

ESTADÍSTICA

Examen parcial

Grau Matemàtiques

Divendres 26 de març de 2021, 9.00-13.00

Podeu fer ús d'una calculadora tradicional, la llista de distribucions que surt a l'atenea del curs i les taules estadístiques. No s'admet l'ús de telèfons mòbils u ordinadors. Separeu en fulls la teoria dels problemes.

TEORIA (4 punts)

- 1) a) En estadística descriptiva, qué vol dir que un estadístic sigui *robust* en presència d'atípics?
Vol dir que tot i que hi hagi atípics (valors extrems) el valor de l'estadístic no es veu influït per aquests i el seu valor no canvia massa. En cas de no ser robust, la presència d'atípics pot fer canviar molt l'estadístic.
- b) Dels següents estadístics descriptius, indica quins són robustos en presència d'atípics?
Mitjana, Mediana, Mitjana retallada, Rang, IQR, Variància, Q1, Percentil 95, Mínim, Coeficient de Variació, Coeficient d'asimetria.
- c) Quan es contrueix un boxplot, de quina manera s'identifiquen les observacions atípiques, tant les moderades com les extremes?
Es consideren valors atípics els valors de la mostra que estan per sota del valor $Q1 - 1.5IQR$ o per sobre de $Q3 + 1.5IQR$. Aquestes dades atípiques són extremes quan estan per sota del valor $Q1 - 3IQR$ o per sobre de $Q3 + 3IQR$, essent $IQR = Q3 - Q1$ el rang interquartílic.

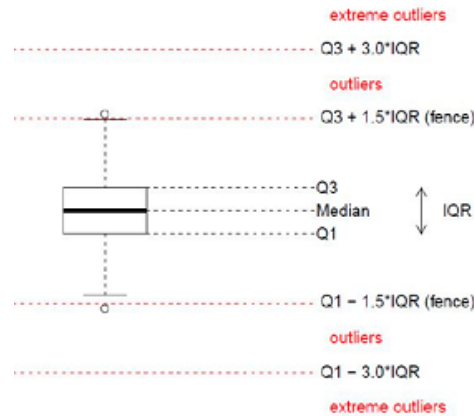


Figura 1: Determinació de valors atípics a partir d'un Box-Plot.

- 2) Sigui $X \sim N(\mu, \sigma^2)$ i $X_1, \dots, X_n$ una m.a.s. d'aquesta variable. En els següents estadístics, els valors de K_1, K_2, K_3 representen constants (no variables aleatòries) que fan que les distribucions d'aquestes expressions siguin conegudes analíticament. En cada cas, indica el valor de la constant i la distribució que segueix l'estadístic:
Nota: Suposarem que en els apartats b) i c) les expressions de numerador i denominador provenen de dues mostres independents de la mateixa variable aleatòria i amb la mateixa mida. D'aquesta manera, podem considerar que ambdues expressions són independents.

a) $K_1 \frac{\bar{X} - \mu}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}}$

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\bar{X} - \mu}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 / (n-1)} / \sqrt{n}} \sim t_{n-1}$$
$$\sqrt{n(n-1)} \frac{\bar{X} - \mu}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}} \sim t_{n-1} \Rightarrow K_1 = \sqrt{n(n-1)}$$

b) $K_2 \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \mu)^2}$

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \chi_{n-1}^2 \quad \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^2} \sim \chi_n^2$$
$$\frac{\sum_{i=1}^n (X_i - \bar{X})^2 / (n-1)}{\sum_{i=1}^n (X_i - \mu)^2 / n} = \frac{n}{n-1} \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \mu)^2} \sim F_{n-1, n} \Rightarrow K_2 = \frac{n}{n-1}$$

c) $K_3 \frac{(\bar{X} - \mu)^2}{\sum_{i=1}^n (X_i - \mu)^2}$

$$\frac{(\bar{X} - \mu)^2}{\sigma^2/n} \sim \chi_1^2 \quad \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^2} \sim \chi_n^2$$

$$\frac{n(\bar{X} - \mu)^2/1}{\sum_{i=1}^n (X_i - \mu)^2/n} = n^2 \frac{(\bar{X} - \mu)^2}{\sum_{i=1}^n (X_i - \mu)^2} \sim F_{1,n} \Rightarrow K_3 = n^2$$

- 3) En un servei telefònic de postvenda, el número d'incidències reportades diàriament es pot modelitzar com una Poisson de paràmetre λ . Es disposa de una mostra $X_1, \dots, X_n$ amb el número d'incidències en n dies. Es vol estimar la probabilitat de que en un dia hi hagi alguna incidència.

- a) Calcula l'estimador màxim-versemblant d'aquesta probabilitat

$$X \sim \text{Poisson}(\lambda) \Rightarrow p = P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-\lambda}$$

Sabem que l'estimador màxim-versemblant de λ és $\bar{X}$. Pel principi d'invariància d'aquest estimador tindrem:

$$\hat{p}_{MV} = 1 - e^{-\hat{\lambda}_{MV}} = 1 - e^{-\bar{X}}$$

- b) Quant val la Informació de Fisher continguda en la mostra pel paràmetre corresponent a aquesta probabilitat?

La Informació de Fisher continguda en la mostra pel paràmetre λ és:

$$I_n(\lambda) = nI_1(\lambda) = nE \left(-\frac{\partial^2}{\partial \lambda^2} \log P(X = x|\lambda) \right) = nE \left(\frac{X}{\lambda^2} \right) = \frac{n}{\lambda}$$

Per tant, la informació de Fisher continguda a la mostra pel paràmetre $p = g(\lambda) = 1 - e^{-\lambda}$ serà:

$$I_n(p) = I_n(g(\lambda)) = \frac{I_n(\lambda)}{g'(\lambda)^2} = \frac{n}{\lambda e^{-2\lambda}}$$

I com $\lambda = -\log(1 - p)$, tenim

$$I_n(p) = -\frac{n}{(1 - p)^2 \log(1 - p)}$$

- 4) En una enquesta per a un estudi de mercat, els resultats del percentatge de persones que comprarien un cert producte es publiquen incorporant el següent text:

Partint dels criteris del mostreig aleatori simple, per a un nivell de confiança del 95.5% (dues sigmes) i per al cas de màxima indeterminació ($p = q = 50$), el marge d'error estadístic per a les dades globals és de +/- 3.2 punts

- a) Qué representa la especificació *dues sigmes* en el càlcul del marge d'error?

Assumirem la distribució normal com resultat de l'aproximació assintòtica per a la distribució binomial(n,p) quan n tendeix a infinit:

$$X \sim B(n, p) \Rightarrow X \approx N(np, np(1 - p)) \Rightarrow \hat{p} = \frac{X}{n} \approx N \left(p, \sigma^2 = \frac{p(1 - p)}{n} \right)$$

Si volem calcular un interval de confiança per a p a partir d'aquesta aproximació, la seva expressió serà:

$$IC_{1-\alpha}(p) = \hat{p} \pm Z_{\alpha/2} \sigma = \hat{p} \pm Z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

Fixada la confiança del 95.5%, l'interval que determina els valors plausibles és +/- 2 sigma, essent sigma la desviació estàndar de la distribució Normal: $Z_{0.0225} = -2$ $Z_{0.9775} = 2$. Per tant, quan es calcula l'interval de confiança al 95.5% el valor crític serà +/-2 i el marge d'error serà +/- 2 sigma

b) Què vol dir *cas de màxima indeterminació*?

Com en l'expressió de la sigma de l'interval de confiança intervé el paràmetre pel qual calculem l'interval, una possible aproximació s'obtindria substituint el valor de p per l'estimació que fem $\hat{p}$. Ara bé, segons sigui aquesta estimació el recobriment de l'interval aproximat podria ser inferior al nominal indicat pel nivell de confiança. Una forma de garantir que el recobriment sigui més gran o igual al nivell de confiança demanat és proposar el valor de p que maximitzi la sigma i per tant, sigui quin sigui el valor de p , l'interval proposat segur que serà d'amplada igual o superior al correcte. El valor de màxima indeterminació es troba quan $p = 1 - p = 0.5$ L'interval serà:

$$IC_{95.5\%}(p) = \hat{p} \pm 2\sqrt{\frac{0.5^2}{n}} = \hat{p} \pm \frac{1}{\sqrt{n}}$$

c) Amb les dades indicades, calcula la mida mostral de l'enquesta

$$\frac{1}{\sqrt{n}} = 0.032 \Rightarrow n = 977$$

PROBLEMA 1: (2.5 punts)

Suposa que les variables aleatòries $Y_1, Y_2, \dots, Y_n$ són tals que:

$$Y_i = \beta x_i + \epsilon_i, \quad (1)$$

on $x_1, x_2, \dots, x_n$ són constants conegudes, i on $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ són una mostra independent i idènticament distribuïda $Normal(0, \sigma^2)$, amb $\beta \in R$ i $\sigma^2 \in [0, \infty)$ desconegudes.

a). Troba un estadístic suficient de dues dimensions per (β, σ^2) .

Donat que $y_i \sim Normal(\beta x_i, \sigma^2)$, la densitat de probabilitat de $Y = (Y_1, \dots, Y_n)$ és:

$$p(Y|\beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{\sum_i (y_i - \beta x_i)^2}{2\sigma^2}\right), \quad (2)$$

$$p(Y|\beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{\beta^2 \sum_i x_i^2}{2\sigma^2}\right) \exp\left(-\frac{1}{2\sigma^2} \sum_i Y_i^2 + \frac{\beta}{\sigma^2} \sum_i x_i Y_i\right), \quad (3)$$

i el teorema de factorització indica que $(\sum_i Y_i^2, \sum_i x_i Y_i)$ és suficient per (β, σ^2) .

b). Troba l'estimador màxim versemblant de β , $\hat{\beta}_1$.

El logaritme de la funció de versemblança és:

$$l_y(\beta, \sigma^2) = Const - \frac{n}{2} \log \sigma^2 - \frac{\beta^2}{2\sigma^2} \sum_i x_i^2 - \frac{1}{2\sigma^2} \sum_i y_i^2 + \frac{\beta}{\sigma^2} \sum_i x_i y_i, \quad (4)$$

i la solució per a qualsevol valor de σ^2 de l'equació:

$$\frac{\partial l_y}{\partial \beta} = \frac{1}{\sigma^2} \sum_i x_i y_i - \frac{\beta}{\sigma^2} \sum_i x_i^2 = 0, \quad (5)$$

és l'estimador màxim versemblant de β ,

$$\hat{\beta}_1 = \hat{\beta}_{MV} = \frac{\sum_i x_i Y_i}{\sum_i x_i^2}. \quad (6)$$

Es pot comprovar que és màxim perquè:

$$\frac{\partial^2 l_y}{\partial^2 \beta} = -\frac{1}{\sigma^2} \sum_i x_i^2 < 0. \quad (7)$$

c). Troba la distribució de l'estimador màxim versemblant de β , $\hat{\beta}_1$.

L'estimador $\hat{\beta}_{MV}$ és una combinació lineal de Y_i , i per tant té una distribució normal,

$$\hat{\beta}_1 = \hat{\beta}_{MV} \sim Normal(E(\hat{\beta}_{MV}) = \beta, \text{Var}(\hat{\beta}_{MV}) = \frac{\sigma^2}{\sum_i x_i^2} = \frac{\sigma^2}{n\bar{x}^2}). \quad (8)$$

Aquest resultat és exacte, i vàlid per tot n . No cal apel·lar a cap convergència asimptòtica.

d). Troba la distribució d'aquests altres dos estimadors de β ,

$$\hat{\beta}_2 = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i}, \quad (9)$$

$$\hat{\beta}_3 = \frac{1}{n} \sum_{i=1}^n \frac{Y_i}{x_i}. \quad (10)$$

Tant $\hat{\beta}_2$ com $\hat{\beta}_3$ són combinacions lineals de Y_i , i per tant tenen una distribució normal.

$$\hat{\beta}_2 \sim \text{Normal}(E(\hat{\beta}_2) = \beta, \text{Var}(\hat{\beta}_2) = \frac{\sigma^2}{n\bar{x}^2}). \quad (11)$$

$$\hat{\beta}_3 \sim \text{Normal}(E(\hat{\beta}_3) = \beta, \text{Var}(\hat{\beta}_3) = \frac{\sigma^2}{n^2} \sum_i \frac{1}{x_i^2}). \quad (12)$$

e). Proposa un quart estimador de β , $\hat{\beta}_4$, si pot ser sense biaix.

Per exemple, tots els estimadors de β llistats a continuació són sense biaix, amb el benentès de que els tres últims de la llista només serveixen si no hi ha rèpliques (valors de x_i repetits).

$$\hat{\beta}_4^c = \frac{\sum_i x_i^c Y_i}{\sum_i x_i^{c+1}}, \quad (13)$$

(observeu que $\hat{\beta}_1 = \hat{\beta}_4^{c=1}$, que $\hat{\beta}_2 = \hat{\beta}_4^{c=0}$ i que $\hat{\beta}_3 = \hat{\beta}_4^{c=-1}$),

$$\hat{\beta}_4 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad (14)$$

$$\hat{\beta}_4 = \frac{1}{n} \sum_{i=1}^n \frac{Y_i - \bar{Y}}{x_i - \bar{x}}, \quad (15)$$

$$\hat{\beta}_4 = \frac{\sum_{i=1}^n \sum_{j=i+1}^n (x_i - x_j)(Y_i - Y_j)}{\sum_{i=1}^n \sum_{j=i+1}^n (x_i - x_j)^2}, \quad (16)$$

$$\hat{\beta}_4 = \frac{2}{n(n-1)} \sum_{i=1}^n \sum_{j=i+1}^n \frac{Y_i - Y_j}{x_i - x_j}, \quad (17)$$

$$\hat{\beta}_4 = \frac{\sum_{i=1}^n \sum_{j=i+1}^n (Y_i - Y_j)}{\sum_{i=1}^n \sum_{j=i+1}^n (x_i - x_j)}. \quad (18)$$

Qualsevol combinació convexa d'aquests estimadors i/o dels tres estimadors de l'enunciat, també és un estimador de β sense biaix. (El teorema de Gauss-Markov indica que $\hat{\beta}_1$ és l'estimador amb mínima variança d'entre tots els estimadors sense biaix de β que, com tots els estimadors d'aquesta llista i com $\hat{\beta}_2$ i $\hat{\beta}_3$, són funcions lineal de $Y_1, \dots, Y_n$). Alternatives més robustes que aquests estimadors lineals de β podrien ser, per exemple,

$$\hat{\beta}_4 = \text{Mediana} \frac{Y_i}{x_i}, \quad (19)$$

$$\hat{\beta}_4 = \frac{1}{n-4} \sum_{i=3}^{n-2} \left(\frac{Y}{x}\right)_{(i)}, \quad (20)$$

tot i que aquests dos estimadors probablement tenen biaix.

f). (Extra) Quin dels tres primers estimadors escolliries, i perquè?

Aplicant la desigualtat de Jensen es comprova que:

$$\text{Var}(\hat{\beta}_1) \leq \text{Var}(\hat{\beta}_2) \leq \text{Var}(\hat{\beta}_3), \quad (21)$$

per tot $(\beta, \sigma^2) \in R \times [0, \infty)$. De fet, Gauss-Markov indica que $\hat{\beta}_1$ és el *Best Linear Unbiased Estimate* de β , fins i tot quan falla la hipòtesi de Normalitat.

PROBLEMA 2: (1.5 punts)

Et diuen que cada una de les cares d'un dau de sis cares ha estat pintada de color blanc o de color negre, però no et diuen quantes de les sis cares són negres. Per estimar el nombre de cares pintades de color negre, i per tant la probabilitat de que quan tiris el dau la cara superior sigui de color negre, el dau es tira 10 vegades, i s'observa el nombre de vegades que la cara superior és negre, Y . El resultat de l'experiment es $y = 7$.

a). Quin és el model estadístic per aquest experiment?

$$M_Y = \{\text{Binomial}(N = 10, \theta_i), \theta_i \in \Omega = \{0, 1/6, 2/6, 3/6, 4/6, 5/6, 1\}\}. \quad (22)$$

b). Quina és la funció de versemblança quan $y = 7$? Dibuixa-la aproximadament.

$$L_{y=7}(\theta_i) = C\theta_i^y(1-\theta_i)^{N-y} = C\theta_i^7(1-\theta_i)^3, \quad \text{per } \theta_i \in \{0, 1/6, 2/6, 3/6, 4/6, 5/6, 1\}. \quad (23)$$

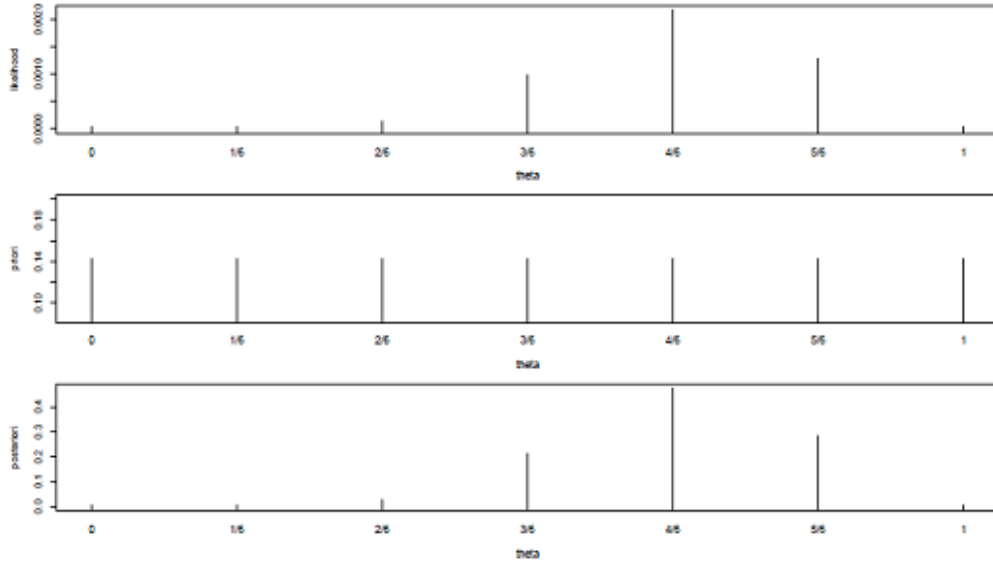


Figura 2: Likelihood, prior and posterior distributions.

c). Tria una distribució a priori per la probabilitat de que al tirar el dau una vegada surti una cara negra, partint de la informació donada fins aquí, i indica com es calcula la distribució a posteriori corresponent si $y = 7$. Dibuixa-les aproximadament.

$$\pi(\theta_i) = \text{Prob}(\theta = \theta_i) = 1/7, \quad (24)$$

$$\pi(\theta_i|Y = 7) = \frac{\pi(\theta_i)p(y = 7|\theta_i)}{\sum_{j=1}^7 \pi(\theta_j)p(y = 7|\theta_j)} = \frac{\pi(\theta_i)L_y(\theta_i)}{\sum_{j=1}^7 \pi(\theta_j)L_y(\theta_j)} = \frac{\theta_i^7(1-\theta_i)^3}{\sum_{j=1}^7 \theta_j^7(1-\theta_j)^3}, \quad (25)$$

per $\theta_i \in \{0, 1/6, 2/6, 3/6, 4/6, 5/6, 1\}$.

d). (Extra) Ara et diuen que per tal de decidir si cada cara s'ha pintat blanca o negra, s'ha tirat una moneda enlaire i s'ha pintat negre si sortia cara, i s'ha pintat blanca si sortia creu. En aquest cas, quina distribució a priori triaries per la probabilitat de que al tirar el dau, et surti una cara negra?

$$\pi(\theta_i) = \text{Prob}(\theta_i = \frac{i-1}{6}) = \binom{6}{i-1} \left(\frac{1}{2}\right)^{i-1} \left(\frac{1}{2}\right)^{6-i+1}, \quad \text{for } i = 1, 2, 3, 4, 5, 6, 7. \quad (26)$$

PROBLEMA 3: (2 punts)

Suposa que les variables aleatòries $X_1, X_2, \dots, X_n$ tenen per densitat de probabilitat:

$$f(x; \theta) = \frac{3\theta^3}{(x + \theta)^4}, \quad (27)$$

on $x > 0$ i $\theta > 0$. (Pista: Comença calculant $E(x + \theta)$ i $E[(x + \theta)^2]$).

a). Troba un estimador de θ pel mètode dels moments, $\hat{\theta}_{MM}$.

$$E(X + \theta) = \frac{3\theta}{2}, \quad (28)$$

$$E(X) = \frac{\theta}{2}, \quad (29)$$

$$\hat{\theta}_{MM} = 2\bar{X}. \quad (30)$$

b). Troba la variança de $\hat{\theta}_{MM}$ i la seva distribució assintòtica. Es un estimador eficient?

$$E((X + \theta)^2) = 3\theta^2 = E(X^2) + 2\theta E(X) + \theta^2 = E(X^2) + 2\theta^2, \quad (31)$$

$$Var(X) = E(X^2) - E(X)^2 = \theta^2 - \frac{\theta^2}{4} = \frac{3\theta^2}{4}, \quad (32)$$

$$Var(\hat{\theta}_{MM}) = 4Var(\bar{X}) = 4 \frac{Var(X)}{n} = \frac{3\theta^2}{n}. \quad (33)$$

$$E(\hat{\theta}_{MM}) = E(2\bar{X}) = \theta. \quad (34)$$

Pel teorema central del límit, per n grans la distribució de $\hat{\theta}_{MM}$ es pot aproximar per:

$$\text{Normal}(\theta, \text{Var}(\hat{\theta}_{MM}) = \frac{3\theta^2}{n}). \quad (35)$$

No és eficient ni assintòticament eficient, perquè la cota de Cramer-Rao, $I_n(\theta)^{-1}$, és menor que $\text{Var}(\hat{\theta}_{MM})$, com es comprova a continuació.

c). Troba la distribució assintòtica de l'estimador màxim versemblant de θ , $\hat{\theta}_{MV}$.

El logaritme de la funció de versemblança per una observació, x_i , és:

$$l_{x_i}(\theta) = \text{Const} + 3 \log \theta - 4 \log(x_i + \theta), \quad (36)$$

$$\frac{dl_{x_i}}{d\theta} = \frac{3}{\theta} - \frac{4}{x_i + \theta}, \quad (37)$$

$$\frac{d^2 l_{x_i}}{d\theta^2} = -\frac{3}{\theta^2} + \frac{4}{(x_i + \theta)^2}, \quad (38)$$

i per tant la informació de Fisher de la mostra de n observacions, X , és:

$$I_X(\theta) = -nE\left[-\frac{3}{\theta^2} + \frac{4}{(x_i + \theta)^2}\right] = \frac{3n}{\theta^2} - \frac{12n}{5\theta^2} = \frac{3n}{5\theta^2}. \quad (39)$$

Per tant, per n grans la distribució de $\hat{\theta}_{MV}$ es pot aproximar per:

$$\text{Normal}(\theta, \text{Var}(\hat{\theta}_{MV}) = \frac{5\theta^2}{3n}), \quad (40)$$

que té una variança assintòtica inferior que $\hat{\theta}_{MM}$.

d). Troba l'estimador màxim versemblant per la mediana poblacional en funció de $\hat{\theta}_{MV}$, i la seva distribució assintòtica.

$$\text{Prob}(X < \text{Med}(X)) = 0.5 = 1 - \frac{\theta^3}{(\text{Med}(X) + \theta)^3}, \quad (41)$$

$$\text{Med}(X) = (\sqrt[3]{2} - 1)\theta = 0.2599\theta, \quad (42)$$

$$\hat{\text{Med}}_{MV} = (\sqrt[3]{2} - 1)\hat{\theta}_{MV} = 0.2599\hat{\theta}_{MV}, \quad (43)$$

i per n grans la seva distribució es pot aproximar per:

$$\text{Normal}(\text{Med}(X), \text{Var}(\hat{\text{Med}}_{MV}) = (\sqrt[3]{2} - 1)^2 \frac{5\theta^2}{3n}). \quad (44)$$

e). (Extra) Se t'acudeix com construir un estimador de θ basat en la mediana mostral?

Iguant la mediana mostral amb la mediana poblacional, per analogia amb el mètode dels moments, s'obté:

$$\hat{\theta}_{**} = \frac{1}{\sqrt[3]{2} - 1} \text{Med}(X_1, \dots, X_n) = 3.847 \text{Med}(X_1, \dots, X_n), \quad (45)$$

que serveix com a estimador de θ .