Seminarul 7

1. i) Estimaţi parametrul necunoscut $p \in (0,1)$ pentru distribuţia binomială a unei caracteristici cercetate: $X \sim Bino(N,p)$, unde $N \in \mathbb{N}^*$ este cunoscut, cu metoda momentelor, respectiv metoda verosimilităţii maxime. Sunt estimatorii obţinuţi nedeplasaţi, respectiv consistenţi?
ii) Într-o urnă sunt bile albe şi negre. Proporţia 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 obţinut: 3, 4, 2, 0, 2, respectiv 1, bile albe. Estimaţi valoarea lui p cu metoda momentelor, respectiv metoda verosimilităţii maxime.
R: i) Fie $X_1, \ldots, X_n$ variabile de selecţie şi $x_1, \ldots, x_n$ date statistice pentru X.
Metoda momentelor: $E(X) = Np = \frac{1}{n}\sum_{i=1}^n X_i \implies \hat{p}(X_1, \ldots, X_n) = \frac{1}{nN}\sum_{i=1}^n X_i$ estimator pentru parametrul necunoscut p.
Metoda verosimilităţii maxime: $P(X = x) = C_N^x p^x (1-p)^{N-x}, x \in \{0, 1, \ldots, N\} \implies L(x_1, \ldots, x_n; p) = \prod_{i=1}^n P(X = x_i) = \prod_{i=1}^n C_N^{x_i} p^{x_i}(1-p)^{N-x_i} = \prod_{i=1}^n C_N^{x_i} \cdot p^{\sum_{i=1}^n x_i}(1-p)^{nN-\sum_{i=1}^n x_i} \implies \ln L(x_1, \ldots, x_n; p) = \sum_{i=1}^n \ln C_N^{x_i} + \sum_{i=1}^n x_i \ln p + (nN - \sum_{i=1}^n x_i)\ln(1-p) \implies \frac{\partial \ln L}{\partial p}(x_1, \ldots, x_n; p) = \frac{1}{p}\sum_{i=1}^n x_i - \frac{1}{1-p}(nN - \sum_{i=1}^n x_i)$. Deci, $\frac{\partial \ln L}{\partial p}(x_1, \ldots, x_n; p) = 0 \implies p = \frac{1}{Nn}\sum_{i=1}^n x_i$; $\frac{\partial^2 \ln L}{\partial p^2}(x_1, \ldots, x_n; p) = -\frac{1}{p^2}\sum_{i=1}^n x_i - \frac{1}{(1-p)^2}\sum_{i=1}^n (N - x_i) < 0 \implies L(x_1, \ldots, x_n; \cdot)$ ia valoarea maximă pentru p găsit mai sus. Estimatorul pentru parametrul necunoscut p este $\hat{p}(X_1, \ldots X_n) = \frac{1}{nN}\sum_{i=1}^n X_i$.
Valoarea estimatorului, pentru ambele metode, este $\hat{p}(x_1, \ldots, x_n) = \frac{1}{nN}\sum_{i=1}^n x_i$.
Deoarece $E(\hat{p}(X_1, \ldots X_n)) = \frac{1}{N}E(X) = \frac{N \cdot p}{N} = p$, estimatorul este nedeplasat. LTNM implică $\hat{p}(X_1, \ldots X_n) \overset{a.s.}{\to} \frac{1}{N}E(X) = p$ (unde considerăm şirul de variabile de selecţie $X_1, \ldots, X_n, \ldots$) 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 funcţia 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. Estimaţi parametrul necunoscut λ cu metoda momentelor, respectiv metoda verosimilităţii maxime. Sunt estimatorii obţinuţi consistenţi?
ii) Durata culorii roşii (în minute) X a unui anumit semafor are funcţia 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 roşii pentru acest semafor: $1, \frac{3}{2}, 3, 2, 3, \frac{5}{2}, 1, 2$. Aplicaţi metoda momentelor, respectiv metoda verosimilităţii maxime, pentru a estima valoarea lui λ, folosind datele furnizate de taximetrist.
R: i) Fie $X_1, \ldots, X_n$ variabile de selecţie şi $x_1, \ldots, x_n$ date statistice pentru X.
Metoda momentelor: $E(X) = \int_0^\infty \lambda^2 x^2 e^{-\lambda x} dx = \frac{2}{\lambda} = \frac{1}{n}\sum_{i=1}^n X_i \implies \hat{\lambda}(X_1, \ldots, X_n) = \frac{2n}{\sum_{i=1}^n X_i}$.
Metoda verosimilităţii maxime: $f_X(x) = \lambda^2 x e^{-\lambda x}, x > 0 \implies L(x_1, \ldots, x_n; \lambda) = \prod_{i=1}^n \lambda^2 x_i e^{-\lambda x_i} \implies \ln L(x_1, \ldots, x_n; \lambda) = 2n \ln \lambda + \sum_{i=1}^n \ln x_i - \lambda \sum_{i=1}^n x_i$. $\frac{\partial \ln L}{\partial \lambda}(x_1, \ldots, x_n; \lambda) = \frac{2n}{\lambda} - \sum_{i=1}^n x_i = 0 \implies \lambda = \frac{2n}{\sum_{i=1}^n x_i}$; $\frac{\partial^2 \ln L}{\partial \lambda^2}(x_1, \ldots, x_n; \lambda) = -\frac{2n}{\lambda^2} < 0 \implies$ estimatorul este $\hat{\lambda}(X_1, \ldots, X_n) = \frac{2n}{\sum_{i=1}^n X_i}$.
Valoarea estimatorului, pentru ambele metode, este $\hat{\lambda}(x_1, \ldots, x_n) = \frac{2n}{\sum_{i=1}^n x_i}$.
LTNM $\implies \frac{1}{n}\sum_{i=1}^n X_i \overset{a.s.}{\to} E(X) = \frac{2}{\lambda}$ (unde considerăm şirul de variabile de selecţie $X_1, \ldots, X_n, \ldots$) $\implies \hat{\lambda}(X_1, \ldots, X_n) = \frac{2}{\frac{1}{n}\sum_{i=1}^n X_i} \overset{a.s.}{\to} \frac{2}{\frac{2}{\lambda}} = \lambda \implies$ 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 populaţie: 71 kg; 68 kg; 77 kg; 69 kg; 65 kg. Presupunem că masa corporală este o caracteristică ce urmează distribuţia normală. Determinaţi intervale de încredere bilaterale cu nivelul de încredere 95% pentru:
a) media masei corporale, ştiind că varianţa masei este 125.
b) media masei corporale, dacă abaterea standard a masei este necunoscută.
c) varianţa masei corporale.
Folosiţi tabelul următor:

$norminv(0.025, 0, 1)$	$tinv(0.025, 4)$	$chi2inv(0.025, 4)$	$chi2inv(0.975, 4)$
-1.96	-2.78	0.48	11.14

R: $n = 5$, $\bar{x}_n = \frac{71+68+77+69+65}{5} = 70$, $\alpha = 5\% = 0{,}05$.
a) $z_{\frac{\alpha}{2}} = \text{norminv}(0.025, 0, 1) = -1{,}96$, $z_{1-\frac{\alpha}{2}} = -z_{\frac{\alpha}{2}}$, $\sigma = \sqrt{125} = 5\sqrt{5}$. Valoarea intervalului de încredere este: $(\bar{x}_n + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x}_n + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}) = (70 - 1{,}96 \cdot 5, 70 + 1{,}96 \cdot 5) = (60{,}2,\ 79{,}8)$.
b) $t_{\frac{\alpha}{2}} = \text{tinv}(0.025, 4) = -2{,}78$, $t_{1-\frac{\alpha}{2}} = -t_{\frac{\alpha}{2}}$, $\tilde{s}_n = \sqrt{\frac{1^2+(-2)^2+7^2+(-1)^2+(-5)^2}{4}} = 2\sqrt{5}$. Valoarea intervalului de încredere este: $(\bar{x}_n - t_{1-\frac{\alpha}{2}} \cdot \frac{\tilde{s}_n}{\sqrt{n}}, \bar{x}_n + t_{1-\frac{\alpha}{2}} \cdot \frac{\tilde{s}_n}{\sqrt{n}}) = (70 - 2{,}78 \cdot 2,\ 70 + 2{,}78 \cdot 2) = (64{,}44,\ 75{,}56)$.
c) $c^2_{\frac{\alpha}{2}} = \text{chi2inv}(0.025, 4) = 0{,}48$, $c^2_{1-\frac{\alpha}{2}} = \text{chi2inv}(0.975, 4) = 11{,}14$. Valoarea intervalului de încredere este: $\left(\frac{(n-1)\bar{s}_n^2}{c^2_{1-\frac{\alpha}{2}}}, \frac{(n-1)\bar{s}_n^2}{c^2_{\frac{\alpha}{2}}}\right) = \left(\frac{80}{11{,}14}, \frac{80}{0{,}48}\right) \approx (7{,}18,\ 166{,}67)$.

4. Într-un sondaj de opinie, suntem interesaţi de proporţia p a persoanelor dintr-un anumit oraş care ar vota candidatul A împotriva candidatului B.
a) Determinaţi un interval de încredere bilateral cu nivelul de încredere 95% pentru p, ştiind că 64 de persoane dintr-un eşantion de 100 de participanţi la sondaj ar vota candidatul A.
b) Estimaţi un număr minim de participanţi la sondaj pentru a obţine un interval de încredere bilateral cu nivelul de încredere 95% de lungime cel mult 5%.

R: a) $n = 100, \bar{x}_n = 0{,}64$, $z_{1-\frac{\alpha}{2}} = 1{,}96$ (a se vedea problema 4.). 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) $0{,}95 \approx P\left(p \in (\bar{X}_n - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}, \bar{X}_n + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}})\right) \le P\left(p \in (\bar{X}_n - 1{,}96\frac{1}{2\sqrt{n}}, \bar{X}_n + 1{,}96\frac{1}{2\sqrt{n}})\right)$.
Căutăm n minim astfel încât: $\frac{0{,}98}{\sqrt{n}} < 2{,}5\% \implies \sqrt{n} > \frac{0{,}98}{0{,}025} = 39{,}2 \implies$ valoarea minimă estimată este 1537.

5. O companie doreşte înlocuirea unui sistem de frânare pentru un anumit tip de maşină cu unul nou, care să reducă semnificativ distanţa de frânare. Media distanţei de frânare pentru vechiul sistem este mai mare sau egală decât 50 m, pentru o viteză de 80 km/h pe ploaie. Folosind tabelul de mai jos, rezolvaţi cerinţele următoare.
i) Ştiind că în urma testării a 100 de maşini cu noul sistem de frânare instalat, pentru o viteză de 80 km/h pe ploaie, s-a constatat că valoarea mediei de selecţie este 49 m şi că valoarea abaterii standard de selecţie este 1 m pentru distanţa de frânare a acestui eşantion, să se construiască un interval de încredere unilateral stâng pentru media distanţei de frânare a noului sistem, iar apoi să se testeze dacă noul sistem este mai perfomant decât cel vechi, cu un nivel de semnificaţie de 1%.

ii) Ştiind că, în condiţiile precizate, abaterea standard a distanţei de frânare a noului sistem de frânare este de 2 m, care este valoarea maximă a mediei distanţei de frânare pentru un eşantion de 100 de maşini testate cu noul sistem de frânare pentru a concluziona, cu ajutorul unui test statistic cu un nivel de semnificaţie de 6%, că noul sistem este mai performant?

x	0.005	0.01	0.03	0.06	0.94	0.97	0.99	0.995
$norminv(x,0,1)$	-2.58	-2.33	-1.88	-1.55	1.55	1.88	2.33	2.58
$tinv(x,99)$	-2.63	-2.36	-1.9	-1.57	1.57	1.9	2.36	2.63
$chi2inv(x,99)$	66.5	69.2	74.3	78.1	121.8	127.1	134.6	139

R: i) Construim un interval de încredere unilateral stâng pentru medie, când varianţa este necunoscută: $n = 100$, $\overline{x}_{100} = 49$, $\widetilde{s}_{100} = 1$, $\alpha = 0{,}01$, $t_\alpha = tinv(\alpha, n-1) = tinv(0.01, 99) = -2{,}36$. Valoarea intervalului de încredere este $(-\infty, \overline{x}_{100} - \frac{\widetilde{s}_{100}}{\sqrt{n}} \cdot t_\alpha) = (-\infty, 49 + \frac{1}{10} \cdot 2{,}36) = (-\infty, 49{,}236)$. $H_0 : m \geq 50$, $H_1 : m < 50$. Aplicăm testul pentru medie, când varianţa este necunoscută: $t = \frac{\overline{x}_{100}-50}{\frac{\widetilde{s}_{100}}{\sqrt{100}}} = \frac{49-50}{\frac{1}{10}} = -10$, $t < t_\alpha \implies$ se respinge H_0, adică se acceptă că sistemul nou este mai performant.
ii) $H_0 : m \geq 50$, $H_1 : m < 50$. Avem: $\alpha = 0{,}06$, $\sigma = 2$, $z_\alpha = norminv(\alpha) = norminv(0{,}06) = -1{,}55$, $z = \frac{\overline{x}_{100}-50}{\frac{2}{10}}$. Aplicând testul pentru medie, când varianţa este cunoscută, respingem ipoteza H_0 în favoarea lui H_1 (i.e. acceptăm că sistemul nou este mai performant) dacă şi numai dacă $z \leq z_\alpha \Leftrightarrow \frac{\overline{x}_{100}-50}{\frac{1}{5}} \leq -1{,}55 \Leftrightarrow \overline{x}_{100} \leq 49{,}69$. Deci valoarea maximă cerută este 49,69.

6. Un provider de internet îşi asigură clienţii că viteza conexiunii la internet este în medie mai mare sau egală decât 250 Mbps între orele 20:00 şi 22:00. Pe de altă parte, providerul susţine că în acest interval orar conexiunea nu este stabilă, având o abatere standard de 40 Mbps. Folosind tabelul de mai jos, rezolvaţi cerinţele următoare.
i) Ştiind că în urma unei selecţii de 100 de clienţi s-a constatat că valoarea mediei de selecţie 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 semnificaţie de 5%.
ii) Care este valoarea maximă a mediei vitezei conexiunii între orele specificate pentru un eşantion de 100 de clienţi, pentru a concluziona, cu ajutorul unui test statistic cu un nivel de semnificaţie de 4%, că media vitezei conexiunii nu este cea pretinsă de provider?

x	0.02	0.025	0.04	0.05	0.95	0.96	0.975	0.98
$tinv(x,99)$	-2.08	-1.98	-1.77	-1.66	1.66	1.77	1.98	2.08
$chi2inv(x,99)$	72.3	73.4	75.8	77	123.2	125	128.4	130
$norminv(x,0,1)$	-2.05	-1.96	-1.75	-1.64	1.64	1.75	1.96	2.05

R: i) Construim un interval de încredere bilateral pentru medie, când varianţa este cunoscută: $n = 100$, $\overline{x}_{100} = 242$, $\sigma = 40$, $\alpha = 0{,}05$, $z_{1-\frac{\alpha}{2}} = norminv(1-\frac{\alpha}{2}) = norminv(0{,}975) = 1{,}96$. Valoarea intervalului de încredere este $(\overline{x}_{100} - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}, \overline{x}_{100} + \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}) = (242 - \frac{40}{10} \cdot 1{,}96, 242 + \frac{40}{10} \cdot 1{,}96) = (234{,}16,\ 249{,}84)$.
$H_0 : m \geq 250$, $H_1 : m < 250$. Aplicăm testul pentru medie, când varianţa este cunoscută: $z_\alpha = norminv(\alpha) = norminv(0{,}05) = -1{,}64$, $z = \frac{\overline{x}_{100}-250}{\frac{\sigma}{\sqrt{100}}} = \frac{242-250}{\frac{40}{10}} = -2$, $z < z_\alpha \implies$ 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: $z_\alpha = norminv(\alpha, 0, 1) = norminv(0{,}04, 0, 1) = -1{,}75$,

$z = \frac{\overline{x}_{100} - 250}{\frac{40}{10}}$. Aplicând testul pentru medie, când varianţa este cunoscută, respingem ipoteza H_0 în favoarea lui H_1 (adică, acceptăm că media vitezei conexiunii nu este cea pretinsă de provider) dacă şi numai dacă $z \leq z_\alpha \Leftrightarrow \frac{\overline{x}_{100} - 250}{\frac{40}{10}} \leq -1{,}75 \Leftrightarrow \overline{x}_{100} \leq 243$. Deci valoarea maximă cerută este 243.