Ciencias de Datos con R: Fundamentos Estadísticos

Ana M. Bianco, Jemina García y Mariela Sued.

Estimación.

Estadística

- Cuándo hacemos estadística no conocemos a F ni el valor del parámetro poblacional de interés.
- Cuándo hacemos estadística queremos hacer una cuenta con la muestra que nos permita estimar el valor θ de inteŕes.

$$\text{Estimador:} \quad \widehat{\theta}_n = \widehat{\theta}_n(X_1 \ldots, X_n)$$

$$\text{Estimación:} \quad \widehat{\theta}_{n,\text{obs}} = \widehat{\theta}_n(x_1 \ldots, x_n),$$

donde $x_1, \ldots, x_n$ representan datos. Valores observados.

Muestra - Datos (Observaciones)

- Muestra (aleatoria simple):

$$X_1, \ldots, X_n \quad \text{Variables aleatorias iid.}$$

- Datos - Observaciones - Valores observados $x_1, \ldots, x_n$: Números.

Datos-Observaciones: son realizaciones de las variables aleatorias

Datos-Observaciones: son los resultados obtenidos al realizar el "experimento"

Estadística

POBLACION $\leftrightarrow F$	MUESTRA $X_1, \ldots X_n$ i.i.d. $X_i \sim F$
Parámetro: Valor asociado de F	Estimador:estadístico para estimar θ
$\theta = \theta(F)$	$\widehat{\theta}_n = \widehat{\theta}_n(X_1, \ldots, X_n)$
θ: valor poblacional	$\widehat{\theta}_n$ NUEVA VARIABLE ALEATORIA

Estadística

Parámetro (en palabras)	Parámetro (en matemática)	Estimador Var. Alea	estimación (en R)
tita	θ	$\widehat{\theta}_n$	$\widehat{\theta}_{n,\text{obs}}$
esperanza	$\mu = \mathbb{E}(X)$	$\widehat{\mu}_n = \overline{X}_n$	`mean(datos)`
desvío	$\sigma = \sqrt{V(X)}$	$\sqrt{S_n^2}$	`sd(datos)`
probabilidad	$p = \mathbb{P}(X \leq 3)$	$\frac{1}{n}\sum_{i=1}^{n} I_{(X_i \leq 3)}$	`mean(datos<=3)`
mediana	$F^{-1}(0.5)$	$X_{(n/2)}$	`median(datos)`
p-quantil	$F^{-1}(p)$	$X_{(pn)}$	`quantile(datos, p)`

Estimación de $F(t)$- *LA empírica*

La LGN demuestra que la frecuencia relativa converge a la probabilidad.

$(X_i)_{i\geq 1}$ i.i.d., $X_i \sim F$

LGN: $$\frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \leq 3\}} \xrightarrow{p} \mathbb{E}(I_{\{X_1 \leq 3\}}) = \mathbb{P}(X_1 \leq 3) = F(3).$$

$$\widehat{F}_n(3) := \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \leq 3\}}$$

Estimación de $F(t)$- *LA empírica*

La LGN demuestra que la frecuencia relativa converge a la probabilidad.

$(X_i)_{i\geq 1}$ i.i.d., $X_i \sim F$

LGN: $$\frac{1}{n}\sum_{i=1}^{n} I_{\{X_i\leq 3\}} \xrightarrow{p} \mathbb{E}(I_{\{X_1\leq 3\}}) = \mathbb{P}(X_1 \leq 3) = F(3).$$

$$\widehat{F}_n(3) := \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i\leq 3\}}$$

$$\widehat{F}_n(t) := \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i\leq t\}}$$

LGN: $$\widehat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i\leq t\}} \xrightarrow{p} \mathbb{E}(I_{\{X_1\leq t\}}) = \mathbb{P}(X_1 \leq t) = F(t).$$

Manos a la obra

Mediciones de gas - Equipo 1

Considere $n = 100$ datos obtenidos al utilizar el equipo 1.

1. Realice un histograma.
2. Calcule el promedio de los datos.
3. Calcule el percentil 0.9 de los datos.
4. Estime la probabilidad de que una medición realizada con este equipo diste de 70 en más de 2 unidades.
5. Repita los ítems anteriores utilizando ahora los primeros $n = 5$ y $n = 30$ datos

Sigue amasando, sigue amasando

Duración de lámparas

Considere los datos de la duración de n lámparas (en meses), para $n \in \{5, 30, 100\}$. En cada caso,

1. realice un histograma.
2. calcule el promedio y el percentil 0.9.
3. estime la probabilidad de que la lámpara dure a lo sumo un año (12 meses).

Mediciones de Gas - Con otro lenguaje

Consideremos las mediciones de gas realizadas por el equipo 1. Sea X_i el resultado de la i-ésima medición, para $i = 1, \ldots, n$.
Asumiremos que $X_1, \ldots, X_n$ son v.a.i.i.d.

1. Indicar cuál cuenta hay que hacer con la muestra $(X_1, \ldots, X_n)$ para estimar $\mu = \mathbb{E}(X_1)$. Es decir, proponer un estimador $\widehat{\mu}_n$ para μ.
2. Considerar $n = 5$ datos correspondientes al equipo 1 y calcular la estimación de μ correspondiente a estos datos. Repetir considerando $n = 30$ y $n = 100$.
3. Sea $q = F^{-1}(0.9)$ con $X_i \sim F$. Indicar cuál cuenta hay que hacer con la muestra $(X_1, \ldots, X_n)$ para estimar q. Es decir, proponer un estimador $\widehat{q}_n$ para q.
4. Sea $p = \mathbb{P}(|X - 70| > 2)$ con $X \sim X_1$. Indicar cuál cuenta hay que hacer con la muestra $(X_1, \ldots, X_n)$ para estimar p. Es decir, proponer un estimador $\widehat{p}_n$ para p.
5. Considerar $n = 5$ datos duraciones de lámparas y calcular la estimación de p correspondiente a estos datos. Repetir considerando $n = 30$ y $n = 100$.

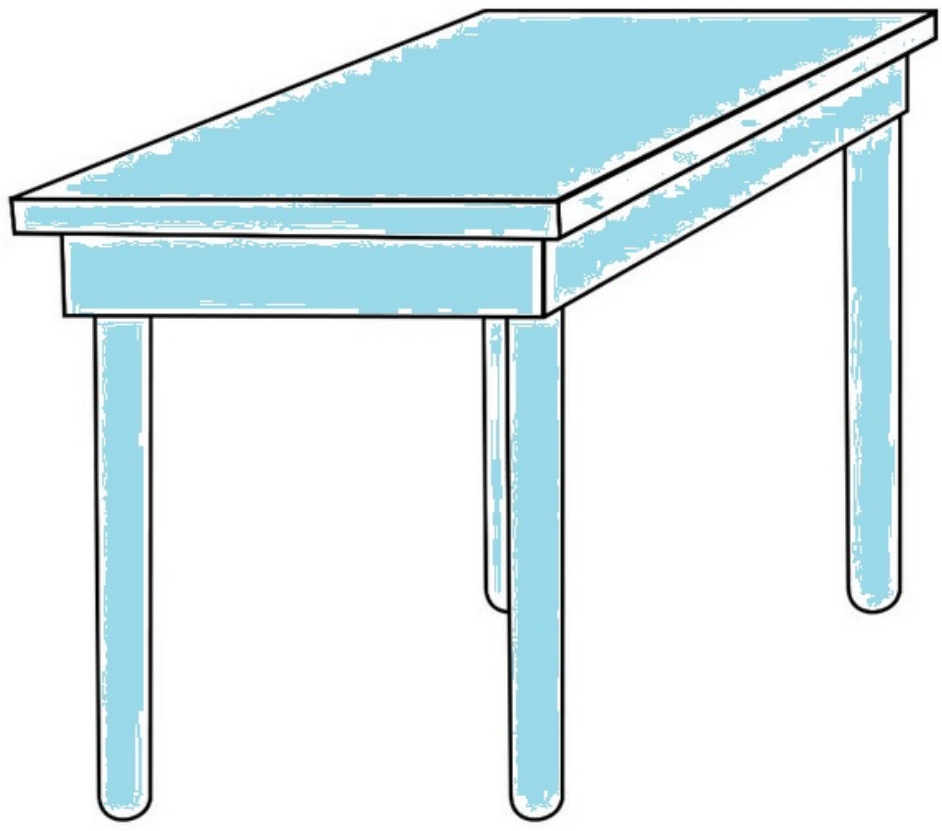
Duración de lámparas - Con otro lenguaje

Consideremos las duraciones de lámparas en meses. Sea X_i la duración de la i-ésima lámpara, para $i = 1, \ldots, n$. Asumiremos que $X_1, \ldots, X_n$ son v.a.i.i.d.

1. Indicar cuál cuenta hay que hacer con la muestra $(X_1, \ldots, X_n)$ para estimar $\mu = \mathbb{E}(X_1)$. Es decir, proponer un estimador $\widehat{\mu}_n$ para μ.
2. Considerar $n = 5$ datos de duraciones de lámparas y calcular la estimación de μ correspondiente a estos datos. Repetir considerando $n = 30$ y $n = 100$.
3. Sea $p = \mathbb{P}(X \leq 12)$ con $X \sim X_1$. Indicar cuál cuenta hay que hacer con la muestra $(X_1, \ldots, X_n)$ para estimar p. Es decir, proponer un estimador $\widehat{p}_n$ para p.
4. Considerar $n = 5$ datos duraciones de lámparas y calcular la estimación de p correspondiente a estos datos. Repetir considerando $n = 30$ y $n = 100$.

Recreo

¿Cuánto mide la mesa?



¿Cuánto mide la mesa?

Estas son las $n = 5$ primeras observaciones:

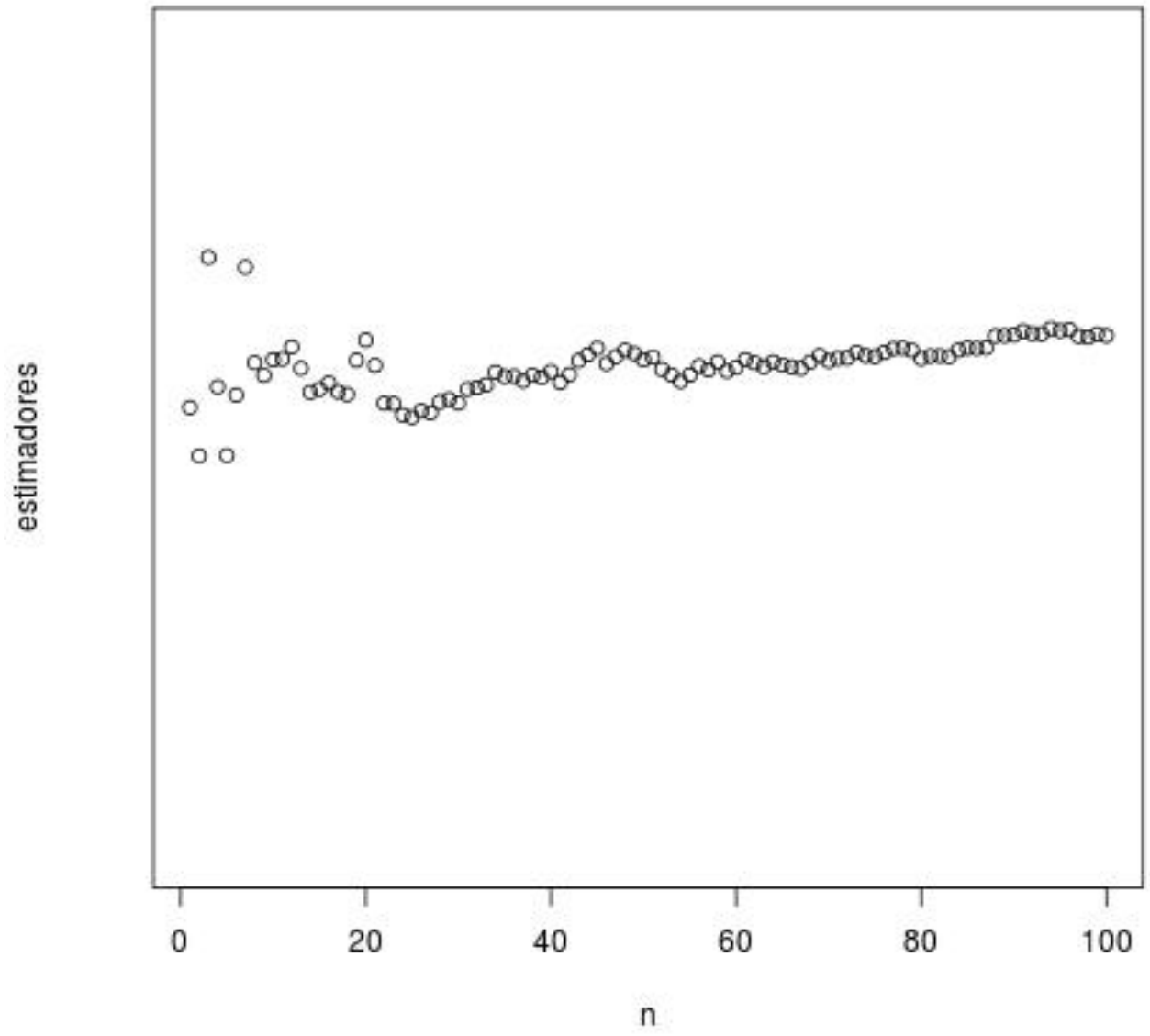
$$1.17, \quad 1.36, \quad 0.15, \quad 2.52, \quad 0.21, \quad 1.78, \quad 2.67$$

¿Qué hacemos? 

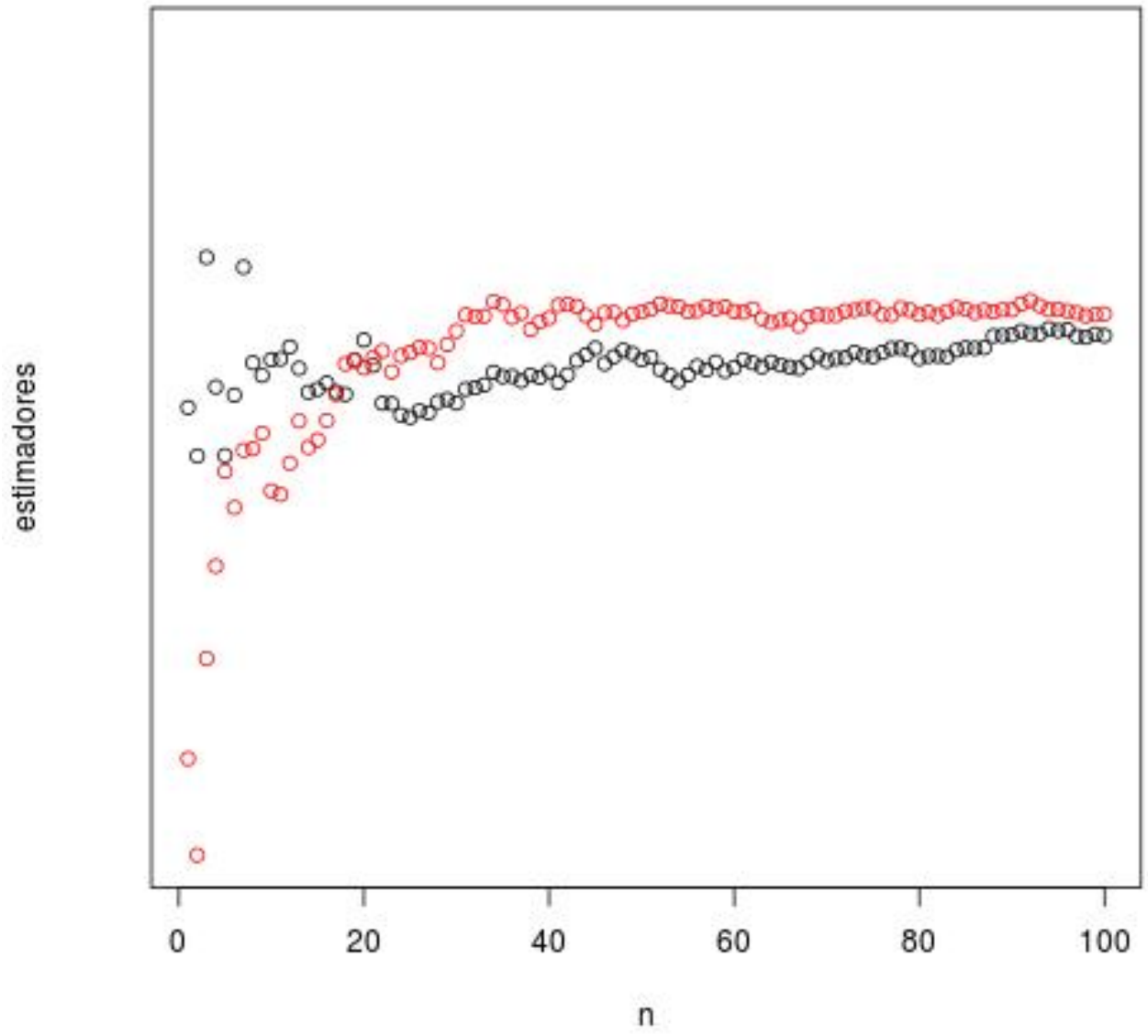
Modelo y Estimadores

Hacemos los puntos 1 y 2 de la Parte 3 de la Guía de actividades

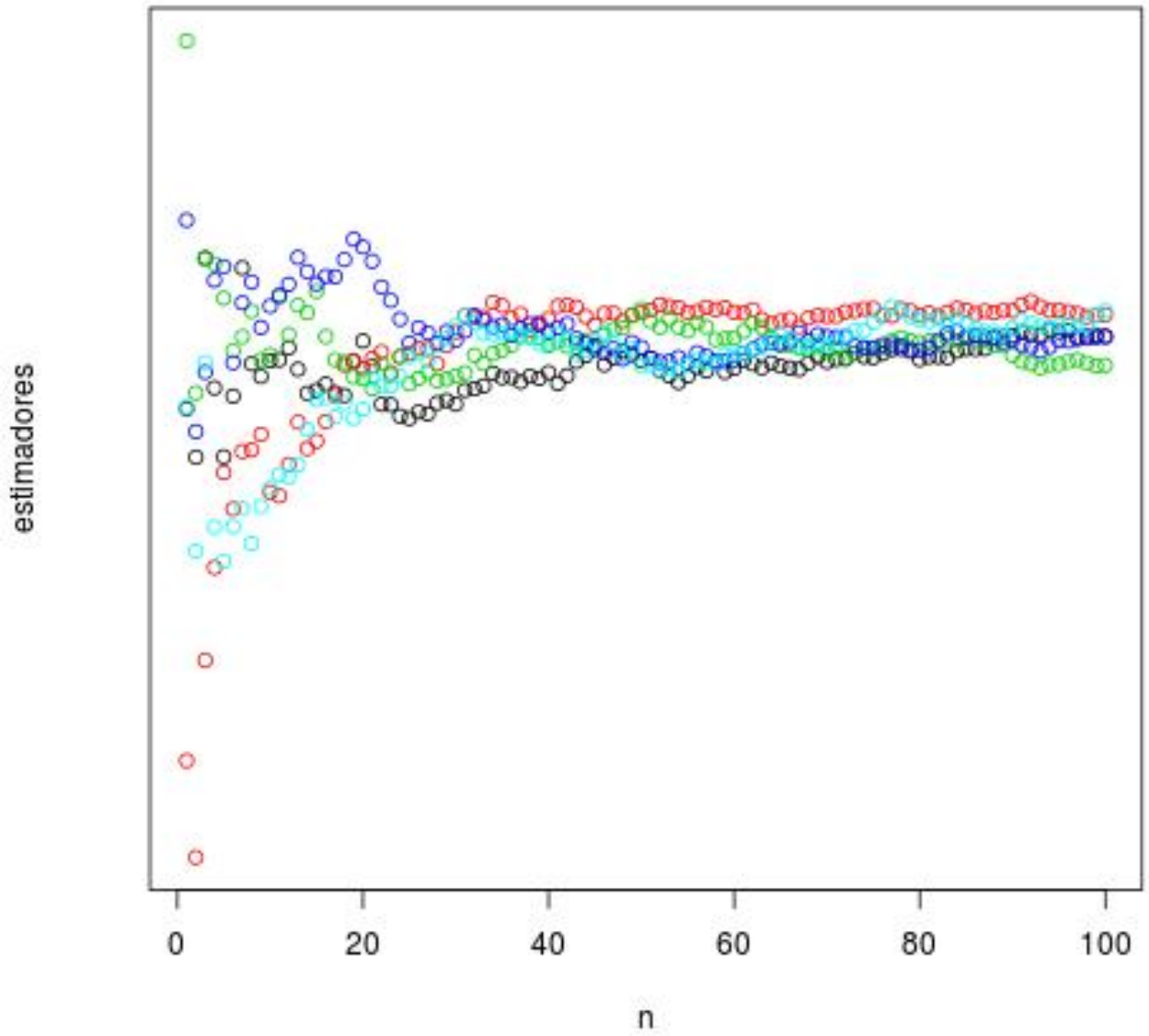
Juan cada vez con más datos. $\widehat{\theta}_n = 2\overline{X}_n$



Juan y Andrea, cada vez con más datos. $\widehat{\theta}_n = 2\overline{X}_n$



Varios, cada vez con más datos. $\widehat{\theta}_n = 2\overline{X}_n$

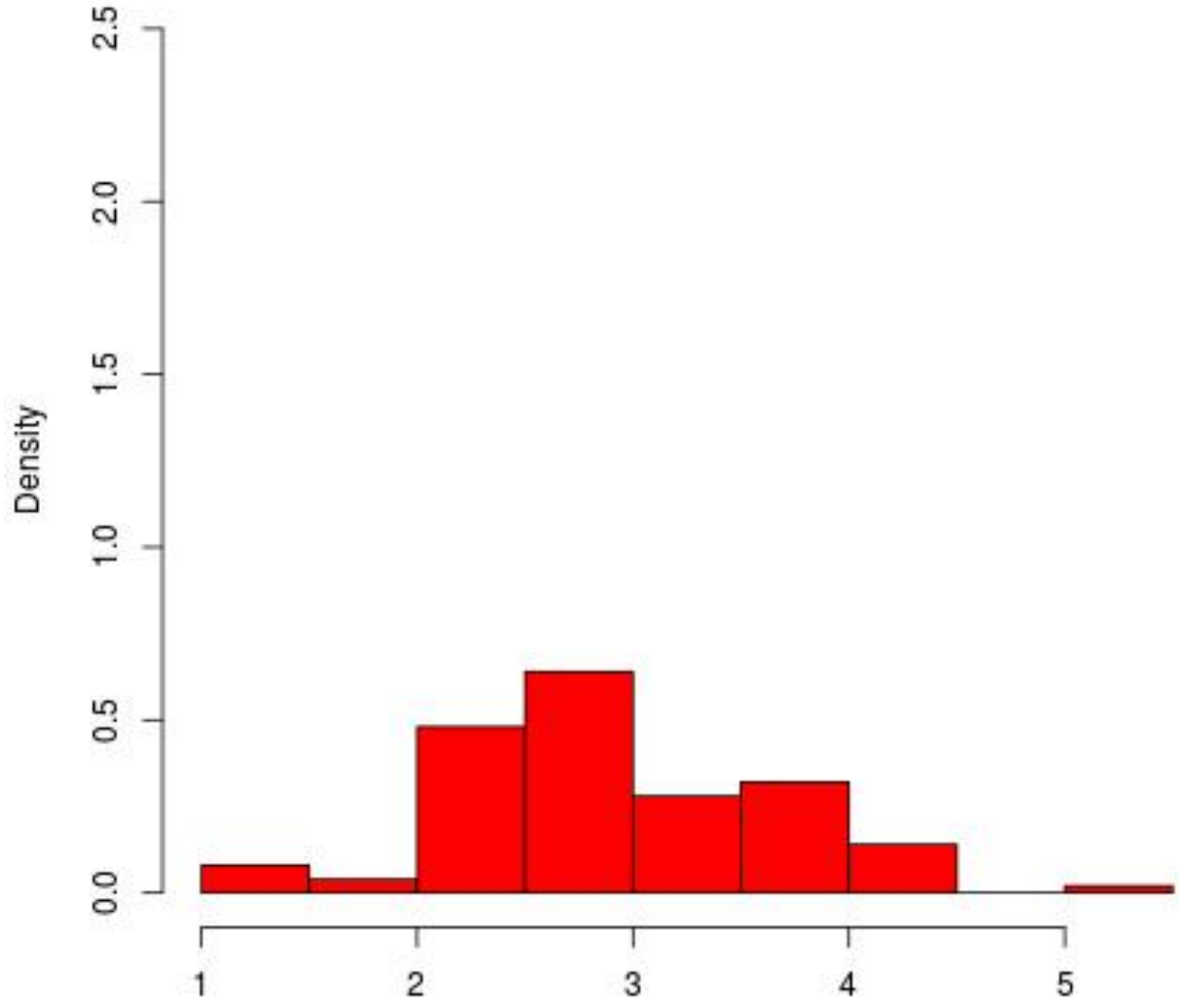


Cada uno con lo suyo. $\widehat{\theta}_n = 2\overline{X}_n$

	Nombre	n=5	n=30	n=50
1	Juan	1.08	3.2	2.96
2	Andrea	2.87	2.95	2.88
3	Flor	3.47	3.2	3.18
4	Gonzalo	3.88	3.23	3.18
5	Paula	3.79	2.93	2.81
6	Agustin	3.01	2.9	2.59
7	Julieta	3.55	3.03	3.01
8	Marina	2.09	2.79	3.1
9	Pablo	4.14	3.41	3.01
10	Enrique	2.65	3.29	3.11
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

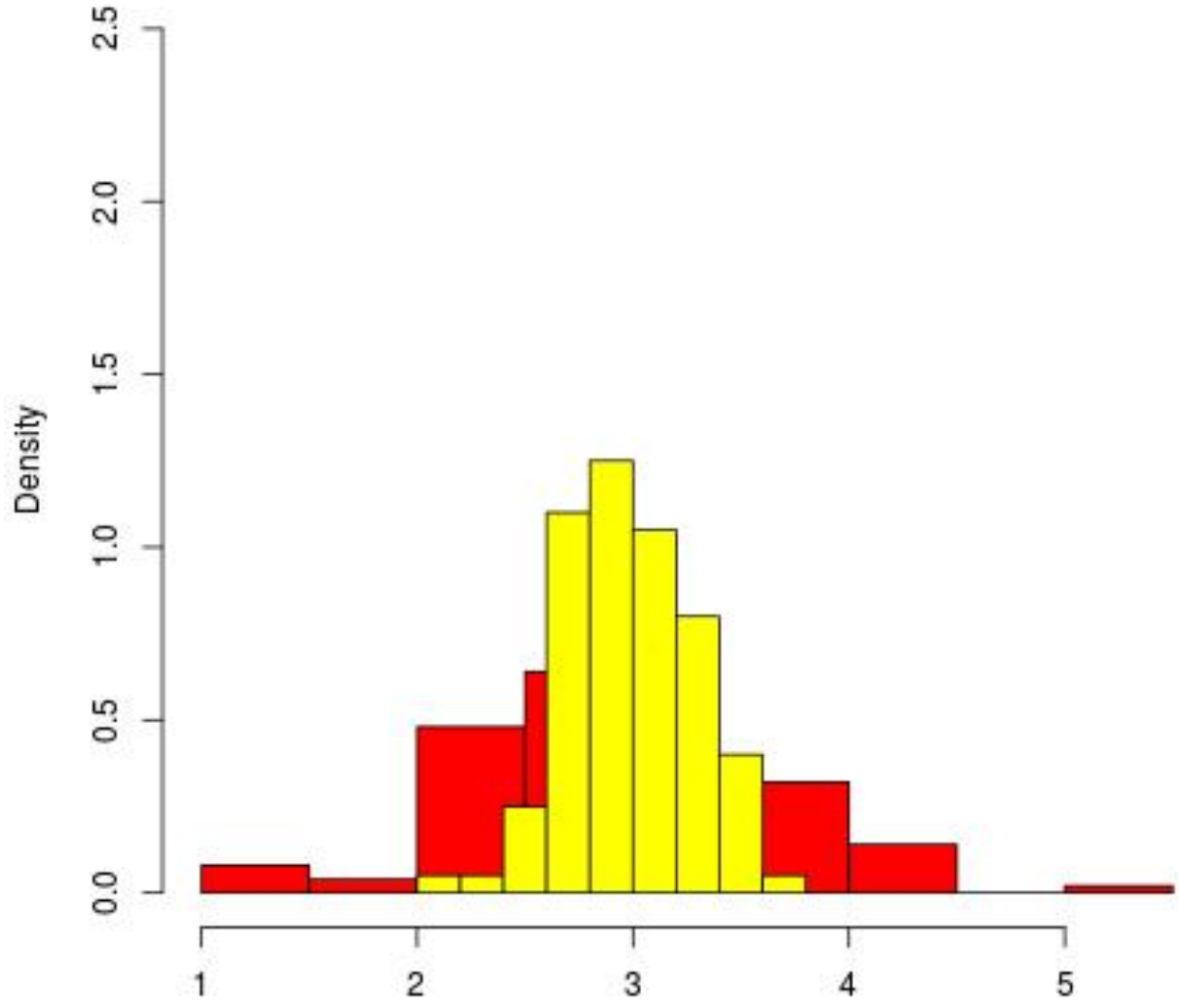
Histogramas de $\widehat{\theta}_n = 2\overline{X}_n$

(empirical) Sampling Distribution of $\widehat{\theta}_n$



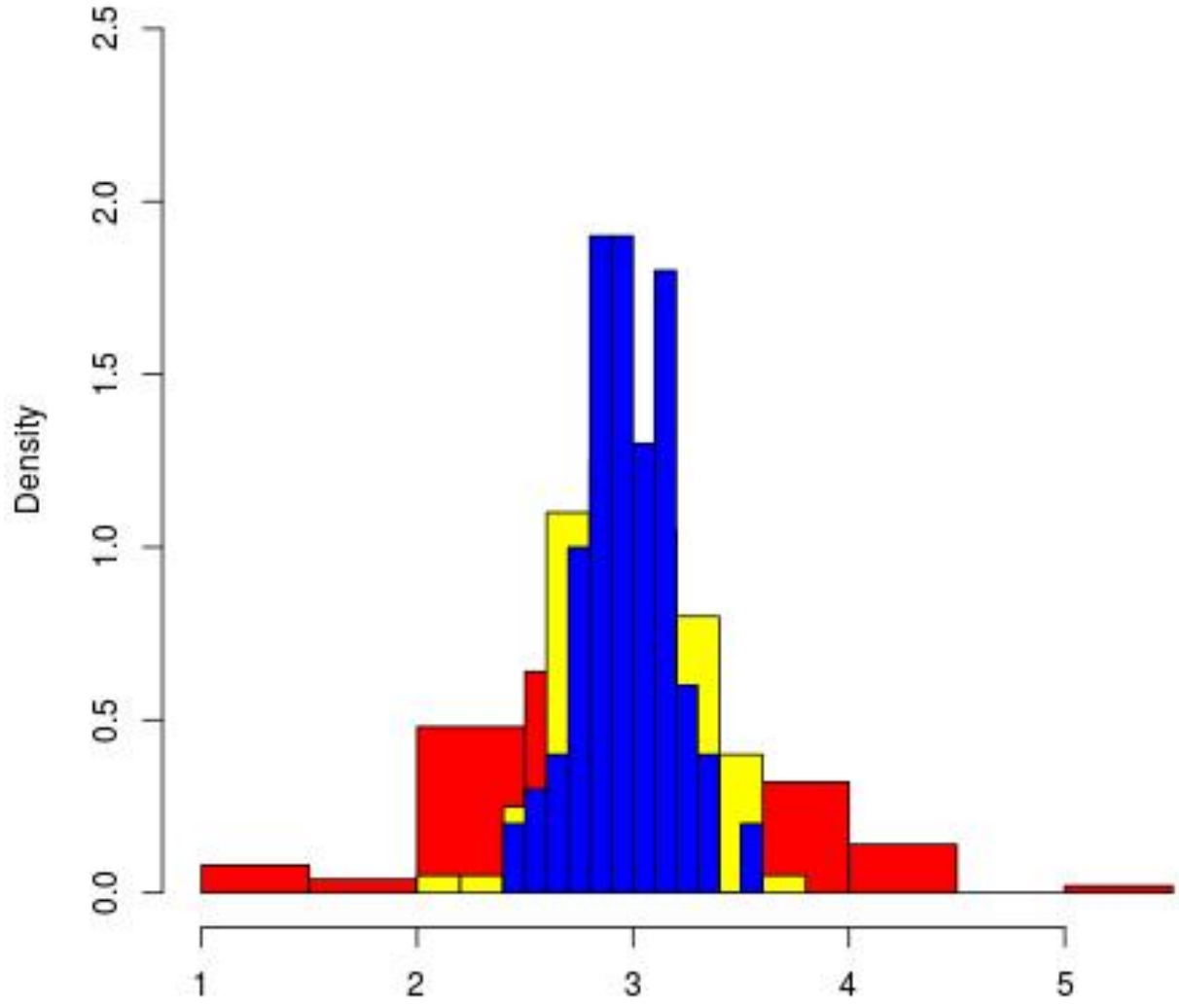
Histogramas de $\widehat{\theta}_n = 2\overline{X}_n$

(empirical) Sampling Distribution of $\widehat{\theta}_n$



Histogramas de $\widehat{\theta}_n = 2\overline{X}_n$

(empirical) Sampling Distribution of $\widehat{\theta}_n$



Estimación

Point estimation refers to providing a single "best guess" of some quantity of interest.
All of statistics. Wasserman

Estimación

Point estimation refers to providing a single "best guess" of some quantity of interest.
All of statistics. Wasserman

- translate(some quantity of interest)= Objeto de interés.
- some quantity of interest: largo de la mesa (θ)
- best guess: Estimador:*cuenta hecha con la muestra*
- best guess: Estimador: Función de la muestra

$$\widehat{\theta}_n \equiv \widehat{\theta}_n(X_1, \ldots, X_n)$$

Estimación

Point estimation refers to providing a single "best guess" of some quantity of interest.
All of statistics. Wasserman

- translate(some quantity of interest)= Objeto de interés.
- some quantity of interest: largo de la mesa (θ)
- best guess: Estimador:*cuenta hecha con la muestra*
- best guess: Estimador: Función de la muestra

$$\widehat{\theta}_n \equiv \widehat{\theta}_n(X_1, \ldots, X_n)$$

- Estimación: Valor del estimador en un conjunto de datos:

$$\widehat{\theta}_n(x_1, \ldots, x_n)$$

Notemos que el estimador ...

$$\widehat{\theta}_n \equiv \widehat{\theta}_n(X_1, \ldots, X_n)$$

- $\widehat{\theta}_n$ es una variable aleatoria.
- $\widehat{\theta}_n$ tiene distribución (siempre).

  Sampling distribution of $\widehat{\theta}_n$*:* $f_{\widehat{\theta}_n}$
- $\widehat{\theta}_n$ tiene (en general) esperanza: $\mathbb{E}(\widehat{\theta}_n)$

Notemos que el estimador ...

$$\widehat{\theta}_n \equiv \widehat{\theta}_n(X_1, \ldots, X_n)$$

- $\widehat{\theta}_n$ es una variable aleatoria.
- $\widehat{\theta}_n$ tiene distribución (siempre).

Sampling distribution of $\widehat{\theta}_n$*:* $f_{\widehat{\theta}_n}$

- $\widehat{\theta}_n$ tiene (en general) esperanza: $\mathbb{E}(\widehat{\theta}_n) = \int u f_{\widehat{\theta}_n}(u) du$
- $\widehat{\theta}_n$ tiene (en general) varianza: $\mathbb{V}(\widehat{\theta}_n)$
- $\widehat{\theta}_n$ tiene (en general) desvío estandar.

$$\mathsf{se} = \mathsf{se}(\widehat{\theta}_n) = \sqrt{\mathbb{V}(\widehat{\theta}_n)} \quad \textit{Standard error of } \widehat{\theta}_n.$$

Veamos el shiny

Consistencia

A medida que aumenta el tamaño n de la muestra, el estimador se aproxima al objeto de interés.

$$\widehat{\theta}_n \longrightarrow \theta \ , \quad \text{cuando } n \to \infty$$

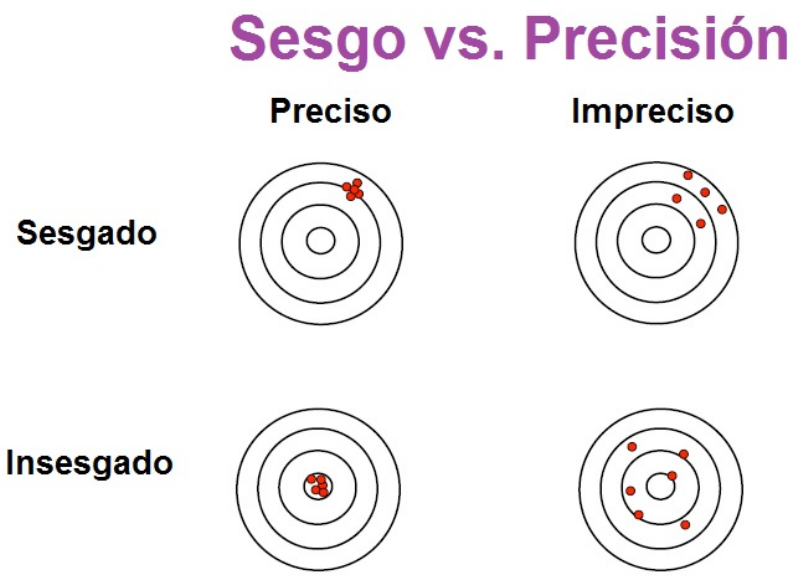
Error cuadrático medio (ECM)

$$\text{ECM} : \mathbb{E}\left\{(\widehat{\theta}_n - \theta)^2\right\}.$$

Lema: Si el ECM de un estimador converge a cero entonces vale la consistencia:

$$\mathbb{E}\left\{(\widehat{\theta}_n - \theta)^2\right\} \longrightarrow 0 \quad \text{implica que} \quad \widehat{\theta}_n \longrightarrow \theta\ .$$

Exactitud (In - Sesgado) - Precisión (Varianza)



Sesgo - Bias

$$\text{Sesgo} : \mathbb{E}(\widehat{\theta}_n) - \theta.$$

Insesgado: *El estimador $\widehat{\theta}_n$ se dice insesgado si su sesgo vale cero*

$$\text{Insesgado} : \mathbb{E}(\widehat{\theta}_n) - \theta = 0$$

En otras palabras, el estimador $\widehat{\theta}_n$ se dice insesgado si su esperanza coincide con el valor de interés que queremos estimar:

$$\text{Insesgado} : \mathbb{E}(\widehat{\theta}_n) = \theta$$

Propidades

Lema: El error cuadrático medio de un estimador se descompone de la siguiente manera:

$$\text{ECM}(\widehat{\theta}_n) = \mathbb{V}(\widehat{\theta}_n) + \left\{\mathbb{E}(\widehat{\theta}_n) - \theta\right\}^2$$

En particular... Si

$$\mathbb{V}(\widehat{\theta}_n) \to 0 \quad \text{y} \quad \mathbb{E}(\widehat{\theta}_n) \to \theta$$

tenemos que ECM converge a cero, y por lo tanto el estimador es consistente:

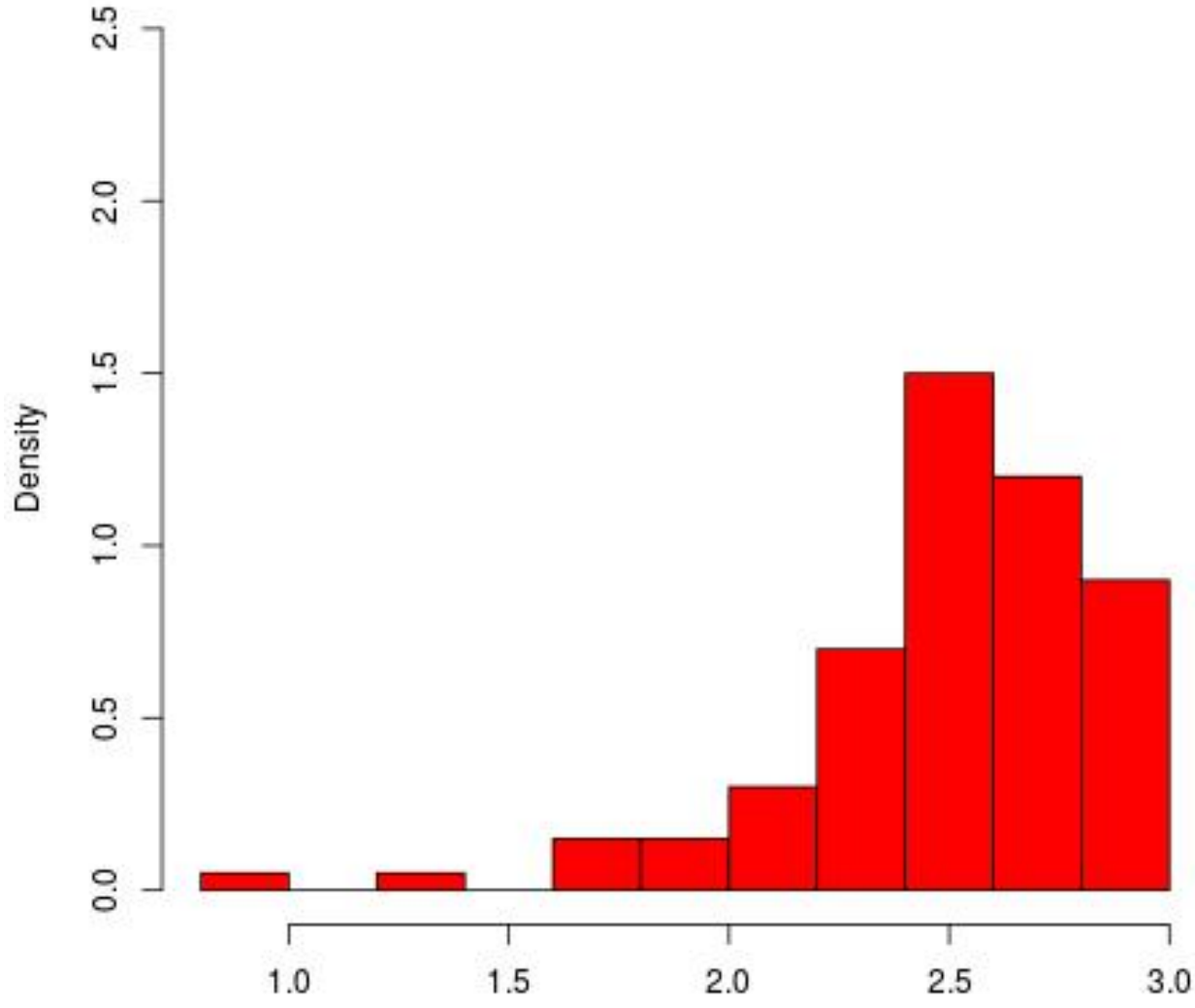
$$\widehat{\theta}_n \longrightarrow \theta$$

Miremos todo en el ejemplo: $\widehat{\theta}_n = 2\bar{X}_n$

- $(X_i)_{i\geq 1}$ i.i.d., $X_i \sim \mathcal{U}[0,\theta]$
- Objeto de interés: θ
- Estimador: $\widehat{\theta}_n = 2\bar{X}_n$
- Distribución de $\widehat{\theta}_n$?
- $\mathbb{E}(\widehat{\theta}_n) = \theta$: Es insesgado
- $\mathbb{V}(\widehat{\theta}_n) = \mathbb{V}(2\bar{X}_n) = 4\mathbb{V}(\bar{X}_n) = 4\frac{\mathbb{V}(X_1)}{n} = 4\frac{\theta^2/12}{n}$
- ECM$(\widehat{\theta}_n) = 0^2 + 4\frac{\theta^2/12}{n}$

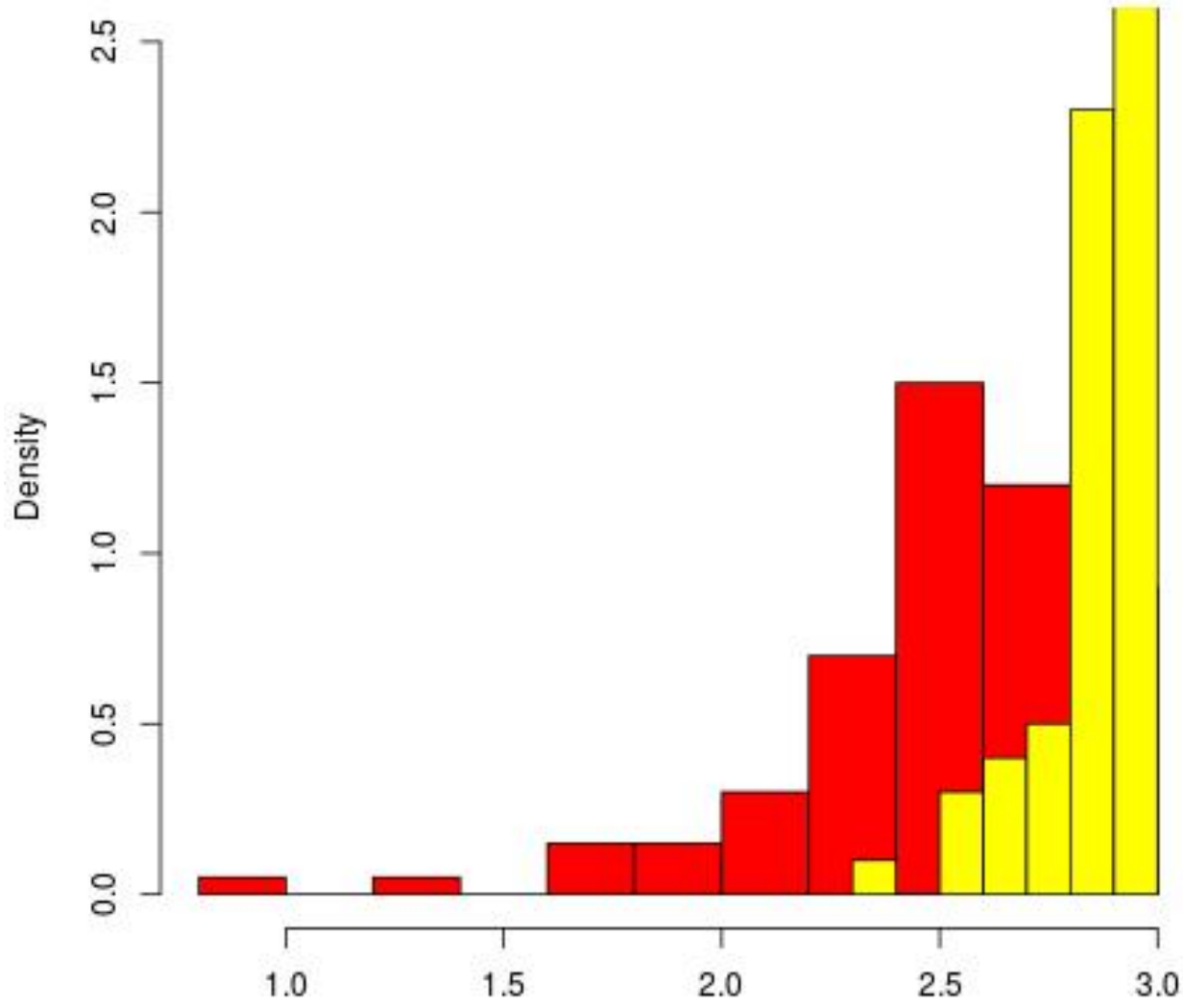
Histogramas de $\widetilde{\theta}_n = \max\{X_1, \ldots, X_n\}$

(empirical) Sampling Distribution of $\widetilde{\theta}_n$



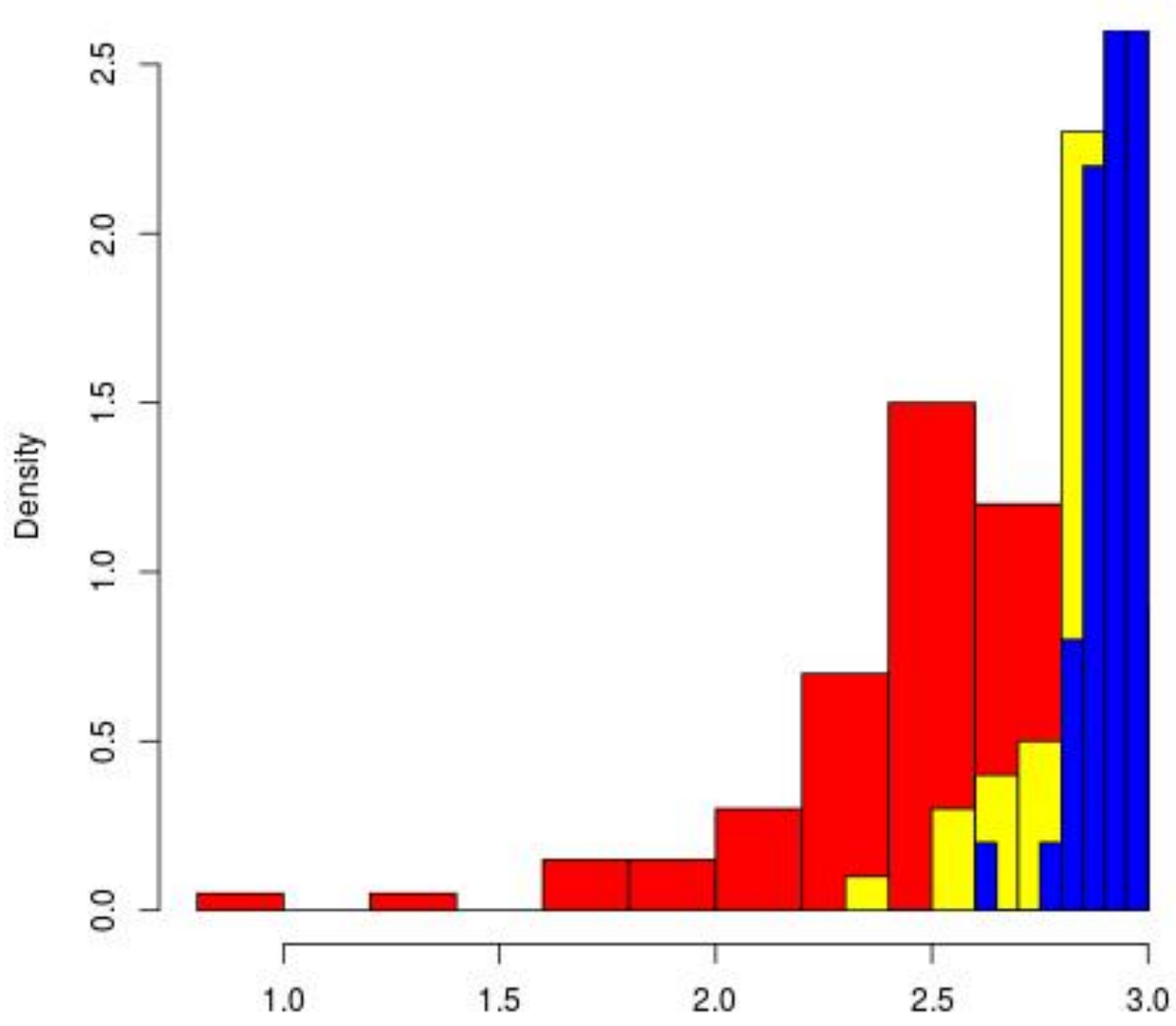
Histogramas de $\widetilde{\theta}_n = \max\{X_1, \ldots, X_n\}$

(empirical) Sampling Distribution of $\widetilde{\theta}_n$



Histogramas de $\widetilde{\theta}_n = \max\{X_1, \ldots, X_n\}$

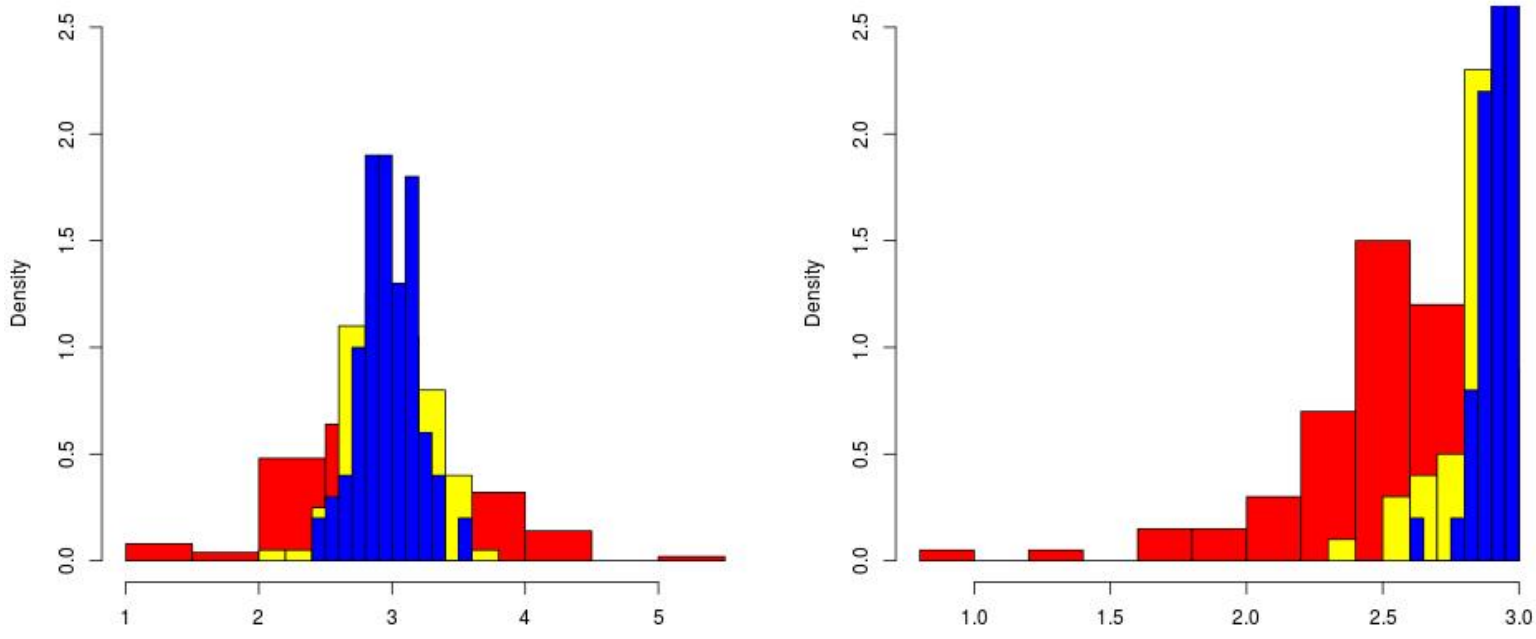
(empirical) Sampling Distribution of $\widetilde{\theta}_n$



Miremos todo ahora para $\widetilde{\theta}_n = \max\{X_1, \ldots, X_n\}$

- $(X_i)_{i \geq 1}$ i.i.d., $X_i \sim \mathcal{U}[0, \theta]$
- Objeto de interés: θ
- Estimador: $\widetilde{\theta}_n = \max\{X_1, \ldots, X_n\}$
- Distribución de $\widetilde{\theta}_n$?
- $\mathbb{E}(\widetilde{\theta}_n)$?
- $\mathbb{V}(\widetilde{\theta}_n)$?
- ECM $(\widetilde{\theta}_n)$

Histogramas de $\widehat{\theta}_n = 2\overline{X}_n$ y de $\widetilde{\theta}_n = \max\{X_1, \ldots, X_n\}$



Vamos guia.

Estadístico: Cuenta hecha con la muestra

$$h(X_1, \ldots, X_n)$$

Estadística

POBLACION $\leftrightarrow F$	MUESTRA $X_1, \ldots X_n$ i.i.d. $X_i \sim F$
Parámetro: Valor asociado de F	Estimador:estadístico para estimar θ
$\theta = \theta(F)$	$\widehat{\theta}_n = \widehat{\theta}_n(X_1, \ldots, X_n)$
θ: valor poblacional	$\widehat{\theta}_n$ NUEVA VARIABLE ALEATORIA

Enfaticemos en que el estimador ... (si, ya la vimos!)

$$\widehat{\theta}_n \equiv \widehat{\theta}_n(X_1, \ldots, X_n)$$

- $\widehat{\theta}_n$ es una variable aleatoria.
- $\widehat{\theta}_n$ tiene distribución (siempre)
- $\widehat{\theta}_n$ tiene (en general) esperanza: $\mathbb{E}(\widehat{\theta}_n)$
- $\widehat{\theta}_n$ tiene (en general) varianza: $\mathbb{V}(\widehat{\theta}_n)$

Consistencia

- $(X_i)_{i\geq 1}$ i.i.d., $X_i \sim F$, $F \in \mathcal{F}$
- $\mathcal{F}$: modelo estadístico.
- $\theta(F)$ objeto de interés definido para cada posible $F \in \mathcal{F}$
- estimador $\widehat{\theta}_n = \widehat{\theta}_n(X_1, \ldots, X_n)$.
- Consistencia:

$$\widehat{\theta}_n(X_1, \ldots, X_n) \longrightarrow \theta(F)$$

cuando $n \to \infty$, $X_i \sim F$, cualquiera sea $F \in \mathcal{F}$

A medida que aumenta el tamaño n de la muestra, el estimador se aproxima al objeto de interés.

$$\widehat{\theta}_n \longrightarrow \theta \ , \quad \text{cuando } n \to \infty$$

Estimación: ejemplo

- $X_1, \ldots, X_n$ i.i.d. $X_i \sim X$.
- Parámetro de interés; $\sigma^2 = \mathbb{E}\{(X - \mu)^2\} = \mathbb{V}(X)$
- ¿Estimador?

Estimación: ejemplo

- $X_1, \ldots, X_n$ i.i.d. $X_i \sim X$.
- Parámetro de interés; $\sigma^2 = \mathbb{E}\{(X - \mu)^2\} = \mathbb{V}(X)$
- ¿Estimador?

$$\widehat{\sigma}_n = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}$$

Estimación: ejemplo

- $X_1, \ldots, X_n$ i.i.d. $X_i \sim X$.
- Parámetro de interés; $\sigma^2 = \mathbb{E}\{(X - \mu)^2\} = \mathbb{V}(X)$
- ¿Estimador?

$$\widehat{\sigma}_n = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n}$$

- $\mathbb{E}(\widehat{\sigma}_n) = (n-1)n^{-1}\sigma^2$
- Estimador (insesgado) de la varianza: $S^2 = S_n^2$

$$S^2 = S_n^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1} \, .$$

- $\mathbb{E}[S^2] = \sigma^2$ (insesgado)
- $S^2 \to \sigma^2$ en probabilidad (consistencia)

Propiedades - si, de nuevo!, pero todas juntas.

- Consistencia

$$\widehat{\theta}_n(X_1, \ldots, X_n) \to \theta(F) \text{ en probabilidad, cuando } X_i \sim F$$

$$\text{abreviado:} \quad \widehat{\theta} \to \theta$$

- Error cuadratico medio: ECM=$\mathbb{E}\{(\widehat{\theta}_n - \theta)^2\}$
- Lema: Si $\mathbb{E}\{(\widehat{\theta}_n - \theta)^2\} \to 0$, entonces $\widehat{\theta}_n \to \theta$
- Sesgo: $\mathbb{E}(\widehat{\theta}_n) - \theta$.
- Estimador insesgado: Sesgo=0: $\mathbb{E}(\widehat{\theta}_n) - \theta$
- Lema: $\mathbb{E}\left\{(\widehat{\theta}_n - \theta)^2\right\} = \mathbb{V}(\widehat{\theta}_n) + \left\{\mathbb{E}(\widehat{\theta}_n) - \theta\right\}^2$
- Si $\mathbb{V}(\widehat{\theta}_n) \to 0$ y $\mathbb{E}(\widehat{\theta}_n) \to \theta$, entonces

$$\mathbb{E}\{(\widehat{\theta}_n - \theta)^2\} \to 0$$