Análisis Estadístico

Bootstrap

San Andrés

May 30, 2022

Cálculo numérico del EMV

Example

Sea $X_1, \ldots X_n \sim \Gamma(\alpha, \lambda)$ esto es

$$f(x; \alpha, \lambda) = \frac{1}{\Gamma(\alpha)} \lambda^{\alpha} x^{\alpha - 1} e^{-\lambda x}, \quad 0 \leq x < +\infty.$$

1. Encontrar los estimadores de momentos para α y λ.
2. Definir en R una funcón que calcula el logaritmo de la verosimilitud de (α, λ) para una muestra dada.
3. Hallar ambos estimadores para los datos del dataset `gamma-arrivals.csv`.
4. Graficar el histograma de la muestra y las densidades estimadas.
5. Para $\alpha = 2$ y $\lambda = 0.5$ Calcular 1000 muestras de tamaño $n = 20$ y graficar los histogramas de ambos estimadores en cada muestra.

Sugerencias: Utilizar las funciones `gamma()`, `rgamma()` y `optim()`.

Función de distribución empírica

Sea $X_1, \dots X_n$ una muestra aleatoria con distribución F llamamos distribución empírica

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,t]}(X_i).$$

Es decir, para cada t

$$F_n(t) = \text{ Proporción de la muestra menor o igual a t.}$$

Función de distribución empírica

Propiedades

Para cada t sabemos la distribución de $F_n(t)$.

$$F_n(t) \sim Bi\left(n, P(X \leq t)\right).$$

Además la distribución empírica es la distribución acumulada de la variable aleatoria Y que tiene como puntos de masa los valores de la muestra:

$$p_Y(t) = \frac{\#\{x_i = t\}}{n}.$$

Distribución empírica

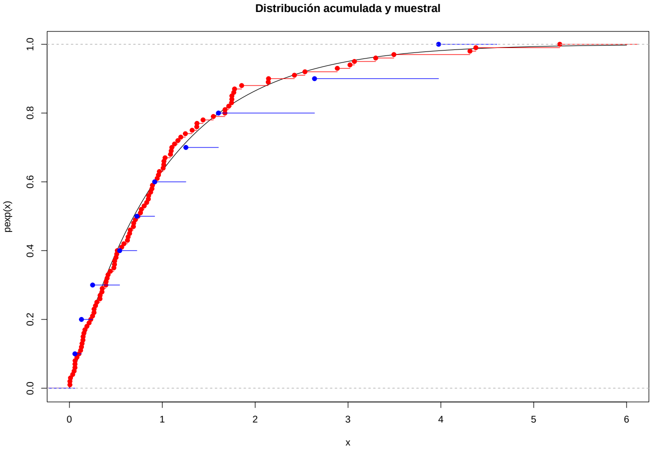


Figure: Función acumulada de $X \sim Exp(1)$ y distribuciones empíricas para $n = 10$ (azul) y $n = 100$ (rojo)

Distribución empírica - Método de plug-in

Método de plug in

A partir de F_n se puede estimar parametros de F.

$$\theta = T(F) \implies \widehat{\theta} = T(F_n).$$

- $\widehat{E(X)} = E(Y) = \frac{1}{n}\sum_{i=1} X_i$
- $\widehat{Var(X)} = Var(Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$
- $F^{-1}(0.9) = F_n^{-1}(0.9)$

A diferencia de momentos y máxima verosimilitud no noecesitamos un modelo paramétrico.

Estimación de la función de distribución

Example (Distribución de una Exponencial)

Estimar de dos manera las función de distibución acumulada de esta muestra $X_1, \ldots X_{10} \sim Exp(\lambda)$.

`2.64, 0.13, 0.73, 1.60, 0.92, 1.25, 0.06, 0.25, 0.54, 3.98`

Como $F(t) = 1 - e^{-\lambda t}$ entonces $\widehat{F} = 1 - e^{-\widehat{\lambda} t}$

Estimación de la función de distribución

Example (Distribución de una Exponencial)

Estimar de dos manera las función de distibución acumulada de esta muestra $X_1, \ldots X_{10} \sim Exp(\lambda)$.

```
2.64, 0.13, 0.73, 1.60, 0.92, 1.25, 0.06, 0.25, 0.54, 3.98
```

Como $F(t) = 1 - e^{-\lambda t}$ entonces $\widehat{F} = 1 - e^{-\widehat{\lambda} t}$

Distribución empírica

Example

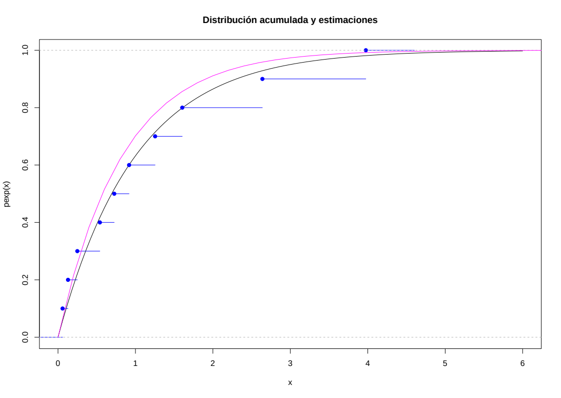


Figure: Función acumulada de $X \sim Exp(1)$ y dos estimaciones de la muestra

Bootstrap

Bootsrap es una herramienta que permite estimar parámetros de la distribución de un estimador utilizando únicamente la información de la muestra.

$$X_1, \dots X_n \sim F$$
$$\widehat{\theta} = g(X_1, \dots X_n)$$

Como sabemos $\widehat{\theta}$ es una variable aleatoria con una distribución.

$$\widehat{\theta} \sim G = g(F)$$

La distribución del estimador depende de la distribución F y de la función g que define al estimador.
¿Cómo estimamos un parámetro $T(G)$)? e.g

$$E(\widehat{\theta}) \quad SD(\widehat{\theta}) \quad G^{-1}(0.9).$$

Bootstrap

Estimación de un parámetro de la distribución del estimador

1. Estimamos el parámetro por plug-in:

$$\widehat{T(G)} = T(\widehat{G}).$$

2. Donde $\widehat{G} = g(\widehat{F})$.
   ¿Cómo llevamos esto a la práctica?

Bootstrap

Estimación del desvío de un estimador

1. $x_1, x_2, \ldots x_n$ es la realización de una muestra de distribución F.
2. $\widehat{F}$ es la estimación de F (por cualquier metodología)
3. Sorteamos una nueva muestra $x_1^\star, x_2^\star, \ldots x_n^\star \sim \widehat{F}$.
4. Calculamos $\widehat{\theta}^\star = g(x_1^\star, \ldots x_n^\star)$
5. Repetimos 3 y 4 N_{boot} veces para obtener una muestra

$$\widehat{\theta}_1^\star, \widehat{\theta}_2^\star, \ldots, \widehat{\theta}_{N_{boot}}^\star \sim \widehat{G}$$

6. Estimamos el parámetro de interes con la muestra bootstrap

Bootstrap

Por ejemplo

$$\widehat{E(\widehat{\theta})} = \frac{1}{n} \sum_{j=1}^{N_{boot}} \widehat{\theta}_j^{\star}$$

$$\widehat{SD(\hat{\theta})} = \sqrt{\frac{1}{n} \sum_{j=1}^{N_{boot}} (\widehat{\theta}_j^{\star} - \overline{\widehat{\theta}^{\star}})^2}$$

$$G\widehat{^{-1}(0.9)} = \text{cuantil}(0.9)\{\widehat{\theta}_1^{\star}, \ldots, \widehat{\theta}_{N_{boot}}^{\star}\}$$

Bootsrap

Bootstrap por resampleo

Para generar una muestra con distribución F_n basta sortear con reposición los elementos de la muestra. (¿Por qué?)

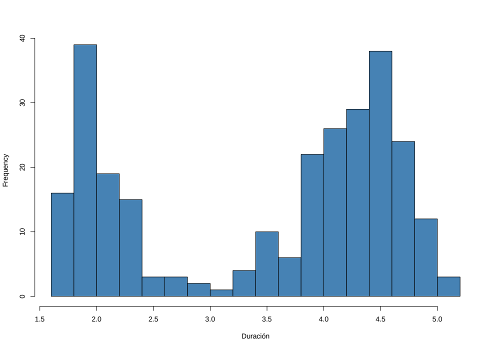
Bootsrap parametrico

Si $F = F_\lambda$ y tenemos $\widehat{\lambda}$ un estimador del parámetro podemos generar muestra bajo la distribución $F_{\widehat{\lambda}}$.

Bootstrap

Example

Old Faithful Se registrarios las duraciones en segundos de 272 erupciones consecutivas del geyser Old Faithful del parque nacional Yellowstone.



Bootsrap

Example

Old Faitfull Vamos a estimar esperanza, desvío y cuantil 0.9. Cómo no tenemos un modelo paramétrico hacemos bootstrap por resampleo

1. Datos: $x_1, \ldots x_{272}$
2. Mediana: $x_{mediana} = 4$
3. Generamos una muestra bootstrap $x_1^\star, \ldots x_{272}^\star \sim F_n$.
4. Calculamos $x_{mediana}^\star$
5. Repetimos 3 y 4 $N_{boot} = 1000$ veces.
6. Calculamos los estimadores plug-in de la muestra.

Bootstrap

Example

Simulación Estimar el desvío estandard mediante bootstrap de los estimadores de máxima verosimilitud y momentos con los datos de `Gamma-arrivals.csv`