Tema 3

Exercițiul 1

a) Nivelul de zgomot al unei maşini de spălat este o v.a. de medie 44 dB şi de abatere standard 5 dB. Admiţand aproximarea normală care este probabilitatea să găsim o medie a zgomotului superioară la 48 dB intr-un eşantion de talie 10 maşini de spălat ?

b) O telecabină are o capacitate de 100 de persoane. Ştiind că greutatea populaţiei (ţarii) este o v.a. de medie 66.3 Kg şi o abatere standard de 15.6 Kg şi presupunand că persoanele care au urcat in telecabină au fost alese in mod aleator din populaţie, care este probabilitatea ca greutatea totală acestora să depăşească 7000 Kg ?

Exercițiul 2

Fie $X_1, \dots, X_n$ un eşantion de talie n dintr-o populaţie de medie μ şi varianţă σ^2. Arătaţi că varianţa varianţei eşantionului este:

$$\mathbb{V}(S^2) = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right)$$

unde $\mu_4 = \mathbb{E}[(X_i - \mu)^4]$ este momentul centrat de ordin 4. Ce revine această formulă in cazul Gaussian (normal) ?

Exercițiul 3

Fie $X_1, \dots, X_n$ un eşantion de talie n dintr-o populaţie de medie μ şi varianţă σ^2. Arătaţi că

$$Cov(\bar{X}, S^2) = \frac{\mu_3}{n}$$

unde $\mu_3 = \mathbb{E}[(X_i - \mu)^3]$ este momentul centrat de ordin 3. Acest rezultat ne arată că cele două statistici sunt asimptotic *necorelate*.

Exercițiul 4

Fie $X_1, X_2, \dots, X_n$ un eșantion de talie n dintr-o populație $\mathcal{U}([0, \theta])$ cu $\theta > 0$ necunoscut.

a) Fie $\hat{\theta}_n = \max\{X_1, \dots, X_n\}$. Determinați funcția de repartiția a lui $\hat{\theta}_n$.

b) Arătați că $\hat{\theta}_n$ este un estimator consistent pentru θ.

c) Arătați că $\hat{\theta}_n$ nu este un estimator nedeplasat pentru θ și construiți un asemenea estimator.

Exercițiul 5

Fie $X_1, \dots, X_n$ un eşantion de talie n cu funcţia de repartiţie $F(x)$ şi densitatea $f(x)$ şi $(Y_1, \dots, Y_n)$ versiunea ordonată crescător a acestuia. Notăm cu $H_k(x)$ şi $h_k(x)$ funcţia de repartiţie şi densitatea v.a. Y_k. Fie $Y_1 = \inf X_i$ şi $Y_n = \sup X_i$.

a) Care este funcţia de repartiţie şi densitatea lui Y_1 şi Y_n ?

b) Care este probabilitatea ca o observaţie dintr-o v.a. de lege $\mathcal{N}(\mu, \sigma^2)$ să depăşească $\mu + 3\sigma$?

c) Dar intr-un eşantion de talie 100 cat este această probabilitate (i.e. probabilitatea ca o observaţie să depăşească $\mu + 3\sigma$)?

d) Dintr-un eşantion de talie 100 dintr-o populaţie repartizată $\mathcal{N}(0, 1)$ ce valoare nu poate fi depăşită cu o probabilitate de 99% ?

e) O societate de analiză a calităţii apei şi a mediului efectuează un sondaj in laboratoarele sale (50 la număr, repartizate pe tot teritoriul Romaniei) pentru a testa dacă efectuează măsurători corecte. Pentru aceasta, serviciul de calitate trimite la fiecare laborator un eşantion de apă care conţine o anumită concentraţie de crom şi le cere să determine această concentraţie de crom. Ţinand cont de fluctuaţiile care apar in prepararea soluţiei, precum şi de imprecizia aparatelor de măsură, societatea presupune că repartiţia concentraţiei de crom (mg/l) este $\mathcal{N}(10, 1)$.

Printre rezultatele obţinute de la laboratoare, două dintre acestea au inregistrat măsurători mai diferite decat celelalte: laboratorul L_1 a inregistrat o concentraţie de 6 mg/l (cea mai mică valoare inregistrată) iar laboratorul L_2 a mă surat o concentraţie de 13 mg/l (cea mai mare dintre măsurători).

Puteţi spune, cu o probabilitate de 99%, că aceste valori sunt coerente sau că valorile obţinute sunt aberante (datorită erorilor de măsurare, de calibrare a aparatelor, etc.) ?

Exercițiul 6

Fie $X \sim B(10, \theta)$ cu $\theta \in (0, 1)$ necunoscut. Fie $\hat{\theta}_1 = \frac{X}{10}$ și $\hat{\theta}_2 = \frac{X+1}{12}$ doi estimatori pentru θ.

a) Calculați $\mathbb{E}_\theta[\hat{\theta}_1]$ și $\mathbb{E}_\theta[\hat{\theta}_2]$.

b) Calculați erorile medii pătratice: $MSE_\theta(\hat{\theta}_1)$ și $MSE_\theta(\hat{\theta}_2)$.

c) Trasați pe același grafic erorile medii pătratice ale celor doi estimatori ca funcții de θ. Pe care dintre cei doi estimatori îl preferați?

Tema 3

Soluții

Exercițiul 1

a) Fie X nivelul de zgomot produs de o maşină de spălat luată la intamplare şi $\bar{X}_{10}$ media unui eşantion de talie 10. Presupunem că aproximarea gaussiană are loc pentru $n = 10$. Avem

$$\bar{X}_{10} \sim \mathcal{N}\left(44, \frac{5^2}{10}\right),$$

de unde $\mathbb{P}(\bar{X}_{10} > 48) \simeq \mathbb{P}\left(Z > \frac{48-44}{5/\sqrt{10}}\right) = \mathbb{P}(Z > 2.53) = 1 - 0.9943 = 0.0057$, unde $Z \sim \mathcal{N}(0, 1)$. Observăm că această proabilitate este foarte mică.

b) Fie X greutatea unei persoane luate la intamplare şi $\bar{X}_{100}$ greutatea media a unui eşantion de 100 de persoane. Aplicand approximarea gaussiană (Teorema Limită Centrală) avem

$$\bar{X}_{100} \simeq \mathcal{N}\left(66.3, \frac{15.6^2}{100}\right),$$

de unde $\mathbb{P}(\bar{X}_{100} > \frac{7000}{100}) \simeq \mathbb{P}\left(Z > \frac{70-66.3}{15.6/\sqrt{100}}\right) = \mathbb{P}(Z > 2.37) = 0.0089$, unde $Z \sim \mathcal{N}(0, 1)$.

Exercițiul 2

Am văzut la curs că

$$(n-1)S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - \frac{1}{n}\left[\sum_{i=1}^{n}(X_i - \mu)\right]^2.$$

Dacă notăm cu $Z_i = X_i - \mu$ atunci observăm că v.a. Z_i sunt i.i.d. iar $\mathbb{E}[Z_i] = 0$, $\mathbb{E}[Z_i^2] = \sigma^2$ şi $\mathbb{E}[Z_i^4] = \mu_4$. Avem că

$$\begin{aligned}(n-1)S^2 &= \sum_{i=1}^{n}(Z_i)^2 - \frac{1}{n}\left(\sum_{i=1}^{n} Z_i\right)^2 = \sum_{i=1}^{n}(Z_i)^2 - \frac{1}{n}\left(\sum_{i=1}^{n} Z_i^2 + 2\sum\sum_{i<j} Z_i Z_j\right)^2 \\ &= \frac{n-1}{n}\sum_{i=1}^{n}(Z_i)^2 - \frac{2}{n}\sum\sum_{i<j} Z_i Z_j\end{aligned}$$

de unde obţinem

$$\begin{aligned}(n-1)^2\mathbb{E}[(S^2)^2] &= \mathbb{E}\left[\left(\frac{n-1}{n}\sum_{i=1}^{n}(Z_i)^2 - \frac{2}{n}\sum\sum_{i<j} Z_iZ_j\right)^2\right]\\ &= \left(\frac{n-1}{n}\right)^2\mathbb{E}\left[\sum_{i=1}^{n} Z_i^4 + 2\sum\sum_{i<j} Z_i^2Z_j^2\right] - \frac{4(n-1)}{n^2}\mathbb{E}\left[\left(\sum_{k=1}^{n} Z_k^2\right)\left(\sum\sum_{i<j} Z_iZ_j\right)\right]\\ &\quad + \frac{4}{n^2}\mathbb{E}\left[\left(\sum\sum_{i<j} Z_iZ_j\right)^2\right] \qquad (\star)\end{aligned}$$

Pentru primul termen din suma de mai sus avem

$$\left(\frac{n-1}{n}\right)^2\mathbb{E}\left[\sum_{i=1}^{n} Z_i^4 + 2\sum\sum_{i<j} Z_i^2Z_j^2\right] = \left(\frac{n-1}{n}\right)^2\left(n\mu_4 + n(n-1)\sigma^4\right).$$

Termenul al doilea din ecuaţia $(\star)$ este 0 deoarece conţine sau termeni de forma $\mathbb{E}[Z_iZ_jZ_k^2]$, cu $i \neq j \neq k$, sau termeni de forma $\mathbb{E}[Z_jZ_k^3]$ cu $j \neq k$.

Pentru ultimul termen avem din ecuaţia $(\star)$ avem

$$\frac{4}{n^2}\mathbb{E}\left[\left(\sum\sum_{i<j} Z_iZ_j\right)^2\right] = \frac{4}{n^2}\mathbb{E}\left[\sum\sum_{i<j} Z_i^2Z_j^2\right] = \frac{2(n-1)}{n}\sigma^4,$$

restul termenilor fiind zero deoarece sunt de forma $\mathbb{E}[Z_i^2Z_jZ_k]$ sau $\mathbb{E}[Z_iZ_jZ_kZ_l]$ cu $i \neq j \neq k \neq l$.

Combinand rezultatele obţinem că

$$\begin{aligned}(n-1)^2\mathbb{V}[S^2] &= \frac{(n-1)^2}{n}\mu_4 + \frac{(n-1)^3}{n}\sigma^4 + 2\frac{n-1}{n}\sigma^4 - (n-1)^2\mathbb{E}[S^2]^2\\ &= \frac{(n-1)^2}{n}\mu_4 + \frac{(n-1)(3-n)}{n}\sigma^4\end{aligned}$$

prin urmare $\mathbb{V}[S^2] = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right)$.

In cazul normal avem că $\mu_4 = 3\sigma^4$ (de ce ?) deci $\mathbb{V}[S^2] = \frac{2\sigma^4}{n-1}$ (vedeţi leagea χ^2).

Exerciţiul 3

Dacă notăm cu $Z_i = X_i - \mu$, atunci $\bar{X} - \mu = \bar{Z}$ şi $\mathbb{E}[\bar{Z}] = 0$. Mai mult,

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(Z_i - \bar{Z})^2 = \sum_{i=1}^{n} Z_i^2 - n\bar{Z}^2$$

prin urmare

$$
\begin{aligned}
Cov(\bar{X}, S^2) = Cov(\bar{X}-\mu, S^2) &= Cov\left(\bar{Z}, \frac{1}{n-1}\left[\sum_{i=1}^{n} Z_i^2 - n\bar{Z}^2\right]\right)\\
&= \frac{1}{n-1}\mathbb{E}\left[\bar{Z}\left(\sum_{i=1}^{n} Z_i^2 - n\bar{Z}^2\right)\right] = \frac{1}{n-1}\left[\mathbb{E}\left[\bar{Z}\left(\sum_{i=1}^{n} Z_i^2\right)\right] - n\mathbb{E}[\bar{Z}^3]\right]
\end{aligned}
$$

Cum

$$
\mathbb{E}\left[\bar{Z}\left(\sum_{i=1}^{n} Z_i^2\right)\right] = \frac{1}{n}\mathbb{E}\left[\left(\sum_{j=1}^{n} Z_j\right)\left(\sum_{i=1}^{n} Z_i^2\right)\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} Z_i^3\right] = \mu_3
$$

şi

$$
\mathbb{E}[\bar{Z}^3] = \frac{1}{n^3}\mathbb{E}\left[\left(\sum_{i=1}^{n} Z_i\right)\left(\sum_{j=1}^{n} Z_j\right)\left(\sum_{k=1}^{n} Z_k\right)\right] = \frac{1}{n^3}\mathbb{E}\left[\sum_{i=1}^{n} Z_i^3\right] = \frac{\mu_3}{n^2}
$$

rezultă că $Cov(\bar{X}, S^2) = \frac{1}{n-1}\left(\mu_3 - \frac{\mu_3}{n}\right) = \frac{\mu_3}{n}$.

Exercițiul 4

a) Observăm că funcția de repartiție pentru $X \sim \mathcal{U}(0,\theta)$ este $F_\theta(x) = \frac{x}{\theta}$ dacă $x \in (0,\theta)$ și $F_\theta(x) = 0$ altfel. Cum $X_1, X_2, \dots, X_n$ sunt i.i.d. $\mathcal{U}(0,\theta)$, funcția de repartiție pentru $\hat{\theta}_n = \max\{X_1, \dots, X_n\}$ este

$$
F_{\hat{\theta}_n}(x) = \mathbb{P}_\theta(\hat{\theta}_n \leq x) = \mathbb{P}_\theta(X_1 \leq x, \cdots, X_n \leq x) = (\mathbb{P}_\theta(X_1 \leq x))^n = \left(\frac{x}{\theta}\right)^n, \quad x \in (0,\theta).
$$

b) Pentru a arăta că $\hat{\theta}_n$ este consistent pentru θ trebuie verificat că $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$. Putem remarca că $\theta \geq \hat{\theta}_n$ deoarece fiecare X_i este strict mai mic decât θ. Pentru $\varepsilon > 0$, avem

$$
\mathbb{P}_\theta(|\hat{\theta}_n - \theta| > \varepsilon) = \mathbb{P}_\theta(\theta - \theta_n > \varepsilon) = \mathbb{P}_\theta(\hat{\theta}_n \leq \theta - \varepsilon) = \left(\frac{\theta - \varepsilon}{\theta}\right)^n
$$

Dacă $\varepsilon < \theta$ atunci membrul drept converge la 0 pentru $n \to \infty$ de unde obținem concluzia. În caz că $\varepsilon > \theta$ atunci membrul drept este egal cu 0 de unde și limita.

c) Pentru a verifica dacă estimatorul $\hat{\theta}_n$ este deplasat trebuie să calculăm $\mathbb{E}_\theta[\hat{\theta}_n]$. Cum funcția de repartiție a lui $\hat{\theta}_n$ este $F_{\hat{\theta}_n}(x) = \left(\frac{x}{\theta}\right)^n$ putem găsi cu ușurință că densitatea este $f_{\hat{\theta}_n}(x) = n\frac{x^{n-1}}{\theta^n}$ pentru $x \in (0,\theta)$ și 0 altfel. Prin urmare

$$
\mathbb{E}_\theta[\hat{\theta}_n] = \int_0^\theta x f_{\hat{\theta}_n}(x)\,dx = n\int_0^\theta \left(\frac{x}{\theta}\right)^n dx \overset{y=x/\theta}{=} n\theta\int_0^1 y^n\,dy = \frac{n\theta}{n+1}.
$$

Cum $\mathbb{E}_\theta[\hat{\theta}_n] \neq \theta$ concluzionăm că estimatorul este deplasat. Dacă definim $\tilde{\theta}_n = \frac{n}{n+1}\theta$, atunci se observă că $\tilde{\theta}_n$ este nedeplasat și cum $\hat{\theta}_n$ era consistent iar $\frac{n}{n+1}$ converge la 1 deducem că $\tilde{\theta}_n$ este un estimator consistent.

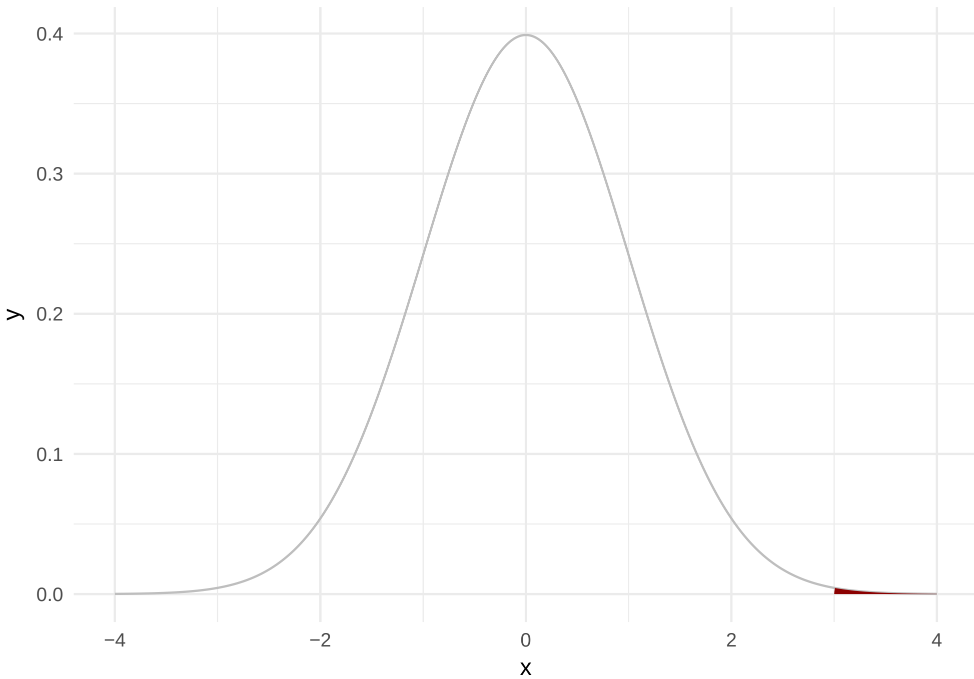
Exercițiul 5

a) Se observă cu uşurinţă că

$$
\begin{aligned}
H_n(x) &= \mathbb{P}(Y_n \leq x) = \mathbb{P}(X_1 \leq x, \cdots, X_n \leq x) \overset{indep.}{=} F(x)^n \\
h_n(x) &= \frac{d}{dx} H_n(x) = n f(x) F(x)^{n-1} \\
H_1(x) &= \mathbb{P}(Y_1 \leq x) = 1 - \mathbb{P}(Y_1 > x) = 1 - \mathbb{P}(X_1 > x, \cdots, X_n > x) \overset{indep.}{=} 1 - (1 - F(x))^n \\
h_1(x) &= \frac{d}{dx} H_1(x) = n f(x) \left(1 - F(x)\right)^{n-1}
\end{aligned}
$$

b) Fie $X \sim \mathcal{N}(\mu, \sigma^2)$. Problema cere să găsim probabilitatea $\mathbb{P}(X > \mu + 3\sigma)$. Avem (vezi porţiunea roşie din figură)

$$\mathbb{P}(X > \mu + 3\sigma) = \mathbb{P}\left(\frac{X - \mu}{\sigma} > 3\right) = 1 - \mathbb{P}\left(\frac{X - \mu}{\sigma} \leq 3\right) = 0.00135$$



c) Fie $X_1, X_2, \ldots, X_n$ un e santion de talie $n = 100$ dintr-o populaţie normală $\mathcal{N}(\mu, \sigma^2)$ şi fie $Z_i = \mathbf{1}_{\{X_i > \mu + 3\sigma\}}$ variabilele Bernoulli care iau valoarea 1 atunci cand $X_i > \mu + 3\sigma$ şi 0 in rest. Problema revine la a determina probabilitatea

$$\mathbb{P}(Z_1 + \cdots Z_n = 1) \overset{i.i.d.}{=} \binom{n}{1} \mathbb{P}(Z_1 = 1)\mathbb{P}(Z_1 = 0)^{n-1} = n\mathbb{P}(X_1 > \mu + 3\sigma)\mathbb{P}(X_1 < \mu + 3\sigma)^{n-1} \simeq 0.11809$$

d) Fie $X_1, X_2, \ldots, X_n$ un e santion de talie $n = 100$ dintr-o populaţie normală $\mathcal{N}(0, 1)$. Problema ne cere să găsim valoarea lui x pentru care probabilitatea $\mathbb{P}(X_1 < x, X_2 < x, \cdots, X_n < x) = 0.99$. Prin urmare vrem să găsim pe x aşa incat $H_n(x) = 0.99$. Din punctul a) avem $H_n(x) = F(x)^n$ deci $x = F^{-1}(\sqrt[n]{0.99}) = 3.7177$.

e) Fie $X_1, X_2, \dots, X_n$ un e santion de talie $n = 50$ dintr-o populaţie normală $\mathcal{N}(10, 1)$ ($n = 50$ reprezintă numărul de laboratoare iar X_i este concentraţia de crom din laboratorul i). Din datele problemei avem că laboratorul 1 a inregistrat cea mai mică valoare (6 mg/l) iar laboratorul 2 a inregistrat cea mai mare valoare (13 mg/l). Problema ne cere să evaluăm probabilitatea

$$\mathbb{P}(Y_1 \leq 6, Y_n \geq 13) = 1 - \mathbb{P}(\{Y_1 > 6\} \cup \{Y_n < 13\}) = 1 - \mathbb{P}(Y_1 > 6) - \mathbb{P}(Y_n < 13) + \mathbb{P}(Y_1 > 6, Y_n < 13).$$

Avem că $\mathbb{P}(Y_1 > 6) = \mathbb{P}(X_1 > 6, \cdots, X_n > 6) = (1 - F(6))^n$ iar $F(6) = \mathbb{P}(X_1 \leq 6) = \mathbb{P}\left(\frac{X_1 - 10}{1} \leq -4\right) \simeq 0.00003$ deci $\mathbb{P}(Y_1 > 6) \simeq 0.99871$.

De asemenea $\mathbb{P}(Y_n < 13) = F(13)^n$ iar cum $F(13) = \mathbb{P}(X_1 \leq 13) = \mathbb{P}\left(\frac{X_1 - 10}{1} \leq 3\right) \simeq 0.9986$ rezultă că $\mathbb{P}(Y_n < 13) \simeq 0.9346$.

In mod similar, $\mathbb{P}(Y_1 > 6, Y_n < 13) = \mathbb{P}(6 < X_1 < 13, \cdots, 6 < X_n < 13) = \mathbb{P}(6 < X_1 < 13)^n$ şi cum $\mathbb{P}(6 < X_1 < 13) = \mathbb{P}(X_1 < 13) - \mathbb{P}(X_1 \leq 6) \simeq 0.9986$ obţinem că $\mathbb{P}(Y_1 > 6, Y_n < 13) \simeq 0.9332$.

In concluzie avem că $\mathbb{P}(Y_1 \leq 6, Y_n \geq 13) \simeq 0.0001$.

Exercițiul 6

a) Cum $\mathbb{E}_\theta[X] = 10\theta$ obținem că $\mathbb{E}_\theta[\hat{\theta}_1] = \theta$ și $\mathbb{E}_\theta[\hat{\theta}_2] = \frac{10\theta+1}{12}$.

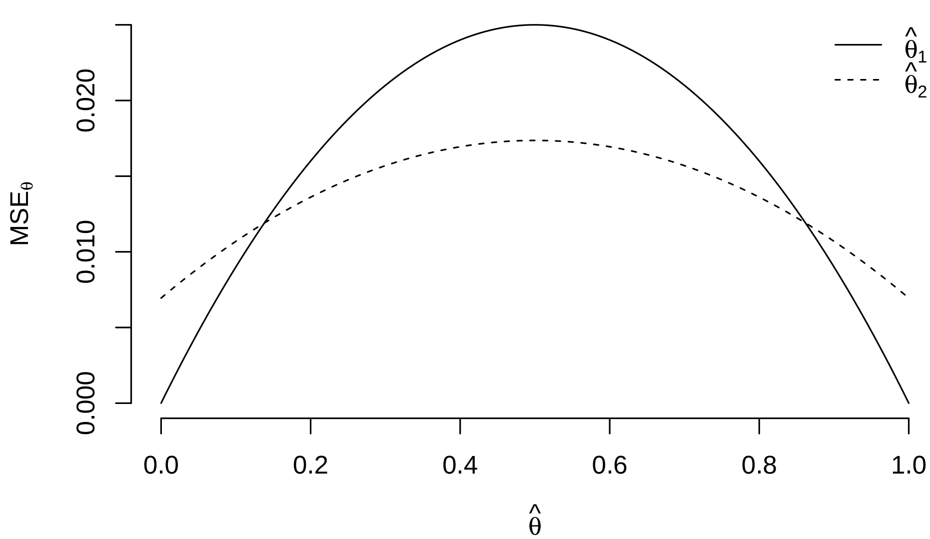
b) Pentru calculul erorii medii pătratice vom folosi următoarea formulă $MSE_\theta(\hat{\theta}) = Var_\theta(\hat{\theta}) + B_\theta(\hat{\theta})^2$. Cum $\hat{\theta}_1$ este un estimator nedeplast rezultă că $B_\theta(\hat{\theta}_1) = 0$ și

$$MSE_\theta(\hat{\theta}_1) = Var_\theta(\hat{\theta}_1) = 10^{-2} Var_\theta(X) = \frac{\theta(1-\theta)}{10}.$$

Pentru $\hat{\theta}_2$ avem $B_\theta(\hat{\theta}_2) = \frac{10\theta+1}{12} - \theta$ de unde

$$MSE_\theta(\hat{\theta}_2) = \frac{Var_\theta(X)}{12^2} + \left(\frac{10\theta+1}{12} - \theta\right)^2 = \frac{6\theta - 6\theta^2 + 1}{144}.$$

c) Avem următoarea figură:



Chiar dacă $\hat{\theta}_1$ este nedeplasat și $\hat{\theta}_2$ este deplasat, niciunul dintre cei doi estimatori nu are eroarea medie pătratică uniform mai mică. Cu toate acestea, eroarea medie pătratică pentru estimatorul $\hat{\theta}_2$ este mai mică decât cea pentru estimatorul $\hat{\theta}_1$ pe aproape toată plaja de valori a lui θ (mai exact pe intervalul $\theta \in \left[\frac{1-\sqrt{\frac{11}{12}}}{2}, \frac{1+\sqrt{\frac{11}{12}}}{2}\right]$). Cum eroarea medie pătratică este mai importantă decât nedeplasarea, recomand folosirea estimatorului $\hat{\theta}_2$.