# Tarea 08: conteo rápido

Rodrigo Alan Garcia Perez

2024-10-09

**Antecedentes** En México, las elecciones tienen lugar un domingo, los resultados oficiales del proceso se presentan a la población una semana después. A fin de evitar proclamaciones de victoria injustificadas durante ese periodo el INE organiza un conteo rápido.

El conteo rápido es un procedimiento para estimar, a partir de una muestra aleatoria de casillas, el porcentaje de votos a favor de cada opción en la boleta.

En 2021 se realizón un conteo rápido para estimar los resultados de la consulta popular 2021.

**Consulta popular**

**Diseño de la muestra** El diseño utilizado es *muestreo estratificado simple*, lo que quiere decir que:

i) se particionan las casillas de la pablación en estratos (cada casilla pertenece a exactamente un estrato), y

ii) dentro de cada estrato se usa *muestreo aleatorio* para seleccionar las casillas que estarán en la muestra.

**Estimación** Una de las metodolgías de estimación, que se usa en el conteo rápido (tanto de elecciones como en consultas), es *estimador de razón combinado*, contruyendo intervalos de 95% de confianza usando el método normal con error estándar bootstrap. En este ejercicio debes construir intervalos usando este procedimiento.

**Para cada opción en la consulta (sí/no/nulos) usarás la muestra del conteo rápido para estimar los resultados de la consulta**.

1. Calcula el estimador de razón combinado, para muestreo estratificado la fórmula es:

$$\hat{p} = \frac{\sum_h \frac{N_h}{n_h} \sum_i Y_{hi}}{\sum_h \frac{N_h}{n_h} \sum_i X_{hi}}$$

donde:

- $\hat{p}$ es la estimación de la proporción de votos que recibió la opción (ej: *sí*).
- $Y_{hi}$ es el número total de votos que recibió *la opción* (ej: *sí*) en la $i$-ésima casillas, que pertence al $h$-ésimo estrato.
- $X_{hi}$ es el número total de votos en la $i$-ésima casilla, que pertence al $h$-ésimo estrato.
- $N_h$ es el número total de casillas en el $h$-ésimo estrato.
- $n_h$ es el número de casillas del $h$-ésimo estrato que se seleccionaron en la muestra.

**Datos**

- Tamaño de los estratos ($N_h$)
- Muestra del conteo rápido usada en la estimación aquí

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dplyr)

estratos_tam <- read_csv("datos/tamaño_estratos.csv")
```

```
## Rows: 300 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): ESTRATO
## dbl (1): N
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
muestra <- read_csv("datos/muestra.csv")
```

```
## Rows: 1745 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): ID, ESTRATO
## dbl (6): SI, NO, NULOS, TOTAL, LISTA_NOMINAL, n
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Vistazo a los datos
glimpse(estratos_tam)
```

```
## Rows: 300
## Columns: 2
## $ ESTRATO <chr> "0101", "0102", "0103", "0201", "0202", "0203", "0204", "0205"~
## $ N       <dbl> 196, 198, 211, 205, 204, 213, 237, 227, 214, 228, 230, 181, 15~
```

```
glimpse(muestra)
```

```
## Rows: 1,745
## Columns: 8
## $ ID            <chr> "010355B01", "010400C01", "010403C02", "010432C01", "010~
## $ ESTRATO       <chr> "0101", "0101", "0101", "0101", "0101", "0101", "0102", ~
## $ SI            <dbl> 26, 7, 56, 64, 104, 49, 24, 64, 130, 121, 26, 26, 64, 93~
## $ NO            <dbl> 0, 1, 0, 3, 1, 1, 1, 3, 0, 1, 0, 1, 0, 2, 1, 1, 0, 3, 1,~
```

```
## $ NULOS         <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0,~
## $ TOTAL         <dbl> 26, 8, 56, 67, 106, 50, 25, 67, 130, 122, 26, 27, 64, 96~
## $ LISTA_NOMINAL <dbl> 1422, 1864, 1533, 1722, 1959, 1900, 1850, 1882, 1965, 19~
## $ n             <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 5,~
```

```r
# Calcular el estimador de razón combinado para la opción "Sí"
estimacion_p <- muestra %>%
  left_join(estratos_tam, by = "ESTRATO") %>%  # Unir las tablas por el estrato
  group_by(ESTRATO) %>%
  summarise(
    N_h = first(N),
    n_h = first(n),
    Y_h = sum(SI),
    X_h = sum(TOTAL)
  ) %>%
  summarise(
    numerador = sum((N_h / n_h) * Y_h),
    denominador = sum((N_h / n_h) * X_h)
  ) %>%
  mutate(estimacion_p = numerador / denominador)

# Mostrar el resultado
estimacion_p$estimacion_p
```

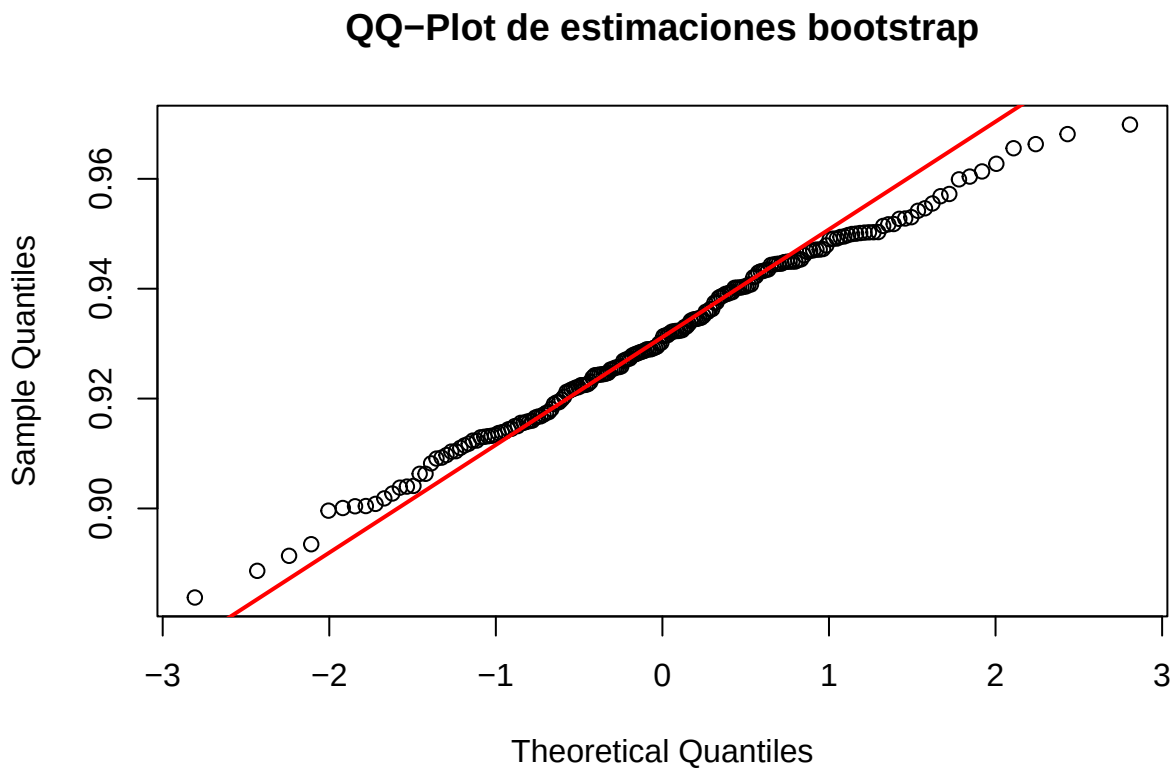
```
## [1] 0.9295214
```

La proporción estimada de votos "Sí" en la consulta es de aproximadamente 92.95%. Esto significa que, según la muestra seleccionada y los ajustes por estratos, alrededor del 93% de los votos emitidos fueron a favor de la opción "Sí".

2. Utiliza **bootstrap** para calcular el error estándar, y reporta tu estimación del error.
   - Genera 200 muestras bootstrap.
   - Recuerda que las muestras bootstrap tienen que tomar en cuenta la metodología que se utilizó en la selección de la muestra original, en este caso implica que para cada remuestra debes tomar muestra aleatoria independiente dentro de cada estrato.

```r
library(ggplot2)

# Funcion para calcular p en una muestra bootstrap

calcular_estimacion_p <- function(muestra_bootstrap, estratos_tam){
  muestra_bootstrap %>%
    left_join(estratos_tam, by = "ESTRATO") %>%
    group_by(ESTRATO) %>%
    summarise(
      N_h = first(N),
      n_h = first(n),
      Y_h = sum(SI),
      X_h = sum(TOTAL)
    ) %>%
    summarise(
      numerador = sum((N_h / n_h) * Y_h),
      denominador = sum((N_h / n_h) * X_h),
      estimacion_p = numerador / denominador
    ) %>%
    pull(estimacion_p)
```

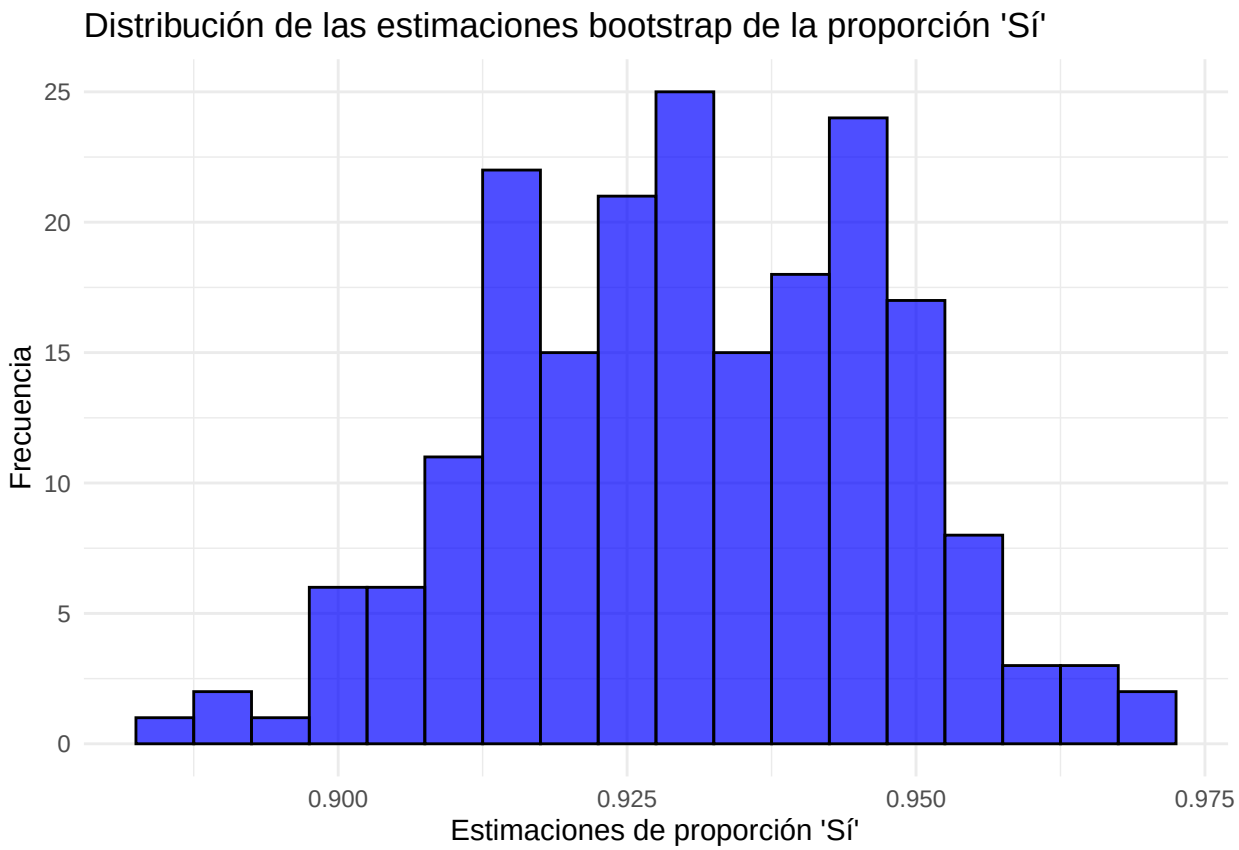
```r
}

# Simulación Bootstrap
set.seed(123)
n_bootstrap <- 200  # Número de muestras bootstrap
bootstrap_estimaciones <- numeric(n_bootstrap)  # Vector para guardar los resultados

for (i in 1:n_bootstrap) {
  # Remuestrear dentro de cada estrato con reemplazo
  muestra_bootstrap <- muestra %>%
    group_by(ESTRATO) %>%
    sample_frac(replace = TRUE)  # Remuestreo con reemplazo dentro de cada estrato

  # Calcular estimación para la muestra bootstrap
  bootstrap_estimaciones[i] <- calcular_estimacion_p(muestra_bootstrap, estratos_tam)
}

# Crear el QQ-Plot
qqnorm(bootstrap_estimaciones, main = "QQ-Plot de estimaciones bootstrap")
qqline(bootstrap_estimaciones, col = "red", lwd = 2)
```




El QQ-Plot sugiere que es razonable asumir que las estimaciones bootstrap se distribuyen de forma aproximadamente normal, con algunas ligeras desviaciones en los valores más extremos.

```r
# Crear el histograma de la distribución de las estimaciones bootstrap
ggplot(data = data.frame(bootstrap_estimaciones), aes(x = bootstrap_estimaciones)) +
  geom_histogram(binwidth = 0.005, fill = "blue", color = "black", alpha = 0.7) +
  labs(title = "Distribución de las estimaciones bootstrap de la proporción 'Sí'",
       x = "Estimaciones de proporción 'Sí'",
       y = "Frecuencia") +
```

```
  theme_minimal()
```




```
# Calcular el error estándar del estimador
error_estandar <- sd(bootstrap_estimaciones)

# Mostrar el resultado
error_estandar
```

```
## [1] 0.01673248
```

```
# theta_hat: Estimación central de la proporción (previamente calculada)
theta_hat <- 0.9295

sd_bootstrap <- error_estandar  # El error estándar calculado con el método bootstrap

# Crear el intervalo de confianza usando 2 * error estándar
IC_normal <- round(c(theta_hat - 2 * sd_bootstrap, theta_hat + 2 * sd_bootstrap), 5)

# Mostrar el intervalo de confianza
IC_normal
```

```
## [1] 0.89604 0.96296
```

El error estándar de 0.0167 indica que la variabilidad en las estimaciones de la proporción de votos “Sí” es baja, lo que refuerza la precisión de la estimación central. El intervalo de confianza de (89.60%, 96.30%) sugiere que, con un 95% de confianza, la proporción real de votos “Sí” en la población se encuentra en ese rango, lo que indica un fuerte apoyo hacia esta opción.