Tarea 05: distribución muestral y Bootstrap

Dalia Camacho

Parte A: Distribución muestral

Consideramos la base de datos primaria, y la columna de calificaciones de español 3o de primaria (esp_3).

```
library(tidyverse)
library(knitr)
primarias <- read_csv("primarias.csv")
set.seed(66454)
```

Seleccina una muestra de tamaño n = 10, 100, 1000. Para cada muestra calcula media y el error estándar de la media usando el principio del plug-in: $\hat{\mu} = \bar{x}$ y $\widehat{se}(\bar{x}) = \hat{\sigma}_{p_n}/\sqrt{(n)}$.

```
Muestra_10   <- sample_n(primarias, 10)
Muestra_100  <- sample_n(primarias, 100)
Muestra_1000 <- sample_n(primarias, 1000)

df <- as.data.frame(cbind(n=c(10,100,1000),
                          mu_plug = c(mean(Muestra_10$esp3),
                                      mean(Muestra_100$esp3),
                                      mean(Muestra_1000$esp3)),
                          se_plug = c(sd(Muestra_10$esp3)/sqrt(10),
                                      sd(Muestra_100$esp3)/sqrt(100),
                                      sd(Muestra_1000$esp3)/sqrt(1000))))
```

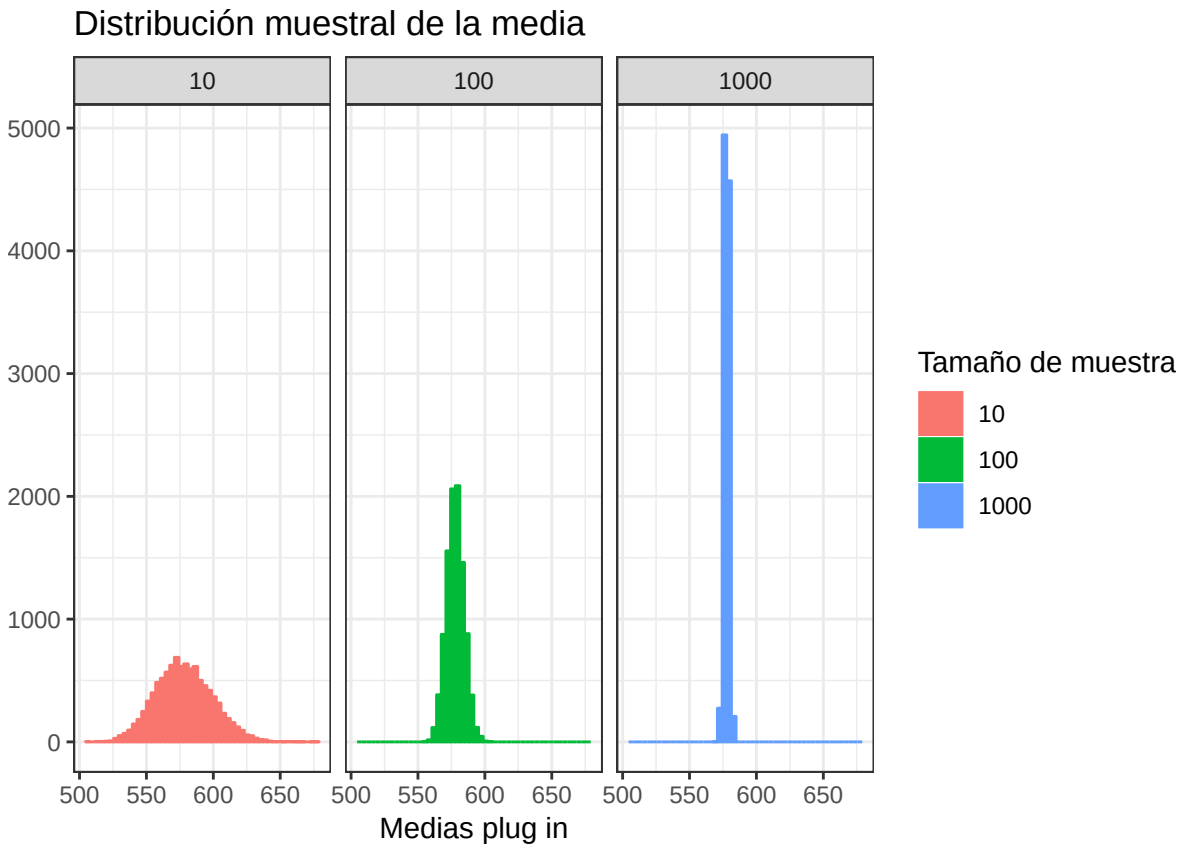
Ahora aproximaremos la distribución muestral, para cada tamaño de muestra n: i) Simula 10,000 muestras aleatorias

ii) Calcula la media en cada muestra

```
for(i in c(10,100,1000)){
  assign(paste0("means_",i),
         unlist(lapply(1:10000, function(j){mean(sample_n(primarias,i)$esp3)})))
}
```

iii) Realiza un histograma de la distribución muestral de las medias (las medias del paso anterior)

```
dfMeans <- as.data.frame(cbind(N=c(rep(10,10000), rep(100,10000), rep(1000,10000)),
                               Means=c(means_10, means_100, means_1000)))
ggplot(dfMeans)+theme_bw()+
  geom_histogram(aes(Means,color=factor(N), fill=factor(N)), bins = 50)+
  facet_wrap(~dfMeans$N)+
  xlab("Medias plug in")+
  ylab("")+
  guides(fill=guide_legend(title="Tamaño de muestra"),
         color=guide_legend(title="Tamaño de muestra"))+
  ggtitle("Distribución muestral de la media")
```



iv) Aproxima el error estándar calculando la desviación estándar de las medias del paso ii)

```
SE_estimado <- c(sd(means_10), sd(means_100), sd(means_1000))
df          <- cbind(df, SE_estimado)
```

Calcula el error estándar de la media para cada tamaño de muestra usando la información **poblacional** (ésta no es una aproximación), usa la fórmula: $se_p(\bar{x}) = \sigma_p/\sqrt{n}$

```
SD   <- sd(primarias$esp3)
SE   <- c(SD/sqrt(10), SD/sqrt(100), SD/sqrt(1000))
df   <- cbind(df, SE)
```

¿Cómo se comparan los errores estándar correspondientes a los distintos tamaños de muestra?

```
kable(df)
```

n	mu_plug	se_plug	SE_estimado	SE
10	579.2580	19.282156	21.184118	21.098200
100	580.2323	6.567649	6.503316	6.671837
1000	574.8605	2.071742	1.756416	2.109820

Al aumentar el tamaño de muestra el error estándar disminuye, debido a que la variabilidad de la media disminuye

Parte B: Bootstrap correlación.

Nuevamente trabaja con los datos primaria, selecciona una muestra aleatoria de tamaño 100 y utiliza el principio del plug-in para estimar la correlación entre la calificación de $y =$ español 3 y la de $z =$ español 6: $\widehat{corr}(y, z) = 0.9$. Usa bootstrap para calcular el error estándar de la estimación.

i) Tomamos la muestra que calculamos en la primera parte de la tarea y estimamos la correlación con el principio del plug-in.

```
cor_plug <- cor(Muestra_100$esp3, Muestra_100$esp6)
```

Con el principio del plug-in la correlación para la muestra es: 0.7996801.

ii) Usamos bootstrap para estimar el error estándar

```
Sim_cor <- 10000 %>%
rerun(sample_n(Muestra_100, replace = TRUE, size = 100)) %>%
map_dbl(~ cor(.x$esp3, .x$esp6))

estES <- sd(Sim_cor)
```

El error estándar para muestras de tamaño 100 es aproximadamente 0.0433653.