EST-24107: Simulación

Profesor: Alfredo Garbuno Iñigo — Otoño, 2023 — *Bootstrap* paramétrico.
Objetivo: En esta sección del curso veremos cómo incorporar mayores supuestos en el modelo de inferencia. Esto con el objetivo de mejorar la estimación por intervalos *bootstrap*. Es decir, para poder reducir la amplitud de los intervalos de confianza. Estudiaremos algunos puntos a considerar que pueden ser vistas como fortalezas y debilidades del método.
Lectura recomendada: Capítulo 6 de Efron and Tibshirani [2]. Sección 13.2 de Chihara and Hesterberg [1].

1. Bootstrap paramétrico

- Supongamos que tenemos una muestra $X_1, \ldots, X_N \overset{\text{iid}}{\sim} \mathbb{P}(x; \theta^\star)$. Es decir, tenemos un `modelo paramétrico` que da lugar a nuestros datos.
- En este tipo de problemas de inferencia suponemos la familia paramétrica

$$\mathcal{P}_\Theta = \{\mathbb{P}(\cdot; \theta) : \theta \in \Theta\} \,, \tag{1}$$

  donde Θ denota el `espacio parametral` (los posibles valores de los parámetros de un modelo).
- En esta tarea no conocemos el valor específico de $\theta^\star$. Por lo tanto, lo tenemos que estimar. Usualmente a través de resolver un problema de optimización

$$\hat{\theta}_{\mathsf{MLE}} = \arg\max_{\theta \in \Theta} \prod_{i=1}^{N} \mathbb{P}(X_i; \theta) \,. \tag{2}$$

  cuya solución llamamos `estimador de máxima verosimilitud`.
- Adicional, nos encantaría poder establecer una cuantificación de la incertidumbre sobre este valor. En particular, reportar

$$\mathsf{ee}\left(\hat{\theta}_{\mathsf{MLE}}\right) = \left(\mathbb{V}(\hat{\theta}_{\mathsf{MLE}})\right)^{1/2} \,. \tag{3}$$

- Para algunos modelos es fácil poder estimarlo, utilizando propiedades asintóticas y/o analíticas de nuestros estimadores (lo ven en el curso de Estadística Matemática).
- Consideremos ejemplos:
  1. Modelo Poisson, $X_1, \ldots, X_N \overset{\text{iid}}{\sim} \mathsf{Poisson}(\lambda)$.
  2. Modelo Bernoulli, $X_1, \ldots, X_N \overset{\text{iid}}{\sim} \mathsf{Bernoulli}(\theta)$.
  3. Modelo uniforme, $X_1, \ldots, X_N \overset{\text{iid}}{\sim} \mathsf{U}(0, \theta)$.
  4. Modelo normal, $X_1, \ldots, X_N \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma^2)$.
- Sin embargo, ¿qué pasa si nuestro estimador no tiene fórmulas cerradas para el cálculo del error estándar? ¿O si nuestro tamaño de muestra no sugiere que los supuestos del TLC se cumplen?

Definición 1.1 (Método *bootstrap* paramétrico)**.** El error estándar estimado para $\hat{\theta}_{\mathsf{MLE}}$ por medio del *bootstrap* paramétrico se calcula como sigue:

1. Se calcula $\hat{\theta}_{\mathsf{MLE}}$ para la muestra observada.
2. Se simula una muestra iid de tamaño N de $X_1^{(b)}, \ldots, X_N^{(b)} \overset{\mathsf{iid}}{\sim} \mathbb{P}(x; \hat{\theta}_{\mathsf{MLE}})$ (muestra *bootstrap*).
3. Se recalcula el estimador de máxima verosimilitud para la muestra *bootstrap*, lo cual denotamos por $\hat{\theta}^{(b)}_{\mathsf{MLE}} = s(X_1^{(b)}, \ldots, X_N^{(b)})$.
4. Se repiten los pasos 2–3 muchas veces ($B = 1{,}000$ – $10{,}000$).
5. Se calcula la desviación estándar de los valores $\hat{\theta}^{(b)}_{\mathsf{MLE}}$ obtenidos. Este es el error estándar estimado para el estimador $\hat{\theta}_{\mathsf{MLE}}$.

Observación (Distribución de remuestreo paramétrica)**.** Considerando el cambio que hacemos de la distribución de remuestreo tenemos lo siguiente:

- El mecanismo de remuestreo cambia $\hat{\mathbb{P}}_N$ por $\mathbb{P}(x; \hat{\theta}_{\mathsf{MLE}})$.
- En espíritu es lo mismo, pero estamos dispuestos a incorporar mayores supuestos en nuestra tarea de inferencia.

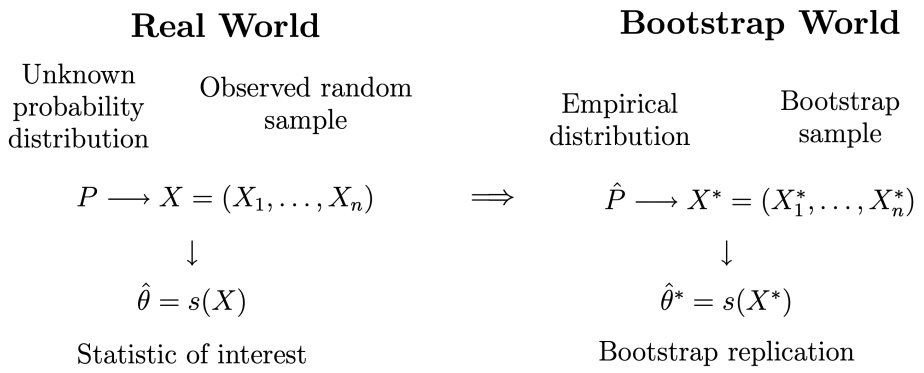

FIGURA 1. *Imagen tomada del material del curso de Cómputo Estadístico de Michael Eichler.*

2. Ejemplo: Datos normales

Como ejercicio, podemos encontrar los estimadores de máxima verosimilitud cuando tenemos una muestra $X_1, \ldots, X_N \overset{\mathsf{iid}}{\sim} \mathsf{N}(\mu, \sigma^2)$ (puedes derivar e igualar a cero para encontrar el mínimo). También podemos resolver numéricamente.

Supongamos que tenemos la siguiente muestra:

```
set.seed(41852)
muestra ← rnorm(150, mean = 1, sd = 2)
```

Para la cual podemos calcular los estimadores de máxima verosimilitud de un modelo normal

```
mle.obs ← broom::tidy(MASS::fitdistr(muestra, "normal")) ▷
  tibble::column_to_rownames("term")
```

```
mle.obs
```

```
     estimate std.error
mean    1.136    0.1502
sd      1.839    0.1062
```

Con esto podemos definir el estimador y el proceso de remuestreo.

```
## paso 1: define el estimador
estimador_mle ← function(datos, modelo = "normal"){
  datos ▷
    MASS::fitdistr(modelo) ▷
    broom::tidy() ▷
    select(-std.error)
}
```

```
## paso 2: define el proceso de remuestreo
paramboot_sample ← function(data){
  rnorm(length(data),
        mean = mle.obs["mean", "estimate"],
        sd = mle.obs["sd", "estimate"])
}
```

```
## paso 3: define el paso bootstrap
paso_bootstrap ← function(id){
  muestra ▷
    paramboot_sample() ▷
    estimador_mle()
}
```

```
## paso 4: aplica bootstrap parametrico
boot_mle ← map_df(1:5000, paso_bootstrap)
```

Las distribuciones son aproximadamente normales. Nótese que esto no siempre sucede, especialmente con parámetros de dispersión como σ. (Examina las curvas de nivel del ejemplo de arriba).

Ahora, supongamos que tenemos una muestra más chica. Repasa los pasos para asegurarte que entiendes el procedimiento:

```
set.seed(4182)
muestra ← rnorm(6, mean = 1, sd = 2)
mle.obs ← broom::tidy(MASS::fitdistr(muestra, "normal")) ▷
  tibble::column_to_rownames("term")
mle.obs
```

```
     estimate std.error
mean   0.3979    0.9794
sd     2.3990    0.6925
```

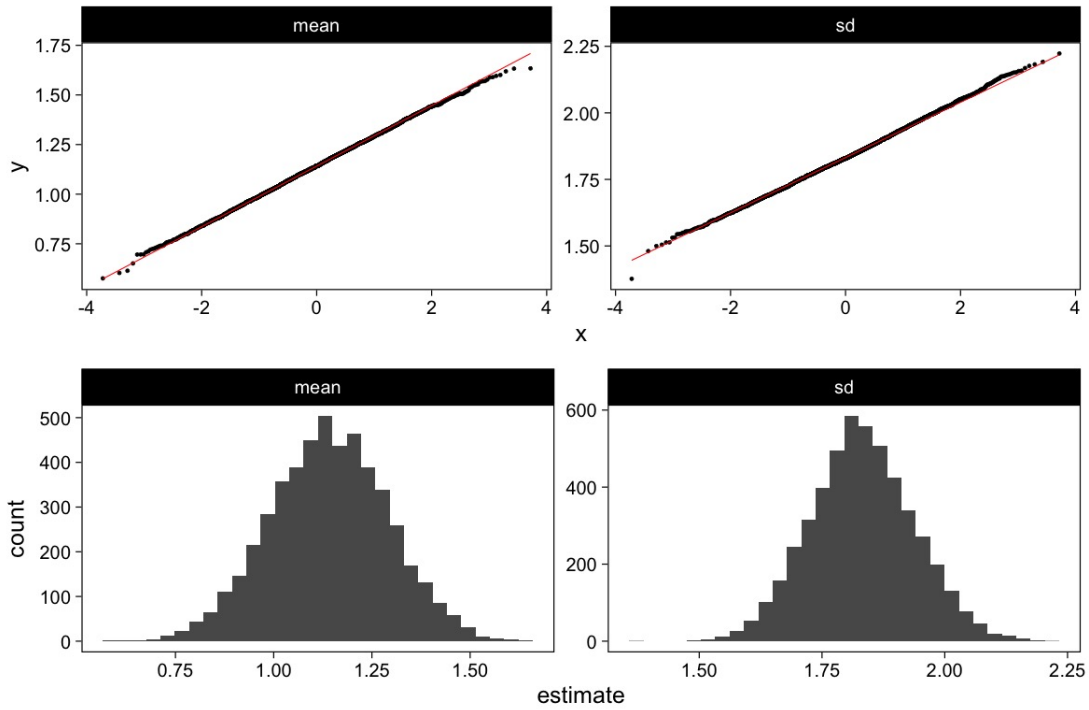

FIGURA 2. *Distribución* bootstrap *para parámetros de media y desviación estandar,* $N = 150$.

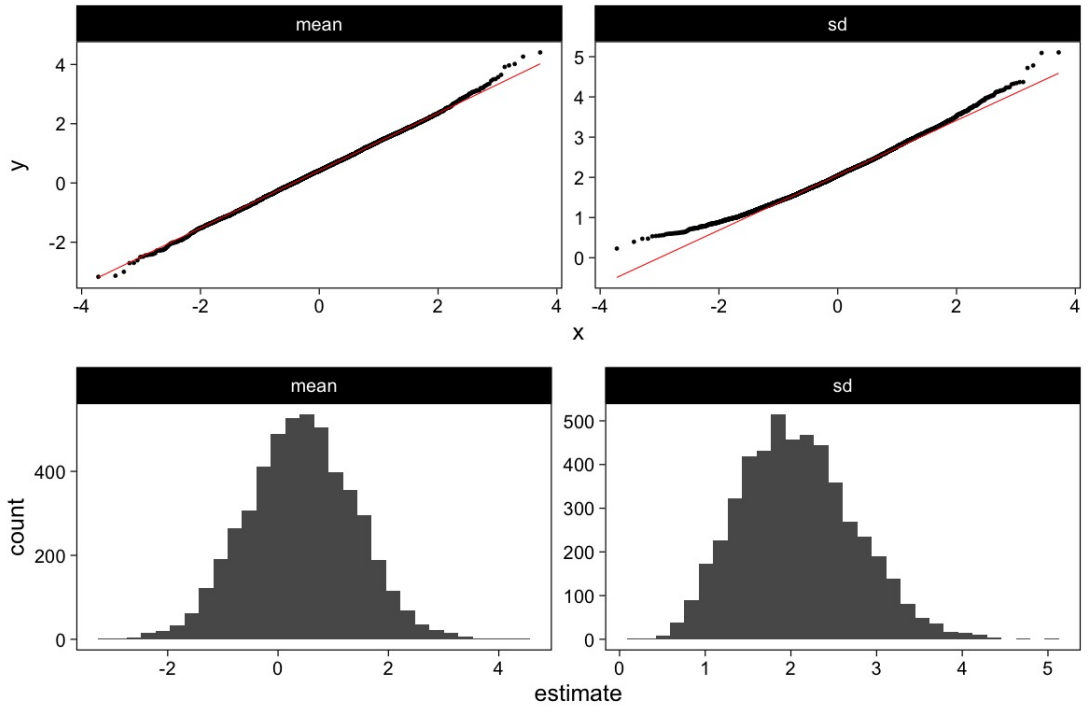

FIGURA 3. *Distribución* bootstrap *para parámetros de media y desviación estandar,* $N = 6$.

```
## paso 4: aplica bootstrap parametrico
boot_mle ← map_df(1:5000, paso_bootstrap)
```

Donde vemos que la distribución de σ tienen sesgo a la derecha, pues en algunos casos obtenemos estimaciones muy cercanas a cero. Podemos usar intervalos de percentiles.

3. Comparación bootstrap paramétrico y no paramétrico

```
propinas ← read_csv("data/propinas.csv",
                    progress = FALSE,
                    show_col_types = FALSE) ▷
  mutate(id = 1:244)
propinas ▷ head()
```

```
# A tibble: 6 × 7
  cuenta_total propina fumador dia   momento num_personas    id
         <dbl>   <dbl> <chr>   <chr> <chr>          <dbl> <int>
1         17.0    1.01 No      Dom   Cena               2     1
2         10.3    1.66 No      Dom   Cena               3     2
3         21.0    3.5  No      Dom   Cena               3     3
4         23.7    3.31 No      Dom   Cena               2     4
5         24.6    3.61 No      Dom   Cena               4     5
6         25.3    4.71 No      Dom   Cena               4     6
```

```
## paso 1: define el estimador
estimador ← function(split, ...){
  muestra ← analysis(split) ▷ group_by(momento)
  muestra ▷
    summarise(estimate = mean(cuenta_total), .groups = 'drop') ▷
    mutate(term = momento)
}
```

```
## paso 2 y 3: remuestrea y calcula estimador
boot_samples ← bootstraps(propinas, strata = momento, 500) ▷
  mutate(res_boot = map(splits, estimador))
## paso 4: construye intervalos de confianza
intervalos_noparam ← boot_samples ▷
  int_pctl(res_boot, alpha = 0.05) ▷
  mutate(across(where(is.numeric), round, 2))
intervalos_noparam
```

```
# A tibble: 2 × 6
  term   .lower .estimate .upper .alpha .method
  <chr>   <dbl>     <dbl>  <dbl>  <dbl> <chr>
1 Cena     19.7      20.8   22.0    0.1 percentile
2 Comida   15.7      17.2   18.7    0.1 percentile
```

Ahora, implementaremos el método *bootstrap* paramétrico.

```
## paso 1: define estimador
estimador_mle_grupos ← function(muestra, modelo = "normal") {
  muestra ▷
    select(momento, cuenta_total) ▷
    group_by(momento) ▷
    nest(data = cuenta_total) ▷
    summarise(mle = map(data, function(x) {
      nobs ← nrow(x)
      unlist(x) ▷
        estimador_mle(modelo = modelo) ▷
        mutate(n = nobs)
    }))
}
```

```
mle.obs ← estimador_mle_grupos(propinas, "normal")
mle.obs ▷ unnest(mle)
```

```
# A tibble: 4 × 4
  momento term  estimate     n
  <chr>   <chr>    <dbl> <int>
1 Cena    mean     20.8    176
2 Cena    sd        9.12   176
3 Comida  mean     17.2     68
4 Comida  sd        7.66    68
```

```
## paso 2: define proceso de remuestreo
param_boot_grupos ← function(estimadores){
  estimadores ▷
    group_by(momento) ▷
    mutate(simulaciones = map(mle, function(m){
      tibble(cuenta_total = rnorm(m$n[1], m$estimate[1], sd = m$estimate[2]))
    })) ▷
    unnest(simulaciones) ▷
    select(-mle) ▷
    ungroup()
}
```

```
## paso 3: paso bootstrap
paso_bootstrap_grupos ← function(id){
  param_boot_grupos(mle.obs) ▷
    estimador_mle_grupos()
}
```

```
## paso 4: aplica bootstrap y presenta intervalos
intervalos_param ← tibble(id = 1:500)▷
  mutate(estimadores = map(id, paso_bootstrap_grupos)) ▷
  unnest(estimadores) ▷
  unnest(mle) ▷
  group_by(momento, term) ▷
  summarise(.lower = quantile(estimate, 0.025),
            .estimate = mean(estimate),
            .upper = quantile(estimate, 0.975),
```

```
            .alpha = .05,
            .method = "percentile (normal)", .groups = "drop") ▷
  filter(term == "mean") ▷ select(-term)
intervalos_param
```

```
# A tibble: 2 × 6
  momento .lower .estimate .upper .alpha .method
  <chr>    <dbl>     <dbl>  <dbl>  <dbl> <chr>
1 Cena      19.6      20.8   22.1    0.1 percentile (normal)
2 Comida    15.3      17.1   18.8    0.1 percentile (normal)
```

```
# A tibble: 2 × 6
  term   .lower .estimate .upper .alpha .method
  <chr>   <dbl>     <dbl>  <dbl>  <dbl> <chr>
1 Cena     19.7      20.8   22.0    0.1 percentile
2 Comida   15.7      17.2   18.7    0.1 percentile
```

```
# A tibble: 1 × 6
  term  .lower .estimate .upper .alpha .method
  <chr>  <dbl>     <dbl>  <dbl>  <dbl> <chr>
1 Cena    17.8      20.8   23.9    0.1 percentile (exponential)
```

El modelo exponencial nos da intervalos mas anchos (mayor incertidumbre) lo cual ilustra que si el modelo paramétrico no es el adecuado, los supuestos adicionales sirven poco para mejorar la estimación de incertidumbre.

4. Ejemplo: Datos de viento

Consideremos los siguientes datos que corresponden datos de producción energética por medio de una turbina de viento. En este caso nos interesa estimar el percentil 10 % pues es lo que esperaríamos que la turbina genere el 90 % de las veces.

```
library(resampledata)
data(Turbine)
Turbine ▷ tibble()
```

Esperamos los problemas usuales de nuestro estimador si utilizáramos el ***bootstrap* no paramétrico**.

```
Turbine ▷
  summarise(estimate = quantile(Production, probs = .1))
```

```
  estimate
1     1817
```

```
library(rsample)
## paso 1: define el estimador
calcula_percentil ← function(split, ...) {
  split ▷
    analysis() ▷
    summarise(estimate = quantile(Production, probs = .1)) ▷
    mutate(term = "Percentil")
}
```

```
nonparam_boot ← bootstraps(Turbine, 1000) ▷
  mutate(resultados = map(splits, calcula_percentil))
```

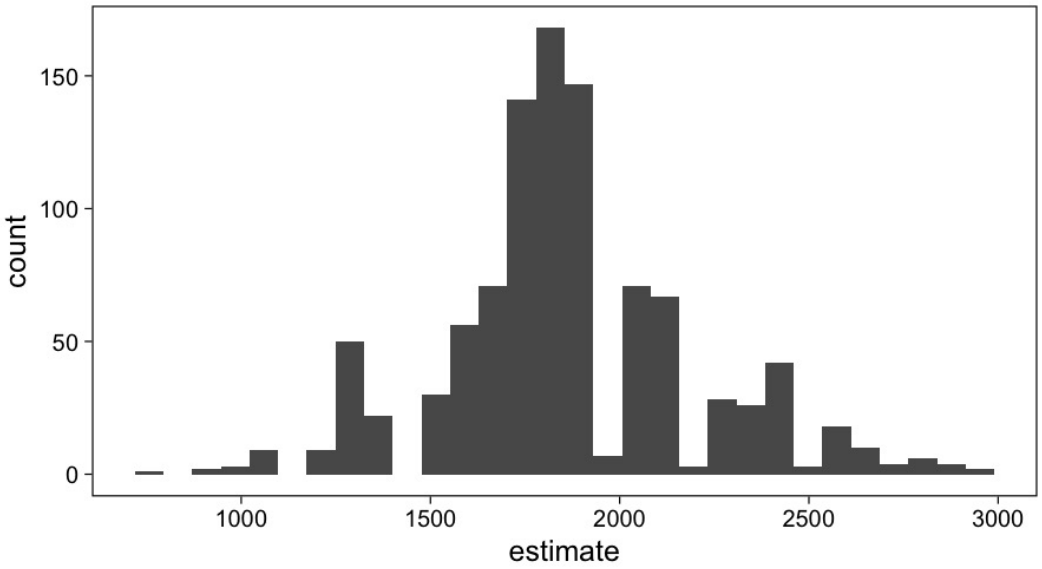

FIGURA 4. *Histograma* bootstrap *de percentil* 10 %.

Si asumimos un modelo $\mathsf{Weibull}(k, \lambda)$ para los datos. Es decir, consideramos $X \sim \mathsf{Weibull}(k, \lambda)$ de tal forma que X tiene función de densidad

$$\pi(x; k, \lambda) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}, & \text{si } x \geq 0\,, \\ 0\,, & \text{si } x < 0\,. \end{cases} \tag{4}$$

Estimando los parámetros obtenemos lo siguiente. Revisa los pasos para asegurarte que queda claro el procedimiento.

```
## paso 1: define el estimador
ajusta_weibull ← function(data){
  tibble(data) ▷
    filter(Production > 0) ▷
    pull(Production) ▷
    MASS::fitdistr("weibull") ▷
    broom::tidy() ▷
    select(-std.error) ▷
    tibble::column_to_rownames("term")
}
```

```

mle.weibull ← ajusta_weibull(Turbine)
mle.weibull
```

```
       estimate
shape     1.283
scale 11795.041
```

```
## paso 2: define el proceso de remuestreo
paramboot_sample ← function(data){
  tibble(Production = rweibull(nrow(data),
                               scale = mle.weibull["scale", "estimate"],
                               shape = mle.weibull["shape", "estimate"])
         )
}
```

```
## paso 1.5: complementa el estimador
extrae_cuantil ← function(params){
  qweibull(scale = params["scale", "estimate"],
           shape = params["shape", "estimate"],
           p = .10) %>%
    tibble(estimate = .)
}
```

```
## paso 3: define el paso bootstrap
paso_bootstrap ← function(id){
  Turbine ▷
    paramboot_sample() ▷
    ajusta_weibull() ▷
    extrae_cuantil()
}
```

```
## paso 4: aplica bootstrap parametrico
param_boot ← map_df(1:1000, paso_bootstrap)
```

Los intervalos de confianza son los siguientes.

```
# A tibble: 1 × 7
  term       .lower .estimate .upper .alpha .method                .length
  <chr>       <dbl>     <dbl>  <dbl>  <dbl> <chr>                    <dbl>
1 Percentil   1261.     1898.  2715.   0.05 percentile (noparam)    1454.
```

```
# A tibble: 1 × 7
  term       .lower .estimate .upper .alpha .method              .length
  <chr>       <dbl>     <dbl>  <dbl>  <dbl> <chr>                  <dbl>
1 Percentil   1699.     2155.  2687.   0.05 percentile (param)     988.
```

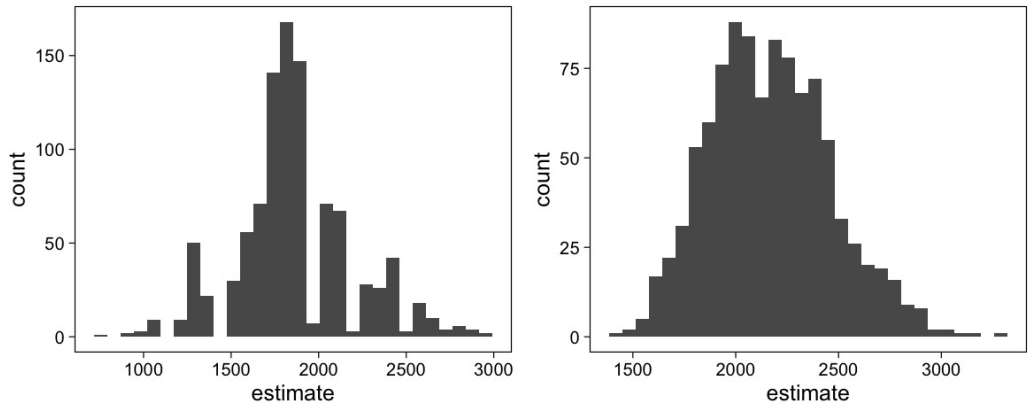

FIGURA 5. *Histogramas* bootstrap *del percentil* 10 %. *En la izquierda utilizando el método no paramétrico, en la derecha utilizando el método paramétrico con distribución Weibull.*

5. El método de momentos

Utilizar máxima verosimilitud **no** es al única manera de poder realizar *bootstrap* paramétrico. Podemos utilizar `el método de momentos`, el cual es otra aplicación directa de la ley de los grandes números.

Definición 5.1 (Método de momentos)**.** Supongamos que queremos estimar k parámetros de un modelo paramétrico $X \sim \mathbb{P}(\cdot;\theta)$. Es decir, queremos realizar inferencia sobre $\theta \in \Theta \subseteq \mathbb{R}^k$. Supongamos que podemos escribir el siguiente sistema de ecuaciones

$$\begin{aligned}\mu_1 &= \mathbb{E}[X] = g_1(\theta_1, \ldots, \theta_k)\,,\\ \mu_2 &= \mathbb{E}[X^2] = g_2(\theta_1, \ldots, \theta_k)\,,\\ &\vdots\\ \mu_k &= \mathbb{E}[X^k] = g_k(\theta_1, \ldots, \theta_k)\,.\end{aligned}$$

Sea $X_1, \ldots, X_N \overset{\text{iid}}{\sim} \mathbb{P}(.;\theta)$ una muestra del modelo probabilístico y denotemos por

$$\hat{\mu}_k = \frac{1}{N}\sum_{n=1}^{N} x_n^k\,, \tag{5}$$

los promedios basados en la muestra. Entonces, el `estimador de momentos` del vector $\theta \in \Theta \subseteq \mathbb{R}^k$ está dado por la solución del sistema de ecuaciones

$$\begin{aligned}\hat{\mu}_1 &= g_1(\hat{\theta}_1, \ldots, \hat{\theta}_k)\,,\\ \hat{\mu}_2 &= g_2(\hat{\theta}_1, \ldots, \hat{\theta}_k)\,,\\ &\vdots\\ \hat{\mu}_k &= g_k(\hat{\theta}_1, \ldots, \hat{\theta}_k)\,.\end{aligned}$$

5.1. `Ejemplo:`

Consideremos los datos $X_1, \ldots, X_N \overset{\text{iid}}{\sim} \mathsf{Gamma}(\alpha, \beta)$ donde tenemos el siguiente sistema de ecuaciones

$$\alpha = \frac{\mathbb{E}(X)^2}{\mathbb{V}(X)}, \qquad \beta = \frac{\mathbb{V}(X)}{\mathbb{E}(X)}.$$

Los cuales podemos estimar utilizando las aproximaciones

$$\mathbb{E}(X^k) \approx \frac{1}{N} \sum_{n=1}^{N} x_n^k. \tag{6}$$

6. Ventajas y desventajas de bootstrap paramétrico

- Ventaja: el *bootstrap* paramétrico puede dar estimadores más precisos e intervalos más angostos y bien calibrados que el no paramétrico, **siempre y cuando el modelo teórico sea razonable.**
- Desventaja: Es necesario decidir el modelo teórico, que tendrá cierto grado de desajuste vs. el proceso generador real de los datos. Si el ajuste es muy malo, los resultados tienen poca utilidad. Para el no paramétrico no es necesario hacer supuestos teóricos.
- Ventaja: el *bootstrap* paramétrico puede ser más escalable que el no paramétrico, pues no es necesario cargar y remuestrear los datos originales, y tenemos mejoras adicionales cuando tenemos expresiones explícitas para los estimadores de máxima verosimilitud (como en el caso normal, donde es innecesario hacer optimización numérica).
- Desventaja: el *bootstrap* paramétrico es conceptualmente más complicado que el no paramétrico, y como vimos arriba, sus supuestos pueden ser más frágiles que los del no paramétrico.

```
sessionInfo()
```

```
R version 4.3.1 (2023-06-16)
Platform: x86_64-apple-darwin20 (64-bit)
Running under: macOS Ventura 13.5.2

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/4.3-x86_64/Resources/lib/
    libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.3-x86_64/Resources/lib/
    libRlapack.dylib;  LAPACK version 3.11.0

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

time zone: America/Mexico_City
tzcode source: internal

attached base packages:
[1] stats     graphics  grDevices datasets  utils     methods   base

other attached packages:
 [1] rsample_1.1.1      resampledata_0.3.1 scales_1.2.1       patchwork_1.1.2
```

```
 [5] lubridate_1.9.2    forcats_1.0.0      stringr_1.5.0      dplyr_1.1.2
 [9] purrr_1.0.1        readr_2.1.4        tidyr_1.3.0        tibble_3.2.1
[13] ggplot2_3.4.2      tidyverse_2.0.0

loaded via a namespace (and not attached):
 [1] utf8_1.2.3         future_1.33.0      generics_0.1.3     renv_1.0.0
 [5] stringi_1.7.12     listenv_0.9.0      hms_1.1.3          digest_0.6.33
 [9] magrittr_2.0.3     grid_4.3.1         timechange_0.2.0   backports_1.4.1
[13] fansi_1.0.4        codetools_0.2-19   cli_3.6.1          rlang_1.1.1
[17] crayon_1.5.2       parallelly_1.36.0  bit64_4.0.5        munsell_0.5.0
[21] withr_2.5.0        tools_4.3.1        parallel_4.3.1     tzdb_0.4.0
[25] colorspace_2.1-0   globals_0.16.2     broom_1.0.5        vctrs_0.6.3
[29] R6_2.5.1           lifecycle_1.0.3    bit_4.0.5          vroom_1.6.3
[33] MASS_7.3-60        furrr_0.3.1        pkgconfig_2.0.3    pillar_1.9.0
[37] gtable_0.3.3       glue_1.6.2         tidyselect_1.2.0   farver_2.1.1
[41] labeling_0.4.2     compiler_4.3.1
```

Referencias

[1] L. M. Chihara and T. C. Hesterberg. *Mathematical Statistics with Resampling and R.* John Wiley & Sons, Inc., Hoboken, NJ, USA, aug 2018. ISBN 978-1-119-50596-9 978-1-119-41654-8. . 1

[2] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap.* Springer US, Boston, MA, 1993. 1