EST-24107: Simulación

Profesor: Alfredo Garbuno Iñigo — Otoño, 2022 — *Bootstrap*.
Objetivo: Que veremos.
Lectura recomendada: Referencia.

1. INTRODUCCIÓN

El remuestreo se refiere a un conjunto de técnicas estadísticas, computacionalmente intensivas, que `estiman` la *distribución de una población* basadas en `muestreo aleatorio con reemplazo`.

Se considera que una muestra aleatoria X1 , X2 , . . . , Xn como si fuera una población finita y se generan muestras aleatorias de la misma muestra para estimar características poblacionales y hacer inferencia de la población muestreada.

Las técnicas de remuestreo permiten calcular medidas de ajuste (en términos de sesgo, varianza, intervalos de confianza, errores de predicción o de algunas otras medidas) a los estimados basados en muestras.

Estas técnicas son usualmente no paramétricas, y varias son tan antiguas como la estadística misma. Por ejemplo, las técnicas de permutación son de Fisher (1935) y Pitmann (1937); la validación cruzadas fue propuesta por Kurtz en 1948, y el Jackknife fue propuesto por Maurice Quenouille en 1949 aunque fue John Tukey en 1958 quién le dio el nombre a la técnica.

1.1. Contexto histórico

Bradley Efron introdujo el Bootstrap en 1979, y sus estudiantes Rob Tibshirani y Trevor Hastie han aportado mucho a la ciencia estadística. Ofrecen un curso en Statistical Learning en la plataforma MOOC de la Universidad de Stanford.

El término 'bootstrapping' se refiere al concepto de "pulling oneself up by one's bootstraps", frase que aparentemente se usó por primera vez en The Singular Travels, Campaigns and Adventures of Varon Munchausen.

1.2. Idea general

El objetivo del remuestreo es estimar alguna característica poblacional, representada por (tal como media, mediana, desviación estándar, coeficientes de regresión, matriz de covarianza, etc.) basado en los datos.

También interesan las propiedades de la distribución de estimador, sin hacer supuestos restrictivos sobre la forma de la distribución de los datos originales.

Para una muestra aleatoria X1 , . . . , Xn , la distribución de remuestreo es la distribución empírica Fn , que asigna probabilidad 1/n a cada una de las observaciones de la muestra.

1.3. Ejemplo

Consideremos una muestra de 6 parejas. La variable de interés es la diferencia del ingreso de los miembros de cada pareja (en miles de pesos al mes).

i	$P_i^{(1)}$	$P_i^{(2)}$	$d_i = P_i^{(1)} - P_i^{(2)}$
1	24	18	6
2	14	17	-3
3	40	35	5
4	44	41	3
5	35	37	-2
6	45	45	0

Definamos θ como el promedio de las diferencias de ingreso poblacional. Podemos estimar θ con

$$\hat{\theta}_n = \frac{6-3+5+3-2+0}{6} = 1.5\,. \tag{1}$$

¿Cómo calculamos la variabilidad de nuestro estimador? Es decir, ¿cómo calculamos la variabilidad de $\hat{\theta}_n$?

1.3.1. ***Ejercicio:*** Escribe la fórmula del error estándar bajo los siguientes supuestos:

1. La diferencia tiene una distribución $d_i \sim \mathsf{N}(\theta, \sigma^2)$.
2. La varianza σ^2 es conocida.

1.4. `Observaciones`:

- Suponer que la diferencia de ingresos es d_i como una variable normal puede no estar *tan* errado. Pues con un número suficiente de muestras podríamos suponer que el resultado del TLC se cumple. Entonces, ¿qué hacemos si no conocemos la distribución de las observaciones?
- Si no conocemos σ^2 lo podemos estimar con la muestra. Por ejemplo, podemos utilizar intervalos de confianza derivados de una distribución t.
- Si nos interesa otro parámetro de la población podemos construir estimadores diferentes. Por ejemplo, nos podría interesar la `mediana` de una población $q_{0.5} = \mathbb{P}^{-1}(1/2)$. Para este caso, podemos estimar dicho parámetro por medio de

$$\hat{q}_{0.5} = \begin{cases} X_{(\frac{n+1}{2})} & \text{si } n \text{ es impar} \\ \frac{X_{(n/2)}+X_{(n/2+1)}}{2} & \text{si } n \text{ es par} \end{cases}\,. \tag{2}$$

En Fig. 1 la estimación de la mediana en distintos grupos acompañados de su estimación de incertidumbre.

2. LA IDEA DEL BOOTSTRAP

Como explicamos, el problema que tenemos ahora es que normalmente sólo tenemos una muestra, así que no es posible calcular las distribuciones de muestreo como hicimos arriba y evaluar qué tan preciso es nuestro estimador. Sin embargo, podemos hacer lo siguiente:

Supongamos que tenemos una muestra $X_1, X_2, \ldots, X_n$ independientes de alguna población desconocida y un estimador $T = t(X_1, \ldots, X_n)$

Mundo poblacional

1. Si tuviéramos la distribución poblacional, simulamos muestras iid para

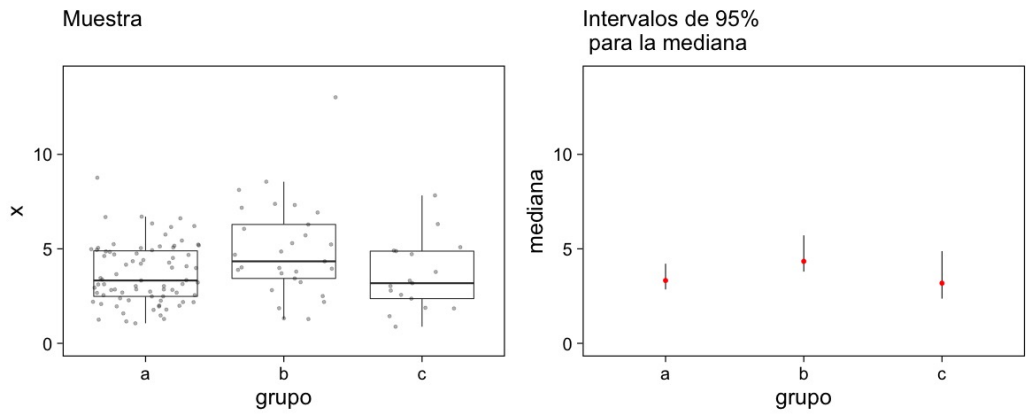


FIGURA 1. *Estimación de mediana (panel izquierdo) con intervalos de incertidumbre (panel derecho).*

aproximar la distribución de muestreo de nuestro estimador, y así entender su variabilidad.

1. Pero **no** tenemos la distribución poblacional.
2. **Sin embargo, podemos estimar la distribución poblacional con nuestros valores muestrales**.

Mundo bootstrap

1. Si usamos la estimación del inciso 3, entonces usando el inciso 1 podríamos

tomar muestras de nuestros datos muestrales, como si fueran de la población, y usando el mismo tamaño de muestra. El muestreo lo hacemos con reemplazo de manera que produzcamos muestras independientes de la misma "población estimada", que es la muestra.

1. Evaluamos nuestra estadística en cada una de estas remuestras.
2. A la distribución resultante le llamamos **distribución bootstrap** o

distribución de remuestreo del estimador.

1. Usamos la distribución bootstrap de la muestra para estimar la variabilidad

en nuestra estimación con **la muestra original**.

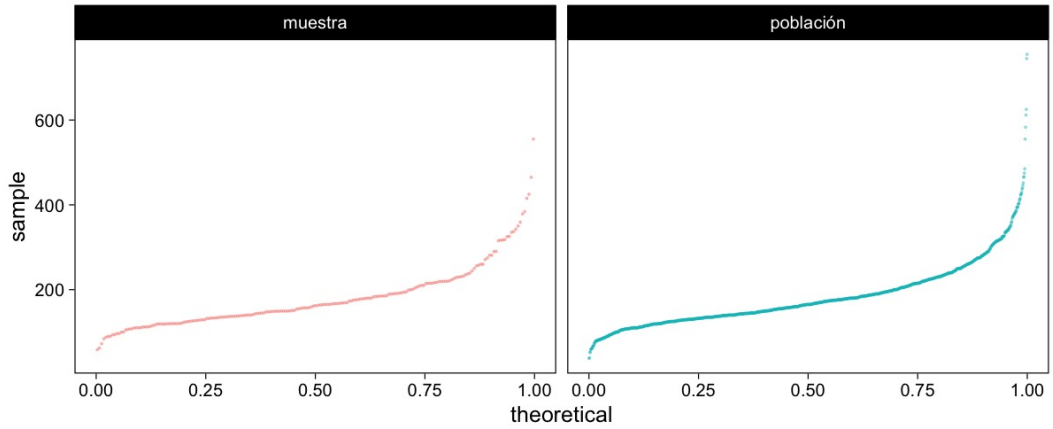
Veamos que sucede para un ejemplo concreto, donde nos interesa estimar la media de los precios de venta de una población de casas. Tenemos nuestra muestra:

```
set.seed(2112)
poblacion_casas ← read_csv("data/casas.csv")
muestra ← sample_n(poblacion_casas, 200, replace = TRUE)
```

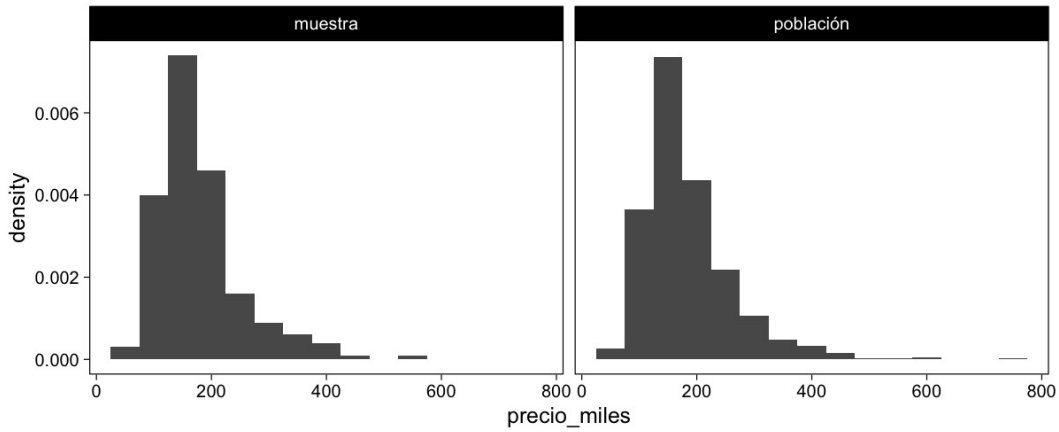
```
mean(muestra$precio_miles)
```

```
[1] 180
```

Esta muestra nos da nuestro estimador de la distribución poblacional. Por ejemplo, podemos fijarnos en un gráfico de cuantiles:



O en histogramas:



Y vemos que la aproximación es razonable en las partes centrales de la distribución.

Ahora supongamos que nos interesa cuantificar la precisión de nuestra estimación de la media poblacional de precios de casas, y usaremos la media muestral para hacer esto. Para nuestra muestra, nuestra estimación puntual es:

```
media ← mean(muestra$precio_miles)
media
```

```
[1] 180
```

Y recordamos que para aproximar la distribución de muestreo podíamos muestrear repetidamente la población y calcular el valor del estimador en cada una de estas muestras. Aquí no tenemos la población, **pero tenemos una estimación de la población**: la muestra obtenida.

Así que para evaluar la variabilidad de nuestro estimador, entramos en el mundo boostrap, y consideramos que la población es nuestra muestra.

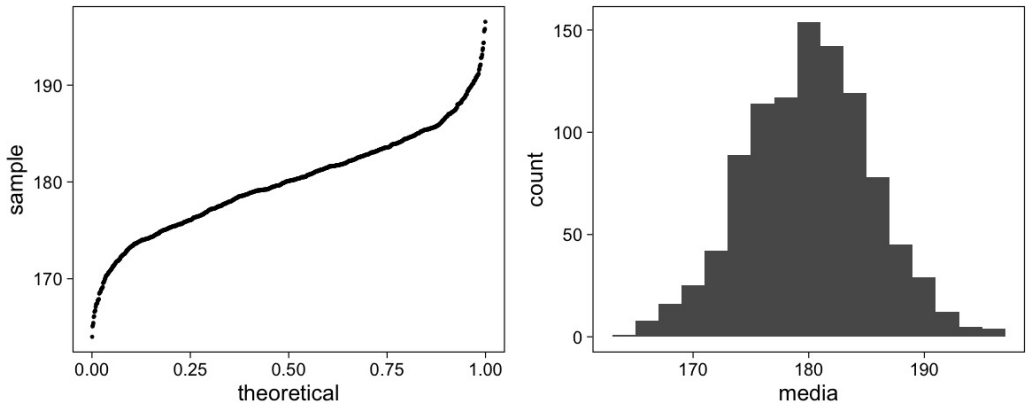
Podemos entonces extraer un número grande de muestras con reemplazo de tamaño 200 **de la muestra**: el muestreo debe ser análogo al que se tomó para nuestra muestra original. Evaluamos nuestra estadística (en este caso la media) en cada una de estas remuestras:

```
media_muestras ← map_dbl(1:1000, ~ muestra %>%
        sample_n(200, replace = T) %>%
```

```
        summarise(media_precio = mean(precio_miles), .groups = "drop") %>%
        pull(media_precio))
media_muestras[1:10]
```

```
[1] 176.4 176.0 175.9 176.4 177.8 186.8 179.2 181.6 175.1 177.1
```

Y nuestra estimación de la distribución de muestreo para la media es entonces:



A esta le llamamos la distribución de remuestreo de la media, que definimos más abajo. Ahora podemos calcular un intervalo de confianza del 90\calculando los cuantiles de esta distribución (no son los cuantiles de la muestra original!):

```
limites_ic ← quantile(media_muestras, c(0.05,  0.95)) %>% round
limites_ic
```

```
 5% 95%
171 189
```

Otra cosa que podríamos hacer para describir la dispersión de nuestro estimador es calcular el error estándar de remuestreo, que estima el error estándar de la distribución de muestreo:

```
ee_boot ← sd(media_muestras)
round(ee_boot, 2)
```

```
[1] 5.4
```

2.0.1. ***Definición:*** Sea $X_1, X_2, \ldots, X_n$ una muestra independiente y idénticamente distribuida, y $T = t(X_1, X_2, \ldots, X_n)$ una estadística. Supongamos que sus valores que obervamos son $x_1, x_2, \ldots, x_n$.

La **distribución de remuestreo** de T es la distribución de $T^* = t(X_1^*, X_2^*, \ldots X_n^*)$, donde cada X_i^* se obtiene tomando al azar uno de los valores de $x_1, x_2, \ldots, x_n$.

Otra manera de decir esto es que la remuestra $X_1^*, X_2^*, \ldots, X_n^*$ es una muestra con reemplazo de los valores observados $x_1, x_2, \ldots, x_n$

2.0.2. Ejemplo: Si observamos la muestra

```
muestra ← sample(1:20, 5)
muestra
```

```
[1] 15 16  8  6  2
```

Una remuestra se obtiene:

```
sample(muestra, size = 5, replace = TRUE)
```

```
[1]  2 15  2 15  2
```

Nótese que algunos valores de la muestra original pueden aparecer varias veces, y otros no aparecen del todo.

2.1. Nota

La idea del bootstrap. La muestra original es una aproximación de la población de donde fue extraída. Así que remuestrear la muestra aproxima lo que pasaría si tomáramos muestras de la población. La **distribución de remuestreo** de una estadística, que se construye tomando muchas remuestras, aproxima la distribución de muestreo de la estadística.

Y el proceso que hacemos es:

2.1.1. Remuestreo para una población: Dada una muestra de tamaño n de una población,

1. Obtenemos una remuestra de tamaño n con reemplazo de la muestra original
2. Repetimos este remuestreo muchas veces (por ejemplo, 10,000).
3. Construímos la distribución bootstrap, y examinamos sus características

(dónde está centrada, dispersión y forma).

3. EL PRINCIPIO DE PLUG-IN

La idea básica detrás del *bootstrap* es el principio de *plug-in* para estimar parámetros poblacionales: si queremos estimar una cantidad poblacional, calculamos esa cantidad poblacional con la muestra obtenida. Es un principio común en estadística.

Por ejemplo, si queremos estimar la media o desviación estándar poblacional, usamos la media muestral o la desviación estándar muestral. Si queremos estimar un cuantil de la población usamos el cuantil correspondiente de la muestra, y así sucesivamente.

En todos estos casos, lo que estamos haciendo es:

- Tenemos una fórmula para la cantidad poblacional de interés en términos de la

distribución poblacional.

- Tenemos una muestra, que usamos para estimar la cantidad poblacional. La

distribución que da una muestra se llama distribución **empírica**.

- Contruimos nuestro estimador .[en]chufando"la distribución empírica de la

muestra en la fórmula del estimador.

En el bootstrap aplicamos este principio simple a la **distribución de muestreo**:

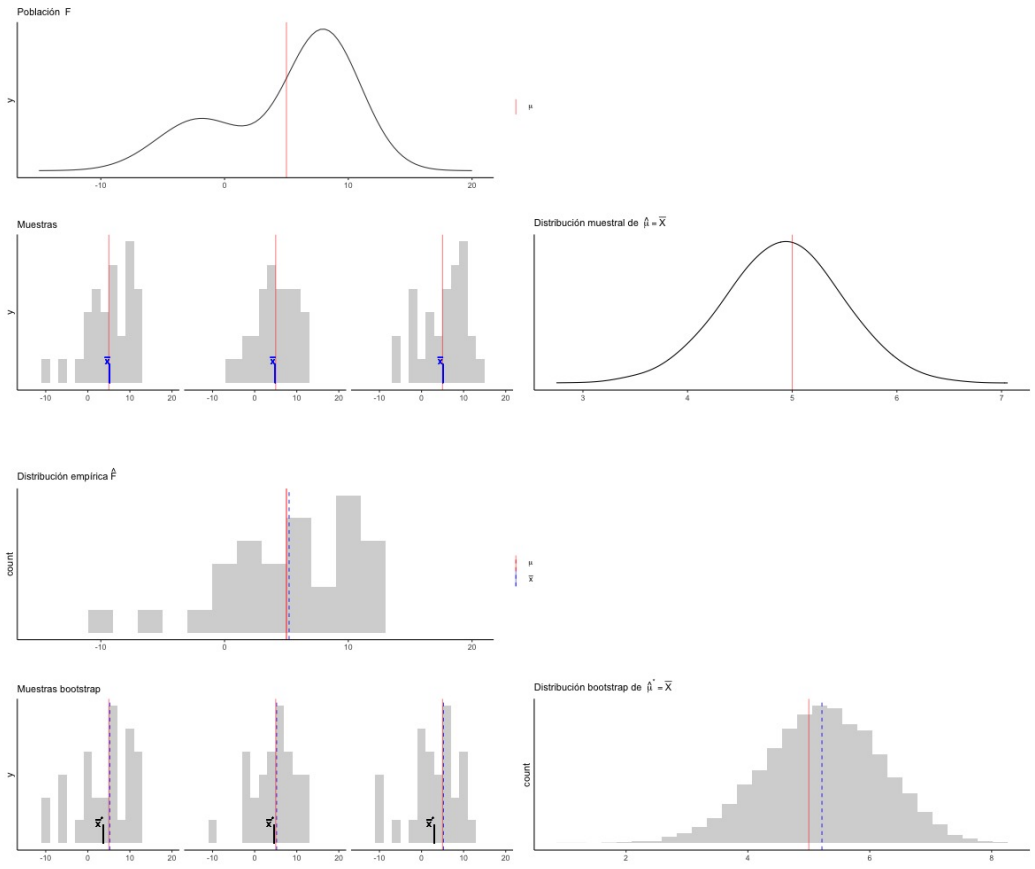
- **Si tenemos la población**, podemos **calcular** la distribución de muestreo de

nuestro estimador tomando muchas muestras de la **población**.

- Estimamos la **poblacion** con la **muestra** y enchufamos en la frase anterior:
- Podemos **estimar** la distribucion de muestreo de nuestro estimador

tomando muchas muestras de la **muestra** (bootstrap).

Nótese que el proceso de muestreo en el último paso **debe ser el mismo** que se usó para tomar la muestra original. Estas dos imágenes simuladas con base en un ejemplo de [1] muestran lo que acabamos de describir:



3.1. Observación:

Veremos ejemplos más complejos, pero nótese que si la muestra original son observaciones independientes obtenidas de la distribución poblacional, entonces logramos esto en las remuestras tomando observaciones con reemplazo de la muestra. Igualmente, las remuestras deben ser del mismo tamaño que la muestra original.

3.1.1. Ejercicio:

- ¿Porqué no funcionaría tomar muestras sin reemplazo? Piensa si hay independencia

entre las observaciones de la remuestra, y cómo serían las remuestras sin reemplazo.

- ¿Por qué no se puede hacer bootstrap si no conocemos cómo se obtuvo la muestra original?

3.2. Observación

Estos argumentos se pueden escribir con fórmulas usando por ejemplo la función de distribución acumulada F de la población y su estimador, que es la función empírica $\hat{F}$, como en [**?**]. Si $\theta = t(F)$ es una cantidad poblacional que queremos estimar, su estimador *plug-in* es $\hat{\theta} = t(\hat{F})$.

3.3. Observación

La distribución empírica $\hat{F}$ es un estimador razonable"de la distribución poblacional $F,$ pues por el teorema de Glivenko-Cantelli ([**?**], o [aquí](https://en.wikipedia.org/wiki/Glivenko-Cantelli_theorem)), $\hat{F}$ converge a F cuando el tamaño de muestra $n \to \infty$, lo cual es intuitivamente claro.

3.4. Ejemplo

En el siguiente ejemplo (tomadores de té), podemos estimar la proporción de tomadores de té que prefiere el té negro usando nuestra muestra:

```
te ← read_csv("data/tea.csv") ▷
  rowid_to_column() ▷
  select(rowid, Tea, sugar)
```

```
te ▷
  mutate(negro = ifelse(Tea == "black", 1, 0)) ▷
  summarise(prop_negro = mean(negro), n = length(negro), .groups = "drop")
```

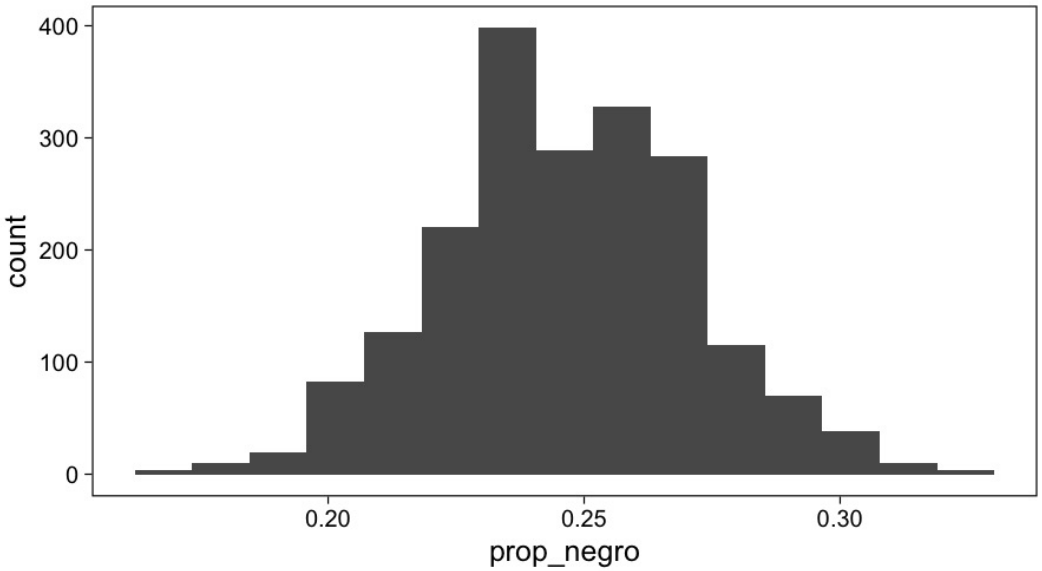
```
# A tibble: 1 × 2
  prop_negro     n
       <dbl> <int>
1      0.247   300
```

¿Cómo evaluamos la precisión de este estimador? Supondremos que el estudio se hizo tomando una muestra aleatoria simple de tamaño 300 de la población de tomadores de té que nos interesa. Podemos entonces usar el boostrap:

```
## paso 1: define el estimador
calc_estimador ← function(datos){
  prop_negro ← datos ▷
    mutate(negro = ifelse(Tea == "black", 1, 0)) ▷
    summarise(prop_negro = mean(negro), n = length(negro), .groups = "drop") ▷
    pull(prop_negro)
  prop_negro
}
```

```
## paso 2: define el proceso de remuestreo
muestra_boot ← function(datos){
  ## tomar muestra con reemplazo del mismo ñtamao
  sample_n(datos, size = nrow(datos), replace = TRUE)
}
```

```
# paso 3: remuestrea y calcula el estimador
prop_negro_tbl ← map_dbl(1:2000,  ~ calc_estimador(muestra_boot(datos = te)))
    %>%
  tibble(prop_negro = .)
```



Y podemos evaluar varios aspectos, por ejemplo dónde está centrada y qué tan dispersa es la distribución bootstrap:

```
prop_negro_tbl ▷
  summarise(media = mean(prop_negro),
            sesgo = mean(prop_negro) - 0.2499,
            ee = sd(prop_negro),
            cuantil_75 = quantile(prop_negro, 0.75),
            cuantil_25 = quantile(prop_negro, 0.25),
            .groups = "drop") ▷
  mutate(across(where(is.numeric), round, 3)) ▷
  pivot_longer(cols = everything())
```

```
# A tibble: 5 × 2
  name         value
  <chr>        <dbl>
1 media        0.246
2 sesgo       -0.003
3 ee           0.025
4 cuantil_75   0.263
5 cuantil_25   0.23
```

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. . 7