Para estimar $F(x)$ usando la información de $X_1, X_2, \dots, X_n$ m.a. de $X \sim F$ se construye la función de distribución acumulada empírica.

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \leq x), \quad \text{para cada } x \in \mathbb{R}$$

en donde

$$\mathbb{1}(X_i \leq x) = \begin{cases} 1, & \text{si } X_i \leq x \\ 0, & \text{e.o.c.} \end{cases}$$

Teorema

Para cada $x \in \mathbb{R}$, se tiene que

1) $\mathbb{E}(\hat{F}_n(x)) = F(x)$

2) $Var(\hat{F}_n(x)) = \dfrac{F(x)(1-F(x))}{n}$

3) $ECM_{F(x)}(\hat{F}_n) = Var(\hat{F}_n(x))$

4) $\hat{F}_n(x) \xrightarrow{P} F(x)$, en otras palabras $\hat{F}_n(x)$ es un estimador consistente para $F(x)$ en todo punto $x \in \mathbb{R}$.

Todo esto nos indica que $\hat{F}_n(x)$ es un buen estimador de $F(x)$, pero tenemos algo aún más ¡poderoso!

Teorema (Glivenko-Cantelli)

Sea $X_1, X_2, \dots, X_n$ una m.a. de $X \sim F(x)$. Para $\varepsilon > 0$

$$\Rightarrow \sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F(x) \right| \xrightarrow{P} 0$$

$\Rightarrow$ $\hat{F}_n(x)$ es un muy buen estimador de $F(x)$

Funcionales estadísticas

¿Por qué es relevante estimar $F(x)$?

Si X es capaz de describir el fenómeno de interés

$\Rightarrow$ una cantidad de interés $\mu = \mathbb{E}(X)$

$$\mu = \mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x)\, dx = \int_{-\infty}^{\infty} x \, dF(x)$$

recordemos que $f(x) = \frac{d}{dx} F(x)$

$\Rightarrow$ ¡Claramente μ y muchos otros parámetros de interés son funciones de $F(x)$!

Un funcional estadístico $T(F)$ es cualquier función de F

$$\mu = \mathbb{E}(X) = \int x \, dF(x)$$

$$\sigma^2 = Var(X) = \int (x-\mu)^2 \, dF(x)$$

La mediana

$$m = F^{-1}(1/2)$$

Definición (estimador plug-in)

El estimador plug-in de $\theta = T(F)$ se define como

$$\hat{\theta}_n = T(\hat{F}_n)$$

en otras palabras, en donde veamos F en el funcional estadístico sustituimos por $\hat{F}_n$

Existe una clase especial de funcionales estadísticos $T(F)$, los llamados funcionales estadísticos lineales. Cualquier funcional estadístico de la forma

$$T(F) = \int g(x) \, dF(x) = \int g(x) f(x) \, dx$$

es un funcional estadístico lineal.

¿Qué tienen de especial estas funcionales estadísticas lineales?

Teorema (estimador plug-in)

Sea $X_1, X_2, \dots, X_n$ una m.a. de $X \sim F(x)$ y

$$\theta = T(F) = \int g(x)\, dF(x)$$

un funcional estadístico lineal

$\Rightarrow$ para obtener un estimador de θ sólo insertamos (plug-in) $\hat{F}_n(x)$ en donde veamos $F(x)$, i.e.

$$\hat{\theta}_n = T(\hat{F}_n) = \int g(x)\, d\hat{F}_n(x)$$

$$= \frac{1}{n} \sum_{i=1}^{n} g(X_i)$$

resultados de teoría de la medida

¡Sí, un promedio!

$\boxed{\text{Ejemplos}}$

1) Media

$$\mu = \mathbb{E}(X) = \int x\, dF(x)$$

$$\Rightarrow \hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

2) Varianza

$$\sigma^2 = Var(X) = \int (x-\mu)\, dF(x)$$

$$= \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

$$= \int x^2\, dF(x) - \mu^2$$

$$\Rightarrow \quad \hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \hat{\mu}_n^2$$

Observen que

$$= \frac{1}{n}\sum (x_i - \hat{\mu}_n)^2$$

$$\frac{1}{n}\sum (x_i - \bar{X}_n)^2 = \frac{1}{n}\sum (x_i^2 - 2\bar{X}_n x_i + \bar{X}_n^2)$$

$$= \frac{1}{n}\sum x_i^2 - \frac{2}{n}\bar{X}_n \sum x_i + \bar{X}_n^2$$

$$= \frac{1}{n}\sum x_i^2 - 2\bar{X}_n^2 + \bar{X}_n^2$$

$$= \frac{1}{n}\sum x_i^2 - \bar{X}_n^2$$

3) Asimetría $\quad K = \dfrac{\mathbb{E}((X-\mu)^3)}{\sigma^3} = \dfrac{\int (x-\mu)^3\, dF(x)}{\sigma^3}$

$$\Rightarrow \hat{K}_n = \frac{\frac{1}{n}\sum (X_i-\hat{\mu}_n)}{\left(\frac{1}{n}\sum (X_i-\hat{\mu}_n)^2\right)^{3/2}}$$

4) Correlación

Sean $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ una m.a. de $(X, Y) \sim F(x,y)$

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{\mathbb{E}\left((X-\mathbb{E}(X))(Y-\mathbb{E}(Y))\right)}{\sigma_X \, \sigma_Y}$$

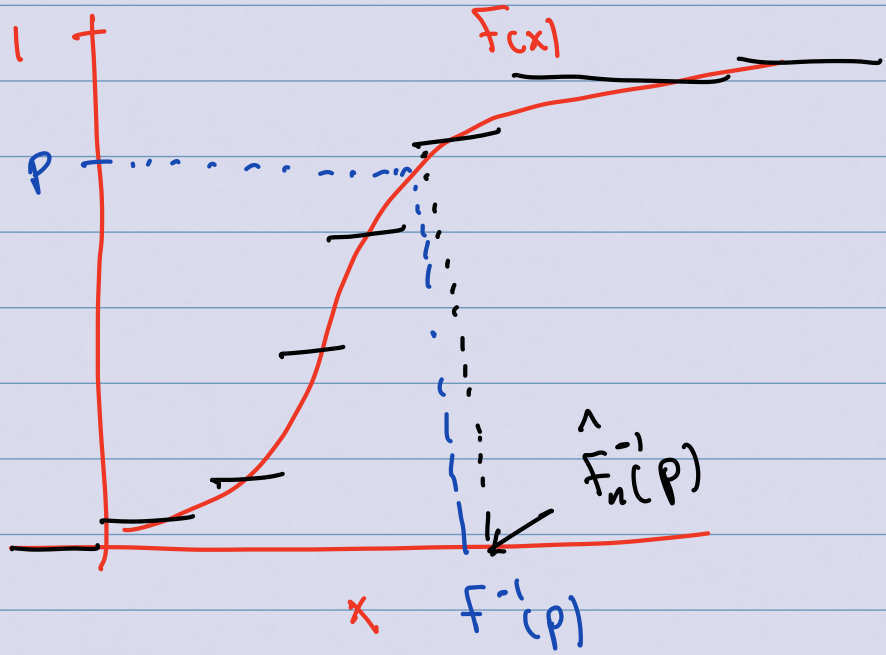
$$\hat{\rho}_n = \frac{\frac{1}{n}\sum (X_i-\hat{\mu}_X)(Y_i-\hat{\mu}_Y)}{\hat{\sigma}_{nX}\,\hat{\sigma}_{nY}}$$

5) Cuantiles (no es un funcional estadístico lineal)

Sea $F(x)$ una función de distribución estrictamente creciente. Para $0<p<1$ el cuantil p-ésimo se define como

$$T(F) = F^{-1}(p) \Rightarrow T(\hat{F}_n) = \hat{F}_n^{-1}(p)$$

en donde $\hat{F}_n^{-1}(p) = \inf\{x : \hat{F}_n(x) \geq p\}$



Ya tenemos una forma de obtener <u>estimadores puntuales</u> de muchos parámetros de una cierta población de interés $X \sim F(x)$

$$\underline{\theta = T(F)}$$

Pero, ¿cómo obtengo estimadores por intervalo?

si $\mu = \mathbb{E}(X)$ y sabemos que $\sigma^2 = Var(X) < \infty$

podemos utilizar directamente el <u>TCL</u>

Ok y si por alguna razón queremos intervalo de confianza para

$$G = \sqrt{Var(X)}$$

$$K = \frac{\mathbb{E}\left((X-\mu)^3\right)}{G^3}$$

Tenemos $\hat{G}_n$ y $\hat{K}_n$ asumiendo normalidad (no sabemos si eso es cierto)

$\Rightarrow$ aun es necesario

$$se(\hat{G}_n) = \sqrt{Var(\hat{G}_n)} = Var\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2}\right)$$

$$se(\hat{K}_n) = \sqrt{Var(\hat{K}_n)}$$

Estimar $\underline{Var(\hat{\theta}_n)}$ cuando $\underline{\hat{\theta}_n}$ es un estimador <u>NO</u> trivial, es uno de los retos más grandes en estadística.

En este caso utilizamos el Bootstrap