a+b)\frac{b^2-a^2}{2}+\frac{(a+b)^2}{4}(b-a)\right)$$

$$=\frac{b^2+ba+a^2}{3}-\frac{a^2+2ab+b^2}{2}+\frac{a^2+2ab+b^2}{4}=\frac{4b^2+4ab+4a^2-3a^2-6ab-3b^2}{12}$$

$$=\frac{b^2-2ba+a^2}{12}=\frac{(b-a)^2}{12}$$

2) $X \sim \mathcal{E}(\lambda)$: $\quad EX = \frac{1}{\lambda}$, $\quad DX = \frac{1}{\lambda^2}$

$$* \; EX = \int_{-\infty}^{+\infty} x \cdot f(x)\,dx = \int_0^{+\infty} x \cdot \lambda \cdot e^{-\lambda x}\,dx = \begin{bmatrix} u = x & du = dx \\ dv = \lambda \cdot e^{-\lambda x}\,dx & v = -e^{-\lambda x} \end{bmatrix}$$

$$= -xe^{-\lambda x}\Big|_0^{+\infty} - \int_0^{+\infty} \frac{\lambda}{\lambda} \cdot (-e^{-\lambda x})\,dx = 0 + \frac{1}{\lambda} \underbrace{\int_0^{+\infty} \lambda \cdot e^{-\lambda x}}_{f(x)}\,dx = \frac{1}{\lambda} \cdot 1 = \frac{1}{\lambda}$$

$$* \; EX^2 = \int_{-\infty}^{+\infty} x^2 \cdot f(x)\,dx = \int_0^{+\infty} x^2 \cdot \lambda \cdot e^{-\lambda x}\,dx = \begin{bmatrix} u = x^2 & du = 2x\,dx \\ dv = \lambda \cdot e^{-\lambda x}\,dx & v = -e^{-\lambda x} \end{bmatrix}$$

$$= -x^2 e^{-\lambda x}\Big|_0^{+\infty} - \int_0^{+\infty} \frac{\lambda}{\lambda} 2x(-e^{-\lambda x})\,dx = 0 + \frac{2}{\lambda} \underbrace{\int_0^{+\infty} x e^{-\lambda x}\,dx}_{EX} = 0 + \frac{2}{\lambda^2}$$

$$\Rightarrow DX = EX^2 - (EX)^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$$

3) $X \sim \gamma(\alpha, \beta)$: $\quad EX = \frac{\alpha}{\beta}$, $\quad DX = \frac{\alpha}{\beta^2}$

$$* \; EX = \int_{-\infty}^{+\infty} x \cdot f(x)\,dx = \int_0^{+\infty} x \cdot \frac{x^{\alpha-1} \beta^{\alpha} e^{-\beta x}}{\Gamma(\alpha)}\,dx = \int_0^{+\infty} \frac{x^{\alpha} \beta^{\alpha} e^{-\beta x}}{\Gamma(\alpha)} \cdot \frac{\beta}{\beta} \cdot \frac{\Gamma(\alpha+1)}{\Gamma(\alpha+1)}\,dx$$

$$= \underbrace{\int_0^{+\infty} \frac{x^{\alpha} \beta^{\alpha+1} e^{-\beta x}}{\Gamma(\alpha+1)}\,dx}_{\text{густина за } \gamma(\alpha+1,\beta)} \cdot \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} \cdot \frac{1}{\beta} = 1 \cdot \frac{1}{\beta} \cdot \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} = \frac{1}{\beta} \cdot \big((\alpha+1)-1\big) = \frac{\alpha}{\beta}.$$

$$* \; EX^2 = \int_{-\infty}^{+\infty} x^2 \cdot f(x)\,dx = \int_0^{+\infty} x^2 \cdot \frac{x^{\alpha-1} \beta^{\alpha} e^{-\beta x}}{\Gamma(\alpha)}\,dx = \int_0^{+\infty} \frac{x^{\alpha+1} \beta^{\alpha} e^{-\beta x}}{\Gamma(\alpha)} \cdot \frac{\beta^2}{\beta^2} \cdot \frac{\Gamma(\alpha+2)}{\Gamma(\alpha+2)}\,dx$$

$$= \underbrace{\int_0^{+\infty} \frac{x^{\alpha+1} \beta^{\alpha+2} e^{-\beta x}}{\Gamma(\alpha+2)}\,dx}_{\text{густина за } \gamma(\alpha+2,\beta)} \cdot \frac{\Gamma(\alpha+2)}{\Gamma(\alpha)} \cdot \frac{1}{\beta^2} = 1 \cdot \frac{1}{\beta^2} \cdot \frac{\frac{\Gamma(\alpha+2)}{\Gamma(\alpha+1)}}{\frac{\Gamma(\alpha)}{\Gamma(\alpha+1)}} = \frac{1}{\beta^2} \cdot \frac{\alpha+1}{\frac{1}{\alpha}} = \frac{\alpha^2+\alpha}{\beta^2}$$

$$\Rightarrow DX = EX^2 - (EX)^2 = \frac{\alpha^2+\alpha}{\beta^2} - \left(\frac{\alpha}{\beta}\right)^2 = \frac{\alpha}{\beta^2}$$

* У наставку су решени задаци са часа.

Домаћи 1: Ако $X \sim \mathcal{E}(\lambda)$, коју расподелу има $\lfloor X \rfloor$?

Решење: $X \sim \mathcal{E}(\lambda) \Rightarrow F_X(x) = \begin{cases} 0, & x \leq 0 \\ 1 - e^{-\lambda x}, & x \geq 0 \end{cases}$

$Y \sim \lfloor X \rfloor \Rightarrow Y$ је дискретна

$$P\{Y = k\} = P\{\lfloor X \rfloor = k\} = P\{k \leq X < k+1\} \overset{X \text{ је непрекидна}}{=} P\{X < k+1\} - P\{X < k\} =$$

$$= F_X(k+1) - F_X(k) = \left(1 - e^{-\lambda(k+1)}\right) - \left(1 - e^{-\lambda k}\right)$$

$$= e^{-\lambda k}\left(1 - e^{-\lambda}\right) = \left(e^{-\lambda}\right)^k \left(1 - e^{-\lambda}\right)$$

Дакле, у питању је нека као геометријска расподела $\mathcal{G}(1 - e^{-\lambda})$.

пошто, по деф. ако $A \sim \mathcal{G}(p)$,
у вероватноћи $P\{A = k\}$, број k је укупан број извођења
док је овде k број „неуспеха"

Задатак 2: Ако X има ф-ју расподеле F на носачу $[a, b]$, коју расподелу има $Y = F(X)$?

Решење: $Y \sim \mathcal{U}[a, b]$: вежбе, 7. задатак.

Задатак 3: Ако $X \sim \mathcal{U}[0,1]$ и F нека ф-ја расподеле апс. непр. сл. вел.
Одредити ф-ју расподеле сл. вел. $Y = F^{-1}(X)$.

Решење: $F_Y(y) = P\{Y \leq y\} = P\{F^{-1}(X) \leq y\} = P\{X \leq F(y)\} = F(F(y))$

$$\Rightarrow F_Y(y) = \begin{cases} 0, & x < 0 \\ F(y), & x \in [0,1] \\ 1, & x > 1 \end{cases}$$

Задатак 4: Ако су $X_1, X_2, X_3 \sim \mathcal{U}[0,\theta]$ независне и $\theta > 0$, одредити F и f за:

а) $X_{(3)} = \max\{X_1, X_2, X_3\}$.

б) $X_{(2)}$ - други по величини у низу X_1, X_2, X_3.

Решење: а) $Y = \max\{X_1, X_2, X_3\}$

$$F_Y(y) = P\{Y \leq y\} = P\{\max\{X_1,X_2,X_3\} \leq y\} = P\{X_1 \leq y,\ X_2 \leq y,\ X_3 \leq y\} =$$

$$\overset{\text{нез.}}{=} P\{X_1 \leq y\} \cdot P\{X_2 \leq y\} \cdot P\{X_3 \leq y\} = F_{X_1}(y) \cdot F_{X_2}(y) \cdot F_{X_3}(y) = F^3(y) \quad \text{(иста расподела)}$$

$$f_Y(y) = F_Y'(y) = 3F^2(y) \cdot f(y)$$

$$\Rightarrow F_Y(y) = \begin{cases} 0 & , \ y < 0 \\ (y/\theta)^3 & , \ y \in [0,\theta] \\ 1 & , \ y > \theta \end{cases} \quad \text{и} \quad f_Y(y) = \begin{cases} 0 & , \ y < 0 \\ y^2/\theta^3 & , \ y \in [0,\theta] \\ 0 & , \ y > \theta \end{cases}$$

б) $F_{X_{(2)}}(y) = P\{X_{(2)} \leq y\} = P\{\text{сва 3 мања} \ \vee \ \text{2 мања}\} = P\{\text{3 мања}\} + P\{\text{2 мања}\}$

$$= F^3(y) + \binom{3}{2} F^2(y)\,(1 - F(y)).$$

деф. Нека су $X_1, X_2, \ldots, X_n$ независне и једнако расподељене сл. вел.
Пермутација таква да $X_{(1)} \leq \ldots \leq X_{(n)}$ је варијациони низ. Тада је $X_{(k)}$ k-та статистика поретка.

Домаћи 2: Ако X_1 има ф-ју расподеле F, одредити расподелу $X_{(k)}$.

Решење: Докажимо индукцијом по k: $f_{X_{(k)}}(x) = \frac{n!}{(n-k)!\,(k-1)!} \cdot F^{k-1}(x)\,(1-F(x))^{n-k} \cdot f(x)$

(БИ) $F_{X_{(1)}}(x) = P\{X_{(1)} \le x\} = 1 - P\{X_{(1)} > x\} =$

$= 1 - P\{X_{(1)} > x, \dots, X_{(n)} > x\} \overset{\text{нез.}}{=} 1 - P\{X_{(1)} > x\} \dots P\{X_{(n)} > x\}$

$= 1 - (1-F(x))^n$ (јер су једнако расподељене)

$\Rightarrow f_{X_{(1)}}(x) = F'_{X_{(1)}}(x) = -n \cdot (1-F(x))^{n-1}(-f(x)) = n(1-F(x))^{n-1} \cdot f(x)$

(ИК) $F_{X_{(k)}}(x) = P\{X_{(k)} \le x\} = 1 - P\{X_{(k)} > x\} =$

$= 1 - P\{$сви већи $\vee$ 1 мањи $\vee \dots \vee$ $k-1$ мањих$\}$

$= 1 - P\{$сви већи$\} - P\{$1 мањи$\} - \dots - P\{k-1$ мањих$\}$

$= 1 - (1-F(x))^n - \binom{n}{1}(1-F(x))^{n-1} F(x) - \dots - \binom{n}{k-1}(1-F(x))^{n-k+1} F^{k-1}(x)$

$\Rightarrow f_{X_{(k)}}(x) = F'_{X_{(k)}}(x) = \left(1 - (1-F(x))^n - \dots - \binom{n}{k-2}(1-F(x))^{n-k+2} F^{k-2}(x)\right)' - \left(\binom{n}{k-1}(1-F(x))^{n-k+1} F^{k-1}(x)\right)'$

$\overset{(\text{ИХ})}{=} \frac{n!}{(n-k+1)!\,(k-2)!} \cdot F^{k-2}(x)\,(1-F(x))^{n-k+1} \cdot f(x) - \frac{n!}{(n-k+1)!\,(k-1)!} \cdot \Big(-(n-k+1)(1-F(x))^{n-k} f(x) F^{k-1}(x)$
$+ (1-F(x))^{n-k+1}(k-1) F^{k-2}(x) f(x)\Big)$

$= \frac{n!}{(n-k+1)!\,(k-2)!} \cdot F^{k-2}(x)\,(1-F(x))^{n-k+1} \cdot f(x) + \frac{n!}{(n-k+1)!\,(k-1)!} \Big((n-k+1)(1-F(x))^{n-k} f(x) F^{k-1}(x)$
$- (1-F(x))^{n-k+1}(k-1) F^{k-2}(x) f(x)\Big)$

$= \frac{n!}{(n-k+1)!\,(k-2)!} F^{k-2}(x)\,(1-F(x))^{n-k} \cdot f(x) \left(1-F(x) + \frac{1}{k-1}\Big((n-k+1)F(x) - (1+F(x))(k-1)\Big)\right)$

$= \frac{n!}{(n-k+1)!\,(k-2)!} F^{k-2}(x)\,(1-F(x))^{n-k} \cdot f(x) \left(\cancel{1-F(x)} + \frac{n-k+1}{k-1} F(x) - \cancel{1+F(x)}\right)$

$= \frac{n!}{(n-k)!\,(k-1)!} \cdot F^{k-1}(x)\,(1-F(x))^{n-k} \cdot f(x)$

1. Увод

Статистика је наука о подацима - бави се њиховим прикупљањем, анализом, презентовањем...
Основни задатак статистичара је да предложи математички модел којим би се подаци адекватно описали.

деф. Популација Ω је скуп јединки чије карактеристике изучавамо.

деф. Обележје $X: \Omega \to \mathbb{R}$ је карактеристика коју проучавамо.

деф. Узорак је подскуп популације на основу ког доносимо закључке о обележју на читавој популацији.

- специјално, уколико се подскуп бира насумично, то је случајан узорак.
- мора бити репрезентативан

Циљ нам је да на основу узорка доносимо закључке о неком конкретном параметру популације.
Тај параметар оцењујемо неком функцијом од чланова узорка.
Та функција се зове статистика. (увешћемо формално у сл. питању)

деф. Даља анализа узорка зависи од типа обележја, па стога уводимо:

* категоричко обележје - изражава се описно, постоје категорије.
  - номинално - не постоји никакво уређење. (крвна група, пол, ...)
  - ординално - постоји уређење. (платни разред, интензитет бола)

* нумеричко обележје - изражава се бројем.
  - дискретно - скуп вредности дискретан. (оцена, број деце, ...)
  - непрекидно - скуп вредности није дискретан. (висина, време чекања, ...)

Битно је знати типове података, како бисмо их боље организовали у анализи.

2. Основни кораци у статистичкој анализи

1) Осмишљавање експеримента;

Пре него што кренемо са прикупљањем података, морамо прецизно одредити циљ истраживања.

2) Узорковање и прикупљање података;

деф. Случајни узорак је узорак у коме сваки од чланова популације има могућност да се нађе у узорку.
Специјално, ако су сви узорци истог обима једнако вероватни, онда је то прост случајан узорак (п.с.у.).

Постоје две основне врсте узорковања: са враћањем $\left(p=\frac{1}{N^n}\right)$ и без враћања $\left(p=\frac{1}{\binom{N}{n}}\right)$.
До даљњег, претпостављамо да је популација велика и да је узорак са враћањем.

Нека је $X:\Omega\to R$ обележје. Тада је прост случајни узорак баш случ. вектор $(X_1,\dots,X_n)$, где су $X_1,\dots,X_n$ независне и једнако расподељене случ. величине.
Са $x_1,\dots,x_n$ означавамо реализован узорак. (регистроване вредности случајних величина)
(X_i - случ. вел. која i-том члану узорка независно додељује вредност x_i при датој расподели)

Свака функција $T:(X_1,\dots,X_n)\to R$ која не зависи од непознатих параметара зове се статистика.

3) Прелиминарна анализа;

Овај корак је важан за проналажење одговарајућег математичког модела.
За то, податке често прикажемо графички.

1° Уколико је у питању категоричко обележје, најчешће се користе: (пример 1)

* табеларни приказ: фреквенција (учесталост) по категоријама; table()
* bar plot: фреквенције се приказују у виду трака на графику; barplot()
* pie chart: фреквенције се приказују у виду исечака на кружном дијаграму; pie()
* итд.

2° Уколико је у питању нумеричко обележје, користимо: (пример 2)

* хистограм: узорак разбијемо на интервале и прикажемо фреквенцију (n_i) за сваки инт.; hist()

  → хистограм апсолутних фреквенција: на y оси фреквенције: n_i
  → хистограм релативних фреквенција: на y оси релативне фреквенције: $\frac{n_i}{n}$
  → хистограм густине: на y оси релативне фреквенције подељене величином интервала: $\frac{n_i}{n\cdot d_i}$

  $$P=P_1+\dots+P_k=\sum^k d_i\cdot\frac{n_i}{n\cdot d_i}=\frac{\sum n_i}{n}=\frac{n}{n}=1.$$

* по стандарду, увек се одреде дескриптивне статистике. (касније)

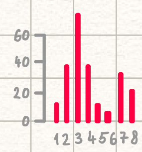
Пример 1: Посматрамо саобраћајне несреће у Калифорнији 2012-2016.
Занима нас шта утиче на несрећу. Посматрамо обележја:
* тип несреће, typeC
* тип пута, typeR
* раскрсница, crossR

	typeC	typeR	crossR
1	4	1	1
2	3	2	1
3	8	1	2
⋮	⋮	⋮	⋮

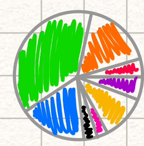
База података

typeC	1	2	3	4	5	6	7	8
	7	34	79	34	5	1	29	11

table()



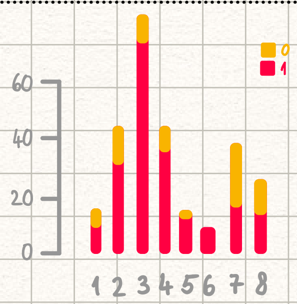
barplot()

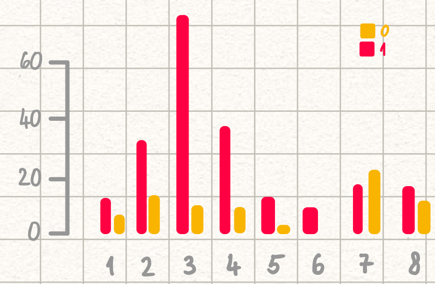


pie()

Чак и када се обележја прикажу одвојено, могу да се донесу неки закључци.
Свакако, некад је потребно приказати два обележја заједно.
То можемо извести нпр. табеларно или на једном bar plot-у:

crossR / typeC	1	2	3	4	5	6	7	8
1	2	11	6	7	1	0	17	4
2	5	23	73	27	4	1	12	7





Пример 2: Цртамо хистограм (густине) узорка који смо генерисали.

Препоруке:
* барем 5 категорија (интервала), по могућности $k = [\log_2 n] + 1$.
* величине категорије:
  1. узорак $X_1, \dots, X_n$ се сортира у варијациони низ;
  2. $R = X_{(n)} - X_{(1)}$ узорачки распон;
  3. $d \approx R/k$ – на веће
* за леву границу првог интервала узети вредност мало мању од минимума, а све границе узети на децималу више од података.
  (да бисмо избегли граничне случајеве)

Када тражимо кандидате за расподелу којом моделирамо посматрано обележје, служимо се хистограмом густине:

```
hist(uzorak, prob = TRUE).
```

Не морамо увек да добијемо расподелу правилног или симетричног облика.
Расподеле могу бити „померене улево" или „померене удесно", „уске", „широке" ...
О тим особинама нам говоре већ поменуте дескриптивне статистике, о којима причамо у наставку.

деф. Мере централне тенденције су:

* очекивана вредност, EX: математичко очекивање;

  Оцена: узорачка средина, $\overline{X} = \frac{X_1 + \dots + X_n}{n}$, тј. њена реализ. вр. $\overline{x} = \frac{x_1 + \dots + x_n}{n}$
  ↳ пример једне статистике
  ↳ у наставку подразумевано мислимо на реализ. вредност

* медијана расподеле, μ: параметар за који $P\{X \leq \mu\} \geq 0.5$ и $P\{X \geq \mu\} \geq 0.5$;

  Оцена: узорачка медијана, $m_e = \begin{cases} X_{(k+1)}, & n = 2k+1 \\ \frac{X_{(k)} + X_{(k+1)}}{2}, & n = 2k \end{cases}$

* мода расподеле: она вредност у којој функција густине (или закон расподеле) достиже максимум.

  Оцена: узорачка мода: вредност која се најчешће појављује у узорку.
  ↳ ово је лоша оцена за апс. непр. расподеле јер нема пуно понављања.

деф. Мере расејања су:

* распон расподеле: скуп тачака у којима је функција густине (или закон расподеле) различита од 0.

  Оцена: узорачки распон, $R = X_{(n)} - X_{(1)}$ (већ смо поменули)

* стандардно одступање расподеле, $\sigma = \sqrt{E(X-EX)^2} = \sqrt{DX}$

  Оцена (за σ^2): узорачка дисперзија, $\overline{S}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$

  или поправљена узорачка дисперзија, $\tilde{S}_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$

* интерквартилно растојање, IQR:

  Прво уводимо α-квантил: то је вредност x ткд. $F(x) = P\{X \leq x\} = \alpha$. $(F^{-1}(\alpha))$

  Специјално: $\alpha = 0.25 \Rightarrow q_1$ - први узорачки квартил;
  $\alpha = 0.75 \Rightarrow q_3$ - трећи узорачки квартил.

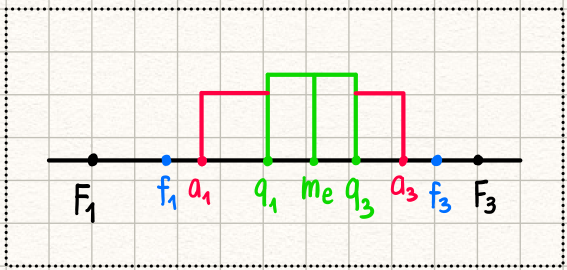
  Њих тражимо тако што нађемо узорачку медијану m_e почетног низа,
  а онда узорачке медијане два добијена подниза. (половине)

  Коначно, $IQR = q_3 - q_1$.

4) Идентификација аутлајера;

деф. **Аутлајер** је члан узорка који се не уклапа у постојећи статистички модел.

Аутлајере никако не треба одбацивати одмах, већ треба испитати утицај те тачке на модел.
Још један начин да представимо податке јесте **box plot** - „кутијасти" дијаграм. boxplot()



boxplot()

q_1, q_3 - први и трећи узорачки квартил;

$f_1 = q_1 - 1.5\ IQR$, $f_3 = q_3 + 1.5\ IQR$;
$F_1 = q_1 - 3\ IQR$, $F_3 = q_3 + 3\ IQR$;

a_1 - први већи од f_1;
a_3 - први мањи од f_3.

Благи аутлајери: између F_1, f_1 и између f_3, F_3.

Прави аутлајери: ван ових граница.

5) Конструкција статистичког модела:

* заључивање о вредностима непознатих параметара;
* тестирање статистичких хипотеза;
* испитивање квалитета модела;

6) Прогноза.

3. Узорачка средина и узорачка дисперзија

* Подсетимо се: узорачка средина $\bar{X}_n = \frac{1}{n}\sum^n X_i$ је оцена за EX.

Теорема 1: $E(\bar{X}_n) = \frac{1}{n}\sum^n EX_i = EX$

Доказ: $E(\bar{X}_n) = E\left(\frac{X_1+\dots+X_n}{n}\right) = \frac{E(X_1+\dots+X_n)}{n} = \frac{EX_1+\dots+EX_n}{n} = \frac{1}{n}\sum EX_i.$

Пошто су сви X_i једнако расподељени $\Rightarrow \frac{1}{n}\sum^n EX_i = \frac{n\cdot EX}{n} = EX.$

Теорема 2: $D(\bar{X}_n) = \frac{1}{n^2}\sum^n D(X_i) = \frac{DX}{n}$

Доказ: $D(\bar{X}_n) = D\left(\frac{X_1+\dots+X_n}{n}\right) = \frac{DX_1+\dots+DX_n}{n^2} = \frac{n\cdot DX}{n^2} = \frac{DX}{n}$

Закључак: Због овога је узорачка средина добра оцена за EX.

Доказ: По Чебишевљевој неједнакости: $P\{|\bar{X}_n - EX| > \varepsilon\} \leq \frac{E(\bar{X}_n - EX)^2}{\varepsilon^2} = \frac{D\bar{X}_n}{\varepsilon^2} = \frac{DX}{n\varepsilon^2}$

Дакле, ако је $DX < +\infty$ и ако $n \to \infty \Rightarrow P\{|\bar{X}_n - EX| > \varepsilon\} \to 0 \Rightarrow$ мало одступање

деф. **Монте-Карло методе** - изводимо експеримент велики број пута и тражимо ср. вредност. На тај начин оцењујемо EX.

* Подсетимо се: узорачка дисперзија $\bar{S}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ је оцена за стандардно одступање.

Теорема 3: $E(n\bar{S}_n^2) = (n-1)\,DX.$

Доказ: $E(n\bar{S}_n^2) = E\left(\sum^n (X_i - \bar{X}_n)^2\right) = E\left(\sum^n X_i^2 - 2\bar{X}_n\sum^n X_i + n\cdot\bar{X}_n^2\right) =$

$$= E\left(\sum^n X_i^2 - 2\cdot n\bar{X}_n^2 + n\bar{X}_n^2\right) = E\left(\sum^n X_i^2 - n\bar{X}_n^2\right) = E\left(\sum^n X_i^2\right) - nE\left(\frac{\sum X_i}{n}\right)^2$$

Приметимо: $EX_iX_j = \begin{cases} EX_i^2 = m^2 + \sigma^2, & \text{за } i=j; \\ EX_i \cdot EX_j = m^2, & \text{за } i \neq j. \end{cases}$ (јер $DX = EX^2 - (EX)^2$)

Такође, $E\bar{X}_n^2 = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n EX_iX_j = \frac{1}{n^2}\sum_{i=1}^n EX_i^2 + \frac{2}{n^2}\sum_{i=1}^{n-1}\sum_{j=i+1}^n EX_iX_j$ (извукли i i, остало симетрично)

$$= \frac{1}{n^2}\cdot n\cdot(m^2+\sigma^2) + \frac{2}{n^2}\cdot\frac{n(n-1)}{2}\cdot m^2 = \frac{m^2+\sigma^2}{n} + \frac{n-1}{n}\cdot m^2$$

Када уврстимо: $E(n\bar{S}_n^2) = E\left(\sum^n X_i^2\right) - nE\left(\frac{\sum X_i}{n}\right)^2 = n\cdot(m^2+\sigma^2) - \left(m^2+\sigma^2+(n-1)m^2\right) = (n-1)\sigma^2$

Напомена: $E(\tilde{S}^2) = \sigma^2$ (исти доказ, само се оно $n-1$ скрати)

4. Емпиријска функција расподеле

Посматрамо обележје X са функцијом расподеле F.
Желимо да оценимо F на основу простог случ. узорка $X_1, X_2, \dots, X_n$.

Како је $F(x) = P\{X \le x\} = E(I\{X \le x\})$, онда је природно оценити F следећом случ. величином:

деф. Емпиријска функција расподеле је $F_n(x) := \dfrac{\sum^n I\{X_i \le x\}}{n}$.

Особине: 1) $E(F_n(x)) = F(x)$;

2) $D(F_n(x)) = \dfrac{F(x)(1-F(x))}{n}$.

Доказ: Приметимо да је случ. вел. $n \cdot F_n(x)$ сума n независних и једнако расп. индикатора. То значи да има биномну $B(n, F(x))$ расподелу.

Тврђење директно следи из овога.

Из ове теореме видимо да што је n веће, то је емпиријска ф-ја све ближа ф-ји расподеле. Ово запажање је садржано у наредној теореми, која је позната и као централна теорема статистике.

Теорема Гливенко-Кантелија:

Нека је $X_1, X_2, \dots, X_n$ п.с.у. из популације са обележјем X са ф-јом расподеле $F(x)$. Даље, нека је $F_n(x)$ одговарајућа емпиријска функција расподеле. Тада:

$$P\{\sup_x |F_n(x) - F(x)| \to 0, \quad \text{кад} \quad n \to \infty\} = 1.$$

Пример: Имамо п.с.у: $-2, -1, 0, 0, 3$.

$$F_n(x) = \begin{cases} 0, & x < -2 \\ 1/5, & -2 \le x < -1 \\ 2/5, & -1 \le x < 0 \\ 4/5, & 0 \le x < 3 \\ 1, & 3 \le x \end{cases}$$

До сада смо се бавили оцењивањем неких параметара популације.
При томе, нисмо имали никакву претпоставку о расподели обележја X (то је било непараметарско оц.).

Шта ако знамо $X \sim \mathcal{N}(m, \sigma^2)$, а не знамо колики су параметри?
Њих желимо да оценимо на основу доступног узорка.

У наставку ћемо показати две методе за то. Једним именом, зову се тачкасте оцене.

5. Метод момената

деф. 1) k-ти теоријски моменат расподеле: EX^k;

2) k-ти узорачки моменат расподеле: $\dfrac{\sum^n X_i^k}{n}$;

3) k-ти теоријски центрирани моменат расподеле: $E(X-EX)^k$;

4) k-ти узорачки центрирани моменат расподеле: $\dfrac{\sum (X_i-\overline{X}_n)^k}{n}$.

теоријски мом.	узорачки мом.	теор. цент. мом.	узор. цент. мом.
EX	$\overline{X}_n$	—	—
EX^2	$\dfrac{\sum^n X_i^2}{n}$	DX	$\overline{S}_n^2$
EX^3	$\dfrac{\sum^n X_i^3}{n}$	$E(X-EX)^3$	$\dfrac{\sum^n (X_i-\overline{X}_n)^3}{n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
EX^k	$\dfrac{\sum^n X_i^k}{n}$	$E(X-EX)^k$	$\dfrac{\sum^n (X_i-\overline{X}_n)^k}{n}$

деф. Метод момената – оцене параметара су решења система који се добија када се изједначе теоријски и узорачки моменти.

Пример: 1) $X \sim \mathcal{P}(\lambda)$. Тражимо оцену параметра λ из Пуасонове расподеле:

$$EX = \lambda = \overline{X}_n \quad \Rightarrow \quad \hat{\lambda} = \overline{X}_n$$

Може и: $DX = \lambda = \overline{S}_n^2 \quad \Rightarrow \quad \tilde{\lambda} = \overline{S}_n^2$

Приметимо: $E\hat{\lambda} = E\overline{X}_n \overset{\boxed{3}T1}{=} EX_1 = EX = \lambda$, $E\tilde{\lambda} = E\overline{S}_n^2 \xrightarrow[n\to\infty]{\boxed{3}T2} \lambda$.
Дакле, ове две оцене се не разликују пуно.

2) $X \sim \mathcal{U}[a,b]$. Тражимо оцену параметара a и b из униформне расподеле:

$$\left.\begin{aligned} EX &= \frac{a+b}{2} = \overline{X}_n \\ DX &= \frac{(b-a)^2}{12} = \overline{S}_n^2 \end{aligned}\right\} \Rightarrow \text{систем: } \begin{aligned} a+b &= 2\overline{X}_n \\ b-a &= \sqrt{12}\,\overline{S}_n \end{aligned} \quad \Rightarrow \quad \begin{aligned} \hat{a} &= \overline{X}_n - \sqrt{3}\,\overline{S}_n \\ \hat{b} &= \overline{X}_n + \sqrt{3}\,\overline{S}_n \end{aligned}$$

3) $X \sim \mathcal{U}[-\theta,\theta]$. Тражимо оцену параметра θ из униформне расподеле:

$EX = 0$ (немамо ништа из овог услова)

$$DX = \frac{(\theta-(-\theta))^2}{12} = \frac{\theta^2}{3} = \overline{S}_n^2 \quad \Rightarrow \quad \hat{\theta} = \sqrt{3}\,\overline{S}_n.$$

6. Метод максималне веродостојности

деф. Метод максималне веродостојности – оцена непознатог параметра (може и вишедимензиони) је вредност која максимизира функцију веродостојности L.

Интуитивно, то је вредност параметра за коју је највероватније да се „деси" баш наш узорак.

деф. Функција веродостојности:

1° дискретно обележје: $L(\theta) := P_\theta\{X_1 = x_1, \dots, X_n = x_n\} \stackrel{\text{п.с.у.}}{=} \prod^n P_\theta\{X_i = x_i\}$

2° непрекидно обележје: $L(\theta) := f_\theta(x_1, \dots, x_n) \stackrel{\text{п.с.у.}}{=} \prod^n f_\theta(x_i)$

Напомена: оцена не мора да постоји, а чак и кад постоји, не мора бити јединствена.

Напомена: често уместо да максимизујемо саму L, максимизујемо $l(\theta) := \log L(\theta)$.
↳ „чува" екстремум

Пример: 1) $X \sim Ber(p)$. Тражимо оцену параметра p из Бернулијеве расподеле:

Тада је обележје дискретно, па је ф-ја веродостојности $L(p) = \prod^n p^{x_i}(1-p)^{1-x_i}$

$l(p) = \log L(p) = \sum^n x_i \cdot \log p + \sum^n (1-x_i) \cdot \log(1-p)$

Ово је диференцијабилно за $p \in (0,1)$, па тражимо $l'(p) = 0$. Добија се $\hat{p} = \overline{X}_n$

Напомена: увек треба и проверити да ли је $\hat{p}$ заиста макс. ф-је веродостојности.

2) $X \sim \mathcal{P}(\lambda)$. Тражимо оцену параметра λ из Пуасонове расподеле:

$L(\lambda) = \prod_{i=1}^n P\{X_i = x_i\} = \prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!} \cdot e^{-\lambda} = \frac{\lambda^{\sum x_i} \cdot e^{-n\lambda}}{\prod x_i!}$

$l(\lambda) = \log L(\lambda) = \sum x_i \cdot \log\lambda - n\lambda - \log(\prod x_i!)$: тражимо λ које максимизира l

$l'(\lambda) = \frac{\sum x_i}{\lambda} - n = 0 \quad \Rightarrow \quad \hat{\lambda} = \frac{\sum x_i}{n} = \overline{X}_n$ (провером, заиста јесте max)
↳ други извод

3) $X \sim \mathcal{U}[0, \theta]$

Сада имамо непрекидно обележје, па: $L(\theta) = \prod_{i=1}^{n} f_\theta(x_i)$

Приметимо: $f_\theta(x_i) = \begin{cases} 0, & x_i \notin [0, \theta] \\ 1/\theta, & \text{иначе} \end{cases} \Rightarrow f_\theta(x_i) = \frac{1}{\theta} \cdot I\{x_i < \theta\}$

Ова ф-ја није диф. по θ, па не можемо радити као у прва два примера.

↳ јер I зависи од θ

$$L(\theta) = \prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} \frac{1}{\theta} \cdot I\{X_i < \theta\} = \frac{1}{\theta^n} I\{X_1 \leq \theta, \dots, X_n \leq \theta\}$$

$$= \frac{1}{\theta^n} I\{X_{(n)} \leq \theta\} = \begin{cases} 0, & \theta < X_{(n)} \\ 1/\theta^n, & \theta > X_{(n)} \end{cases}$$

Како за $\theta \uparrow$ важи $\frac{1}{\theta^n} \downarrow$ $\Rightarrow$ максимизирамо L за најмање могуће θ.

Дакле, $\hat{\theta} = X_{(n)}$.

7. Особине оцена

До сада смо показали два метода, а има их још.
Поставља се питање како одабрати оцену? Због тога уводимо неке особине за оцене.

* деф. Нека је $\hat{\theta}_n$ оцена непознатог параметра θ на основу узорка $X_1, X_2, \dots, X_n$.

Уколико је $E(\hat{\theta}_n) = \theta$, оцена $\hat{\theta}_n$ је непристрасна.

Уколико је $\lim_{n\to\infty} E(\hat{\theta}_n) = \theta$, оцена $\hat{\theta}_n$ је асимптотски непристрасна.

Пример: За параметар θ из $X \sim \mathcal{U}[0,\theta]$ смо добили следеће две оцене: 1) $\hat{\theta}_n = 2\bar{X}_n$ $\boxed{5}$
2) $\tilde{\theta}_n = X_{(n)}$ $\boxed{6}$

1) $\hat{\theta}_n$ је непристрасна: $E(\hat{\theta}_n) = E(2\bar{X}_n) = 2EX = 2\cdot\frac{\theta}{2} = \theta$.

2) $\tilde{\theta}_n$ је асимптотски непристрасна: почињемо као у домаћем у $\boxed{0}$, тј. тражимо $f_{X_{(n)}}$.

$$F_{X_{(n)}}(x) = P\{X_{(n)} \le x\} = P\{X_1 \le x, \dots, X_n \le x\} = \prod P\{X_i \le x\} = F^n(x) = \frac{x^n}{\theta^n}$$

$$\Rightarrow f_{X_{(n)}}(x) = F'_{X_{(n)}}(x) = \frac{n\cdot x^{n-1}}{\theta^n}, \quad \text{за} \quad x \in [0,\theta].$$

Тада: $$E(\tilde{\theta}_n) = EX_{(n)} = \int_{-\infty}^{+\infty} x\cdot\frac{nx^{n-1}}{\theta^n}\,dx = \frac{n}{\theta^n}\int_0^\theta x^n\,dx = \frac{n}{\theta^n}\cdot\frac{x^{n+1}}{n+1}\Big|_0^\theta = \frac{n}{n+1}\cdot\theta \xrightarrow{n\to\infty} \theta.$$

* деф. Уколико је $\forall \varepsilon > 0$ $\lim_{n\to\infty} P\{|\hat{\theta}_n - \theta| > \varepsilon\} = 0$, оцена $\hat{\theta}_n$ је постојана.

Напомена: Довољно је проверити услов: $\lim_{n\to\infty} E(\hat{\theta}_n - \theta)^2 = 0$. (због Чебишевљеве неједнакости)

Ако је оцена непристрасна, овај услов је екв. са: $\lim_{n\to\infty} D(\hat{\theta}_n) = 0$. (**)

Пример: Радимо исти пример од раније:

1) $\hat{\theta}_n$ је постојана: Пошто смо утврдили да је непристрасна, можемо да проверимо (**)

$$D\hat{\theta}_n = D(2\bar{X}_n) = 4\cdot D\left(\frac{X_1+\dots+X_n}{n}\right) = \frac{4}{n^2}\cdot nDX_1 = \frac{4}{n}\cdot\frac{\theta^2}{12} = \frac{\theta^2}{3n} \xrightarrow{n\to\infty} 0.$$

2) $\tilde{\theta}_n$ је постојана:

$$E(\tilde{\theta}_n - \theta)^2 = E\tilde{\theta}_n^2 - 2\theta\cdot E\tilde{\theta}_n + \theta^2 \overset{(*)}{=} \theta^2\cdot\frac{n}{n+2} - 2\theta^2\cdot\frac{n}{n+1} + \theta^2$$

$$= \theta^2\,\frac{n^2+n-2n^2-4n+n^2+3n+2}{(n+2)(n+1)} = \frac{2\theta^2}{(n+2)(n+1)} \xrightarrow{n\to\infty} 0.$$

$$(*)\ E\tilde{\theta}_n^2 = EX_{(n)}^2 = \int_{-\infty}^{+\infty} x^2\cdot\frac{nx^{n-1}}{\theta^n}\,dx = \frac{n}{\theta^n}\int_0^\theta x^{n+1}\,dx = \theta^2\cdot\frac{n}{n+2}$$

* Претпоставимо да имамо две постојане оцене $\hat{\theta}_n$ и $\hat{\hat{\theta}}_n$ за неки параметар θ.
За коју да се определимо? У томе нам помаже наредни критеријум:

деф. Нека су $\hat{\theta}_n$ и $\hat{\hat{\theta}}_n$ две оцене параметра θ.
Кажемо да је $\hat{\theta}_n$ боља у средњеквадратном од $\hat{\hat{\theta}}_n$, уколико за свако θ важи:

$$E(\hat{\theta}_n-\theta)^2 < E(\hat{\hat{\theta}}_n-\theta)^2.$$

Пример: По претх. примерима, $\tilde{\theta}_n$ је боља у средњеквадратном од $\hat{\theta}_n$.

* У наставку ћемо да опишемо како емпиријски закључујемо о особинама које смо увели.

Алгоритам којим добијамо оцене средњеквадратног одступања:

1) Генеришемо узорак $x=(x_1, \dots, x_n)$ из расподеле $F(\theta)$;
2) На основу узорка одредимо $\hat{\theta}_n = \hat{\theta}_n(x)$;
3) Кораке 1 и 2 поновимо N пута: тако добијемо низ оцена $\hat{\theta}_n^{(1)}, \dots, \hat{\theta}_n^{(N)}$;
4) Одредимо квадратно одступање за сваку од N оцена: тако добијемо низ $(\hat{\theta}_n^{(1)}-\theta)^2, \dots, (\hat{\theta}_n^{(N)}-\theta)^2$;
5) Средњеквадратно одступање $E(\hat{\theta}-\theta)^2$ оцењујемо са: $\dfrac{\sum^{N}(\hat{\theta}_n^{(i)}-\theta)^2}{N}$

Аналогно поступамо и ако користимо неко друго растојање $d(\hat{\theta}, \theta)$ уместо средњеквадратног.

Једина разлика је што у 5) оцењујемо са $\dfrac{\sum d(\hat{\theta}_n^{(i)}, \theta)}{N}$.

8. Интервалне оцене параметара

деф. Нека су L_n и U_n статистике такве да: $\to P\{L_n \leq U_n\} = 1;$
$\to P\{L_n < \theta < U_n\} = \beta.$

Интервал (L_n, U_n) је $\beta\%$ двострани интервал поверења за параметар θ.

деф. Ако једна граница није случајна величина, онда имамо једнострани интервал поверења.

деф. Параметар β називамо ниво поверења.
Углавном узимамо вредности 90%, 95%, 99%.

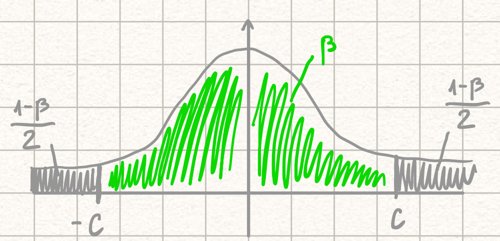
* Поставља се питање како у општем случају конструишемо интервале поверења?
Одговор: потребно је да нађемо функцију од узорка у којој се налази и наш непознат параметар, али чија расподела не зависи од самог параметра.

деф. Статистика из претходне реченице назива се стожерна величина, у ознаци T.

Пример: Нека нам је непознат параметар $m = EX$. Такође, рецимо да имамо велик узорак.

8.2.1 апрокс.

Можемо узети $T = \dfrac{\bar{X}_n - m}{\frac{\tilde{S}_n}{\sqrt{n}}}$ (због великог n то можемо апрокс. са $\mathcal{N}(0,1)$, по ЦГТ).



$$\Phi(c) = \frac{1-\beta}{2} + \beta = \frac{1+\beta}{2} \Rightarrow c = \Phi^{-1}\left(\frac{1+\beta}{2}\right).$$

$$P\{|T| < c\} = \beta \Rightarrow P\left\{\left|\frac{\bar{X}_n - m}{\frac{\tilde{S}_n}{\sqrt{n}}}\right| < c\right\} = \beta$$

$$\Rightarrow P\left\{\bar{X}_n - c\cdot\frac{\tilde{S}_n}{\sqrt{n}} < m < \bar{X}_n + c\cdot\frac{\tilde{S}_n}{\sqrt{n}}\right\} = \beta.$$

Тада је добијени интервал: $\left(\bar{X}_n - c\cdot\frac{\tilde{S}_n}{\sqrt{n}},\ \bar{X}_n + c\cdot\frac{\tilde{S}_n}{\sqrt{n}}\right)$, где $c = \Phi^{-1}\left(\frac{1+\beta}{2}\right)$.

Напомена: за $\beta = 0.95$ је $c \approx 1.96$.

8.1. Закључивање у моделу са биномном $(1,p)$ расподелом

Јасно, у овом случају параметар који оцењујемо је p.

* Најчешће користимо $T = \dfrac{\bar{X}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \overset{\text{ЦГТ}}{\sim} \mathcal{N}(0,1)$.

Одредимо интервал поверења за ово T:

$$P\{|T| < c\} = \beta = P\{|T|^2 < c^2\}$$

$$\beta = P\left\{\left(\frac{\bar{X}_n - p}{\sqrt{\frac{p(1-p)}{n}}}\right)^2 < c^2\right\} = P\left\{\bar{X}_n^2 - 2p\bar{X}_n + p^2 < c^2\frac{p}{n} - c^2\frac{p^2}{n}\right\}$$

$$\beta = P\left\{ p^2\left(\frac{c^2}{n}+1\right) - p\left(\frac{c^2}{n} + 2\bar{X}_n\right) + \bar{X}_n < 0 \right\}.$$

Дакле наш интервал је $p \in (\hat{p}_1, \hat{p}_2)$, где су $\hat{p}_1$ и $\hat{p}_2$ решења квадратне једначине.

Напомена: морамо додатно водити рачуна о скупу допустивих вредности: $p \in [0,1]$.

* За велико n (≥ 100) и када p није блиско ни 0 ни 1, (ни $n\hat{p}$ ни $n(1-\hat{p})$ није мало, праг је 5) можемо и $D\bar{X}_n = \frac{DX}{n} = \frac{p(1-p)}{n}$, оценити својом оценом макс. верод. $\frac{\bar{X}_n(1-\bar{X}_n)}{n}$

Тада користимо $T = \dfrac{\bar{X}_n - p}{\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}} \overset{\text{ЦГТ}}{\sim} \mathcal{N}(0,1)$.

На исти начин као у примеру, интервал који добијамо је:

$$\left(\bar{X}_n - c\cdot\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}},\ \bar{X}_n + c\cdot\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right). \qquad \left(c = \Phi^{-1}\left(\frac{1+\beta}{2}\right)\right)$$

Ову формулу користимо и када желимо да одредимо приближан обим узорка који ће нам обезбедити да нам дужина интервала буде мања од неке унапред задате вредности.

Наиме, дужина интервала је $d = 2c\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. (пошто не знамо n, не знамо ни $\hat{p}$)

Међутим, пошто је $\hat{p}(1-p) \leq 1/4$, онда можемо да нађемо n ткд. $\frac{c}{\sqrt{n}} < d_0$. (убацимо $1/4$ изнад)

8.2. Закључивање у моделу са нормалном расподелом

Све параметре тражимо користећи следећу теорему:

Теорема 1: Нека је $X_1, \dots, X_n$ п.с.у. из $\mathcal{N}(m, \sigma^2)$ расподеле. Тада:

1) $\dfrac{\overline{X}_n - m}{\frac{\sigma}{\sqrt{n}}}$ има $\mathcal{N}(0,1)$ расподелу;

2) $\dfrac{(n-1)\tilde{S}_n^2}{\sigma^2} = \dfrac{n\overline{S}_n^2}{\sigma^2}$ има χ^2_{n-1} расподелу;

3) $\overline{X}_n$ и $\tilde{S}_n^2$ су независне случајне величине;

4) $\dfrac{\sqrt{n}(\overline{X}_n - m)}{\tilde{S}_n}$ има t_{n-1} расподелу.

Идеја доказа: 1) ЦГТ

2) $\dfrac{(n-1)\tilde{S}_n^2}{\sigma^2} = \dfrac{\sum^n (X_i - \overline{X}_n)^2}{\sigma^2} = \sum^n \left(\dfrac{X_i - \overline{X}_n}{\sigma}\right)^2 = \sum^n \left(\dfrac{X_i - m}{\sigma}\right)^2 \overset{\text{алт. деф.}}{\sim} \chi^2_n$

3) изостављамо

4) Из 1) $\Rightarrow Z = \dfrac{\overline{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$.

Из 2) $\Rightarrow Y = \dfrac{(n-1)\tilde{S}_n^2}{\sigma^2} \sim \chi^2_{n-1}$

Из 3) $\Rightarrow Z, Y$ су независне.

$\overset{\text{алт. деф.}}{\Longrightarrow} \dfrac{Z}{\sqrt{\frac{Y}{n-1}}} \sim t_{n-1}$.

Када то распишемо, добијамо баш $\dfrac{\sqrt{n}(\overline{X}_n - m)}{\tilde{S}_n}$.

8.2.1. Интервал поверења за m

1° σ^2 нам је познато: користимо $T = \dfrac{\overline{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$.

Добијамо двострани интервал тако што одређујемо c ткд. $P\{|T| < c\} = \beta \Rightarrow c = \Phi^{-1}\left(\frac{1+\beta}{2}\right)$.

Интервал добијамо из: $|T| \le c \Leftrightarrow \overline{X}_n - c \cdot \frac{\sigma}{\sqrt{n}} < m < \overline{X}_n + c \cdot \frac{\sigma}{\sqrt{n}}$.

2° σ^2 нам није познато: користимо $T = \dfrac{\overline{X}_n - m}{\frac{\tilde{S}_n}{\sqrt{n}}} \sim t_{n-1}$

Добијамо двострани интервал тако што одређујемо c ткд. $P\{|T| < c\} = \beta \Rightarrow c = F^{-1}_{t_{n-1}}\left(\frac{1+\beta}{2}\right)$.

Интервал добијамо из: $|T| \le c \Leftrightarrow \overline{X}_n - c \cdot \frac{\tilde{S}_n}{\sqrt{n}} < m < \overline{X}_n + c \cdot \frac{\tilde{S}_n}{\sqrt{n}}$.

Напомена: као и код биномне, у оба случаја имамо дужину интервала d.
Зато можемо опет одредити n ткд. $d \le D_0$.

8.2.2. Интервал поверења за σ^2

Користимо $T = \frac{(n-1)\tilde{S}_n^2}{\sigma^2} \sim \chi_{n-1}$

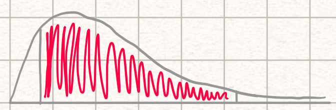
а) двострани:

Пошто χ^2_{n-1} није симетрична (иако се за велико n може апроксимирати нормалном), двострани интервал не можемо наћи као у претх. случајевима.
Уобичајено се прави тако што га одвојимо једнако од обе границе.

Другим речима, тражимо C_1 и C_2 ткд. $P\{T<C_1\} = \frac{1-\beta}{2}$ и $P\{T<C_2\} = \frac{1-\beta}{2}$. $(P\{C_1<T<C_2\}=\beta)$

$\Rightarrow C_1 = F^{-1}_{\chi^2_{n-1}}\left(\frac{1-\beta}{2}\right)$, $C_2 = F^{-1}_{\chi^2_{n-1}}\left(\frac{1+\beta}{2}\right) = F^{-1}_{\chi^2_{n-1}}\left(\frac{1-\beta}{2}+\beta\right)$

па је двострани интервал добијамо из: $C_1<T<C_2 \iff \frac{(n-1)\tilde{S}_n^2}{C_2} < \sigma^2 < \frac{(n-1)\tilde{S}_n^2}{C_1}$.



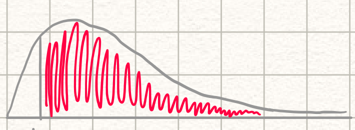
б) једнострани:

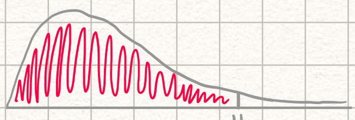
1° $P\{T<C\} = \beta$: тада је $C = F^{-1}_{\chi^2_{n-1}}(\beta) \Rightarrow P\{\sigma^2 > \frac{(n-1)\tilde{S}_n^2}{C}\} = \beta$

Дакле, добијамо интервал облика $(L_n, +\infty)$, где је $L_n = \frac{(n-1)\tilde{S}_n^2}{C}$, $C = F^{-1}_{\chi^2_{n-1}}(\beta)$.

2° $P\{T>C\} = \beta$: тада је $C = F^{-1}_{\chi^2_{n-1}}(1-\beta) \Rightarrow P\{\sigma^2 < \frac{(n-1)\tilde{S}_n^2}{C}\} = \beta$.

Дакле, добијамо интервал облика $(0, U_n)$, где је $U_n = \frac{(n-1)\tilde{S}_n^2}{C}$, $C = F^{-1}_{\chi^2_{n-1}}(1-\beta)$





8.3. Закључивање у моделу са Пуасоновом расподелом

Јасно, у овом случају параметар који оцењујемо је λ.

* Најчешће користимо $T = \frac{\bar{X}_n - \lambda}{\sqrt{\frac{\lambda}{n}}} \overset{\text{ЦГТ}}{\sim} \mathcal{N}(0,1)$.

Остатак аналогно $B(1,p)$ (тј. $\boxed{8}$.1)

9. Интервалне оцене у случају два узорка

Теорема 1: Нека су $X_1, \dots, X_{n_1}$ и $Y_1, \dots, Y_{n_2}$ два независна п.с.у. из $\mathcal{N}(m_1, \sigma_1^2)$ и $\mathcal{N}(m_2, \sigma_2^2)$ редом.

1) $\dfrac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ има $\mathcal{N}(0,1)$ расподелу;

2) Ако је $\sigma_1^2 = \sigma_2^2 = \sigma^2$: $\dfrac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (m_1 - m_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ има $t_{n_1+n_2-2}$ расподелу, $S^2 = \dfrac{(n_1-1)\tilde{S}_{n_1}^2 + (n_2-1)\tilde{S}_{n_2}^2}{n_1+n_2-2}$;

3) Ако је $\sigma_1^2 \neq \sigma_2^2$: $\dfrac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (m_1 - m_2)}{\sqrt{\frac{\tilde{S}_{n_1}^2}{n_1} + \frac{\tilde{S}_{n_2}^2}{n_2}}}$ има t_ν расподелу, ν је реализ. вр. $\dfrac{\left(\frac{\tilde{S}_{n_1}^2}{n_1} + \frac{\tilde{S}_{n_2}^2}{n_2}\right)^2}{\frac{\left(\frac{\tilde{S}_{n_1}^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{\tilde{S}_{n_2}^2}{n_2}\right)^2}{n_2-1}}$

4) $\dfrac{\tilde{S}_{n_1}^2/\sigma_1^2}{\tilde{S}_{n_2}^2/\sigma_2^2}$ има Фишерову F_{n_1-1, n_2-1} расподелу.

9.1. Интервал за количник дисперзија (нормална)

Користимо $T = \dfrac{\tilde{S}_{n_1}^2/\sigma_1^2}{\tilde{S}_{n_2}^2/\sigma_2^2} \sim F_{n_1-1, n_2-1}$.

Како Фишерова расподела није симетрична, интервал тражимо као у 8.2.2. а):

Нађемо C_1 и C_2 т.к.д. $P\{T < C_1\} = \frac{1-\beta}{2}$ и $P\{T < C_2\} = \frac{1+\beta}{2}$. $(P\{C_1 < T < C_2\} = \beta)$

$\Rightarrow C_1 = F^{-1}_{F_{n_1-1, n_2-2}}\left(\frac{1-\beta}{2}\right)$, $C_2 = F^{-1}_{F_{n_1-1, n_2-2}}\left(\frac{1+\beta}{2}\right)$.

па је двострани интервал добијамо из: $C_1 < T < C_2 \Leftrightarrow C_1 \cdot \dfrac{\tilde{S}_{n_2}^2}{\tilde{S}_{n_1}^2} < \dfrac{\sigma_2^2}{\sigma_1^2} < C_2 \cdot \dfrac{\tilde{S}_{n_2}^2}{\tilde{S}_{n_1}^2}$

9.2. Интервал за $m_1 - m_2$ (нормална)

1° σ_1^2 и σ_2^2 су нам познати: користимо $T = \dfrac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (m_1 - m_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1)$.

Добијамо двострани интервал тако што одређујемо C ткд. $P\{|T| < C\} = \beta \Rightarrow C = \Phi^{-1}\left(\frac{1+\beta}{2}\right)$.

Интервал добијамо из: $|T| < C \Leftrightarrow \overline{X}_{n_1} - \overline{Y}_{n_2} - C \cdot \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}} < m_1 - m_2 < \overline{X}_{n_1} - \overline{Y}_{n_2} + C \cdot \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$

2° σ_1^2 и σ_2^2 нам нису познати: Морамо да водимо рачуна о томе да ли је $\sigma_1^2 = \sigma_2^2$.

1. начин: оценимо их и видимо да ли су (приближно) једнаке.

2. начин: - одредимо интервал поверења за $\dfrac{\sigma_2^2}{\sigma_1^2}$.
   - ако је 1 у њему, сматрамо да су дисперзије једнаке.

Бирамо одговарајућу стожерну величину из теореме (2) или 3))

Интервал одређујемо на стандардан начин:

- Нађемо C ткд. $P\{|T| < C\} = \beta$ (водимо рачуна о расподели)
- Из услова $|T| < C$ добијемо интервал.

9.3. Интервал за $p_1 - p_2$ (биномна)

Нека су $X: \begin{pmatrix} 0 & 1 \\ 1-p_1 & p_1 \end{pmatrix}$ и $Y: \begin{pmatrix} 0 & 1 \\ 1-p_2 & p_2 \end{pmatrix}$. Уз то, имамо два узорка обима n_1 и n_2.

За велико n, по ЦГТ и нашој теореми: $T = \dfrac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (p_1 - p_2)}{\sqrt{\dfrac{\overline{X}_{n_1}(1-\overline{X}_{n_1})}{n_1} + \dfrac{\overline{Y}_{n_2}(1-\overline{Y}_{n_2})}{n_2}}} \sim \mathcal{N}(0,1)$.

Интервал одређујемо на стандардан начин.

10. Тестирање статистичких хипотеза

* До сада смо се бавили оцењивањем параметара.
На даље се бавимо проблемом тестирања статистичких хипотеза о вредностима параметара.

деф. Нулта хипотеза H_0 је наша почетна хипотеза, она од које крећемо.

деф. Алтернативна хипотеза H_1 је хипотеза која се прихвата уколико одбацујемо H_0.

деф. Тест статистика T је она статистика на основу чије реализоване вредности доносимо закључак.

деф. Уколико реализ. вр. тест статистике упадне у критичну област W, хипотезу одбацујемо.

Последњи „састојак" статистичког теста је вероватноћа грешке коју допуштамо.

Напомена: Оно што желимо да покажемо стављамо у алтернативну хипотезу H_1.
Тестирање вршимо да би се H_0 одбацила у корист прихватања H_1.

* Приликом статистичког закључивања, могуће је направити грешке.

H_0	тачна	нетачна
прихватамо	+	— (2)
одбацујемо	— (1)	+

деф. Грешка прве врсте је одбацивање тачне нулте хипотезе. $(\Leftrightarrow T \in W \mid H_0)$
Грешка друге врсте је прихватање нетачне нулте хипотезе.

Пример: Суди се оптуженом човеку: ако се докаже да је крив, иде под мач.
H_0 - оптужени невин; H_1 - оптужени крив.

Грешка прве врсте $\Rightarrow$ невин човек страда;
Грешка друге врсте $\Rightarrow$ кривац на слободи.

У статистичком тесту не можемо истовремено да контролишемо обе грешке.

деф. Вероватноћа грешке прве врсте се ограничава пре тестирања.
То ограничење зовемо ниво значајности теста: $P_{H_0}\{T \in W\} \leq \alpha$. (најчешће $\alpha \in \{0.1, 0.05, 0.01\}$)

деф. Мера теста је $\sup P_{H_0}\{T \in W\}$.
Јасно, мера не може бити већа од α. Заправо, често је једнака баш α, па се тако и означава.

деф. Вероватноћа грешке друге врсте означава се са β. (Напомена: $\alpha \uparrow$ $\beta \downarrow$)

деф. Моћ теста је вероватноћа да се одбаци H_1: то је $1-\beta$. (исто што и $P_{H_1}\{T \in W\}$)

* Основни кораци у тестирању:

1) Поставимо нулту и алтернативну хипотезу;

2) Одредимо критичну област W ткд. је ниво значајности теста баш α;

Потребно је познавање расподеле тест статистике под нултом хипотезом.
Ако је расподела иста увек кад важи H_0, онда се расподела може оценити Монте Карло методом.

Алгоритам:
1) Генеришемо узорак из расподеле одређене нултом хипотезом;
2) Одредимо реализ. вр. тест статистике за тај узорак;
3) Поновимо ова два N пута:
   добијамо низ $T_n^{(1)}, \dots, T_n^{(N)}$ који одређује емпиријску ф-ју расподеле F_N тест стат.;
4) Одредимо емпиријски W ткд. $\hat{P}\{T_n \in W\} = \alpha$:

1° Ако је $W = \{T_n \geq C\}$ $\Rightarrow$ $C = F_N^{-1}(1-\alpha)$;
2° Ако је $W = \{T_n \leq C\}$ $\Rightarrow$ $C = F_N^{-1}(\alpha)$;
3° Ако је $W = \{|T_n| \geq C\}$ $\Rightarrow$ $C = F_N^{-1}(1-\frac{\alpha}{2})$;
4° Ако је $W = \{T_n \leq C_1\} \cup \{T_n \geq C_2\}$ $\Rightarrow$ наместимо да је са обе стране по $\frac{\alpha}{2}$.

3) Одредимо вредност тест статистике и видимо да ли упада у критичну област.

Напомена: Моћ расте сразмерно са обимом узорка.

Напомена: Како је моћ теста $P_{H_1}\{T \in W\}$, треба нам и расподела тест стат. ако важи H_1.
Ако то не знамо, можемо га оценити Монте-Карло методама: $P_{H_1}\{T \in W\} = E\,I(T_n \in W) = \frac{S}{N}$ (број упада)

деф. **р-вредност теста** је најмање α за које ћемо, на основу датог узорка, одбацити H_0.

(Ако је $p < \alpha$, одбацујемо H_0; иначе прихватамо)

Најчешће се преко овога врши тестирање.

1° $W = \{T_n \leq C\}$ $\Rightarrow$ $p = P_{H_0}\{T_n \leq \hat{T}_n\}$; ($\hat{T}_n$ — реализ. вр. тест стат.)

2° $W = \{T_n \geq C\}$ $\Rightarrow$ $p = P_{H_0}\{T_n \geq \hat{T}_n\}$;

3° $W = \{|T_n| \geq C\}$ $\Rightarrow$ $p = P_{H_0}\{|T_n| \geq |\hat{T}_n|\}$;

4° $W = \{T_n \leq C_1\} \cup \{T_n \geq C_2\}$ $\Rightarrow$ $p = 2 \cdot \max(P_{H_0}\{T_n \leq \hat{T}_n\}, P_{H_0}\{T_n \geq \hat{T}_n\})$.
(под претпоставком да су делови W такви да имају исту вероватноћу)

Тестови могу бити:

1) параметарски: када је расподела узорка (или условна расподела узорка) позната до на непознат параметар; тестирамо хипотезу о вредностима параметара.

* тестови у нормалном моделу;
* тестови у биномном моделу.

2) непараметарски

* тестови о параметрима популације;
* тестови сагласности с расподелом;
* тестови симетрије; (да ли је расподела симетрична)
* тестови независности 2 или више обележја;
* тестови о једнакој расподељености два узорка.

11. Тестови у нормалном моделу

Нека обележје X има нормалну $\mathcal{N}(m, \sigma^2)$ расподелу и имамо п.с.у. $X_1, \dots, X_n$.

11.1. $H_0: m = m_0$

Пошто је $\overline{X}_n$ тачкаста оцена за m, ако се $\overline{X}_n$ превише разликује од m_0, има смисла одбацити H_0.

* Који тест статистику користимо?

1° Ако је σ^2 познато: користимо $T_n = \dfrac{\overline{X}_n - m_0}{\frac{\sigma}{\sqrt{n}}}$; (ако важи H_0, по $\boxed{8}$ T1.1 $\Rightarrow T_n \sim \mathcal{N}(0,1)$)

2° Ако је σ^2 непознато: користимо $T_n = \dfrac{\overline{X}_n - m_0}{\frac{\overline{S}_n}{\sqrt{n}}}$. (ако важи H_0, по $\boxed{8}$ T1.4 $\Rightarrow T_n \sim t_{n-1}$)

Дакле, под H_0 знамо расподеле ових статистика.

* Сада формирамо критичну област (она зависи од алт. хипотезе - испитујемо само за неке H_1)

1° Ако је σ^2 познато:

а) $H_1: m \neq m_0 \quad \Rightarrow \quad W = \{|T_n| \geq c\}, \quad c = \Phi^{-1}(1-\frac{\alpha}{2})$;

б) $H_1: m < m_0 \quad \Rightarrow \quad W = \{T_n \leq c\}, \quad c = \Phi^{-1}(\alpha)$;

в) $H_1: m > m_0 \quad \Rightarrow \quad W = \{T_n \geq c\}, \quad c = \Phi^{-1}(1+\frac{\alpha}{2})$.

2° Ако је σ^2 непознато:

а) $H_1: m \neq m_0 \quad \Rightarrow \quad W = \{|T_n| \geq c\}, \quad c = F^{-1}_{t_{n-1}}(1-\frac{\alpha}{2})$;

б) $H_1: m < m_0 \quad \Rightarrow \quad W = \{T_n \leq c\}, \quad c = F^{-1}_{t_{n-1}}(\alpha)$;

в) $H_1: m > m_0 \quad \Rightarrow \quad W = \{T_n \geq c\}, \quad c = F^{-1}_{t_{n-1}}(1+\frac{\alpha}{2})$.

Напомена: Тестирање је еквивалентно са прављењем $(1-\alpha)\%$ интервала поверења и провером да ли m_0 припада том интервалу.

Другим речима, интервал поверења је „инвертована" критична област. Важи и обрнуто.

* Одредимо сада моћ теста за $H_0: m = m_0$ и $H_1: m \neq m_0$.

Нека је $M(\theta)$ моћ теста када је $m = \theta$.

$$M(\theta) = P_\theta\{|\frac{\overline{X}_n - m_0}{\frac{\sigma}{\sqrt{n}}}| > c\} = 1 - P_\theta\{|\frac{\overline{X}_n - m_0}{\frac{\sigma}{\sqrt{n}}}| \leq c\} = 1 - P_\theta\{m_0 - \frac{c \cdot \sigma}{\sqrt{n}} \leq \overline{X}_n \leq m_0 + \frac{c \cdot \sigma}{\sqrt{n}}\}$$

$$= 1 - P_\theta\{\frac{m_0 - \frac{c\cdot\sigma}{\sqrt{n}} - \theta}{\frac{\sigma}{\sqrt{n}}} \leq \frac{\overline{X}_n - \theta}{\frac{\sigma}{\sqrt{n}}} \leq \frac{m_0 + \frac{c\cdot\sigma}{\sqrt{n}} - \theta}{\frac{\sigma}{\sqrt{n}}}\}$$

$$= 1 - P_\theta\{-c + \frac{m_0 - \theta}{\frac{\sigma}{\sqrt{n}}} \leq \frac{\overline{X}_n - \theta}{\frac{\sigma}{\sqrt{n}}} \leq c + \frac{m_0 - \theta}{\frac{\sigma}{\sqrt{n}}}\}.$$

$$M(\theta) = 1 - \phi\left(c + \frac{m_0 - \theta}{\frac{\sigma}{\sqrt{n}}}\right) + \phi\left(-c + \frac{m_0 - \theta}{\frac{\sigma}{\sqrt{n}}}\right).$$

1° $\theta = m_0 \Rightarrow M(\theta) = 1 - (1 - 2 \cdot \frac{\alpha}{2}) = \alpha$ (то и треба да важи, јер тада важи H_0)

2° $\theta > m_0 \Rightarrow M(\theta) = 1 + \left[\phi(c) - \phi\left(c + \frac{m_0 - \theta}{\frac{\sigma}{\sqrt{n}}}\right)\right] - \phi(c) - \left[\phi(-c) - \phi\left(-c + \frac{m_0 - \theta}{\frac{\sigma}{\sqrt{n}}}\right)\right] + \phi(-c)$

$= \alpha + A - B$

Због облика густине $\mathcal{N}(0,1) \Rightarrow A \geq B \Rightarrow M(\theta) \geq \alpha$

3° $\theta < m_0 \Rightarrow M(\theta) = 1 - \left[\phi\left(c + \frac{m_0 - \theta}{\frac{\sigma}{\sqrt{n}}}\right) - \phi(c)\right] - \phi(c) + \left[\phi\left(-c + \frac{m_0 - \theta}{\frac{\sigma}{\sqrt{n}}}\right) - \phi(-c)\right] + \phi(-c)$

$= \alpha - D + E$

Због облика густине $\mathcal{N}(0,1) \Rightarrow E \geq D \Rightarrow M(\theta) \geq \alpha$

Напомена: за $\theta \neq m_0$ важи $\lim_{n\to\infty} M(\theta) = 1$

11.2. $H_0: \sigma^2 = \sigma_0^2$

Користимо $T_n = \dfrac{(n-1)\tilde{S}_n^2}{\sigma_0^2}$ (ако важи H_0, по $\boxed{8}$ Т1.2 $\Rightarrow$ $T_n \sim \chi^2_{n-1}$)

Како је $\tilde{S}_n$ оцена за σ^2, велике вредности T_n упућују да је $\sigma^2 > \sigma_0^2$, а мале $\sigma^2 < \sigma_0^2$.

а) $H_1: \sigma^2 \neq \sigma_0^2 \Rightarrow W = \{\overset{\alpha/2}{T_n \leq c_1}\} \cup \{\overset{\alpha/2}{T_n \geq c_2}\}$, $c_1 = F^{-1}_{\chi^2_{n-1}}\left(\frac{\alpha}{2}\right)$, $c_2 = F^{-1}_{\chi^2_{n-1}}\left(1-\frac{\alpha}{2}\right)$;

б) $H_1: \sigma^2 < \sigma_0^2 \Rightarrow W = \{T_n \leq c\}$, $c = F^{-1}_{\chi^2_{n-1}}(\alpha)$;

в) $H_1: \sigma^2 > \sigma_0^2 \Rightarrow W = \{T_n \geq c\}$, $c = F^{-1}_{\chi^2_{n-1}}(1-\alpha)$.

Пример: през. 8, слајд 20 (одличан пример)

11.3. Случај два узорка

Нека су $X_1, \dots, X_{n_1}$ и $Y_1, \dots, Y_{n_2}$ два независна узорка за обележја $X \sim \mathcal{N}(m_1, \sigma_1^2)$ и $Y \sim \mathcal{N}(m_2, \sigma_2^2)$

1° Ако су σ_1^2, σ_2^2 познати:

Користимо тест статистику $T_{n_1,n_2} = \dfrac{\overline{X}_{n_1} - \overline{Y}_{n_2} - m_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ (ако важи H_0, по $\boxed{9}$T1.1 $\Rightarrow T_{n_1,n_2} \sim \mathcal{N}(0,1)$)

а) $H_1: m_1 - m_2 \neq m_0 \Rightarrow W = \{ |T_{n_1,n_2}| \geq c \}$, $c = \Phi^{-1}(1 - \frac{\alpha}{2})$;
б) $H_1: m_1 - m_2 > m_0 \Rightarrow W = \{ T_{n_1,n_2} \geq c \}$, $c = \Phi^{-1}(1 - \alpha)$;
в) $H_1: m_1 - m_2 < m_0 \Rightarrow W = \{ T_{n_1,n_2} \leq c \}$, $c = \Phi^{-1}(\alpha)$.

2° Ако су σ_1^2, σ_2^2 непознати:

* Прво тестирамо да ли важи $H_0: \sigma_1^2 = \sigma_2^2$

Нека је $\frac{\sigma_1^2}{\sigma_2^2} = A$. (дакле $\sigma_1^2 = \sigma_2^2 \Leftrightarrow A = 1$)

Овде користимо $T_{n_1,n_2} = \dfrac{\tilde{S}^2_{n_1}}{\tilde{S}^2_{n_2}} \cdot \dfrac{1}{A}$ (ако важи H_0, по $\boxed{9}$T1.4 $\Rightarrow T_{n_1,n_2} \sim \mathcal{F}_{n_1-1,\, n_2-1}$)

а) $H_1: \frac{\sigma_1^2}{\sigma_2^2} \neq A \Rightarrow W = \{ T_{n_1,n_2} \leq c_1 \} \cup \{ T_{n_1,n_2} \geq c_2 \}$, $c_1 = F_{\mathcal{F}}^{-1}(\frac{\alpha}{2})$, $c_2 = F_{\mathcal{F}}^{-1}(1 - \frac{\alpha}{2})$;
(јер је $\mathcal{F}$ асиметр.)

б) $H_1: \frac{\sigma_1^2}{\sigma_2^2} < A \Rightarrow W = \{ T_{n_1,n_2} \leq c \}$, $c = F_{\mathcal{F}}^{-1}(\alpha)$;

в) $H_1: \frac{\sigma_1^2}{\sigma_2^2} > A \Rightarrow W = \{ T_{n_1,n_2} \geq c \}$, $c = F_{\mathcal{F}}^{-1}(1-\alpha)$.

Напомена: за овај „међутест" узимамо мало веће α него иначе (нпр. 0.1 или 0.2).

2_1° $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Користимо $T_{n_1,n_2} = \dfrac{\overline{X}_{n_1} - \overline{Y}_{n_2} - m_0}{S \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ (ако важи H_0, по $\boxed{9}$T1.2 $\Rightarrow T_{n_1,n_2} \sim t_{n_1+n_2-2}$)

а) $H_1: m_1 - m_2 \neq m_0 \Rightarrow W = \{ |T_{n_1,n_2}| \geq c \}$, $c = F_t^{-1}(1 - \frac{\alpha}{2})$;
б) $H_1: m_1 - m_2 > m_0 \Rightarrow W = \{ T_{n_1,n_2} \geq c \}$, $c = F_t^{-1}(1 - \alpha)$;
в) $H_1: m_1 - m_2 < m_0 \Rightarrow W = \{ T_{n_1,n_2} \leq c \}$, $c = F_t^{-1}(\alpha)$.

2_2° $\sigma_1^2 \neq \sigma_2^2$

Користимо $T_{n_1,n_2} = \dfrac{\overline{X}_{n_1} - \overline{Y}_{n_2} - m_0}{\sqrt{\frac{\tilde{S}^2_{n_1}}{n_1} + \frac{\tilde{S}^2_{n_2}}{n_2}}}$ (ако важи H_0, по $\boxed{9}$T1.3 $\Rightarrow T_{n_1,n_2} \sim t_\nu$)

а) $H_1: m_1 - m_2 \neq m_0 \Rightarrow W = \{ |T_{n_1,n_2}| \geq c \}$, $c = F_{t_\nu}^{-1}(1 - \frac{\alpha}{2})$;
б) $H_1: m_1 - m_2 > m_0 \Rightarrow W = \{ T_{n_1,n_2} \geq c \}$, $c = F_{t_\nu}^{-1}(1 - \alpha)$;
в) $H_1: m_1 - m_2 < m_0 \Rightarrow W = \{ T_{n_1,n_2} \leq c \}$, $c = F_{t_\nu}^{-1}(\alpha)$.

Пример: през. 8, слајд 30

11.4. Спарени тест

Могуће је да се деси да посматрана обележја нису независна и имамо п.с.у. парова $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Нека је $EX_i = m_1$, $EY_i = m_2$.

Претпоставимо да знамо и да $D_i = X_i - Y_i$ има $\mathcal{N}(m_D, \sigma_D^2)$ расподелу, при чему је σ_D^2 непознато.

Тестирамо $H_0: m_D := m_1 - m_2 = m_0$ против оних стандардних алтернативних хипотеза.

То радимо као у $\boxed{11}$. 1.2°: $T_n = \dfrac{\bar{D}_n - m_0}{\frac{\tilde{S}_n}{\sqrt{n}}}$ (ако важи H_0, по $\boxed{8}$ Т1.4 $\Rightarrow T_n \sim t_{n-1}$)

Пример: през. 8, слајд 34

$\boxed{12.}$ Тестови у биномном моделу

12.1. $H_0: p = p_0$

Нека је $X_1, \dots, X_n$ п.с.у. са $\mathcal{B}(1,p)$ расподелом. Тестирамо $H_0: p = p_0$.

1° Ако имамо велики узорак: користимо $T_n = \dfrac{\bar{X}_n - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ (ако важи H_0, по ЦГТ $\Rightarrow T_n \sim \mathcal{N}(0,1)$)

а) $H_1: p \neq p_0 \Rightarrow W = \{|T_n| \geq c\}$, $c = \Phi^{-1}(1-\frac{\alpha}{2})$;

б) $H_1: p < p_0 \Rightarrow W = \{T_n \leq c\}$, $c = \Phi^{-1}(\alpha)$;

в) $H_1: p > p_0 \Rightarrow W = \{T_n \geq c\}$, $c = \Phi^{-1}(1+\frac{\alpha}{2})$.

Пример: през. 8, слајд 37

2° Ако немамо велики узорак: користимо $S_n = X_1 + \dots + X_n$ (ако важи $H_0 \Rightarrow S_n \sim \mathcal{B}(n,p)$)

а) $H_1: p \neq p_0 \Rightarrow W = \{S_n \leq c_1\} \cup \{S_n \geq c_2\}$,

где $\sum_{i=0}^{c_1} \binom{n}{i} p_0^i (1-p_0)^{n-i} \leq \frac{\alpha}{2}$, $\sum_{i=0}^{c_1+1} \binom{n}{i} p_0^i (1-p_0)^{n-i} > \frac{\alpha}{2}$,

$\sum_{i=c_2}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i} \leq \frac{\alpha}{2}$, $\sum_{i=c_2-1}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i} > \frac{\alpha}{2}$.

б) $H_1: p < p_0 \Rightarrow W = \{S_n < c\}$,

где $\sum_{i=0}^{c} \binom{n}{i} p_0^i (1-p_0)^{n-i} \leq \alpha$, $\sum_{i=0}^{c+1} \binom{n}{i} p_0^i (1-p_0)^{n-i} > \alpha$.

в) $H_1: p > p_0 \Rightarrow W = \{S_n > c\}$,

где $\sum_{i=c}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i} \leq \alpha$, $\sum_{i=c-1}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i} > \alpha$.

Пример: през. 8, слајд 41

12.2. Случај два узорка

Нека су $X_1, \dots, X_{n_1}$ и $Y_1, \dots, Y_{n_2}$ два независна узорка.
Тестирамо $H_0: p_1 - p_2 = p_0$.

Користимо тест статистику $T_{n_1,n_2} = \dfrac{\bar{X}_{n_1} - \bar{Y}_{n_2} - p_0}{\sqrt{\underbrace{\frac{\bar{X}_{n_1}(1-\bar{X}_{n_1})}{n_1}}_{\text{оцена за } \frac{p_1(1-p_1)}{n_1}} + \underbrace{\frac{\bar{Y}_{n_2}(1-\bar{Y}_{n_2})}{n_2}}_{\text{оцена за } \frac{p_2(1-p_2)}{n_2}}}}$ (ако важи H_0, по $\boxed{9}$ Т1.1 $\Rightarrow T_{n_1,n_2} \sim \mathcal{N}(0,1)$)

Пример: през. 8, слајд 43

13. Непараметарски тестови

Оно што је заједничко за све тестове до сад је да смо приликом конструкције тест статистика знали њену расподелу под нултом хипотезом.

деф. **Непараметарски тестови** се користе када:

1) Желимо да тестирамо хипотезу о неком параметру популације без претпоставке о расподели.
   Корисно када имамо мали узорак, па не можемо да користимо нормалну расподелу.

2) Желимо да тестирамо сагласност са неком расподелом.

3) Желимо да тестирамо независност два обележја.

13.1. Тест знакова

Тестирамо H_0: $m_e = m_{e0}$, где је m_e медијана расподеле коју има обележје X.

јер је тачно пола веће од медијане

Због деф. медијане, природно долазимо до $T_n = \sum_{i=1}^{n} I\{X_i > m_{e0}\}$ (ако важи $H_0 \Rightarrow T_n \sim \mathcal{B}(n, \frac{1}{2})$)

За велико n (већ од $n > 10$) можемо узети и $T_n^* = \dfrac{T_n - n/2}{\sqrt{\frac{n}{4}}}$ (ако важи $H_0 \Rightarrow T_n^* \sim \mathcal{N}(0,1)$)

Користимо „центрирану верзију": $T_n^c = \sum_{i=1}^{n} I\{X_i > m_{e0}\} - \frac{n}{2}$ (по H_0 би требало да је бр. чл. узорка који су мањи од m_e исти као бр. већих)

а) H_1: $m_e \neq m_{e0}$ $\Rightarrow$ $W = \{|T_n^c| \geq c\}$, односно $W = \{|T_n^*| \geq c\}$;

б) H_1: $m_e < m_{e0}$ $\Rightarrow$ $W = \{T_n^c \leq c\}$, односно $W = \{T_n^* \leq c\}$;

в) H_1: $m_e > m_{e0}$ $\Rightarrow$ $W = \{T_n^c \geq c\}$, односно $W = \{T_n^* \geq c\}$.

Напомена: тест знакова је алтернатива за $\boxed{11}$.1.

13.2. Спарени тест

Посматрамо 2-димензионо обележје (X, Y), али без претпоставке да $D = X - Y$ има нормалну расподелу.
Имамо узорак $(X_1, Y_1), \ldots, (X_n, Y_n)$. Правимо низ: $D_i = X_i - Y_i$.

Тестирамо H_0 да је медијана расподеле $D = X - Y$ има неку фиксну вредност. (нпр. 0)
Зато примењујемо тест знакова на узорак $D_1, \ldots, D_n$.
Ако D има симетричну расподелу, ово се своди на H_0 да је разлика очекивања једнака фиксној вр.

Напомена: ово је алтернатива за $\boxed{11}$.4. (тестирамо једнакост очекивања два обележја из $\mathcal{N}$)

13.3. Вилкоксонов тест

Знамо да је расподела обележја X апс. непрекидна и симетрична.

Тестирамо H_0: $m = m_0$

Приметимо: због симетричности, ако важи $H_0 \Rightarrow X - m_0$ има исту расподелу као $m_0 - X$.
деф. Ранг елемента је његов редни број по величини у том узорку.

Користимо $T_n = \sum_{i=1}^{n} R_i \cdot I\{X_i - m_0 \geq 0\}$, где је R_i ранг елемента $|X_i - m_0|$ у узорку $|X_1 - m_0|, \dots, |X_n - m_0|$.

Уз то, већ за $n > 12$ можемо апроксимирати: $T^* = \dfrac{T_n - ET_n}{\sqrt{DT_n}} \sim \mathcal{N}(0,1)$ (због ЦГТ)

Лема: 1) $ET_n = \dfrac{n(n+1)}{4}$; 2) $DT_n = \dfrac{n(n+1)(2n+1)}{24}$.

Доказ: Приметимо да ако важи H_0, тада T_n има исту расподелу као $T_n' = \sum_{i=1}^{n} I_i$,
где су I_i међусобно независне случ. вел. за које $P\{I_i = i\} = P\{I_i = 0\} = 0.5$ (због сим.)

1) $ET_n = \sum^n EI_i = \sum^n \frac{i}{2} = \frac{n(n+1)}{4}$; (јер $I: \begin{pmatrix} 0 & i \\ 1/2 & 1/2 \end{pmatrix}$, па $EI_i = \frac{i}{2}$)

2) $DT_n = \sum^n DI_i = \sum^n \frac{i^2}{4} = \frac{n(n+1)(2n+1)}{24}$. (јер $I^2: \begin{pmatrix} 0 & i^2 \\ 1/2 & 1/2 \end{pmatrix}$, па $DI_i = \frac{i^2}{2} - \frac{i^2}{4} = \frac{i^2}{4}$)

13.4. Вилкоксонов тест за два независна узорка

Имамо два независна обележја X и Yткд. важи $X = Y + c$ (иста расподела до на парам. локације)
Имамо и узорке $X_1, \dots, X_n, Y_{n+1}, \dots, Y_{n+m}$.
Тестирамо H_0: $c = 0$ (тј. $X = Y$)

Користимо $T = \sum_{i=1}^{n} R_i$, где је R_i ранг i-тог елем. из узорка $X_1, \dots, X_n$ у обједињеном узорку.

Пример: први узорак: 1, 3, 5, 8; други узорак: 0, 1, 3, 4, 7.
обједињени узорак (сортиран): 0, 1, 1, 3, 3, 4, 5, 7, 8.
одговарајући рангови: 1, 2.5, 2.5, 4.5, 4.5, 6, 7, 8, 9.
$\hat{T}_9 = 2.5 + 4.5 + 7 + 9$

Ову тест статистику можемо записати и као $T_{n,m} = \sum_{i=1}^{n+m} R_i \cdot z_i$, где је $z_i = \begin{cases} 1, & \text{ако је из првог узорка} \\ 0, & \text{иначе} \end{cases}$

Ако важи H_0 и нема међусобног понављања $\Rightarrow ET = \dfrac{n(n+m+1)}{2}$, $DT = \dfrac{n \cdot m \cdot (n+m+1)}{12}$. (доказ: стр. 68)

Наравно, опет користимо: $T^* = \dfrac{T - ET}{\sqrt{DT}} \sim \mathcal{N}(0,1)$ (по ЦГТ)

$\boxed{14.}$ Тестови сагласности са расподелом

Овде посматрамо непараметарске тестове у којима проверавамо да ли је неки модел исправан. Тачније, тестирамо H_0: $F = F_0$. (F - функција расподеле)

Подсетимо се Гливенко-Кантелијеве теореме (из $\boxed{4}$):

Нека је $X_1, X_2, \ldots, X_n$ п.с.у. из популације са обележјем X са ф-јом расподеле $F(x)$. Даље, нека је $F_n(x)$ одговарајућа емпиријска функција расподеле. Тада:

$$P\{\sup_x |F_n(x) - F(x)| \to 0, \quad \text{кад} \quad n \to \infty\} = 1.$$

Дакле, ако је $F_n(x) - F(x)$ значајно различито од нуле, онда треба одбацити H_0: $F = F_0$.

14.1. Тест Колмогоров-Смирнова

Користимо: $D_n := \sup_x |F_n(x) - F_0(x)|$.

Напомена: Ако важи H_0, расподела D_n не зависи од F_0.

Доказ: $D_n = \sup_x |F_n(x) - F_0(x)| = \sup_{y = F_0(x)} |F_n(F_0^{-1}(y)) - y| = |\frac{1}{n}\sum_{i=1}^{n} I\{X_i \le F_0^{-1}(y)\} - y|$

$= |\frac{1}{n}\sum_{i=1}^{n} I\{F_0(X_i) \le y\} - y|$

$= |\frac{1}{n}\sum_{i=1}^{n} I\{U_i \le y\} - y|$, где $U_i \sim \mathcal{U}[0,1]$ (по $\boxed{0}$ Зад 2)

За лакше одређивање вр. тест статистике, користимо други запис $D_n = \max_{1 \le k \le n} \left(|F_0(X_i) - \frac{k-1}{n}|, |\frac{k}{n} - F_0(X_i)|\right)$.
Ово важи јер је $F_n(x)$ део по део константна, а F_0 непрекидна (проверавамо промене у тачкама скока).

Критична област: $W = \{D_n \geq c\}$. (због Гливенко-Кантелија)

14.2. Тест Крамер-фон Мизеса

Користимо: $\omega_n^2 := \int_{-\infty}^{+\infty} (F_n(x) - F_0(x))^2 \, dF_0(x)$. (ако важи H_0, вредности ω_n^2 су блиске нули)

Критична област: $W = \{\omega_n^2 > c\}$.

14.3. Тест Андерсон-Дарлинга

Користимо: $A_n := \int_{-\infty}^{+\infty} \frac{(F_n(x) - F_0(x))^2}{F_0(x) \cdot (1 - F_0(x))} \, dF_0(x)$. (ако важи H_0, вредности A_n су блиске нули)

Критична област: $W = \{A_n > c\}$.

* Шта ако желимо да тестирамо H_0: $F = F_0(\theta)$, где је θ непознат параметар? Само адаптирамо претх.

θ оценимо (најбоље методом макс. веродостојности) $\Rightarrow$ добијамо $\hat{\theta}$.
То $\hat{\theta}$ користимо за рачунање тест статистике.

Проблем: Расподела тест статистике под H_0 није иста као кад се параметар не оцењује.
Такође, може и зависити од F_0.

Срећом: Уколико непознати параметар осликава скалирање или локацију, (нпр. $\mathcal{N}, \mathcal{E}$...)
тада расподела т.с. не зависи од оцењених параметара $\Rightarrow$ можемо емп. оценити.

14.4. χ^2 - тест

Могу се примењивати и када X није апсолутно непрекидна случајна величина.

1) Поделимо цео скуп вредности сл. величине X у k дисјунктних категорија. (по процени)
M_j – број елемената у j-тој категорији.

Напомена: $M_j \sim \mathcal{B}(n, p_j)$, при чему, ако важи H_0 $\Rightarrow$ $p_j = P_{H_0}\{X$ је у j-тој категорији$\}$

Дакле, $EM_j = n \cdot p_j$.

2) Користимо: $T_n := \sum_{j=1}^{k} \frac{(M_j - np_j)^2}{np_j}$ (ако важи H_0, по $\boxed{8}$ T1.2 $\Rightarrow$ $T_n \sim \chi^2_{k-1}$)

Напомене: 1) Ако F_0 зависи од непознатих параметара θ, они се оцене тако да минимизирају T_n.
На основу њих одређујемо EM_j.

2) Мора да важи: $np_j \geqslant 5$ $(n\hat{p}_j \geqslant 5)$, иначе треба спојити неке категорије.

15. Тестови једнаке расподељености два узорка

* Може се користити Вилкоксонов тест за два независна узорка. ($\boxed{13}$.4)

* Осим тога, можемо правити тестове аналогне класичним тестовима сагласности (само сад за 2 узорка).

H_0: $F = G$ (F, G - функције расподела обележја X и Y)

F_{n_1}, G_{n_2} - емп. ф-је расподела.
$N = n_1 + n_2$ - величина обједињеног узорка.

Оцењујемо $F-G$ са $F_{n_1} - G_{n_2}$:

1° Колмогоров-Смирнов тест: $D_{n_1,n_2} := \sup_{x\in\mathbb{R}} |F_{n_1}(x) - G_{n_2}(x)|$

за $\sqrt{\frac{n_1 n_2}{N}} D_{n_1,n_2}$ је нађена гранична расподела под H_0

2° Крамер-фон Мизесов тест: $CM_{n_1,n_2} := \int_{-\infty}^{+\infty} (F_{n_1}(x) - G_{n_2}(x))^2 \, dH_N(x)$ (H_N - емп. ф-ја за обједињени)

за $\frac{n_1 n_2}{N} CM_{n_1,n_2}$ је нађена гранична расподела под H_0

16. Тестови независности два обележја

16.1. χ^2 тест независности

Имамо п.с.у. $(X_1, Y_1), \dots, (X_n, Y_n)$

Тестирамо H_0: обележја X, Y су независна. (тј. $\forall A, B \quad P\{X \in A, Y \in B\} = P\{X \in A\} \cdot P\{Y \in B\}$)

1) Поделимо узорак у $K \cdot L$ категорија (K за X и L за Y)
   M_{ij} - бр. елем. узорка чија се X компонента налази у i-тој, а Y компонента у j-тој категорији.

   Напомена: $M_{ij} \sim \mathcal{B}(n, p_{ij})$, где је $p_{ij} = P\{X \in A_i, Y \in B_j\}$.

2) Користимо: $T_n := \sum_{i=1}^{K} \sum_{j=1}^{L} \frac{(M_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}}$. (ако важи $H_0 \Rightarrow T_n \sim \chi^2_{(K-1)(L-1)}$, асимптотски)

   Напомена: Ако важи $H_0 \Rightarrow p_{ij} = p_{i\cdot} \cdot p_{\cdot j}$, где $p_{i\cdot} = \{X \in A_i\}$, $p_{\cdot j} = \{Y \in B_j\}$ - маргиналне вероватноће

   Њих тада исто можемо да оценимо: $\hat{p}_{i\cdot} = \frac{\sum^{L} M_{ij}}{n}$, $\hat{p}_{\cdot j} = \frac{\sum^{K} M_{ij}}{n}$

3) Критична област: $W = \{T_n \geq c\}$.

Напомена: Мора да важи: $np_{ij} \geq 5$ $(n\hat{p}_{ij} \geq 5)$, иначе треба спојити неке категорије.

16.2. Пирсонов и Спирманов тест некорелисаности

Опет је H_0: обележја X, Y су независна.

Из УУВ знамо: Коваријација случ. вел. X и Y је $cov(X,Y) := E((X-EX)(Y-EY))$.

Коефицијент корелације је $\rho := \frac{cov(X,Y)}{\sqrt{DX}\cdot\sqrt{DY}}$.

Напомена: X, Y независни $\overset{\Rightarrow}{\not\Leftarrow}$ $\rho = 0$: X, Y некорелисане;
X, Y зависни $\Leftrightarrow$ $\rho = \pm 1$: X, Y корелисане;

деф. Пирсонов коеф. корелације је оцена за ρ: $\hat{\rho}_n := \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X}_n)^2}\cdot\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y}_n)^2}}$.

Лема 1: Ако важи H_0 и ако $X, Y \sim \mathcal{N}$, онда $T_n = \hat{\rho}_n\sqrt{\frac{n-2}{1-\hat{\rho}_n^2}} \sim t_{n-2}$.

Критична област:

Због напомене, ако су вредности $\hat{\rho}_n$ блиске ± 1, то упућује на јаку корелисаност, док вредности $\hat{\rho}_n$ блиске 0 упућују на некорелисаност

Баш зато, ако сумњамо на корелисаност, критична област ће бити: $W_{+1} = \{T_n \geq c\}$ или $W_{-1} = \{T_n \leq c\}$.
У супротном, критична област ће бити облика: $W_0 = \{|T_n| \geq c\}$.

Ово је лепо, али смета претпоставка $X, Y \sim \mathcal{N}$. Како тога да се решимо?

деф. Спирманов коеф. корелације је Пирсонов коеф. примењен на статистике ранга, тј. $(R_1, \dots, R_n)$ и $(S_1, \dots, S_n)$.

$$\hat{r}_n := \frac{\sum_{i=1}^{n}(R_i - \bar{R}_n)(S_i - \bar{S}_n)}{\sqrt{\sum_{i=1}^{n}(R_i-\bar{R}_n)^2}\cdot\sqrt{\sum_{i=1}^{n}(S_i-\bar{S}_n)^2}}.$$

Лема 2: 1) За велико n, када нема понављања у узорку и ако су обележја независна, важи:

$$T_n := \frac{\hat{r}_n}{\sqrt{\frac{1}{n-1}}} \sim \mathcal{N}(0,1)$$

2) Већ за $n > 10$, можемо користити следећу статистику:

$$T_n := \hat{r}_n\sqrt{\frac{n-2}{1-\hat{r}_n^2}} \sim t_{n-2}$$

Напомена: У пракси, може и са понављањем, само за рангове истих елем. узимамо ср. вред.

17. Регресиони модели

Идеја је да моделујемо зависност између две или више случајних величина.

деф. Регресиона функција је $f(X) := E(Y|X)$. (X може бити и вишедимензиона).
У том случају, Y зовемо зависна променљива, док X зовемо предиктор.

деф. Регресиони модел има за циљ да моделује зависност између случ. величина.
Један пример за то је адитивни регр. модел: $Y = f(X) + \varepsilon$, ε - нека случ. вел. независна од X
(најчешће $\varepsilon \sim \mathcal{N}(0, \sigma^2)$)

Дакле, наш циљ је да моделирамо зависност, тј. $f(x)$. Зато можемо X сматрати познатим.
Имамо две могућности:

1° претпоставимо функционалну зависност која зависи од неких параметара и да њих оценимо;
2° непараметарски оценимо саму функцију.

Бирамо прву могућност: то зовемо проста линеарна регресија.

Имамо узорак $(X_1, Y_1), \dots, (X_n, Y_n)$.
Модел који желимо да применимо је: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, где је $\{\varepsilon_i\}$ низ случ. вел. ткд:

1) $E(\varepsilon_i) = 0$, $\forall i$;
2) $E(\varepsilon_i \varepsilon_j) = 0$, $\forall i \neq j$;
3) $D(\varepsilon_i) = \sigma^2 < +\infty$;
4) X_i, ε_i су независне.

Због ових услова, важи: $EY_i = \beta_0 + \beta_1 X_1$; $E(Y_i|X_i) = \beta_0 + \beta_1 X_1$;
Сада оцењујемо непознате параметре β_0, β_1. То ћемо урадити тзв. методом најмањих квадрата:

Тражимо β_0 и β_1 које минимизирају функцију: $S(\beta_0, \beta_1) := \sum_{i=1}^{n} (Y_i - (\beta_1 X_i + \beta_0))^2$.
То постижемо решавањем следећег система:

$$\frac{\partial S(\beta_0,\beta_1)}{\partial \beta_1} = \sum_{i=1}^{n} -2X_i(Y_i - \beta_1 X_i - \beta_0) = 0 \iff \sum^n Y_i X_i - \beta_0 \sum^n X_i - \beta_1 \sum^n X_i^2 = 0; \quad (1)$$

$$\frac{\partial S(\beta_0,\beta_1)}{\partial \beta_0} = \sum_{i=1}^{n} -2(Y_i - \beta_1 X_i - \beta_0) = 0 \iff \sum^n Y_i - \beta_0 \cdot n - \beta_1 \sum^n X_i = 0;$$

$$\overset{/:n}{\iff} \bar{Y}_n - \beta_0 - \beta_1 \bar{X}_n = 0;$$

$$\iff \beta_0 = \bar{Y}_n - \beta_1 \bar{X}_n; \quad (2)$$

Уврстимо (2) у (1): $\sum^n Y_i X_i - \beta_1 \sum^n X_i^2 - \bar{Y}_n \sum^n X_i + \beta_1 \bar{X}_n \sum^n X_i = 0 \overset{/:n}{\Longrightarrow} \frac{\sum^n Y_i X_i}{n} - \beta_1 \frac{\sum^n X_i^2}{n} - \bar{Y}_n \cdot \bar{X}_n + \beta_1 \cdot \bar{X}_n^2 = 0;$

Коначно, добијамо: $\hat{\beta}_1 = \dfrac{\frac{\sum Y_i X_i}{n} - \bar{Y}_n \cdot \bar{X}_n}{\frac{\sum X_i^2}{n} - \bar{X}_n^2} = \dfrac{\sum Y_i X_i - n \bar{Y}_n \bar{X}_n}{\sum X_i^2 - n \bar{X}_n^2} = \dots = \sum^n Y_i \cdot \dfrac{(X_i - \bar{X}_n)}{n \cdot \bar{S}_X^2}$;

$$\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n = \dots = \sum^n \frac{Y_i}{n}\left(1 - \frac{\bar{X}_n (X_i - \bar{X}_n)}{\bar{S}_X^2}\right).$$

Закључак: Наша оцењена регресиона функција је: $Y = \hat{\beta}_0 + \hat{\beta}_1 X$.
↳ познато

Напомена: Ова права садржи тачку $(\bar{X}_n, \bar{Y}_n)$.
Одавде видимо да овај приступ моделирању има за циљ да добро опише тачке близу просека.

Испитајмо особине управо добијених оцена:

1) $\hat{\beta}_1$: а) јесте непристрасна:

$$E(\hat{\beta}_1) = \sum^n EY_i \cdot \frac{(x_i - \bar{x}_n)}{n \cdot \bar{s}_x^2} = \sum^n (\beta_0 + \beta_1 x_i) \cdot \frac{(x_i - \bar{x}_n)}{n \cdot \bar{s}_x^2} = \beta_0 \cdot \sum^n \frac{(x_i - \bar{x}_n)}{n \cdot \bar{s}_x^2} + \beta_1 \cdot \frac{\sum x_i^2 - n \cdot \bar{x}_n^2}{n \cdot \bar{s}_x^2} = 0 + \beta_1 \cdot \frac{n \cdot \bar{s}_x^2}{n \cdot \bar{s}_x^2} = \beta_1.$$

б) јесте постојана:

$$D(\hat{\beta}_1) = \sum^n DY_i \cdot \frac{(x_i - \bar{x}_n)^2}{n^2 \cdot \bar{s}_x^4} = \sum^n \sigma^2 \cdot \frac{(x_i - \bar{x}_n)^2}{n^2 \cdot \bar{s}_x^4} = \sigma^2 \cdot \frac{n \bar{s}_x^2}{n^2 \bar{s}_x^4} = \frac{\sigma^2}{n \cdot \bar{s}_x^2}.$$

2) $\hat{\beta}_0$: а) јесте непристрасна:

$$E(\hat{\beta}_0) = E(\bar{Y}_n - \beta_1 \bar{X}_n) = \sum^n \frac{EY_i}{n} - \beta_1 \bar{X}_n = \sum^n \frac{\beta_1 X_i + \beta_0}{n} - \beta_1 \cdot \bar{X}_n = \beta_0$$

б) јесте постојана:

$$D(\hat{\beta}_0) = D\left(\sum^n \frac{Y_i}{n}\left(1 - \frac{\bar{x}_n (x_i - \bar{x}_n)}{\bar{s}_x^2}\right)\right) = \frac{1}{n^2} \cdot \sum^n \sigma^2 \left(1 - \frac{\bar{x}_n (x_i - \bar{x}_n)}{\bar{s}_x^2}\right)^2$$

$$= \frac{\sigma^2}{n^2} \sum^n \left(1 - \frac{2\bar{x}_n (x_i - \bar{x}_n)}{\bar{s}_x^2} + \frac{\bar{x}_n^2 (x_i - \bar{x}_n)^2}{\bar{s}_x^4}\right) = \frac{\sigma^2}{n^2}\left(n + \frac{\bar{x}_n^2 \, n \bar{s}_x^2}{\bar{s}_x^4}\right) = \frac{\sigma^2}{n^2}\left(1 + \frac{\bar{x}_n^2}{\bar{s}_x^2}\right).$$

Напомена: **Уколико додатно претпоставимо да су грешке модела $\{\varepsilon_i\}$ низ незав. случ вел. са $\mathcal{N}(0, \sigma^2)$,**

1) **добијене оцене се поклапају са оценама које добијамо и методом максималне веродостојности;**

2) **можемо одредити расподелу добијених оцена.**

Доказ: 1) Приметимо да ако $\varepsilon_i \sim \mathcal{N}(0, \sigma^2) \Rightarrow Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i)$.

Зато је функција веродостојности: $L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2}}$

Њен логаритам је $l(\beta_0, \beta_1, \sigma^2) = -n \log(\sqrt{2\pi}) - \frac{n}{2} \log(\sigma^2) - \frac{S(\beta_0, \beta_1)}{2\sigma^2}$.

Одавде је јасно да вредности које максимизују l су баш оне које миним. $S(\beta_0, \beta_1)$.
Осим тога, добијамо и оцену за σ^2:

$$\tilde{\sigma}_n^2 = \frac{S(\hat{\beta}_0, \hat{\beta}_1)}{n} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n}.$$

2) $\hat{\beta}_0$ и $\hat{\beta}_1$ су лин. комб. нормално расподељених случ. вел.

Зато су њихове расподеле редом $\mathcal{N}(E(\hat{\beta}_0), D(\hat{\beta}_0))$ и $\mathcal{N}(E(\hat{\beta}_1), D(\hat{\beta}_1))$.
Како смо $E\hat{\beta}_i$ и $D\hat{\beta}_i$ већ израчунали горе, ако стандардизујемо, добијамо:

$$\frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{n} \cdot \bar{s}_x}} \sim \mathcal{N}(0,1) \quad \text{и} \quad \frac{\hat{\beta}_0 - \beta_0}{\frac{\sigma}{\sqrt{n}} \sqrt{1 + \frac{\bar{x}_n^2}{\bar{s}_x^2}}} \sim \mathcal{N}(0,1)$$

Из претх, jасно je да можемо узети баш ове статистике да бисмо направили инт. поверења за β_0, β_1.
или да тестирамо имаjу ли баш неку вр.

Ипак, нећемо баш то радити. Покажимо зашто:

Пример: Тестирамо H_0: $\beta_1 = 0$ (ако je тачно $\Rightarrow$ предиктор нема утицаjа, jер онда $Y_i = \beta_0 + 0 \cdot X_i + \varepsilon_i$)

Ако желимо да искористимо $T_n := \dfrac{\hat{\beta}_1}{\frac{\sigma}{\sqrt{n} \cdot \bar{S}_X}}$ ($\sim \mathcal{N}(0,1)$ ако je H_0 тачно)

Проблем: σ^2 je непознат $\Rightarrow$ морамо да га оценимо.

Можемо ли да узмемо нпр. ону оцену из претх. доказа?
Не, зато што je то параметар коjи представља дисперзиjу грешака. ($\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$)
Због тога, ова оцена ниjе непристрасна. (добиjе се $E(\tilde{\sigma}_n^2) = \frac{n-2}{n} \sigma^2$)

Ево „правог" поступка:

Узмемо оцену за σ^2 коjа jесте непристрасна, нпр. $\hat{\sigma}_n^2 = \frac{n}{n-2} \tilde{\sigma}_n^2$. („поправили" стару)

Узмемо статистику: $T_n := \dfrac{\hat{\beta}_1}{\frac{\hat{\sigma}}{\sqrt{n} \cdot \bar{S}_X}} \sim t_{n-2}$ (ако je H_0 тачно) (иста, само оцењено σ)

Пример: аналогно се врши тестирање H_0: $\beta_0 = 0$.

18. Оцена регресионе функције

Ако је вредност предиктора x_0, онда је оцена рег. ф-је у тој тачки је $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$. (закључак на првој страни 17)

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = \frac{1}{n}\sum Y_i - \bar{x}_n \cdot \sum Y_i \cdot \frac{(x_i - \bar{x}_n)}{n \cdot \bar{s}_x^2} + x_0 \cdot \sum Y_i \cdot \frac{(x_i - \bar{x}_n)}{n \cdot \bar{s}_x^2}$$

$$\hat{Y}_0 = \sum Y_i \cdot \left(\frac{1}{n} + (x_0 - \bar{x}_n) \cdot \frac{(x_i - \bar{x}_n)}{n \cdot \bar{s}_x^2}\right)$$

Дакле, $\hat{Y}_0$ може да се представи као лин. комб. нормално расподељених случ. вел.

$$E(\hat{Y}_0) = E(\hat{\beta}_0) + E(\hat{\beta}_1) x_0 = \beta_0 + \beta_1 x_0;$$

$$D(\hat{Y}_0) = \sum\left(\frac{1}{n} + (x_0 - \bar{x}_n) \cdot \frac{(x_i - \bar{x}_n)}{n \cdot \bar{s}_x^2}\right)^2 \cdot \sigma^2 = \ldots = \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x}_n)^2}{n \bar{s}_x^2}\right)$$

Следи да важи:
$$\frac{\hat{Y}_0 - EY_0}{\hat{\sigma}_n \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x}_n)^2}{n \bar{s}_x^2}}} \sim t_{n-2}$$

Због тога, можемо направити β% интервал поверења за средњу вредност зависне променљиве:

$$\left(\hat{Y}_0 - C \cdot \hat{\sigma}_n \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x}_n)^2}{n \bar{s}_x^2}},\quad Y_0 + C \cdot \hat{\sigma}_n \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x}_n)^2}{n \bar{s}_x^2}}\right), \quad C = F^{-1}_{t_{n-2}}\left(\frac{1+\beta}{2}\right)$$ (t_{n-2} симетрична)

Напомена: Интервал је најужи када узмемо $x_0 = \bar{x}_n$.
Такође, за $n \to \infty$, дужина интервала $\to 0$ (зато што је $\hat{Y}_0$ постојана оцена за $\beta_0 + \beta_1 x_0$).

Приметимо и следеће:

Како је $Y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0$, добијамо:
* $E(\hat{Y}_0 - Y_0) = 0$
* $D(\hat{Y}_0 - Y_0) = D(\hat{Y}_0) + D(\varepsilon_0) = \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x}_n)^2}{n \bar{s}_x^2}\right)^2 + \sigma^2$

Следи да важи:
$$\frac{\hat{Y}_0 - Y_0}{\hat{\sigma}_n \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x}_n)^2}{n \bar{s}_x^2}}} \sim t_{n-2}.$$

Због тога, можемо направити β% интервал поверења за вредност зависне променљиве:

$$\left(\hat{Y}_0 - C \cdot \hat{\sigma}_n \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x}_n)^2}{n \bar{s}_x^2}},\quad Y_0 + C \cdot \hat{\sigma}_n \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x}_n)^2}{n \bar{s}_x^2}}\right), \quad C = F^{-1}_{t_{n-2}}\left(\frac{1+\beta}{2}\right)$$

Напомена: Овај интервал је шири од оног горе.
Осим тога, када $n \to \infty$, дужина интервала $\not\to 0$

19. Квалитет регресионог модела

Сада ћемо показати како да установимо колико добро наш модел описује посматрану зависност.

деф. Резидуал i-те обзервације је $e_i := Y_i - (\beta_0 + \hat{\beta}_i X_i)$. $\quad (= Y_i - \hat{Y}_i)$

деф. Уведимо ознаке: 1) $SSE := \sum_{i=1}^{n} e_i^2$;

2) $SSR := \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}_n)^2$;

3) $SSTO := \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2$.

Лема 1: $SSTO = SSR + SSE$. (укупан варијабилитет = објашњен методом + грешка)

Доказ: $\sum^{n} (\hat{Y}_i - \bar{Y}_n) e_i = \sum^{n} (\hat{\beta}_0 + \hat{\beta}_i X_i) e_i = \hat{\beta}_1 \sum^{n} X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0.$

$$SSTO = \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2 = \sum^{n} (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y}_n)^2 = \underbrace{\sum^{n} (Y_i - \hat{Y}_i)^2}_{e_i^2} + \underbrace{\sum^{n} (\hat{Y}_i - \bar{Y}_n)^2}_{SSR} + \underbrace{2 \sum^{n} (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}_n)}_{0 \text{ (горе)}} = SSE + SSR.$$

деф. Природно, мера за квалитет модела је коефицијент детерминације: $R^2 := 1 - \frac{SSE}{SSTO} = \frac{SSR}{SSTO}$

Напомена: Овај коеф. се прво израчуна на тзв. тест подацима. (као историјски код акција)

Ово није једина мера, постоје и друге.
Свакако, ако нам је циљ предикција, R^2 је добра мера.