Probabilidad y estadística

Clase 4

Estadística

¿Qué es la estadística?

Ahora lo que tenemos son observaciones (realizaciones) de las variables aleatorias.

El objetivo es poder hacer algún tipo de inferencia a partir de los valores observados.

Muestra aleatoria

Supongamos que tenemos un experimento aleatorio relacionado con una v.a. X. La v.a. X representa un observable del experimento aleatorio.

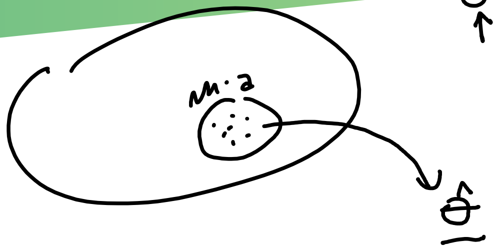
Los valores de X son la población de estudio, y el objetivo es saber como se comporta esa población.

Una muestra aleatoria de tamaño n, es una sucesión de n v.a **independientes** $\underline{X}_n = (X_1, \ldots, X_n)$, tal que $X_i \sim X$

Estimadores puntuales

Estimadores puntuales

$\underline{X}$ población , $F_\theta(x)$



Def: Dada una muestra aleatoria $\underline{X}_n$, un estimador $\hat{\theta}$ de un parámetro θ es una función de la muestra aleatoria que provee un valor aproximado del parámetro o característica desconocido.

$\hat{\theta} = \hat{\theta}(\underline{X}_n)$ es una v.a

Bondades de los estimadores

$X \sim F_\theta(x)$

el valor poblacional

m.a.

$\int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \left(\hat{\theta}(\overbrace{x_1, \dots x_n}^{\text{m.a.}}) - \theta\right) f(x_1) f(x_2) \dots dx_1 \dots$

es una v.a.

fijo

- **Def:** $\hat{\theta}$ es un estimador **insesgado** para θ si $B := \mathbb{E}[\hat{\theta} - \theta] = 0 \; \boxed{\forall \theta}$
En caso contrario diremos que es **sesgado**. A B se lo conoce como **sesgo**

1) asumo θ fijo, 2) $X_1, \dots X_n$ es m.a. con $X_i \sim X \sim F_\theta(x)$

2) calculo $E[\hat{\theta}(X_1, \dots X_n)] - \theta = B$

$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i$ es un estimador de μ

$X \sim N(\mu, 1)$

Def: $\hat{\theta}$ es un estimador **asintóticamente insesgado** para θ si

$$\lim_{n\to\infty} \mathbb{E}[\hat{\theta} - \theta] = 0 \;\; \forall\, \theta$$

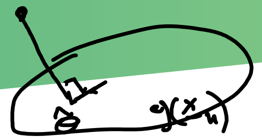
linealidad

$B = E[\hat{\mu}] - \mu \overset{\text{linealidad}}{=} \frac{1}{n} \sum \overbrace{E X_i}^{\mu} - \mu$

$\overset{\text{iid}}{=} \frac{1}{n} \sum_{i=1}^{n} \mu - \mu = 0$

Bondades de los estimadores

$\langle X, Y\rangle = E[X \cdot Y] \to$ Espacio de Hilbert



Def: El error cuadrático medio (ECM) como $\mathbb{E}[(\hat{\theta}-\theta)^2] = \langle \hat{\theta}-\theta, \hat{\theta}-\theta\rangle$

$= \underbrace{\|\hat{\theta}-\theta\|^2}_{\text{error}}$

$E[((\hat{\theta} - E\hat{\theta}) + (E\hat{\theta} - \theta))^2]$

$= E[(\hat{\theta} - E\hat{\theta})^2] + 2\underbrace{(E[\hat{\theta}] - E\hat{\theta})}_{0}(E\hat{\theta} - \theta)$

$+ (\underbrace{E\hat{\theta} - \theta}_{B})^2$

Obs: El ECM se puede descomponer como:

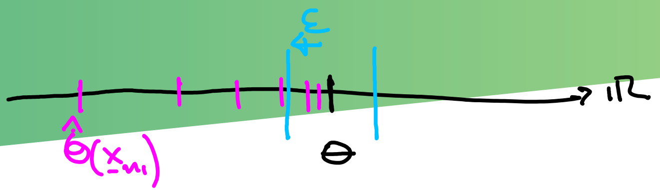
compromiso sesgo - varianza $\leftarrow$

$$ECM = \underbrace{var(\hat{\theta})}_{varianza} + \underbrace{B(\hat{\theta})^2}_{sesgo}$$

donde $var(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]$ y $B = \mathbb{E}[\hat{\theta} - \theta]$

Def: Un estimador $\theta^*(\underline{X}_n)$ es **óptimo** (en media cuadrática) si $\underline{ECM(\theta^*) \leq ECM(\hat{\theta})}$ para todo $\hat{\theta}(\underline{X}_n)$

Bondades de los estimadores



Def: Dada una sucesión de estimadores $\hat{\theta}_n$ de θ, diremos que $T = \hat{\theta}$ es (débilmente) consistente si

$$\forall\, \varepsilon > 0, \mathbb{P}(|\hat{\theta} - \theta| > \varepsilon) \xrightarrow[n\to\infty]{} 0$$

Teorema: Si $var(\hat{\theta}) \to 0$ y $\underbrace{\mathbb{E}[\hat{\theta}] \to \theta}_{B=0}$, entonces $\hat{\theta}$ es consistente.

Def: Un estimador es consistente en media cuadrática si

$$\lim_{n\to\infty} \underbrace{ECM(\hat{\theta}_n)}_{var(\hat\theta) + B^2(\hat\theta)} = 0, \forall\theta$$

$var(\hat\theta) \to 0$, $B^2(\hat\theta) \to 0$

Ejercicio 1

Se desea estimar la media de una variable con distribución N(μ,9) a partir del promedio de n realizaciones. Analizar las bondades de las que goza dicho estimador.

$X \sim N(\mu, 9)$

$\underline{X}_n$ es una m.a. de tamaño n de X ($X_1, \dots, X_n$ son iid y $X_i \sim X$)

$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i$, $\cdot\, E\,\hat{\mu} = \mu \Rightarrow \hat{\mu}$ es insesgado.

$Var(\hat{\mu}) = Var\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} Var\left(\sum X_i\right) \overset{iid}{=} \frac{1}{n^2} \sum_{i=1}^{n} \overbrace{Var\, X_i}^{9} = \frac{1}{n^2} \cdot n \cdot 9 = \frac{9}{n} \xrightarrow[n \to \infty]{} 0$

$\therefore$ el promedio es consistente

Algunos comportamientos límite

LGN + TCL

Ley (débil) de los grandes números

Sea $\underline{X} = (X_1, \ldots, X_n)$ una muestra aleatoria, con X_i i.i.d. con media μ y varianza σ^2. Para $\epsilon > 0$ se tiene que

promedio

$$\mathbb{P}(|\bar{X}_n - \mu| > \varepsilon) \xrightarrow[n \to \infty]{} 0$$

Es decir, a medida que crece la cantidad de muestras, el promedio $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ tiende a la media real de la distribución.

Teorema central del límite

Sea $\underline{X} = (X_1, \ldots, X_n)$ una muestra aleatoria, con X_i i.i.d. con media μ y varianza $\sigma^2 < \infty$. Sea $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, luego

converge en distribución

$$Z_n = \frac{\bar{X}_n - \mu}{\sqrt{\mathbb{V}(\bar{X}_n)}} = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \rightsquigarrow Z$$

con $Z \sim \mathcal{N}(0,1)$, o equivalentemente:

fun de distr. de la Normal (0,1)

$$\lim_{n\to\infty} \underline{\mathbb{P}(Z_n \leq z)} = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

Fun. de distr. de promedio estandarizado.

Es decir, $\bar{X}_n$ tiene una distribución aproximadamente Normal, con media μ y variancia σ^2/n . Para n finito, podemos aproximar

$$\mathbb{P}\left(\frac{\sum_{i1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \leq z\right) \approx \Phi(z), \qquad \mathbb{P}\left(\sqrt{n}\frac{\bar{X}_i - \mu}{\sigma} \leq z\right) \approx \Phi(z),$$

Enfoque frecuentista

Estadístico suficiente → Reducción de dimensiones.

Def: Dada una muestra aleatoria $\underline{X}_n$, un estadístico es cualquier función $T_n = T(\underline{X}_n)$

- **Def:** Sea una muestra aleatoria $\underline{X}_n$, cuya distribución es $\boxed{F_\theta(\underline{x})}, \theta \in \Theta$ → esp. de parámetros., se dice que $T = r(\underline{X}_n)$ es un estadístico suficiente para θ si $F_{\underline{X}|T=t}(\underline{x})$ no depende de θ.

- **Teorema de factorización:** Diremos que $T = r(\underline{X}_n)$ es un est. suficiente para θ sii existen funciones h y g tales que:

$$\boxed{f_\theta(\underline{x})} = \underbrace{g(r(\underline{x}), \theta)}\underbrace{h(\underline{x})}$$

Ejercicio 2

1. Hallar un estadístico suficiente para el parámetro de una v.a. con distribución exponencial $\rightarrow \sum x_i$
2. Hallar un estadístico suficiente para el parámetro de una v.a. con distribución de Bernoulli $\rightarrow \sum x_i$
3. Hallar un estadístico suficiente para el parámetro de una v.a. con distribución Uniforme en el intervalo (0,θ)

1) $$f_{\underline{x}}(\underline{x}) \underset{\text{indep.}}{=} \prod_{i=1}^{n} f_{x_i}(x_i) \underset{\text{id.}}{=} \prod_{i=1}^{n} \lambda e^{-\lambda x_i} \mathbb{1}\{x_i \geq 0\} = \underbrace{\lambda^n e^{-\lambda \overbrace{\sum x_i}^{=T(\underline{x}_n)}}}_{g(T,\theta)} \underbrace{\prod_{i=1}^{n} \mathbb{1}\{x_i \geq 0\}}_{h}$$

2) $$f_{\underline{x}}(\underline{x}) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = \underbrace{p^{\overbrace{\sum x_i}^{T}} (1-p)^{n-\sum x_i}}_{g}$$

$X_i \sim U(0,\theta)$

$$f_{\underline{X}}(\underline{x}) = \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{1}\{0 \le x_i \le \theta\} = \frac{1}{\theta^n} \mathbb{1}\{\min(x_i) \ge 0\} \mathbb{1}\{\max(x_i) \le \theta\}$$

$$g(T,\theta)$$

$$\therefore \boxed{T(\underline{X}_n) = \max(X_i)}$$

Método de máxima verosimilitud

Método de Máxima Verosimilitud

Def: Diremos que $\hat{\theta}(\underline{X})$ es un Estimador de Máxima Verosimilitud (MLE) si se cumple que:

$$f(\underline{X}, \hat{\theta}) = \max_{\theta} f_{\theta}(\underline{X})$$

La idea es que si observé una determinada muestra, entonces esta debería tener alta probabilidad de ocurrir, por lo tanto busco el θ que maximiza esa probabilidad de ocurrencia

Método de Máxima Verosimilitud

Def: Definimos la función de verosimilitud como

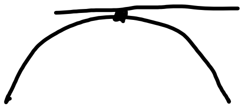
$L(\theta) = f(\underline{x}, \theta)$ (vista como función de θ) luego,

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta).$$

$\theta \in \Theta \rightarrow$ espacio de parámetros

Si el soporte de X no depende θ, Θ es un conjunto abierto y $f_\theta(x)$ es derivable respecto de θ, entonces para hallar el EMV puedo hallar θ tal que

$$\left\{ \frac{\partial \, ln(L(\theta))}{\partial \, \theta} = 0 \right\}$$



Ejercicio 3

La probabilidad de acertar a un blanco es p. Se realizan 10 tiros independientes, en los cuales se observaron 4 aciertos. A partir este valor observado estimar el valor de p por MV.

$X \sim Ber(p)$, $0 \leq p \leq 1$.

$\underline{X}_n$ con $n = 10$; $T = \sum_{i=1}^{n} x_i = 4$

$$\mathcal{L}(p) = \prod_{i=1}^{10} p^{x_i} (1-p)^{1-x_i} = p^{\sum x_i} (1-p)^{10 - \sum x_i} = p^4 (1-p)^6$$

$$\ell(p) := \log \mathcal{L}(p) = 4 \log p + 6 \log(1-p)$$

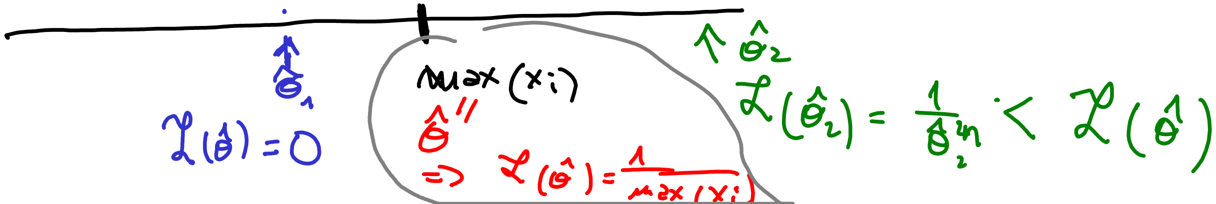
$$\frac{d}{dp} \ell(p) = \frac{4}{p} - \frac{6}{1-p} = 0 \Rightarrow 4(1-p) = 6p \Rightarrow 4 = 10p \Rightarrow \boxed{\hat{p} = 0,4}$$

Ejercicio 4

1. Sea X~U(0, θ). Hallar el EMV para θ basado en una muestra de tamaño n. $\Rightarrow \hat{\theta} = \max(x_i)$
2. Si en una muestra de tamaño 3 se observaron los valores (0.756, 1.1, 3,11), hallar el valor estimado de θ. $\Rightarrow \hat{\theta} = 11$

$X \sim \text{Unif}(0,\theta)$ y $\underline{X}_n$ es una m.a. con $X_i \sim X$.

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{1}\{0 \le x_i \le \theta\} = \frac{1}{\theta^n} \mathbb{1}\{\min(x_i) \ge 0\} \mathbb{1}\{\underbrace{\max(x_i) \le \theta}\}$$



Principio de invariancia

Supongamos que ahora queremos estimar por máxima verosimilitud a $\lambda = q(\theta)$.

una función general

Teorema: Si $\hat{\theta}$ es MLE de θ, entonces $\hat{\lambda} = q(\hat{\theta})$.

¿Por qué es útil? Por ejemplo, podría querer estimar una probabilidad de la v.a. X, que en general no puedo porque desconozco el parámetro de la distribución.

$P(X \in A) = \int_A f_\theta(x)\,dx = q(\theta) \Rightarrow \hat{P}(X \in A) = q(\hat{\theta}) = \int_A f_{\hat{\theta}}(x)\,dx$

Ejercicio 5

Siguiendo el ejercicio 3, estimar la probabilidad de que se necesiten al menos 2 tiros para observar el primer acierto.

Habíamos visto que el MLE de p es $\hat{p} = 0,4$.

$$Y \sim Geom(p)$$

$$P(Y \geq 2) = 1 - P(Y < 2) = 1 - P(Y = 1) = 1 - p$$

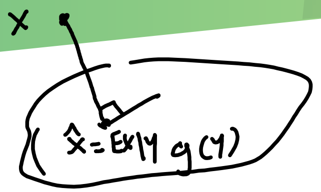
Por T. de invarianza funcional,

$$\hat{P}(Y \geq 2) = 1 - \hat{p} = 1 - 0,4 = 0,6$$

Estimadores de cuadrados mínimos

Estimador de cuadrados mínimos

$\langle x, y \rangle = E[XY] \Rightarrow$ E.H.



Es común querer estimar el valor de una v.a. X a partir de una medición Y. Ejemplo: Y es una versión ruidosa de X.

Buscamos un estimador $\hat{X}$ de X tal que tenga mínimo error cuadrático medio

$\hat{X}(Y)$

$$\mathrm{ECM} = \mathbb{E}[(X - \hat{X})^2 \mid Y] = \|X - \hat{X}\|^2 \mid Y$$

Observar que se corresponde con la distnacia asociada al p.i. canónico para v.a.

Estimador de mínimos cuadrados

En otras palabras, queremos $\hat{X} = g^*(Y)$ tal que

$$\mathbb{E}[(X - \hat{X})^2 \mid Y] \leq \mathbb{E}[(X - g(Y))^2 \mid Y] \qquad \forall g(Y) \text{ (medible)}$$

¿Quién era $\hat{X}$? $\boxed{\hat{X} = \mathbb{E}[X|Y]}$ $\longrightarrow$ es general

Idea de demostración: [Ejercicio]

$\Rightarrow g(Y) = c$

1. Probar que el mejor estimador constante es $\mathbb{E}[X]$
2. Probar que el mejor estimador condicional es $\mathbb{E}[X|Y = y]$.
3. Dejar que Y tome todos los valores posibles (i.e. reemplazo y por Y), recupero la esperanza condicional.

1) $$\frac{d}{dc} E\left[(X-c)^2\right] = E\left[2(X-c)\right] = 0$$

$$\Rightarrow \quad E[X] = c \, .$$

2) $$\frac{d}{dg(y)} E\left[(X - \underbrace{g(y)}_{\text{es cte}})^2 \,\middle|\, Y=y\right] = E\left[2(X-g(y)) \,\middle|\, Y=y\right] = 0$$

$$\Rightarrow \quad E\left[X \,\middle|\, Y=y\right] = g(y)$$

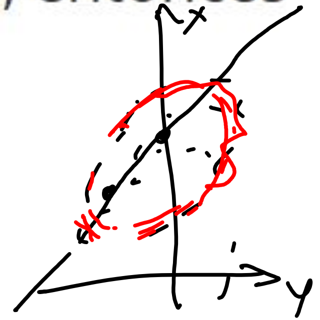
3) $$g(Y) = E\left[X \,\middle|\, Y\right] .$$

Mínimos cuadrados: caso lineal

(x) $\hat{x} = \frac{cov(x,y)}{var(y)}\left(Y - E[Y]\right) + E[X]$

necesito $f_{X|Y} = \frac{f_{XY}}{f_Y}$

A veces obtener $\mathbb{E}[X|Y]$ puede ser muy complicado, entonces nos restringimos a los estimadores lineales.



Asumimos modelo lineal →

Buscamos a, b tq $\mathbb{E}[(X - \overbrace{(aY + b)}^{g(y)})^2 \mid Y]$ sea mínima.

(x) Resulta que $a = \frac{cov(X,Y)}{var(Y)}$ y $b = \frac{cov(X,Y)}{var(Y)}\mathbb{E}[Y] + \mathbb{E}[X]$

Obs: Si se asume que X e Y son conjuntamente gaussianas, el estimador de mínimos cuadrados (la esperanza condicional) coincide con el estimador de mínimos cuadrados asumiendo un modelo lineal y puede obtenerse a partir del modelo condicional de $X \mid Y$

∴ en el caso gaussiano $E[X|Y]$ es lineal ⇒ el estimador lineal es óptimo.

es lineal

$(x_1, x_2) \sim N\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right) \Rightarrow x_1|x_2 \sim N\left(\boxed{E[x_1|x_2]}, var(x_1|x_2)\right)$

$E[(X-EX)(Y-EY)] = E[XY] - EX\,EY \cancel{- EX\,EY} \cancel{+ EX\,EY}$

Ejercicio 6

$E[(Y-\hat{Y})^2 \mid X]$

Sea X~U(0,1) e Y=X². Hallar la mejor aproximación lineal de Y basada en X. Comparar con la mejor estimación de Y basada en X.

$$\hat{Y} = \frac{Cov(X,Y)}{Var(X)}(X - E[X]) + E[Y].$$

$$Cov(X,Y) = E_{XY}[(X-EX)(Y-EY)] = E[XY] - \overset{1/2}{E[X]}\,E[Y] \qquad \longleftrightarrow \quad Var(X) = E(X^2) - (E[X])^2$$

$$E_{XY}[XY] = E_X[E_Y[XY \mid X]] = E[X\,E[Y \mid X]] = E[X\,X^2] = E[X^3] = \int_0^1 x^3\,dx = \frac{x^4}{4}\Big|_0^1 = \frac{1}{4}$$

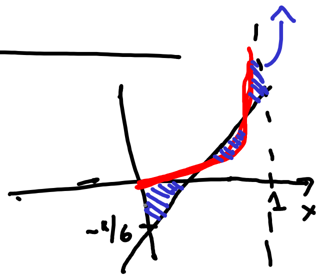
$$E[Y] = E[E[Y \mid X]] = E[X^2] = \int_0^1 x^2\,dx = \frac{x^3}{3}\Big|_0^1 = \frac{1}{3}$$

$$Var(X) = \frac{1}{12}$$

$$Cov(XY) = \frac{1}{4} - \frac{1}{2}\cdot\frac{1}{3} = \frac{1}{12}$$

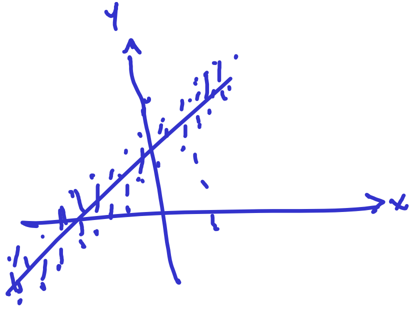
$$\boxed{\hat{Y} = \frac{1/12}{1/12}\cdot\left[X - 1/2\right] + \frac{1}{3} = X - 1/6}$$

se minimiza



Regresión lineal (OLS)

x, e y son multivariadas.



Contamos con n observaciones conjuntas $(y_1, x_1), (y_2, x_2), \cdots, (y_n, x_n)$
el modelo de regresión lineal es

es lo aleatorio

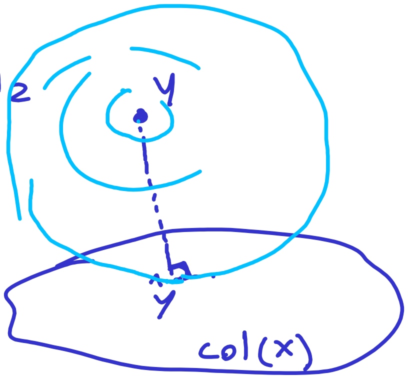
$$\boxed{\underbrace{Y}_{n\times q} = \underbrace{X}_{n\times p+1}\underbrace{\beta}_{p+1\times q} + \varepsilon} \quad \text{con} \quad \mathbb{E}(\varepsilon_i \mid X_i) = 0.$$

donde

Matrices de datos.

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}, \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\min \|Y - \hat{Y}(x)\|_2 = \min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$



El estimador de cuadrados mínimos de $\hat{\beta}$ está dado por

$$\boxed{\hat{\beta} = (X^\top X)^{-1} X^T Y}$$

Obs: Si se asume que $\varepsilon_i \sim \boxed{\mathcal{N}(0, \sigma^2)}$ y son i.i.d., este estimador coincide con el estimador de máxima verosimilitud. → $\underline{Y} | \underline{X} \sim \mathcal{N}\left(\beta_0 + \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_q \end{pmatrix}^T X, \sigma^2 I\right)$

Bibliografía

Bibliografía

- Notas de estadística. Graciela Boente y Víctor Yohai
- “All of Statistics : A Concise Course in Statistical Inference”, Larry Wasserman, Springer-Verlag New York Inc.
- “Mathematical statistics with applications”, Mendenhall, W., Scheaffer, R. L., & Wackerly, D. D. Boston: Duxbury Press.