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: $E(X) = Np = \frac{1}{n}\sum_{i=1}^n x_i \implies \hat{p}(x_1, \ldots, x_n) = \frac{1}{nN}\sum_{i=1}^n x_i$ valoarea estimată 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. Valoarea estimată pentru parametrul necunoscut p este $\hat{p}(x_1, \ldots x_n) = \frac{1}{nN}\sum_{i=1}^n x_i$.
Estimatorul 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 $\hat{p}(X_1, \ldots X_n) \overset{a.s.}{\to} \frac{1}{N}E(X) = p$ (am considerat 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: $f_X(x) = \lambda^2 x e^{-\lambda x}, x > 0 \implies L(x_1, \ldots, x_n; \lambda) = \prod_{i=1}^n \lambda^2 x_i e^{-\lambda x_i} \implies \ln L(x_1, \ldots, x_n; \lambda) = 2n \ln \lambda + \sum_{i=1}^n \ln x_i - \lambda \sum_{i=1}^n x_i$. $\frac{\partial \ln L}{\partial \lambda}(x_1, \ldots, x_n; \lambda) = \frac{2n}{\lambda} - \sum_{i=1}^n x_i = 0 \implies \lambda = \frac{2n}{\sum_{i=1}^n x_i}$; $\frac{\partial^2 \ln L}{\partial \lambda^2}(x_1, \ldots, x_n; \lambda) = -\frac{2n}{\lambda^2} < 0 \implies$ valoarea estimatorului este $\hat{\lambda}(x_1, \ldots, x_n) = \frac{2n}{\sum_{i=1}^n x_i}$. Estimatorul 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 estimată este $\hat{\lambda}(1, \frac{3}{2}, 3, 2, 3, \frac{5}{2}, 1, 2) = 1$.

Pentru rezolvarea următoarelor probleme de statistică, folosiți următoarele valori numerice calculate în Python:

[1]:
```
from scipy.stats import norm, t, chi2
```

[2]:
```
norm.ppf([0.025,0.05,0.95,0.975])
```

[2]:
```
array([-1.95996398, -1.64485363,  1.64485363,  1.95996398])
```

[3]:
```
t.ppf([0.025,0.05,0.95,0.975],4), t.ppf([0.025,0.05,0.95,0.975],99)
```

[3]:
```
(array([-2.77644511, -2.13184678,  2.13184678,  2.77644511]),
 array([-1.98421695, -1.66039116,  1.66039116,  1.98421695]))
```

[4]:
```
chi2.ppf([0.025,0.05,0.95,0.975],2), chi2.ppf([0.025,0.05,0.95,0.975],4)
```

[4]:
```
(array([0.05063562, 0.10258659, 5.99146455, 7.37775891]),
 array([ 0.48441856,  0.71072302,  9.48772904, 11.14328678]))
```

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.
a) Ştiind că varianţa/dispersia masei corporale este 20, determinaţi un interval de încredere bilateral cu nivelul de încredere 95% pentru media masei corporale, apoi testați, cu probabilitatea de risc 5%, ipoteza că media masei corporale este 75 kg.
b) Ştiind că varianţa/dispersia masei corporale este necunoscută, determinaţi un interval de încredere bilateral cu nivelul de încredere 95% pentru media masei corporale, apoi testați, cu probabilitatea de risc 5%, ipoteza că media masei corporale este 75 kg.
c) Determinaţi un interval de încredere bilateral cu nivelul de încredere 95% pentru varianţa masei corporale, apoi testați, cu probabilitatea de risc 5%, ipoteza că abaterea standard a masei corporale este 10 kg.

R: $n = 5$, $\bar{x}_n = \frac{71+68+77+69+65}{5} = 70$, $\alpha = 5\% = 0{,}05$. Fie μ media masei corporale.
a) $z_{1-\frac{\alpha}{2}} = \texttt{norm.ppf}(0.975) \approx 1{,}96$, $\sigma = \sqrt{20} = 2\sqrt{5}$.
Valoarea intervalului de încredere este: $(\bar{x}_n - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x}_n + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}) \approx (70 - 1{,}96 \cdot 2,\ 70 + 1{,}96 \cdot 2) = (66{,}08,\ 73{,}92)$. Deoarece 75 nu aparţine intervalului, respingem $H_0: \ \mu = 75$ și acceptăm $H_1: \ \mu \neq 75$.
b) $t_{1-\frac{\alpha}{2}} = \texttt{t.ppf}(0.975, 4) \approx 2{,}78$, $s_n = \left(\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x}_n)^2\right)^{\frac{1}{2}} = \sqrt{\frac{1^2+(-2)^2+7^2+(-1)^2+(-5)^2}{4}} = 2\sqrt{5}$.
Valoarea intervalului de încredere este: $(\bar{x}_n - t_{1-\frac{\alpha}{2}} \cdot \frac{s_n}{\sqrt{n}}, \bar{x}_n + t_{1-\frac{\alpha}{2}} \cdot \frac{s_n}{\sqrt{n}}) \approx (70 - 2{,}78 \cdot 2,\ 70 + 2{,}78 \cdot 2) = (64{,}44,\ 75{,}56)$. Deoarece 75 aparţine intervalului, acceptăm $H_0: \ \mu = 75$ și respingem $H_1: \ \mu \neq 75$.
c) $c_{\frac{\alpha}{2}} = \texttt{chi2.ppf}(0.025, 4) \approx 0{,}48$, $c_{1-\frac{\alpha}{2}} = \texttt{chi2.ppf}(0.975, 4) \approx 11{,}14$.
Valoarea intervalului de încredere este: $\left(\frac{(n-1)s_n^2}{c_{1-\frac{\alpha}{2}}}, \frac{(n-1)s_n^2}{c_{\frac{\alpha}{2}}}\right) \approx \left(\frac{80}{11{,}14}, \frac{80}{0{,}48}\right) \approx (7{,}18,\ 166{,}67)$. Deoarece $100 \in \left(\frac{80}{11{,}14}, \frac{80}{0{,}48}\right) \Leftrightarrow 48 < 80 < 111{,}4$, acceptăm $H_0: \ \sigma = 10$ și respingem $H_1: \ \sigma \neq 10$.

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. 576 din participanții la sondaj au declarat că ar vota

candidatul A, iar 324 din participanți au declarat că ar vota cu candidatul B.
a) Determinaţi un interval de încredere bilateral cu nivelul de încredere 95% pentru p.
b) Testați cu probabilitatea de risc 5% ipoteza că $p = 0{,}6$.

R: $n = 576 + 324 = 900, \bar{x}_n = \frac{576}{900} = 0{,}64$, $\alpha = 5\% = 0{,}05$, $z_{1-\frac{\alpha}{2}} = \texttt{norm.ppf}(0.975) \approx 1{,}96$.
a) Valoarea intervalului de încredere este: $(\bar{x}_n - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}}, \bar{x}_n + z_{1+\frac{\alpha}{2}}\sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}}) \approx (0{,}64 - 1{,}96\sqrt{\frac{0{,}64\cdot 0{,}36}{900}}, 0{,}64 + 1{,}96\sqrt{\frac{0{,}64\cdot 0{,}36}{900}}) = (0{,}64 - 1{,}96 \cdot \frac{0{,}8\cdot 0{,}6}{30},\ 0{,}64 + 1{,}96 \cdot \frac{0{,}8\cdot 0{,}6}{30}) \approx (0{,}61,\ 0{,}67)$.
b) Aplicăm testul pentru proporție. $p_0 = 0.6$, $np_0(1-p_0) \geq 10$, $z = \frac{\bar{x}_n - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0{,}04}{\sqrt{\frac{0{,}24}{900}}} = \sqrt{6} \approx 2{,}45 \implies |z| \geq z_{1-\frac{\alpha}{2}} \implies$ se respinge $H_0 : p = 0.6$, adică se acceptă $H_1 : p \neq 0.6$.

5. Un provider de internet îşi asigură clienţii că viteza conexiunii la internet este în medie 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. 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.
i) Să se construiască un interval de încredere unilateral stâng cu nivelul de încredere 95% pentru media vitezei conexiunii.
ii) Să se testeze, cu nivelul de semnificaţie 5%, dacă media vitezei este cea pretinsă de provider.

R: $n = 100$, $\overline{x}_{100} = 242$, $\sigma = 40$, $\alpha = 0{,}05$.

i) $z_\alpha = \texttt{norm.ppf}(0.05) \approx -1{,}64$. Valoarea intervalului de încredere unilateral este $(-\infty,\ \overline{x}_{100} - \frac{\sigma}{\sqrt{n}} \cdot z_\alpha) \approx (-\infty,\ 242 + \frac{40}{10} \cdot 1{,}64) = (-\infty,\ 248{,}56)$. Deoarece viteza conexiunii nu poate fi negativă, avem intervalul $[0,\ 248{,}56)$.
ii) $H_0 : \mu = 250$, $H_1 : \mu \neq 250$. Aplicăm testul pentru medie, când varianţa este cunoscută: $z_{1-\frac{\alpha}{2}} = \texttt{norm.ppf}(0.975) \approx 1{,}96$, $z = \frac{\overline{x}_{100} - 250}{\frac{\sigma}{\sqrt{100}}} = \frac{242-250}{\frac{40}{10}} = -2$, $|z| \geq z_{1-\frac{\alpha}{2}} \implies$ se respinge H_0, adică se acceptă că media vitezei conexiunii la internet nu este cea pretinsă de provider.

6. 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. 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.
i) Să se construiască un interval de încredere unilateral drept cu nivelul de încredere 95% pentru media distanței de frânare a noului sistem.
ii) Să se testeze cu probabilitatea de risc 5% dacă noul sistem de frânare este diferit de cel vechi.

R: $n = 100$, $\overline{x}_n = 49$, $s_n = 1$, $\alpha = 0{,}05$.

i) $t_{1-\alpha} = \texttt{t.ppf}(0.95, 99) \approx 1{,}66$. Valoarea intervalului de încredere unilateral este $(\overline{x}_{100} - \frac{s_n}{\sqrt{n}} \cdot t_{1-\alpha}, \infty) \approx (49 - 1{,}66,\ \infty) = (47{,}34,\ \infty)$.
ii) $H_0 : \mu = 50$, $H_1 : \mu \neq 50$. Aplicăm testul pentru medie, când varianţa este necunoscută: $t_{1-\frac{\alpha}{2}} = \texttt{t.ppf}(0.975, 99) \approx 1{,}98$, $t = \frac{\overline{x}_{100} - 50}{\frac{s_n}{\sqrt{n}}} = \frac{49-50}{\frac{1}{10}} = -10$, $|t| \geq t_{1-\frac{\alpha}{2}} \implies$ se respinge H_0, adică se acceptă că sistemul nou este diferit de cel vechi.

7. Un sondaj de opinie are în vedere studiul a două caracteristici: F, genul de film preferat, și M, genul de muzică preferat. S-au obținut următoarele date statistice:

Film \ Muzică	Rock	Pop
Acțiune	25	45
Comedie	26	39
Dramă	39	26

Testați, cu nivelul de semnificație 5%, dacă F și M sunt independente.

R: Aplicăm testul pentru independența a două caracteristici discrete.
$r = 3$, $s = 2$, $n_{1\cdot} = 70$, $n_{2\cdot} = 65$, $n_{3\cdot} = 65$, $n_{\cdot 1} = 90$, $n_{\cdot 2} = 110$, $n = 200$,
$x = \frac{\left(25-\frac{70\cdot 90}{200}\right)^2}{\frac{70\cdot 90}{200}} + \frac{\left(45-\frac{70\cdot 110}{200}\right)^2}{\frac{70\cdot 110}{200}} + \frac{\left(26-\frac{65\cdot 90}{200}\right)^2}{\frac{65\cdot 90}{200}} + \frac{\left(39-\frac{65\cdot 110}{200}\right)^2}{\frac{65\cdot 110}{200}} + \frac{\left(39-\frac{65\cdot 90}{200}\right)^2}{\frac{65\cdot 90}{200}} + \frac{\left(26-\frac{65\cdot 110}{200}\right)^2}{\frac{65\cdot 110}{200}} \approx 6{,}12,$
$c_{1-\alpha} = \texttt{chi2.ppf}(1-\alpha, (r-1)(s-1)) = \texttt{chi2.ppf}(0.95, 2) \approx 6 \implies x \geq c_{1-\alpha} \implies$ se respinge H_0 : "F și M sunt independente" și se acceptă H_1 : "F și M sunt dependente".

8. Într-o urnă sunt 100 de bile. Fiecare bilă este colorată cu una din culorile: roșu, albastru, verde. În urma extragerii a 120 bile, cu returnare, s-au obținut 34 de bile roșii, 55 de bile albastre și 31 de bile verzi. Folosind un test statistic cu nivelul de semnificaţie 5%, se poate afirma că în urnă sunt 25 de bile roșii, 50 de bile albastre și 25 de bile verzi?

R: Aplicăm testul χ^2 de concordanță.
$n = 34 + 55 + 31 = 120$, $p_r = \frac{25}{100}$, $p_a = \frac{50}{100}$, $p_v = \frac{25}{100}$, $v = \frac{(34-120\cdot 0{,}25)^2}{120\cdot 0{,}25} + \frac{(55-120\cdot 0{,}5)^2}{120\cdot 0{,}5} + \frac{(31-120\cdot 0{,}25)^2}{120\cdot 0{,}25} \approx 1{,}48,$
$c_{1-\alpha} = \texttt{chi2.ppf}(0.95, 2) \approx 6$. $v < c_{1-\alpha} \implies$ se acceptă H_0 : $p_r = \frac{25}{100}$, $p_a = \frac{50}{100}$, $p_v = \frac{25}{100}$.