Curs 17

Media, varianța și momentele empirice

Fie $\underbrace{X_1, X_2, \dots, X_n}$ $\sim f_\theta$, $\theta \in \Theta \subseteq \mathbb{R}^k$

indep. și identic repartizate cu densitatea (resp. fct. de masă) f_θ

Exp: a) discrete — Binomiale $(\underbrace{n, p}_{\theta})$

Bernoulli $(\underbrace{p}_{\theta})$

Geometrică $(\underbrace{p}_{\theta})$

Poisson $(\underbrace{\lambda}_{\theta})$

b) continue — Exponențiala $(\underbrace{\lambda}_{\theta})$

Normală $(\underbrace{\mu, \sigma^2}_{\theta})$

media eșantionului (empirică) $\rightarrow T_n = \overline{X_n} = \frac{X_1 + X_2 + \dots + X_n}{n}$

varianța empirică $\rightarrow V_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X_i})^2$

varianța eșantionului $\rightarrow S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X_i})^2$

moment empiric de ordin k $\rightarrow M_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$

moment empiric centrat de ordin k $\rightarrow M_k' = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X_i})^k$

(P) Dacă $X_1, X_2, \dots, X_n \sim f_\theta$ de medie μ și dispersie σ^2 atunci:

a) $E[X_n] = \mu$

b) $Var(\overline{X_n}) = \sigma^2 / n$

c) $E[V_n^2] = \frac{n-1}{n} \sigma^2$ și $E[S_n^2] = \sigma^2$

Dem:

a) $E[\overline{X_n}] = E\left[\frac{X_1 + X_2 + \dots + X_n}{n}\right] = \frac{1}{n} E[X_1 + X_2 + \dots + X_n]$

$= \frac{1}{n} (\underbrace{E[X_1]}_{\mu} + \underbrace{E[X_2]}_{\mu} + \dots + \underbrace{E[X_n]}_{\mu}) = \frac{1}{n} \cdot n \, E[X_1] = \mu$

b) $Var(\overline{X_n}) = Var\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \frac{1}{n^2} Var(X_1 + X_2 + \dots + X_n)$

$= \frac{1}{n^2} [Var(X_1) + Var(X_2) + \dots + \underbrace{Var(X_n)}_{\sigma^2}] = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}$

c) $\mathbb{E}[V_n^2] = ?$

$$V_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n}(\underbrace{X_i - \mu + \mu - \overline{X}_n}_{(X_i-\mu)-(\overline{X}_n-\mu)})^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[(X_i-\mu)^2 - 2(X_i-\mu)(\overline{X}_n-\mu) + (\overline{X}_n-\mu)^2\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{2}{n}\sum_{i=1}^{n}(X_i-\mu)(\overline{X}_n-\mu) + \frac{1}{n}\sum_{i=1}^{n}(\overline{X}_n-\mu)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{2}{n}(\overline{X}_n-\mu)\sum_{i=1}^{n}(X_i-\mu) + (\overline{X}_n-\mu)^2$$

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu) = \overline{X}_n - \mu$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{\cancel{n}}{\cancel{n}}2(\overline{X}_n-\mu)^2 + (\overline{X}_n-\mu)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - (\overline{X}_n-\mu)^2$$

$$\mathbb{E}[X_n^2] = \mathbb{E}\left[1/n\sum_{i=1}^{n}(X_i-\mu)^2 - (\overline{X}_n-\mu)^2\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\underbrace{(X_i-\mu)^2}_{Var(X_i)}] - \mathbb{E}[\underbrace{(\overline{X}_n-\mu)^2}_{Var(\overline{X}_n)}]$$

$$= \sigma^2 - \sigma^2/n = \frac{n-1}{n}\sigma$$

În mod similar: $\mathbb{E}[S_n^2] = \mathbb{E}\left[\frac{n}{n-1}V_n^2\right] = \sigma^2$

Obs: $\mathbb{E}(\ ) \longrightarrow \frac{1}{n}\sum_{i=1}^{n}(\ )$ (LNM)

din prob. X $\longrightarrow$ în statistică x_i

Obs: $\boxed{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}_n^2}$

$Var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

Def: Considerăm un eşantion de volum n de perechi de observaţii $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Se numeşte covarianţă empirică:

$\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n)$ ← cov. empirică   ($\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]$) ← cov. teoretică

Coeficientul de corelaţie liniară empiric

$$\frac{\sum_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2\sum_{i=1}^{n}(y_i - \overline{y}_n)^2}} \qquad \left(\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}\right)$$

Cazul în care populaţia este normală:

$X_1, X_2, \dots, X_n \sim N(\mu, \sigma^2)$

(P) Fie $X_1, X_2, \dots, X_n$ un eșantion de volum n dintr-o populație normală $N(\mu, \sigma^2)$ și fie $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ și $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$!

Atunci:

a) $\overline{X}_n \sim N(\mu, \sigma^2/n)$

b) $\overline{X}_n$ și S_n^2 sunt independente

c) $\frac{(n-1)S_n^2}{\sigma^2}$ este repartizată hi pătrat (χ^2) cu n-1 grade de libertate

Repartiții derivate din repartiția normală

1) Repartiția χ^2 (hi-pătrat)

Def: Dacă $Z_1, Z_2, \dots, Z_\nu$ sunt v.a. i.i.d. de rep. comună $N(0,1)$, atunci v.a. $Z = Z_1^2 + Z_2^2 + \dots + Z_\nu^2$ este repartizată χ^2 cu ν grade de libertate și se notează $\boxed{Z \sim \chi^2(\nu)}$

Obs: Dacă $Z \sim \chi^2(\nu)$, atunci $E[Z] = \nu$, $Var(Z) = 2\nu$

Obs: Dacă $Z_1 \sim \chi^2(\nu_1)$ și $Z_2 \sim \chi^2(\nu_2)$, iar $Z_1 \perp\!\!\!\perp Z_2$, atunci $Z_1 + Z_2 \sim \chi^2(\nu_1 + \nu_2)$

Obs: Cum $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1)$ și $Var(\chi^2(n-1)) = 2(n-1)$ găsim că

$Var\left(\frac{n-1}{\sigma^2} S_n^2\right) = 2(n-1) \Rightarrow \boxed{Var(S_n^2) = \frac{2\sigma^4}{n-1}}$

2) Repartiția Student

Def: Fie $Z \sim N(0,1)$ și $Q \sim \chi^2(\nu)$ astfel încât Z și Q sunt independente. Atunci v.a. $T = \frac{Z}{\sqrt{\frac{Q}{\nu}}}$ este rep. Student cu ν grade de libertate și se ~~notează~~ notează $\boxed{T \sim t(\nu)}$

(P) Aplicația fundamentală a rep. Student.

Fie $X_1, X_2, \dots, X_n$ sunt eșantionul de volum n dintr-~~un~~ o populație normală $N(\mu, \sigma^2)$. Atunci: $\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim t(n-1)$

Obs: $X_1, X_2, \dots, X_n \sim N(\mu, \sigma^2) \Rightarrow \overline{X}_n \sim N(\mu, \sigma^2/n)$

$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

μ, σ^2 - necunoscute

Scanned with CamScanner

$$\frac{\bar{X}_n-\mu}{S_n/\sqrt{n}} = \frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}\cdot \sigma/S_n = \frac{\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{S_n^2}{\sigma^2}}} = \frac{\overbrace{\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}}^{\sim N(0,1)}}{\sqrt{\frac{\underbrace{\frac{(n-1)S_n^2}{\sigma^2}}_{\sim \chi^2(n-1)}}{n-1}}} \sim t(n-1)$$

3) Repartiția lui Fisher - Snedecor

Def: Fie U și V două v.a. indep. a.î. $U\sim\chi^2(\nu_1)$ și $V\sim\chi^2(\nu_2)$. Atunci:

$F=\dfrac{U/\nu_1}{V/\nu_2}$ este rep. Fisher - Snedecor cu ν_1 grade de libertate la numărător și ν_2 grad de libertate la numitor

Se notează: $F\sim F(\nu_1,\nu_2)$

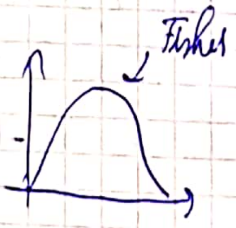
Obs: Dacă $F\sim F(\nu_1,\nu_2)$ at. $1/F\sim F(\nu_2,\nu_1)$

Obs: Apare atunci când studiem raportul a două sume de pătrate de v.a. normale independente.

Fie $X_1,X_2,\dots,X_{n_1}\sim N(\mu_1,\sigma_1^2)$ și $Y_1,Y_2,\dots,Y_{n_2}\sim N(\mu_2,\sigma_2^2)$

Știm $\dfrac{(n_1-1)S_{n_1}^2}{\sigma_1^2}\sim\chi^2(n_1-1)$

$\dfrac{(n_2-1)S_{n_2}^2}{\sigma_2^2}\sim\chi^2(n_2-1)$



Atunci: $$\frac{\frac{(n_1-1)S_{n_1}^2}{\sigma_1^2}\Big/ (n_1-1)}{\frac{(n_2-1)S_{n_2}^2}{\sigma_2^2}\Big/ (n_2-1)} = \frac{\frac{\chi^2(n_1-1)}{n_1-1}}{\frac{\chi^2(n_2-1)}{n_2-1}}\sim F(n_1-1,n_2-1)$$

, deci: $\boxed{\dfrac{S_{n_1}^2}{S_{n_2}^2}\cdot\dfrac{\sigma_2^2}{\sigma_1^2}\sim F(n_1-1,n_2-1)}$

• Estimarea punctuală:

Fie $X_1,X_2,\dots,X_n\sim f_\theta,\ \theta\in\underbrace{\Theta}_{\text{spațiul parametrilor}}\subseteq\mathbb{R}^2$

Problema estimării punctuale constă în găsirea unei funcții care depinde de obs. (eșantion) și care să aproximeze (estimeze) cât mai „bine" parametrul necunoscut

Def: Fie $X_1,X_2,\dots,X_n\sim f_\theta,\ \theta\in\Theta\subseteq\mathbb{R}^2$. S.n. eșantion punctual al lui θ o statistică $\hat{\theta}_n=\hat{\theta}_n(X_1,X_2,\dots,X_n)$ care ia valori în Θ

Obs: În general vom considera că Θ este un interval din $\mathbb{R}$ cu excepția cazului populației normale în care $\Theta \in \mathbb{R}^2$

Obs: Cum repartiția eșantionului $X_1, X_2, \ldots, X_n$ depinde de θ și rep. estimatorului $\hat{\theta}_n$ va depinde de θ.

Proprietățile ale estimatorilor

1) Nedeplasare (în medie este egal cu parametrul pe care îl estimează)

Def: S.n. deplasarea estimatorului $\hat{\theta}_n$ față de θ cantitatea

$b_\theta(\hat{\theta}_n) = E_\theta[\hat{\theta}_n] - \theta$ (b_θ (b_θ - vine de la bias)

Spunem că estimatorul $\hat{\theta}_n$ este nedeplasat dacă $b_\theta(\hat{\theta}_n) = 0, \forall \theta$

sau $E_\theta(\hat{\theta}_n] = \theta, \forall \theta$

Exp: $X_1, X_2, \ldots, X_n$ un eșantion dintr-o pop. de medie μ și dispersie σ^2

Fie $\overline{X}_n$ și S_n^2 media respectiv varianța eșantionului

Am văzut $\left.\begin{array}{l} E[\overline{X}_n] = \mu \\ E[S_n^2] = \sigma^2 \end{array}\right\} \Rightarrow \overline{X}_n$ și S_n^2 sunt estimatori nedeplasați (pt μ și σ^2)

Dacă luăm în loc de S_n^2, V_n^2, atunci:

$E[V_n^2] = \frac{n-1}{n}\sigma^2 \Rightarrow b_{\sigma^2}(V_n^2) = \frac{n-1}{n}\sigma^2 - \sigma^2 = -1/n\,\sigma^2$

Obs: Să presupunem că $\hat{\theta}_n = \begin{cases} \theta + 1000, & \text{prob } 1/2 \\ \theta - 1000, & \text{prob } 1/2 \end{cases}$

$P_\theta(\hat{\theta}_n = \theta - 1000) = P_\theta(\hat{\theta}_n = \theta + 1000) = 1/2$

$E_\theta[\hat{\theta}_n] = \theta, \forall \theta \Rightarrow \hat{\theta}_n$ este nedeplasat

Obs: În general, dacă $\hat{\theta}_n$ este un estimator nedeplasat pt. θ, iar g este o funcție oarecare at. $g(\hat{\theta}_n)$ ~~nu~~ este un estimator nedeplasat pt $g(\theta)$

Exp: Fie $X \sim Pois(\theta)$ și $\eta = e^{-2\theta}$

Dacă $\hat{\theta}_1 = X$ at. $\hat{\theta}_1$ este nedeplasat pentru θ; dar ce se întâmplă pentru $\hat{\eta} = e^{-2X}$?

Dar dacă considerăm $\hat{\eta}_2 = (-1)^X$?

2) Consistență (pt ca estimatorul să fie mai aproape de parametrul pe care îl estimează)

Def: Spunem că estimatorul $\hat{\theta}_n$ este consistent pt θ dacă $\hat{\theta}_n \xrightarrow{P} \theta$

Reamintim: Fie $(X_n)_n$ și X o v.a., atunci $X_n \xrightarrow{P} X$, dacă $\forall \varepsilon > 0$

$$\lim_n P(|X_n - X| > \varepsilon) = 0 \quad (LNM)$$

Obs: Intuitiv avem că repartiția lui $\hat{\theta}_n$ se concentrează în jurul lui θ

Exp: Media eșantionului este un ~~estimator~~ estimator consistent pt μ.

$\overline{X_n} \xrightarrow{P} \mu$ (deci! - din LNM)

Obs! Varianța eșantionului este un estimator consistent pentru σ^2 — varianța populației

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X_n})^2 = \frac{n}{n-1} V_n^2$$

$$V_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X_n})^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \overline{X_n}^2$$

Din LNM: $\left.\begin{array}{l} \frac{1}{n} \sum_{i=1}^{n} X_n^2 \xrightarrow{P} E[X_1^2] \\ \overline{X_n} \xrightarrow{P} E[X_1] \end{array}\right\} \Rightarrow V_n^2 \xrightarrow{P} \underbrace{E[X_1^2] - E[X_1]^2}_{\sigma^2}$

$\frac{n}{n-1} \to 1$ pt $n \to \infty$

Altfel $S_n^2 = \frac{n}{n-1} \cdot V_n^2 \xrightarrow{P} \sigma^2$

Exp: ~~$S_n^2 = \frac{n}{n-1} V_n^2 \xrightarrow{P}$~~ Fie $X_1, X_2, \ldots, X_n \sim N(0,1)$

$$\hat{\theta}_n = \begin{cases} 17, & n < 10^{100} \\ \overline{X_n}, & \text{altfel} \end{cases}$$

$a_n = P_\theta(|\hat{\theta}_n - \theta| > \varepsilon)$

$b_n = P_\theta(|\overline{X_n} - \theta| > \varepsilon) \to 0$ pt $n \to \infty$ (LNM) și în plus $a_n = b_n$ ~~$\forall n$~~ $\forall n \geq 10^{100} = N$

$\Rightarrow a_n \to 0 \Rightarrow \hat{\theta}_n \xrightarrow{P} \theta$ (este consistent)