XS3310 Teoría Estadística

I Semestre 2021

Escuela de Estadística

2021-05-17

class: center, middle

¿Qué hemos visto hasta ahora?

Todo sobre estimadores puntuales + pivotes e intervalos de confianza.

¿Qué vamos a discutir hoy?

Bootstrap

Bootstrap

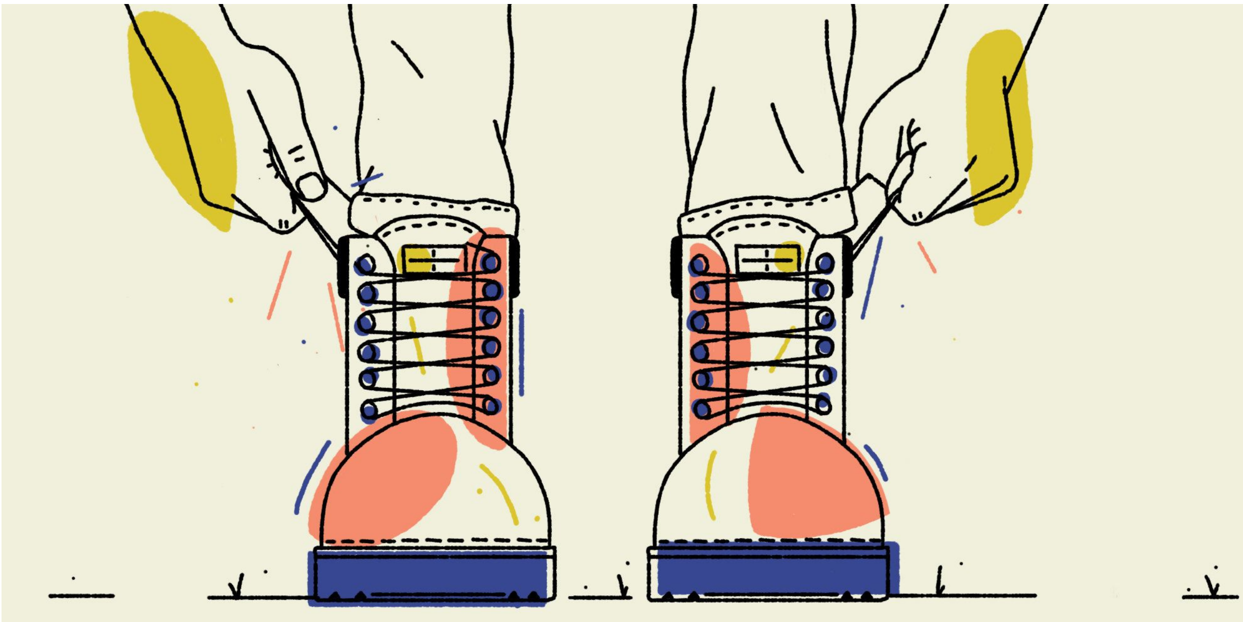
- La inferencia frecuentista se basa en modelos y supuestos. En muchos casos, las expresiones acerca de la exactitud (tales como el error estándar) están basadas en teoría asintótica, y por lo tanto no deberían usarse con muestras pequeñas.
- En otros casos, no estamos usando teoría asintótica, pero no sabemos cómo hacer una suposición acerca de la distribución poblacional, debido a que la muestra no se parece a ninguna forma conocida.
- Una alternativa "moderna" es el método de bootstrap, introducida por Efron así casi 40 años (1979). Bootstrap es un método de remuestreo que es computacionalmente intensivo, y que es aplicable a una gran variedad de casos, incluyendo aquellos en los que los supuestos son más realistas.

Visualmente: https://seeing-theory.brown.edu/frequentist-inference/es.html

---

Bootstrap

¿De dónde viene la expresión?



https://www.huffpost.com/entry/pull-yourself-up-by-your-bootstraps-nonsense_n_5b1ed024e4b0bbb7a0e037d4

Dr. Bradley Efron

https://www.youtube.com/embed/Cx5pgZCdDGM

Principios de Bootstrap

- Si no existe información acerca de la distribución, en la muestra observada podemos encontrar información acerca de la distribución subyacente. Por lo tanto, re-muestrear la muestra es la mejor forma de acercarnos a lo que obtendríamos si se pudiera la oportunidad de re-muestrear de la distribución poblacional.
- Suponga que una muestra $X = (X_1, \ldots, X_n)^T$ es utilizada para estimar un parámetro θ. Sea $\hat{\theta} = s(X)$ un estadístico para estimar el parámetro θ. Para hacer inferencia acerca de θ, nos interesa la distribución muestral de $\hat{\theta}$, o ciertos aspectos acerca de esa distribución: la exactitud de nuestra estimación, el intervalo de confianza, etc. En muchas aplicaciones, la distribución muestral de $\hat{\theta}$ no se puede encontrar.
- Si conociéramos la distribución poblacional P, podríamos sacar muestras $X^{(b)}, b = 1, \ldots, B$ de P usando métodos de Monte Carlo para estimar la distribución muestral del estimado. Sin embargo, si F es desconocido, entonces bootstrap sugiere que podemos aproximar ese muestreo re-muestreando nuestra muestra original. Así, podemos encontrar la distribución *empírica* del estimador.

https://seeing-theory.brown.edu/frequentist-inference/es.html

---

Distribución Empírica

Para una muestra $X_1, \ldots, X_n$ de variables aleatorias con valores reales, independientes con distribución P, definimos la distribución $\hat{P}$ como:

$\hat{P}(A) = \frac{1}{n} \sum_{i=1}^{n} 1_A(X_i)$ para $A \subseteq \mathbb{R}$.

$\hat{P}$ es la distribución empírica de la muestra X. $\hat{P}$ puede pensarse como una distribución que pone masa $1/n$ en cada observación X_i (para valores que ocurren más de una vez la masa será un múltiplo de $1/n$). Entonces, $\hat{P}$ es una distribución de probabilidad discreta con un espacio efectivo de muestreo $X_1, \ldots, X_n$.

Puede demostrarse que $\hat{P}$ es el estimador máximo verosimil no paramétrico de P, lo cual justifica que podamos estimar P con $\hat{P}$ sin tener otra información acerca de P (como por ejemplo si P pertenece a una familia paramétrica).

Distribución Empírica

Resultados teóricos

Sea $A \subseteq \mathbb{R}$ (tal que $P(A)$ está definido), entonces tenemos: $\hat{P}(A) \xrightarrow{d} P(A)$ cuando $n \to \infty$.

De forma alternativa, podemos ver este resultado como una consecuencia directa de La Ley de los Grandes Números, ya que:

$$n\hat{P}(A) = \sum_{i=0}^{n} 1_A(X_i) \sim Bin(n, P(A))$$

por lo que $\hat{P}(A)$ tiende a su valor esperado $P(A)$ cuando $n \to \infty$.

El teorema de Glivenko-Cantelli formaliza este resultado:

$$\sup_{A \in I} |\hat{P}(A) - P(A)| \to 0 \quad \text{si} \quad n \to \infty$$

donde I es el conjunto de intervalos en $\mathbb{R}$. En otras palabras, la distribución P(A) puede ser aproximada por $\hat{P}(A)$ igual de bien para toda $A \in I$.

Distribución Empírica

Muestras de una distribución empírica $\hat{P}$

Suponga que queremos una muestra iid de $\hat{P}$: $X^* = (X_1^*, \ldots, X_n^*)^T$. Como mencionamos antes, $\hat{P}$ pone masa $1/n$ en cada observación X_i. Entonces, cuando muestreamos de $\hat{P}$, la observación i-ésima X_i en la muestra original puede ser seleccionada con probabilidad $1/n$. Esto nos lleva al siguiente proceso:

- Seleccione $i_1, \ldots, i_n$ independientemente de una distribución uniforme en $1, \ldots, n$.
- Ahora haga $X_j^* = X_{i_j}$ y $X^* = (X_1^*, \ldots, X_n^*)^T$.

En otras palabras, saque una muestra aleatoria con reemplazo de la muestra original $X_1, \ldots, X_n$.

El Principio de Bootstrap

- $X = (X_1, \ldots, X_n)^T$ es una muestra aleatoria de una distribución P.
- $\theta = t(P)$ es algún parámetro de la distribución.
- $\hat{\theta} = s(X)$ es un estimador para θ.

La distribución muestral de $\hat{\theta}$ es entonces estimada por su equivalente de bootstrap:

$$\hat{P}(\hat{\theta} \in A) = P^*(\hat{\theta} \in A)$$

El Principio de Bootstrap

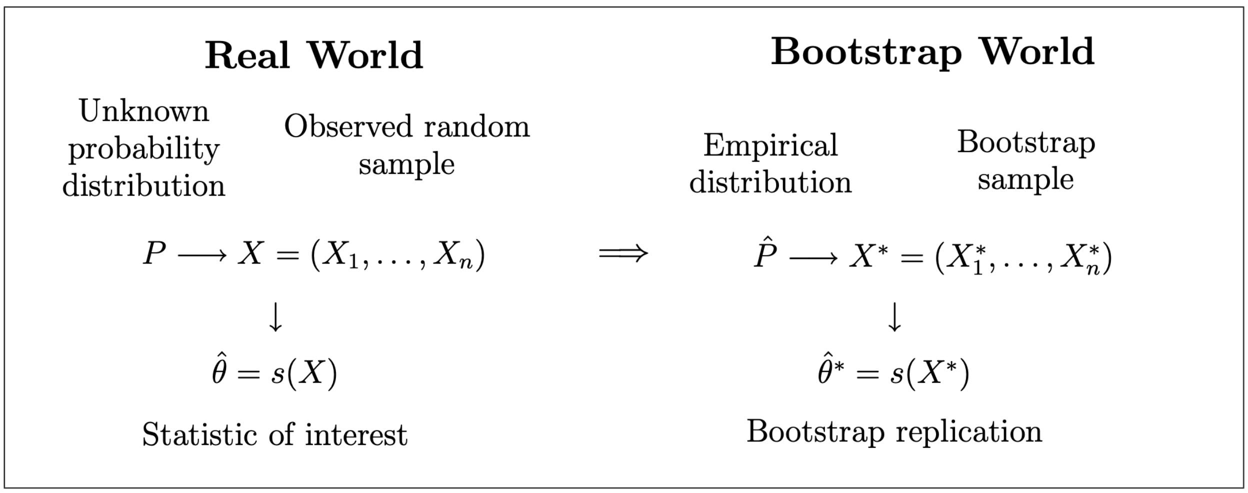


Figure 1: Diagrama

Ejemplo concreto

```r
CommuteAtlanta <- read.csv2("data/CommuteAtlanta.csv")
```

City	Age	Distance	Time	Sex
Atlanta	19	10	15	M
Atlanta	55	45	60	M
Atlanta	48	12	45	M
Atlanta	45	4	10	F
Atlanta	48	15	30	F
Atlanta	43	33	60	M

La aproximación de Monte Carlo

- En algunas ocasiones la forma de la distribución poblacional es conocida, pero la evaluación de la distribución exacta de la distribución muestral no es calculable.
- El procedimiento consiste en:
  - Escoja B muestras bootstrap independientes $X^{*(1)}, \dots, X^{*(B)}$ de $\hat{P}$: $X_1^{*(b)}, \dots, X_n^{*(b)} \sim_{iid} \hat{P}$ para $b = 1, \dots, B$.
  - Evalúe las repeticiones de bootstrap: $\hat{\theta}^{*(b)} = s(X^{*(b)})$.
  - Estime la distribución muestral de θ con la distribución empírica de las repeticiones bootstrap: $\hat{\theta}^{*(1)}, \dots, \hat{\theta}^{*(B)}$:

$$\hat{P}(\hat{\theta}(A)) = \frac{1}{B}\sum_{b=1}^{B} 1_A(\hat{\theta}^{*(b)})$$

para conjuntos apropiados de $A \subseteq \mathbb{R}^p$ (si $\hat{\theta} \in \mathbb{R}^p$).

Pero, ¿y si solo queremos una cantidad de esa distribución muestral? pues hay fórmulas para calcularlas directamente.

Bootstrap para calcular errores estándar

Sea $\hat{\theta}$ un estimador de θ y suponga que queremos conocer el error estándar de $\hat{\theta}$. Un error estándar estimado de bootstrap se puede obtener con el siguiente algoritmo:

- Escoja B muestras bootstrap independientes $X^{*(1)}, \dots, X^{*(B)}$ de $\hat{P}$: $X_1^{*(b)}, \dots, X_n^{*(b)} \sim_{iid} \hat{P}$ para $b = 1, \dots, B$.
- Evalúe las repeticiones de bootstrap: $\hat{\theta}^{*(b)} = s(X^{*(b)})$.
- Estime los errores estándar con la desviación estándar de las B repeticiones:

$$\hat{s}_{boot} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}^{*(b)} - \hat{\theta}^{*(\cdot)}\right)^2}$$

donde $\hat{\theta}^{*(\cdot)} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*(b)}$.

Estimadores usuales

```r
x <- CommuteAtlanta$Time
(n <- length(x))
```

```
## [1] 500
```

```r
(Tn <- var(x))
```

```
## [1] 429.2484
```

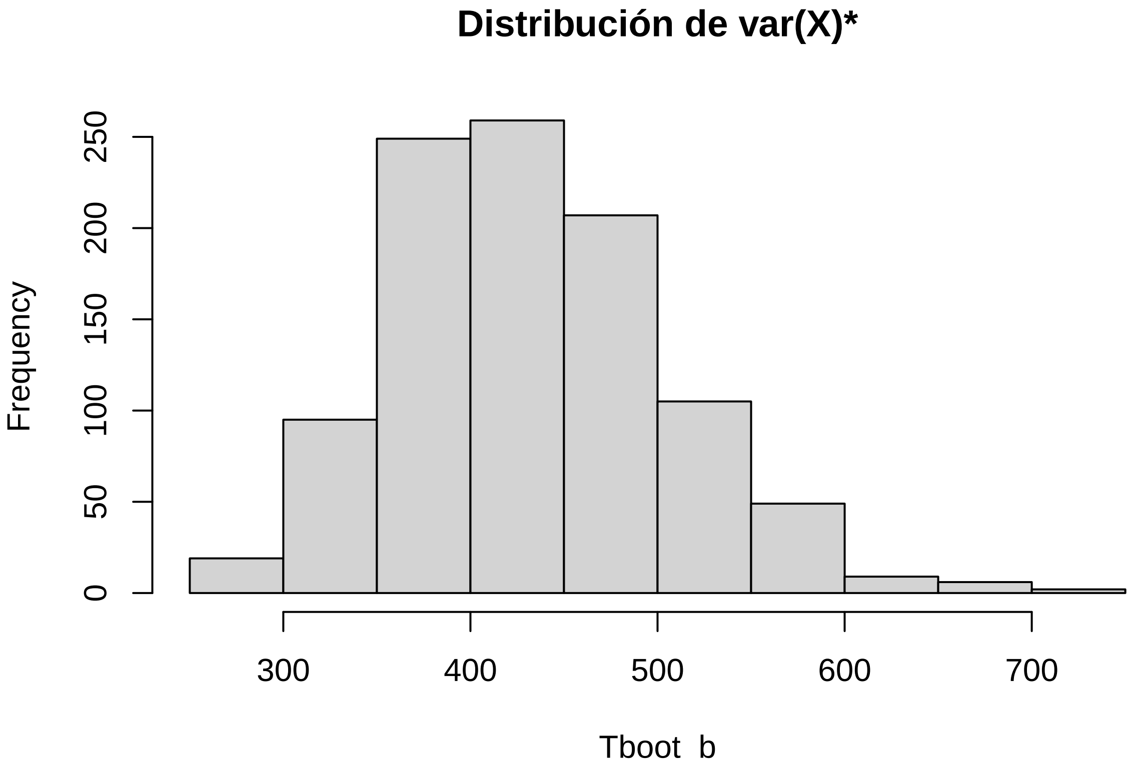
Muestra bootstrap

```r
B <- 1000
Tboot_b <- NULL
for(b in 1:B) {
  xb <- sample(x, size = n, replace = TRUE)
  Tboot_b[b] <- var(xb)
}
Tboot_b[1:10]
```

```
##  [1] 429.0391 492.5339 332.8533 438.5389 474.3159 387.1645 354.5111 551.7446
##  [9] 436.7800 399.9257
```

Distribución $\hat{P}$

```r
hist(Tboot_b, main= "Distribución de var(X)*")
```



Bootstrap para calcular el sesgo

Suponga que queremos estimar un parámetro $\theta = t(P)$ con el estadístico $\hat{\theta} = s(X)$. El sesgo de un estimador $\hat{\theta}$ está definido como:

$$bias(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Si sustituimos P por la distribución empírica $\hat{P}$, entonces obtenemos el estimado bootstrap del sesgo:

$$\widehat{bias(\hat{\theta})} = bias^*(\hat{\theta}^*) = E(\hat{\theta}^*) - \hat{\theta}$$

donde $\hat{\theta}$ es el estimador empírico de la muestra.

Cálculo de estadísticos bootstrap

```
(Tboot <- mean(Tboot_b))
```

```
## [1] 432.196
```

```
(Vboot <- var(Tboot_b))
```

```
## [1] 5408.924
```

```
(sdboot <- sqrt(Vboot))
```

```
## [1] 73.54539
```

El sesgo bootstrap es

```
mean(Tboot_b) - Tn
```

```
## [1] 2.947608
```

Bootstrap para calcular el intervalo de confianza

Si tenemos las repeticiones bootstrap $\hat{\theta}^{*(1)}, \ldots, \hat{\theta}^{*(B)}$, podemos estimar la distribución muestral de $\hat{\theta}$. A partir de esto, podemos construir intervalos de confianza para θ. Hay cuatro opciones: IC estándar, IC bootstrap t, IC percentiles, IC percentiles corregido por sesgo.

- IC normal: Utilizamos el resultado del TLC para decir que $\hat{\theta}$ es distribuido aproximadamente normal con media θ y variancia $s(\hat{\theta})^2$. Entonces, un IC $(1-\alpha)$ aproximado para θ está dado por:

$$\hat{\theta} \pm z_{\alpha/2}\hat{s}_{boot}(\hat{\theta})$$

Cálculo de IC normal

```
(z <- qnorm(1 - 0.05 / 2))
```

```
## [1] 1.959964
```

```
c(Tn - z * sdboot, Tn + z * sdboot)
```

```
## [1] 285.1021 573.3947
```

Bootstrap para calcular el intervalo de confianza

- IC bootstrap studentizado: Utilizando el mismo resultado anterior, pero ahora usando $\hat{s}_X(\hat{\theta})$ como estimador de $s(\hat{\theta})$ basado en la muestra X. De las muestras bootstrap $X^{*(b)}$ se calcula:

$$Z^{*(b)} = \frac{\hat{\theta}^{*(b)} - \hat{\theta}}{\hat{s}_{X^*}(\hat{\theta})}$$

De los valores $Z^{*(b)}$, podemos estimar el valor crítico $z_{\alpha/2}$ como $\hat{z}_{\alpha/2}$ tal que:

$$\frac{1}{B}\sum_{b=1}^{B} 1_{[Z^{*(b)} \leq \hat{z}_\alpha]} \approx \alpha$$

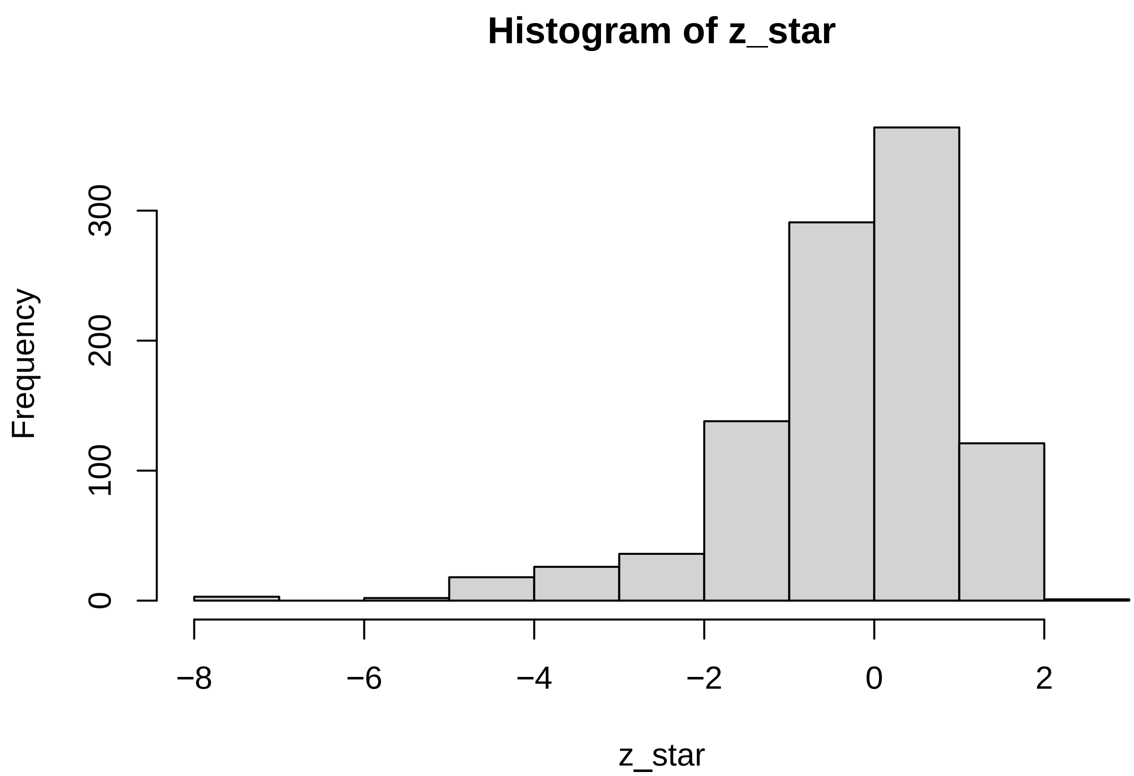
Entonces:

$$\left[\hat{\theta} - \hat{z}_{1-\alpha/2} s(\hat{\theta}), \hat{\theta} - \hat{z}_{\alpha/2} s(\hat{\theta})\right]$$

Cálculo de IC bootstrap studentizado

```
B <- 1000
Tboot_b <- NULL
Tboot_bm <- NULL
sdboot_b <- NULL
for (b in 1:B) {
  xb <- sample(x, size = n, replace = TRUE)
  Tboot_b[b] <- var(xb)
  for (m in 1:B) {
    xbm <- sample(xb, size = n, replace = TRUE)
    Tboot_bm[m] <- var(xbm)
  }
  sdboot_b[b] <- sd(Tboot_bm)
}
z_star <- (Tboot_b - Tn) / sdboot_b
```

```
hist(z_star)
```



```r
c(Tn - quantile(z_star, 1 - 0.05 / 2) * sdboot,
  Tn - quantile(z_star, 0.05 / 2) * sdboot)
```

```
##    97.5%     2.5%
## 311.7205 719.2377
```

Bootstrap para calcular el intervalo de confianza pivotales

El intervalo de confianza pivotal de tamaño $1-\alpha$ es

$$\left(2\widehat{\theta}_n - \widehat{\theta}_L, 2\widehat{\theta}_n - \widehat{\theta}_U^*\right)$$

donde

$$\hat{P}^*(\hat{\theta}^* \leq \hat{\theta}_L) = \frac{1}{B}\sum_{b=1}^{B} 1[\hat{\theta}^{*(b)} \leq \hat{\theta}_L] \approx 1-\alpha/2$$

$$\hat{P}^*(\hat{\theta}^* \leq \hat{\theta}_U) = \frac{1}{B}\sum_{b=1}^{B} 1[\hat{\theta}^{*(b)} \leq \hat{\theta}_U] \approx \alpha/2$$

El $2\hat{\theta}$ corrige el error por sesgo.

La prueba de este resultado está en *All of nonparametric statistics* de Larry Wassermann, p.32.

IC pivotal

```
c(2 * Tn - quantile(Tboot_b, 1 - 0.05 / 2),
  2 * Tn - quantile(Tboot_b, 0.05 / 2))
```

```
##    97.5%     2.5%
## 263.6263 551.5090
```

Referencias:

- UC3M - español
- Chicago - inglés
- Efron, B.; Tibshirani, R. (1993). An Introduction to the Bootstrap. Boca Raton, FL: Chapman & Hall/CRC. ISBN 0-412-04231-2.

Práctica de Intervalos de Confianza

1. La vida útil de cierto aparato de aire acondicionado sigue una distribución de Rayleigh, cuya función de densidad viene dada por la fórmula:

$f(x|\theta) = \frac{x}{\theta^2} \exp\left(\frac{-x^2}{2\theta^2}\right) 1_{(x>0)}$

Suponga que $X_1, X_2, \ldots, X_n$ es una muestra aleatoria correspondiente a la vida útil de n aparatos de aire acondicionado:

a) Determine un estadístico suficiente para θ.

b) Considere el pivote $\frac{1}{\theta^2} \sum_{j=1}^{n} X_j^2$ para construir un intervalo de confianza para θ con una confianza del $(1-\alpha)\%$.

c) ¿Cuál es la relación entre el estimador de máxima verosimilitud obtenido en b) con la estimación por intervalo obtenido en c).

d) Considere la muestra aleatoria de $n = 15$ datos de una distribución $U(0,1)$ que se ofrece, para simular una muestra aleatoria de 15 datos de una distribución de Rayleigh con $\theta = 10$. Encuentre un intervalo de confianza del 95% para estimar θ.

Práctica de Intervalos de Confianza

```r
data <- c(0.466, 0.589, 0.097, 0.809, 0.214, 0.315, 0.971, 0.298, 0.005, 0.126,
          0.019, 0.553, 0.385, 0.232, 0.989)
```

2. Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una población de Poisson con parámetro λ.

a) Utilice la Desigualdad de Cramer-Rao, y la información de Fisher para demostrar que $\bar{X}$, es un estimador de variancia mínima para estimar λ.

b) Demuestre que la variable $U = \frac{\bar{X}-\lambda}{\sqrt{\bar{X}/n}}$ tiene distribución que converge a una normal estándar.

c) Utilice la variable U del inciso anterior, como pivote para construir un intervalo de confianza para λ con probabilidad del 95%.

Práctica de Intervalos de Confianza

3. Si $Y_1, Y_2, \ldots, Y_n$ corresponden a una muestra aleatoria de una distribución gamma con parámetros α desconocido y β desconocido.

a) Demuestre que la variable $U = \frac{2\sum_{j=1}^{n} Y_j}{\beta}$ puede ser utilizada como pivote para estimar el valor de β y construya un intervalo de confianza de $1-\alpha$ para estimar el valor β.

b) Por teorema del límite central, la variable aleatoria $Z = \frac{\bar{X}-E(\bar{X})}{\sqrt{Var(\bar{Y})}}$ tiene distribución que converge a una $N(0,1)$. Supongamos que n es suficientemente grande, determine la variable aleatoria Z vinculada con este problema, que puede ser utilizada como pivote para estimar el valor de β. Construya un intervalo de confianza de para estimar el valor β.

Práctica de Intervalos de Confianza

c) Considere la siguiente muestra aleatoria que pertenece a una distribución gamma con $\alpha = 3$:

```r
data <- c(66.8, 26.6, 8.7, 25.9, 17.0, 17.4, 9.2, 19.6, 27.8, 33.3)
```

Utilice los resultados obtenidos en a) y b) para determinar dos intervalos de confianza del 95% para estimar β, uno para cada método. Compare los resultados. ¿A qué atribuye las diferencias?

---

class: center, middle

¿Qué discutimos hoy?

Bootstrap: concepto, ejemplos y definiciones. IC utilizando bootstrap