0.1 Estimadors i Reducció de Dades

Primer defnim notació bàsica per esclarir-nos. Donada aquesta frase:

$Y_1, ..., Y_n$ és una mostra, definida per $Y \sim f(\boldsymbol{y}, \theta)$

Notation 0.1.1.

- Y_i és una variable aleatòria representant la mostra observada. No són encara dades, sinó com serien aquestes generades.
- Y és un vector aleatòri, és a dir, una variable aleatòria que representa la distribució, és a dir, una mostra de la distribució.
- $Y \sim f(\boldsymbol{y}, \theta)$ Y segueix una distribució amb densitat $f(\boldsymbol{y}, \theta)$, i $\boldsymbol{y}$ són les observacions $(y_1, ..., y_n)$ (és a dir, dades conegudes d'avantmà) depenent del parametre θ.

Definition 0.1.2 (Estadístic). Un **Estadístic** és una funció de la mostra $Y = (Y_1, ..., Y_n)$ que no depen del parametre θ

Sobre θ, és un parametre que modifica la distribució de les dades. Si és conegut, sabem exactament com la població Y es comporta. En general però, no s'acostuma a saber, i és l'objectiu d'aquest curs *estimar* θ correctament.

Per estimar un parametre, hem de fer tres distincions conceptuals:

Definition 0.1.3 (Parametre a estimar). El Parametre a Estimar (o Estimand) és la quantitat que no coneixem i a la que volem donar-li valor, el que seria θ.

Definition 0.1.4 (Estimador). Un **Estimador** és una funció que, donada la mostra, ens dona el valor de θ.

Definition 0.1.5 (Estimació). La **estimació** d'un parametre és el valor que s'obté amb l'estimador corresponent.

És a dir, si $X = (X_1, ..., X_n) \sim N(\mu, \sigma^2)$ amb variància desconeguda i volem trobar la mitjana:

- Parametre a estimar: $\theta = \mu$
- Estimador: $\bar{X}$ és la funció que estima θ
- Estimand: El valor de $\bar{X} = \sum_{i=i}^{n} \frac{X_i}{n} = \mu$

Definition 0.1.6 (Estimador Puntual). Un estimador puntual $\hat{\theta}$ de θ és qualsevol funció mesurable de la mostra $(Y_1, ..., Y_n)$ que no depèn de θ. Concretament, qualsevol estadístic és un estimador puntual.

Per exemple, $\hat{\theta} = \hat{\theta}(Y_1, ..., Y_n) = \frac{1}{n} \sum_{i=1}^{n} Y_i$ és una variable aleatòria.

Definition 0.1.7 (Estimació Puntual). Una estimació puntual $\hat{\theta}$ de θ és el resultat numèric del procés d'estimació, basat en els valors observats $(y_1, ..., y_n)$.

Per exemple, $\hat{\theta} = \hat{\theta}(y_1, ..., y_n) = \frac{1}{n} \sum_{i=1}^{n} y_i$ és un valor real. La diferència clau és que:

- L'estimador puntual treballa amb variables aleatòries $(Y_1, ..., Y_n)$
- L'estimació puntual treballa amb valors observats $(y_1, ..., y_n)$

Tot i que ambós es denotin com $\hat{\theta}$, representen conceptes diferents: una funció (estimador puntual) vs. un nombre (estimació puntual)

0.2 Versemblança

Definition 0.2.1 (Versemblança). Donada una FMP (FDP) $f(\boldsymbol{y}|\theta)$ d'una mostra $Y = (Y_1, ..., Y_n)$ i les observacions $\boldsymbol{y} = (y_1, ..., y_n)$, la funció de versemblança de θ és

$$L(\theta|\boldsymbol{y}) = f(\boldsymbol{y}|\theta) = \prod_{i=1}^{n} f_Y(y_i \mid \theta)$$

Versemblança vs Probabilitat

Aquí l'exemple de Poisson on es fan la mateixa pregunta a la inversa però obtens la mateixa resposta.

La diferència entre la probabilitat i la versemblança és només el teu coneixement a priori.

- Quina és la probabilitat de que $Y = 3$ si $Y \sim$ Poiss (2)? És a dir, ja sabem que $\lambda = 2$, i volem trobar com de probable és la nostra observació.
- Quina és "la probabilitat" de que les meves dades segueixen una Possion amb parametre $\lambda = 2$ si la observació és $Y = 3$? No sabem la distribució, concretament és el que volem estimar.

La versemblança ens respon la pregunta que volem contestar quan busquem un estimador: tenint les dades $\boldsymbol{y}$ quin parametre θ fa més versemblants les observacions? Com podem saber quins estimadors *maximitzen la versemblança*?

0.3 Estimadors de Màxima Versemblança (MLE)

Recordem l'estimador del parametre θ depen de les dades Y_i, i per tant, amb dades diferents seguint el mateix proces cada vegada, cosa no molt desitjable. En essència, tenim un vector aleatòri $(\theta_1, ..., \theta_n)$ per cada vegada realitzem l'experiment. Denominarem $\hat{\theta}_n$ a aquest vector, i tot sovint ometrem el subindex d'n, ja que no sempre importa la dimensió.

Per tant, $\hat{\theta}$ és l'estimador de θ desconegut.

Definition 0.3.1 (MLE). Siguin $\boldsymbol{y} \in \mathcal{Y}$ mostres de dades. Direm que el MLE $\hat{\theta}$ de θ és el valor de Θ que maximitza la versemblança, és a dir:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta|\boldsymbol{y})$$

Proposition 0.3.2 (). Si $\hat{\theta}$ és el MLE de θ, aleshores per a qualsevol funció g bijectiva, el MLE de $g(\theta)$ és $g\left(\hat{\theta}\right)$ **TODO**: COMPROVAR QUE AIXÒ ÉS AIXÍ, NO TINC REFERÈNCIES

En general sempre es vol trobar l'estimador de màxima versemblança (*Maximum Likelihood Estimator*) ja que tenen propietats òptimes i son bons estimadors. Per trobar bons estimadors necessitem definir:

Definition 0.3.3 (Biaix). El biaix d'un estimador $\hat{\theta}$ del parametre θ és:

$$B_\theta\left(\hat{\theta}\right) = E_\theta\left(\hat{\theta}\right) - \theta.$$

Direm que aquest estimador és no esbiaixat si $B_\theta\left(\hat{\theta}\right) = 0 \Rightarrow E_\theta\left(\hat{\theta}\right) = \theta$

La interpretació del biaix d'un estimador és d'equivocar-se sistemàticament de la mateixa manera de θ.

Dues de les propietats necessaries que ha de tenir un estimador per ser considerat bo són:

1. Biaix petit: en general l'estimador no s'equivoca, i si ho fa, el més petit possible. Concretament, si és no esbiaixat millor.
2. Variància petita: denotarem la variància $V_\theta\left(\hat{\theta}\right)$

Aquestes dues propietats es condensen en l'expressió del **mean square error**

Definition 0.3.4 (Mean Square Error (MSE)). La mitjana de l'error quadrat (MSE) és

$$\mathrm{MSE}_\theta\big(\hat{\theta}\big) = E_\theta\Big(\big(\hat{\theta}-\theta\big)^2\Big)$$

Proposition 0.3.5 (Càlcul del MSE).

$$\mathrm{MSE}_\theta\big(\hat{\theta}\big) = V_\theta\big(\hat{\theta}\big) + B_\theta\big(\hat{\theta}\big)^2$$

La demo està feta a l'annex pàgina 1

A tall de reflexió, no sempre el millor estimador és un no esbiaixat, com per exemple l'exemple 1.4 dels apunts de la Lupe. Allà es calcula el MLE d'una mostra seguint una normal, i es veu que el MSE de l'estimador no esbiaixat no és el mateix que el MLE per la variància.

Tot i l'exemple anterior, la intuïció de cercar el millor estimador $\hat{\theta}_n$ d'entre tots els no esbiaixats és correcta, ja que trobar propietats d'entre el conjunt d'estimador que minimitzen l'MSE però son no esbiaxiats és difícil. Per tant, essent $\tau(\theta)$ una reparametrització, definim $\mathcal{C}_\theta = \{W : E_\theta(W) = \tau(\theta)\}$, on W representa l'estimador (el que fins ara ha estat $\hat{\theta}_n$) del parametre θ d'entre tots els estimadors no esbiaixats.

Definition 0.3.6 (Uniform Minimum Variance Unbiased Estimator (UMVUE)). Un estimador W^* és el millor estimador de $\tau(\theta)$ si

1. $E_\theta(W^*) = \tau(\theta) \forall \theta \in \Theta$ (equivalenment $W^* \in \mathcal{C}_\theta$)
2. $\forall W \in \mathcal{C}_\theta : W \neq W^*$ compleix $V_\theta(W^*) \leq V_\theta(W) \forall \theta \in \Theta$

0.4 Estadístics Suficients

En aquest apartat es tractarà el process de reducció de dades i la manera de trobar-ne un representat adequat.

Definition 0.4.1 (Sufficient Statistic). Sigui T un estadístic. S'anomena a T estadístic suficient si

1. la distribució condicional de la mostra $\boldsymbol{X}$ donat el valor $T(X)$ no depen de θ.

És a dir, sent $g(t)$ funció no depenent de θ
1. Si la llei d'$\boldsymbol{X}$ és discreta: $P(X = x|T = t;\theta) = g(t)$
2. Si la llei és continua: $f_{\boldsymbol{X}}(\boldsymbol{x}|T = t;\theta) = g(t)$

Un estadístic T és suficient per a θ (abreujat com a SSt a partir d'ara) si tot a inferència sobre el parametre θ depen de la mostra $\boldsymbol{X} = (X_1, ..., X_n)$ només segons el valor de $T(\boldsymbol{X})$

El raonament darrera d'aquesta definició és que ens permet definir el **Principi de Suficència**: donats dos conjunts de dades $\boldsymbol{x}, \boldsymbol{y}$ tals que $T(x) = T(y)$, aleshores tota inferència de θ ha de ser la mateixa per a tots dos conjunts de dades. És a dir, T és SSt per als dos, significa que les dades "es comporten igual" i per tant, només usant T hem d'obtenir les mateixes conclusions tant per $\boldsymbol{x}$ com per $\boldsymbol{y}$.

Nota: L'estadístic suficient NO és únic, poden haver-hi més d'un que complieixin les condicions

Aleshores, el que interessa és trobar un mètode efectiu per determinar si un estadístic depen o no del paràmetre, i el següent teorema ens hi dona resposta:

Theorem 0.4.2 (Factoritzation Theorem). Sigui $f_{\boldsymbol{X}}(\boldsymbol{x}|\theta)$ la distirbució conjunta de $X = (X_1, ..., X_n)$. L'estadístic $T(X)$ és SSt de θ sii existeixen $g(t;\theta)$ i $h(x)$ tal que $\forall \boldsymbol{x} \in \mathcal{X}, \theta \in \Theta$ podem escriue

$$f_{\boldsymbol{X}}(\boldsymbol{x}|\theta) = g(T(X);\theta)h(x)$$

0.5 Teoremes de Rao-Blackwell i Lehman-Scheffé

Aquí entrariem ara a CramerRao i a lehman sheffé per trobar l'UMVUE i la lògica que segueix, però per el midterm no ho escriuré en detall ja que és massa teòric per l'aplicació que se l'hi dona (literalment aplicar-ho)

0.6 Teorema i Cota de Cramer Rao

Per operar amb prestesa la Cramer Rao bound, necessitem primer entendre (i operar) amb els següents conceptes sobre la **informació de Fisher**

Intuïció La informació de fisher és "el contrari" de la variància, a més informació de fisher tingui X VA, més sabrem sobre la distrbució que segueix X. Considerem el següent exemple:

Example 0.6.1 (Informació de Fisher). Considerem una variable aleatòria X que representa les notes dels alumnes d'un país de la Unió Europea, on el nostre objectiu és determinar de quin país provenen aquestes notes. Assumim que disposem d'una mostra de mida equivalent i suficientment gran per a tots els països, de manera que $X{\sim}\mathcal{N}(\mu, \sigma^2)$. Paradoxalment, quan σ^2 és molt petita (la variància s'apropa a zero), la informació de Fisher també tendeix a zero. Per entendre això, considerem un exemple pràctic: A Alemanya, les notes van de l'1 al 5, mentre que a Catalunya van del 0 al 10. Suposem que tenim una mostra amb $\sigma^2 = 0$ i $\mu = 5$. En aquest cas, tots els estudiants han obtingut exactament un 5, i no podem determinar l'origen de les dades: podria ser tant un excel $\cdot$ lent a Alemanya com un aprovat just a Catalunya. D'altra banda, quan la variància augmenta (és a dir, σ^2 creix), obtenim més informació sobre la distribució subjacent de les dades. Això ens permet identificar millor el sistema d'avaluació utilitzat i, per tant, el país d'origen de les notes.

Per tant, la informació de Fisher a més gran (proporcional a la variància) més insight ens dona sobre les dades

Denotarem amb $l(\theta|\boldsymbol{x}) = \log L(\theta|\boldsymbol{x})$, ja que s'usa constantment en practicament tots els casos.

Definition 0.6.2 (Puntuació (Score Function)). Anomenem la Puntuació (score function) a la primera derivada de $l(\theta|\boldsymbol{x})$

$$S_{\boldsymbol{X}}(\theta) = S(\theta|\boldsymbol{X}) = \frac{\partial}{\partial\theta} l(\theta|\boldsymbol{X})$$

Definition 0.6.3 (Informació de Fisher). La informació de fisher que X té sobre θ és

$$i_{\boldsymbol{X}}(\theta) = \frac{\partial}{\partial\theta} S(\theta|\boldsymbol{X}) = -\frac{\partial^2}{\partial^2} l(\theta|x)$$

Definition 0.6.4 (Observed Fisher's Information). És el mateix que la informació de fisher però amb $i_{\boldsymbol{X}}\big(\hat{\theta}_{\text{MLE}}(\boldsymbol{x})\big)$.

Definition 0.6.5 (Expected Fisher's Information). La informació esperada de fisher que X té sobre θ és

$$I_{\boldsymbol{X}}(\theta) = E_\theta\big(i_{X(\theta)}\big)$$

En general, $I_{\boldsymbol{X}}(\theta)$ ser la més útil i la que dona més informació, és també a vegades molt difícil de calcular a la pràctica i s'acaba usant $i_{\boldsymbol{X}}(\theta)$