Tema 5

Soluții

Exercițiul 1

Considerăm T_1 și T_2, doi estimatori nedeplasați ai parametrului θ de variențe V_1 și respectiv V_2. Fie T_3 estimatorul

$$T_3 = \alpha T_1 + (1-\alpha) T_2.$$

a) Arătați că estimatorul T_3 este nedeplasat.

b) Determinați constanta α pentru care estimatorul T_3 are varianța minimă.

c) Presupunând că ipotezele teoremei Rao-Cramer sunt verificate, este posibil ca ambii estimatori T_1 și T_2 să fie eficienți?

a) Pentru a arăta că T_3 este un estimator nedeplasat este suficient să verificăm că $\mathbb{E}[T_3] = \theta$. Observăm că

$$\mathbb{E}[T_3] = \mathbb{E}[\alpha T_1 + (1-\alpha)T_2] = \alpha\mathbb{E}[T_1] + (1-\alpha)\mathbb{E}[T_2] = \alpha\theta + (1-\alpha)\theta = \theta.$$

b) Dacă notăm cu $V_3 = V_3(\alpha) = Var(T_3)$, atunci

$$V_3(\alpha) = Var\left(\alpha T_1 + (1-\alpha)T_2\right) = \alpha^2 Var(T_1) + (1-\alpha)^2 Var(T_2) = \alpha^2 V_1 + (1-\alpha)^2 V_2.$$

Pentru a determina minimul funcției $V_3(\alpha)$ rezolvom ecuația $\frac{dV_3}{d\alpha} = 0$ de unde găsim că $2\alpha V_1 - 2(1-\alpha)V_2 = 0$, prin urmare $\alpha = \frac{V_2}{V_1+V_2}$ este punct critic. Cum $\frac{d^2 V_3}{d\alpha^2} = 2(V_1+V_2) > 0$, deci $V_3(\alpha)$ este convexă, deducem că valoarea minimă se atinge pentru $\alpha = \frac{V_2}{V_1+V_2}$ și aceasta este $V_3 = \frac{V_1 V_2}{V_1+V_2}$.

c) Să presupunem prin reducere la absurd că ambii estimatori T_1 și T_2 sunt eficienți, ceea ce implică atingerea bornei Rao-Cramer, i.e.

$$V_1 = V_2 = \frac{1}{nI_1(\theta)}.$$

Cum din punctul b) am găsit că estimatorul T_3, de dispersie minimă are dispersia $V_3 = \frac{V_1 V_2}{V_1+V_2}$, deducem că

$$V_3 = \frac{V_1 V_2}{V_1+V_2} = \frac{V_1^2}{2V_1} = \frac{V_1}{2} < \frac{1}{nI_1(\theta)},$$

ceea ce contrazice teorema Rao-Cramer și prin urmare T_1 și T_2 nu pot fi simultan eficienți.

Exercițiul 2

Fie $X_1, X_2, \dots, X_n$ un eșantion de talie n dintr-o populație $f_\theta(x)$ dată de:

a) $f_\theta(x) = e^{-\theta}\frac{\theta^x}{x!}$, $x = 0, 1, 2\dots$, $\theta > 0$

b) $f_\theta(x) = \frac{1}{\sqrt{2\pi\theta}}e^{-\frac{x^2}{2\theta}}$, $x \in \mathbb{R}$, $\theta > 0$

c) $f_\theta(x) = \frac{1}{\theta^\alpha \Gamma(\alpha)}x^{\alpha-1}e^{-\frac{x}{\theta}}$, $x > 0$, $\theta > 0$ iar $\alpha > 0$ cunoscut

Pentru fiecare caz în parte determinați un estimator pentru θ și studiați calitățile acestuia (deplasare, consistență, eficiență).

a) Folosind metoda verosimilității maxime avem că funcția de verosimilitate este

$$L(\theta|x_1,\dots,x_n) = \prod_{i=1}^{n} f_\theta(x_i) = e^{n\theta}\frac{\theta^{\sum_{i=1}^n x_i}}{x_1!x_2!\cdots x_n!}$$

de unde, pentru a găsi maximul, rezolvăm ecuația $\frac{\partial \log(L)}{\partial\theta} = -n + \frac{\sum_{i=1}^n x_i}{\theta} = 0$ care conduce la $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}_n$. Cum $\frac{\partial^2 \log(L)}{\partial\theta^2} = -\frac{\sum_{i=1}^n x_i}{\theta^2} < 0$, $\forall\theta > 0$ concluzionăm că $\hat{\theta}_n = \bar{X}_n$ este estimatorul de verosimilitate maximă.

Cum pentru o variabilă aleatoare $X \sim Pois(\theta)$ avem că $\mathbb{E}[X] = Var(X) = \theta$, deducem că $\mathbb{E}[\hat{\theta}_n] = \theta$, deci $\hat{\theta}_n$ este un estimator nedeplasat, și $Var(\hat{\theta}_n) = \frac{\theta}{n}$.

Observăm că $MSE_\theta(\hat{\theta}_n) = \frac{\theta}{n} \overset{n\to\infty}{\longrightarrow} 0$ din relația $MSE_\theta(\hat{\theta}_n) = Var(\hat{\theta}_n) + b_\theta(\hat{\theta}_n)^2$, de unde găsim că $\hat{\theta}_n$ este un estimator consistent pentru θ.

Pentru a verifica dacă este sau nu eficient trebuie să calculăm Informația lui Fisher care conduce la:

$$I_n(\theta) = nI_1(\theta) = -n\mathbb{E}\left[\frac{\partial^2 \log f_\theta(X)}{\partial\theta^2}\right] = -n\mathbb{E}\left[-\frac{X}{\theta^2}\right] = \frac{n}{\theta} = \frac{1}{Var(\hat{\theta}_n)}.$$

Deoarece dispersia estimatorului $\hat{\theta}_n$ este egală cu marginea din inegalitatea Rao-Cramer deducem că estimatorul este eficient.

b) Oberservăm că densitatea din problemă corespunde cu a unei normale $\mathcal{N}(0, \sqrt{\theta})$ iar funcția de verosimilitate este

$$L(\theta|x_1,\dots,x_n) = \prod_{i=1}^{n} f_\theta(x_i) = \left(\frac{1}{\sqrt{2\pi\theta}}\right)^n e^{-\frac{1}{2\theta}\sum_{i=1}^n x_i^2}.$$

Rezolvând ecuația $\frac{\partial \log(L)}{\partial\theta} = 0$ găsim că $-\frac{n}{2\theta} + \frac{1}{2\theta^2}\sum_{i=1}^n x_i^2$ de unde $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i^2$ (momentul empiric de ordin doi). Proprietățile acestui estimator au fost văzute la curs.

c) Observăm că densitatea de repartiție $f_\theta(x) = \frac{1}{\theta^\alpha\Gamma(\alpha)}x^{\alpha-1}e^{-\frac{x}{\theta}}$ corespunde unei repartiții Gamma. Prin metoda verosimilității maxime avem că funcția de verosimilitate este

$$L(\theta|x_1,\dots,x_n) = \prod_{i=1}^{n} f_\theta(x_i) = \frac{1}{\theta^{n\alpha}\Gamma(\alpha)^n}\left(\prod_{i=1}^n x_i\right)^{\alpha-1} e^{-\frac{1}{\theta}\sum_{i=1}^n x_i}$$

de unde logaritmul ei este

$$l(\theta|x_1,\dots,x_n) = \log L(\theta|x_1,\dots,x_n) = -n\alpha\log\theta - n\log\Gamma(\alpha) + (\alpha-1)\sum_{i=1}^n \log x_i - \frac{1}{\theta}\sum_{i=1}^n x_i$$

și rezolvând ecuația $\frac{\partial \log(L)}{\partial \theta} = 0$ găsim că $-\frac{n\alpha}{\theta} + \frac{\sum_{i=1}^{n} x_i}{\theta^2} = 0$ de unde $\hat{\theta}_n = \frac{\bar{X}_n}{\alpha}$ iar din faptul că funcția $\frac{\partial \log(L)}{\partial \theta}$ este pozitivă la stânga soluției și negativă la dreapta concluzionăm că $\hat{\theta}_n$ este estimatorul de verosimilitate maximă.

Pentru a determina calitățile acestui estimator vom folosi proprietățile repartiției Gamma și anume că $\mathbb{E}[X] = \alpha\theta$ iar $Var(X) = \alpha\theta^2$. Prin urmare avem că

$$\mathbb{E}[\hat{\theta}_n] = \theta \quad \text{și} \quad Var(\hat{\theta}_n) = \frac{\theta^2}{n\alpha}$$

ceea ce arată că estimatorul $\hat{\theta}_n$ este nedeplasat și consistent.

Informația lui Fisher pentru eșantionul $X_1, X_2, \dots, X_n$ de talie n este $I_n(\theta) = nI_1(\theta)$ și cum

$$I_1(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial \log f_\theta(X)}{\partial \theta}\right)^2\right] = -\mathbb{E}_\theta\left[\frac{\partial^2 \log f_\theta(X)}{\partial \theta^2}\right]$$

iar $\frac{\partial^2 \log f_\theta(X)}{\partial \theta^2} = \frac{\alpha}{\theta^2} - \frac{2X}{\theta^3}$ găsim că

$$I_1(\theta) = \frac{\alpha}{\theta^2}.$$

Prin urmare $I_n(\theta) = \frac{n\alpha}{\theta^2}$ de unde deducem că estimatorul $\hat{\theta}_n$ este eficient.

Exercițiul 3

Dintr-un total de 100 de persoane chestionate, 51 au declarat că vor vota cu candidatul Bugs Bunny la următoarele alegeri parlamentare. Dați un interval de încredere de nivel 95% pentru proporția p, de intenții de vot pentru acest candidat în populație. Aceeași întrebare dacă sondajul ar fi avut loc pentru un eșantion de 1000 de persoane. Câți electori ar trebui întrebați pentru a avea o precizie de cel puțin 2%?

Pentru a determina un interval de încredere de nivel de încredere $1-\alpha$ pentru proporția p de intenții de vot pentru candidatul Bugs Bunny la alegerile parlamentare să observăm că suntem în contextul unui interval de încredere pentru o proporție. Fie $X_1, \dots, X_n$ variabilele aleatoare care descriu intenția de vot a celor $n = 100$ de canditați eșantionați, cu $X_i \sim \mathcal{B}(p)$.

Am văzut că un interval de încredere, de tip Wald, pentru p este dat de

$$IC_1^{1-\alpha}(p) = \left[\hat{p}_n - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}, \hat{p}_n + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right]$$

unde $\hat{p}_n = \bar{X}_n$ este estimatorul de verosimilitate maximă a lui p iar $z_{1-\frac{\alpha}{2}}$ este cuantila de ordin $1-\frac{\alpha}{2}$ a repartiției normale standard. Înlocuind cu valorile din ipoteza problemei, $n = 100$, $\alpha = 0.05$ și respectiv $\sum_{i=1}^{n} X_i = 51$, găsim că

$$IC_1^{0.95}(p) = [0.412, 0.607].$$

În cazul în care am fi avut un eșantion de talie $n = 1000$ și 510 alegători (pentru a păstra proporția $\hat{p}_{1000} = 0.51$) ar fi declarat că îl preferă pe candidatul Bugs Bunny atunci intervalul ar fi devenit

$$IC_1^{0.95}(p) = [0.479, 0.540].$$

Am văzut la curs că un alt interval de încredere pentru p, de nivel de încredere $1-\alpha$, se obține rezolvând după p inecuația din interiorul probabilității

$$\mathbb{P}\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\hat{p}_n - p}{\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}} \leq z_{1-\frac{\alpha}{2}}\right) \approx 1-\alpha$$

care conduce la

$$IC_2^{1-\alpha}(p) = \left[\frac{\hat{p}_n + \frac{z^2_{1-\frac{\alpha}{2}}}{2n}}{1+\frac{z^2_{1-\frac{\alpha}{2}}}{n}} - \frac{1}{1+\frac{z^2_{1-\frac{\alpha}{2}}}{n}}\sqrt{\frac{z^2_{1-\frac{\alpha}{2}}}{n}\hat{p}_n(1-\hat{p}_n) + \frac{z^4_{1-\frac{\alpha}{2}}}{4n^2}}, \frac{\hat{p}_n + \frac{z^2_{1-\frac{\alpha}{2}}}{2n}}{1+\frac{z^2_{1-\frac{\alpha}{2}}}{n}} + \frac{1}{1+\frac{z^2_{1-\frac{\alpha}{2}}}{n}}\sqrt{\frac{z^2_{1-\frac{\alpha}{2}}}{n}\hat{p}_n(1-\hat{p}_n) + \frac{z^4_{1-\frac{\alpha}{2}}}{4n^2}}\right]$$

iar numeric pentru $n = 100$

$$IC_2^{0.95}(p) = [0.413, 0.605]$$

iar pentru $n = 1000$

$$IC_2^{0.95}(p) = [0.479, 0.540].$$

Exercițiul 4

Un producător de becuri anunță că durata medie a becurilor pe care le produce este de 170 de ore. Pentru a verifica această afirmație, un corp de control al protecției consumatorilor extrage aleator un eșantion de 100 de becuri dintr-un lot de fabricație și, după experimentare, constată că eșantionul are o durată medie de viață de 158 de ore cu o abatere standard de 30 de ore. Dacă presupunem că durata de viață a becurilor urmează o lege normală, putem deduce din această investigație că afirmația producătorului este falsă?

Pentru a construi un interval de încredere de nivel de încredere $1-\alpha$ pentru media μ a unei populații normale de abatere standard necunoscută σ vom folosi statistica $T_n = \sqrt{n}\frac{\bar{X}_n - \mu}{S_n}$ care știm că este repartizată t-Student cu $n-1$ grade de libertate.

Un interval de încredere bilateral pentru μ este

$$IC_1^{1-\alpha}(\mu) = \left[\bar{x}_n - t^{n-1}_{1-\frac{\alpha}{2}}\frac{s_n}{\sqrt{n}}, \bar{x}_n + t^{n-1}_{1-\frac{\alpha}{2}}\frac{s_n}{\sqrt{n}}\right]$$

care pentru $n = 100$, $\alpha = 0,01$, $\bar{x}_n = 158$ și $s_n = 30$ devine

$$IC_1^{0.99}(\mu) = \left[158 - 2.626\frac{30}{\sqrt{100}}, 158 + 2.626\frac{30}{\sqrt{100}}\right] = [150.12, 165.87]$$

și cum valoarea anunțată de producător este în afara acestui interval cocluzionăm că reclama este falsă.

Dacă ne interesăm la un interval de încredere de nivel de încredere $1-\alpha$ unilateral spre stânga atunci am avea că

$$\mathbb{P}\left(T_n > t\right) = \mathbb{P}\left(\sqrt{n}\frac{\bar{X}_n - \mu}{S_n} > t\right) = 1-\alpha$$

de unde

$$IC_2^{1-\alpha}(\mu) = \left[0, \bar{x}_n - t_\alpha^{n-1}\frac{s_n}{\sqrt{n}}\right] = \left[0, 158 + 2,364\frac{30}{\sqrt{100}}\right] = [0, 165, 09]$$

prin urmare afirmația producătorului este falsă.

Exercițiul 5

Pentru a estima precizia unui termometru, s-au realizat $n = 100$ de măsurători independente a temperaturii dintr-un lichid menținut la temperatura constantă de 20 de grade Celsius. Observațiile $x_1, x_2, \dots, x_{100}$ au condus la valoarea $\sum_{i=1}^{100} x_i^2 = 40011$. Construiți un interval de încredere de nivel de încredere de 95% pentru precizia termometrului, măsurată prin varianța σ^2 a măsurătorilor.

Cum măsurătorile pot fi presupuse ca un eșantion de talie n, $X_1, X_2, \dots, X_n$ dintr-o populație $\mathcal{N}(\mu = 20, \sigma^2)$ considerăm ca estimator al varianței estimatorul

$$\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \mu^2$$

iar din $\frac{n\hat{\sigma}_0^2}{\sigma^2} \sim \chi^2(n)$ deducem că un interval de încredere bilateral pentru σ^2 de nivel de încredere $1-\alpha$ este

$$IC_1^{1-\alpha}(\sigma^2) = \left[\frac{n\hat{\sigma}_0^2}{\chi^2_{n,1-\frac{\alpha}{2}}}, \frac{n\hat{\sigma}_0^2}{\chi^2_{n,\frac{\alpha}{2}}}\right].$$

Înlocuind cu valorile din ipoteză, $n = 100$, $\alpha = 0.05$, $\mu = 20$ și $\sum_{i=1}^{100} x_i^2 = 40011$ găsim că

$$IC_1^{0.95}(\sigma^2) = \left[\frac{100\hat{\sigma}_0^2}{\chi^2_{100,0.975}}, \frac{100\hat{\sigma}_0^2}{\chi^2_{100,0.025}}\right] = \left[\frac{100\cdot 0.11}{129.56}, \frac{100\cdot 0.11}{74.22}\right] = [0.0849, 0.1482].$$

Dacă ne uităm după un interval de încredere unilateral, de nivel de încredere $1-\alpha$, la stânga atunci

$$IC_2^{1-\alpha}(\sigma^2) = \left[0, \frac{n\hat{\sigma}_0^2}{\chi^2_{n,\alpha}}\right] = \left[0, \frac{100\cdot 0.11}{77.929}\right] = [0, 0.1411].$$

Exercițiul 6

Numărul de blocaje de trafic mai mari de un minut de pe linia tramvaiului 41, pe parcursul unei zile, se presupune că urmează o repartiție Poisson de medie necunoscută și ne propunem să estimăm acest parametru plecând de la un eșantion de talie 200 (s-au urmărit blocajele pe parcursul a 200 de zile). Momentele empirice calculate pe acest eșantion au condus la $\bar{x}_{200} = 3$ și $s_{200}^2 = 3.2$. Determinați un interval de încredere de nivel de încredere de 95% pentru media numărului de blocaje.

Fie $X_1, X_2, \dots, X_n$ eșantionul de talie $n = 200$ din populația $Pois(\theta)$ care descriu numărul de blocaje de trafic mai mari de un minut pe linia tramvaiului 41. Pentru a găsi un interval de încredere pentru θ să ne aducem aminte că estimatorul de verosimilitate maximă pentru θ este $\bar{X}_n$ și aplicând Teorema Limită Centrală avem

$$\sqrt{n}\frac{\bar{X}_n - \theta}{\sqrt{\theta}} \xrightarrow{d} \mathcal{N}(0,1).$$

Un interval de încredere de tip Wald pentru θ, de nivel de încredere $1-\alpha$, este obținut prin înlocuirea la numitor a lui θ cu estimatorul său $\bar{X}_n$, deci

$$\mathbb{P}\left(-z_{1-\frac{\alpha}{2}} \leq \sqrt{n}\frac{\bar{X}_n - \theta}{\sqrt{\bar{X}_n}} \leq z_{1-\frac{\alpha}{2}}\right) \approx 1-\alpha$$

de unde

$$IC_{Wald}^{1-\alpha}(\theta) = \left[\bar{X}_n - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}_n}{n}}, \bar{X}_n + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}_n}{n}}\right].$$

Înlocuind cu valorile din ipoteză, $n = 200$, $\alpha = 0.05$ și $\bar{x}_{200} = 3$ obținem $IC_{Wald}^{0.95}(\theta) = [2.759, 3.240]$.

Un alt interval de încredere pentru θ, de nivel de încredere $1-\alpha$, se poate obține rezolvând inecuația din probabilitate

$$\mathbb{P}\left(-z_{1-\frac{\alpha}{2}} \leq \sqrt{n}\frac{\bar{X}_n - \theta}{\sqrt{\theta}} \leq z_{1-\frac{\alpha}{2}}\right) \approx 1-\alpha$$

după θ.

În acest caz găsim că

$$IC_2^{1-\alpha}(\theta) = \left[\bar{X}_n + \frac{z_{1-\frac{\alpha}{2}}^2}{2n} - \sqrt{\frac{\bar{X}_n z_{1-\frac{\alpha}{2}}^2}{n} + \frac{z_{1-\frac{\alpha}{2}}^4}{4n^2}}, \bar{X}_n + \frac{z_{1-\frac{\alpha}{2}}^2}{2n} + \sqrt{\frac{\bar{X}_n z_{1-\frac{\alpha}{2}}^2}{n} + \frac{z_{1-\frac{\alpha}{2}}^4}{4n^2}}\right]$$

și înlocuind cu valorile din ipoteză rezultă $IC_2^{0.95}(\theta) = [2.769, 3.249]$.