

Seminarul 7

1. i) Estimați parametrul necunoscut $p \in (0, 1)$ pentru distribuția binomială a unei caracteristici cercetate: $X \sim \text{Bino}(N, p)$, unde $N \in \mathbb{N}^*$ este cunoscut, cu metoda momentelor, respectiv metoda verosimilității maxime. Sunt estimatorii obținuți nedeplasați, respectiv consistenți?

ii) Într-o urnă sunt bile albe și negre. Proporția de bile albe $p \in (0, 1)$ este necunoscută. În urma a $n = 6$ serii a câte $N = 5$ extrageri cu returnarea bilei extrase în urnă s-au obținut: 3, 4, 2, 0, 2, respectiv 1, bile albe. Estimați valoarea lui p cu metoda momentelor, respectiv metoda verosimilității maxime.

R: i) Fie X_1, \dots, X_n variabile de selecție și x_1, \dots, x_n date statistice pentru X .

Metoda momentelor: $E(X) = Np = \frac{1}{n} \sum_{i=1}^n X_i \implies \hat{p}(X_1, \dots, X_n) = \frac{1}{nN} \sum_{i=1}^n X_i$ estimator pentru parametrul necunoscut p .

Metoda verosimilității maxime: $P(X = x) = C_N^x p^x (1-p)^{N-x}, x \in \{0, 1, \dots, N\} \implies L(x_1, \dots, x_n; p) = \prod_{i=1}^n P(X = x_i) = \prod_{i=1}^n C_N^{x_i} p^{x_i} (1-p)^{N-x_i} = \prod_{i=1}^n C_N^{x_i} \cdot p^{\sum_{i=1}^n x_i} (1-p)^{nN - \sum_{i=1}^n x_i} \implies \ln L(x_1, \dots, x_n; p) = \sum_{i=1}^n \ln C_N^{x_i} + \sum_{i=1}^n x_i \ln p + (nN - \sum_{i=1}^n x_i) \ln(1-p) \implies \frac{\partial \ln L}{\partial p}(x_1, \dots, x_n; p) = \frac{1}{p} \sum_{i=1}^n x_i - \frac{1}{1-p} (nN - \sum_{i=1}^n x_i)$. Deci, $\frac{\partial \ln L}{\partial p}(x_1, \dots, x_n; p) = 0 \implies p = \frac{1}{nN} \sum_{i=1}^n x_i$; $\frac{\partial^2 \ln L}{\partial p^2}(x_1, \dots, x_n; p) = -\frac{1}{p^2} \sum_{i=1}^n x_i - \frac{1}{(1-p)^2} \sum_{i=1}^n (N - x_i) < 0 \implies L(x_1, \dots, x_n; \cdot)$ ia valoarea maximă pentru p găsit mai sus. Estimatorul pentru parametrul necunoscut p este $\hat{p}(X_1, \dots, X_n) = \frac{1}{nN} \sum_{i=1}^n X_i$.

Valoarea estimatorului, pentru ambele metode, este $\hat{p}(x_1, \dots, x_n) = \frac{1}{nN} \sum_{i=1}^n x_i$.

Deoarece $E(\hat{p}(X_1, \dots, X_n)) = \frac{1}{N} E(X) = \frac{Np}{N} = p$, estimatorul este nedeplasat. LTNM implică $\hat{p}(X_1, \dots, X_n) \xrightarrow{a.s.} \frac{1}{N} E(X) = p$ (unde considerăm șirul de variabile de selecție X_1, \dots, X_n, \dots) deci estimatorul este consistent.

ii) Valoarea estimatorului este $\hat{p}(3, 4, 2, 0, 2, 1) = \frac{12}{6 \cdot 5} = 40\%$.

2. i) O caracteristică cercetată X are funcția de densitate

$$f_X(x) = \begin{cases} \lambda^2 x e^{-\lambda x}, & x > 0, \\ 0, & x \leq 0, \end{cases}$$

unde $\lambda > 0$ este fixat. Estimați parametrul necunoscut λ cu metoda momentelor, respectiv metoda verosimilității maxime. Sunt estimatorii obținuți consistenți?

ii) Durata culorii roșii (în minute) X a unui anumit semafor are funcția de densitate f_X dată mai sus, cu parametrul $\lambda > 0$ necunoscut. Un taximetrist (curios din fire) a observat următoarele durate (în minute) ale culorii roșii pentru acest semafor: 1, $\frac{3}{2}$, 3, 2, 3, $\frac{5}{2}$, 1, 2. Aplicați metoda momentelor, respectiv metoda verosimilității maxime, pentru a estima valoarea lui λ , folosind datele furnizate de taximetrist.

R: i) Fie X_1, \dots, X_n variabile de selecție și x_1, \dots, x_n date statistice pentru X .

Metoda momentelor: $E(X) = \int_0^\infty \lambda^2 x^2 e^{-\lambda x} dx = \frac{2}{\lambda} = \frac{1}{n} \sum_{i=1}^n X_i \implies \hat{\lambda}(X_1, \dots, X_n) = \frac{2n}{\sum_{i=1}^n X_i}$.

Metoda verosimilității maxime: $f_X(x) = \lambda^2 x e^{-\lambda x}, x > 0 \implies L(x_1, \dots, x_n; \lambda) = \prod_{i=1}^n \lambda^2 x_i e^{-\lambda x_i} \implies \ln L(x_1, \dots, x_n; \lambda) = 2n \ln \lambda + \sum_{i=1}^n \ln x_i - \lambda \sum_{i=1}^n x_i$. $\frac{\partial \ln L}{\partial \lambda}(x_1, \dots, x_n; \lambda) = \frac{2n}{\lambda} - \sum_{i=1}^n x_i = 0 \implies \lambda = \frac{2n}{\sum_{i=1}^n x_i}$; $\frac{\partial^2 \ln L}{\partial \lambda^2}(x_1, \dots, x_n; \lambda) = -\frac{2n}{\lambda^2} < 0 \implies$ estimatorul este $\hat{\lambda}(X_1, \dots, X_n) = \frac{2n}{\sum_{i=1}^n X_i}$.

Valoarea estimatorului, pentru ambele metode, este $\hat{\lambda}(x_1, \dots, x_n) = \frac{2n}{\sum_{i=1}^n x_i}$.

LTNM $\implies \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{a.s.} E(X) = \frac{2}{\lambda}$ (unde considerăm șirul de variabile de selecție X_1, \dots, X_n, \dots) $\implies \hat{\lambda}(X_1, \dots, X_n) = \frac{2}{\frac{1}{n} \sum_{i=1}^n X_i} \xrightarrow{a.s.} \frac{2}{\frac{2}{\lambda}} = \lambda \implies$ estimatorul este consistent.

ii) Valoarea estimatorului este $\hat{\lambda}(1, \frac{3}{2}, 3, 2, 3, \frac{5}{2}, 1, 2) = 1$.

3. Considerăm următoarele date statistice pentru masa corporală a persoanelor dintr-o anumită populație: 71 kg; 68 kg; 77 kg; 69 kg; 65 kg. Presupunem că masa corporală este o caracteristică ce urmează distribuția normală. Determinați intervale de încredere bilaterale cu nivelul de încredere 95% pentru:

- a) media masei corporale, știind că varianța masei este 125.
- b) media masei corporale, dacă abaterea standard a masei este necunoscută.
- c) varianța masei corporale.

R: $n = 5$, $\bar{x}_n = \frac{71+68+77+69+65}{5} = 70$, $\alpha = 5\% = 0,05$.

a) $z_{1-\frac{\alpha}{2}} = \text{norminv}(0.975, 0, 1) \approx 1,96$, $\sigma = \sqrt{125} = 5\sqrt{5}$.

Valoarea intervalului de încredere este: $(\bar{x}_n - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x}_n + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}) = (70 - 1,96 \cdot 5, 70 + 1,96 \cdot 5) = (60,2, 79,8)$.

b) $t_{1-\frac{\alpha}{2}} = \text{tinv}(0.975, 4) \approx 2,78$, $s_n = \left(\frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x}_n)^2 \right)^{\frac{1}{2}} = \sqrt{\frac{1^2 + (-2)^2 + 7^2 + (-1)^2 + (-5)^2}{4}} = 2\sqrt{5}$.

Valoarea intervalului de încredere este: $(\bar{x}_n - t_{1-\frac{\alpha}{2}} \cdot \frac{s_n}{\sqrt{n}}, \bar{x}_n + t_{1-\frac{\alpha}{2}} \cdot \frac{s_n}{\sqrt{n}}) = (70 - 2,78 \cdot 2, 70 + 2,78 \cdot 2) = (64,44, 75,56)$.

c) $c_{\frac{\alpha}{2}} = \text{chi2inv}(0.025, 4) \approx 0,48$, $c_{1-\frac{\alpha}{2}} = \text{chi2inv}(0.975, 4) \approx 11,14$.

Valoarea intervalului de încredere este: $\left(\frac{(n-1)s_n^2}{c_{1-\frac{\alpha}{2}}}, \frac{(n-1)s_n^2}{c_{\frac{\alpha}{2}}} \right) = \left(\frac{80}{11,14}, \frac{80}{0,48} \right) \approx (7,18, 166,67)$.

4. Un provider de internet își asigură clienții că viteza conexiunii la internet este în medie mai mare sau egală decât 250 Mbps între orele 20:00 și 22:00. Pe de altă parte, providerul susține că în acest interval orar conexiunea nu este stabilă, având o abatere standard de 40 Mbps.

i) Știind că în urma unei selecții de 100 de clienți s-a constatat că valoarea mediei de selecție este 242 Mbps pentru viteza conexiunii între orele specificate, să se construiască un interval de încredere bilateral pentru media vitezei conexiunii, iar apoi să se testeze dacă media vitezei este cea pretinsă de provider, cu un nivel de semnificație de 5%.

ii) Care este valoarea maximă a mediei vitezei conexiunii între orele specificate pentru un eșantion de 100 de clienți, pentru a concluziona, cu ajutorul unui test statistic cu un nivel de semnificație de 4%, că media vitezei conexiunii nu este cea pretinsă de provider?

R: i) Construim un interval de încredere bilateral pentru medie, când varianța este cunoscută: $n = 100$, $\bar{x}_{100} = 242$, $\sigma = 40$, $\alpha = 0,05$, $z_{1-\frac{\alpha}{2}} = \text{norminv}(1 - \frac{\alpha}{2}, 0, 1) = \text{norminv}(0.975, 0, 1) = 1,96$. Valoarea intervalului de încredere este $(\bar{x}_{100} - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}, \bar{x}_{100} + \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}) = (242 - \frac{40}{10} \cdot 1,96, 242 + \frac{40}{10} \cdot 1,96) = (234,16, 249,84)$.

$H_0 : m \geq 250$, $H_1 : m < 250$. Aplicăm testul pentru medie, când varianța este cunoscută: $z_{\alpha} = \text{norminv}(\alpha, 0, 1) = \text{norminv}(0.05) = -1,64$, $z = \frac{\bar{x}_{100} - 250}{\frac{\sigma}{\sqrt{100}}} = \frac{242 - 250}{\frac{40}{10}} = -2$, $z < z_{\alpha} \implies$ se respinge H_0 , adică se acceptă că media vitezei conexiunii la internet nu este cea pretinsă de provider.

ii) $H_0 : m \geq 250$, $H_1 : m < 250$. Avem: $z_{\alpha} = \text{norminv}(\alpha, 0, 1) = \text{norminv}(0.04, 0, 1) = -1,75$, $z = \frac{\bar{x}_{100} - 250}{\frac{40}{10}}$. Aplicând testul pentru medie, când varianța este cunoscută, respingem ipoteza H_0 în favoarea lui H_1 (adică, acceptăm că media vitezei conexiunii nu este cea pretinsă de provider) dacă și numai dacă $z \leq z_{\alpha} \Leftrightarrow \frac{\bar{x}_{100} - 250}{\frac{40}{10}} \leq -1,75 \Leftrightarrow \bar{x}_{100} \leq 243$. Deci valoarea maximă cerută este 243.

5. Într-un sondaj de opinie, suntem interesați de proporția p a persoanelor dintr-un anumit oraș care ar vota candidatul A împotriva candidatului B .

a) Determinați un interval de încredere bilateral cu nivelul de încredere 95% pentru p , știind că 64

de persoane dintr-un eșantion de 100 de participanți la sondaj ar vota candidatul A.

b) Demonstrați că 1600 de participanți la sondaj sunt suficienți pentru a obține un interval de încredere bilateral cu nivelul de încredere 95% de lungime cel mult 5%.

R: a) $n = 100$, $\bar{x}_n = 0,64$, $z_{1-\frac{\alpha}{2}} = 1,96$. Valoarea intervalului de încredere este: $(\bar{x}_n - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}}, \bar{x}_n + z_{1+\frac{\alpha}{2}} \sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}}) = (0,64 - 1,96 \sqrt{\frac{0,64 \cdot 0,36}{100}}, 0,64 + 1,96 \sqrt{\frac{0,64 \cdot 0,36}{100}}) = (0,64 - 1,96 \frac{0,8 \cdot 0,6}{10}, 0,64 + 1,96 \frac{0,8 \cdot 0,6}{10}) = (0,54592, 0,73408)$.

b) Lungimea intervalului de încredere $(\bar{X}_n - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}, \bar{X}_n + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}})$ este

$$2z_{1-\frac{\alpha}{2}} \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} = 1,96 \cdot \frac{2\sqrt{\bar{X}_n(1-\bar{X}_n)}}{40} \leq 1,96 \cdot \frac{X_n + (1-\bar{X}_n)}{40} = \frac{1,96}{40} = 0,049 < 5\%,$$

unde am folosit inegalitatea mediilor: $\sqrt{ab} \leq \frac{a+b}{2}, \forall a, b \geq 0$.

6. O companie dorește înlocuirea unui sistem de frânare pentru un anumit tip de mașină cu unul nou, care să reducă semnificativ distanța de frânare. Media distanței de frânare pentru vechiul sistem este mai mare sau egală decât 50 m, pentru o viteză de 80 km/h pe ploaie.

i) În urma testării a 100 de mașini cu noul sistem de frânare instalat, pentru o viteză de 80 km/h pe ploaie, s-a constatat că valoarea mediei de selecție este 49 m și că valoarea abaterii standard de selecție este 1 m pentru distanța de frânare a acestui eșantion. Folosind intervale de încredere unilaterale, să se testeze dacă noul sistem este mai performant decât cel vechi, cu un nivel de semnificație de 1%.

ii) Știind că, în condițiile precizate, abaterea standard a distanței de frânare a noului sistem de frânare este de 2 m, care este valoarea maximă a mediei distanței de frânare pentru un eșantion de 100 de mașini testate cu noul sistem de frânare pentru a concluziona, cu ajutorul unui test statistic cu un nivel de semnificație de 6%, că noul sistem este mai performant?

R: i) Vom testa $H_0 : m \geq 50$, $H_1 : m < 50$, testul pentru medie, când varianța este necunoscută (Student test); folosim datele $n = 100$, $\bar{x}_{100} = 49$, $s_{100} = 1$, $\alpha = 0,01$, $m_0 = 50$, $t_\alpha = \text{tinv}(\alpha, n-1) = \text{tinv}(0,01, 99) = -2,36$. Valoarea intervalului de încredere corespunzător acestui test este $(-\infty, \bar{x}_{100} - \frac{s_{100}}{\sqrt{n}} \cdot t_\alpha) = (-\infty, 49 + \frac{1}{10} \cdot 2,36) = (-\infty, 49,236)$. În această problemă concretă de fapt intervalul de încredere este $(0, 49,236)$.

Cum $m_0 = 50$ nu aparține acestui interval numeric, se respinge H_0 , adică se acceptă că sistemul nou este mai performant.

ii) $H_0 : m \geq 50$, $H_1 : m < 50$. Avem: $\alpha = 0,06$, $\sigma = 2$, $z_\alpha = \text{norminv}(\alpha, 0, 1) = \text{norminv}(0,06, 0, 1) = -1,55$, $z = \frac{\bar{x}_{100} - 50}{\frac{2}{\sqrt{100}}}$. Aplicând testul pentru medie, când varianța este cunoscută, respingem ipoteza H_0 în favoarea lui H_1 (i.e. acceptăm că sistemul nou este mai performant) dacă și numai dacă $z \leq z_\alpha \Leftrightarrow \frac{\bar{x}_{100} - 50}{\frac{2}{\sqrt{100}}} \leq -1,55 \Leftrightarrow \bar{x}_{100} \leq 49,69$. Deci valoarea maximă cerută este 49,69.