

Entrega 1: Creación de un Estimador de la media poblacional

Daniel Aramburu, Guillermo Palomo, Jorge Salas y Marc Pastor

9 de Octubre de 2020

1. Introducción

El objetivo de este trabajo es crear un Estadístico para estimar la media poblacional de una distribución continua determinada (que no sea la Distribución Normal ni la Exponencial) y comparar su rendimiento con la media muestral (que es el estadístico más usado para aproximar la media poblacional). En nuestro caso hemos optado por la distribución Beta (con α y $\beta = 2$) debido a que su función de densidad es simétrica y centrada, y nos parece que es sencillo encontrar un Estadístico fiable para aproximar su media poblacional.

2. Distribución Beta

La Distribución Beta es una distribución continua que depende de dos parámetros (α y β) y que toma valores en el intervalo $[0,1]$. Debido a que solo está definida en $[0,1]$ es una distribución muy usada para modelizar la probabilidad de que ocurra un evento, aunque también es usada para describir datos empíricos (debido a la variedad de formas que puede adoptar en función de los valores que tomen sus parámetros) y para modelar la fiabilidad de un sistema. Su función de densidad es distinta de cero solo cuando $0 < x < 1$ y es la siguiente:

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$$

También se puede escribir como:

$$f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}$$

Dónde la función Beta es la siguiente:

$$B(\alpha, \beta) = \frac{\Gamma(\alpha) \Gamma(\beta)}{\Gamma(\alpha + \beta)} = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx$$

Propiedades de la función Beta

La función beta cumple las siguientes propiedades (las usaremos para encontrar su función generadora de momentos, etc):

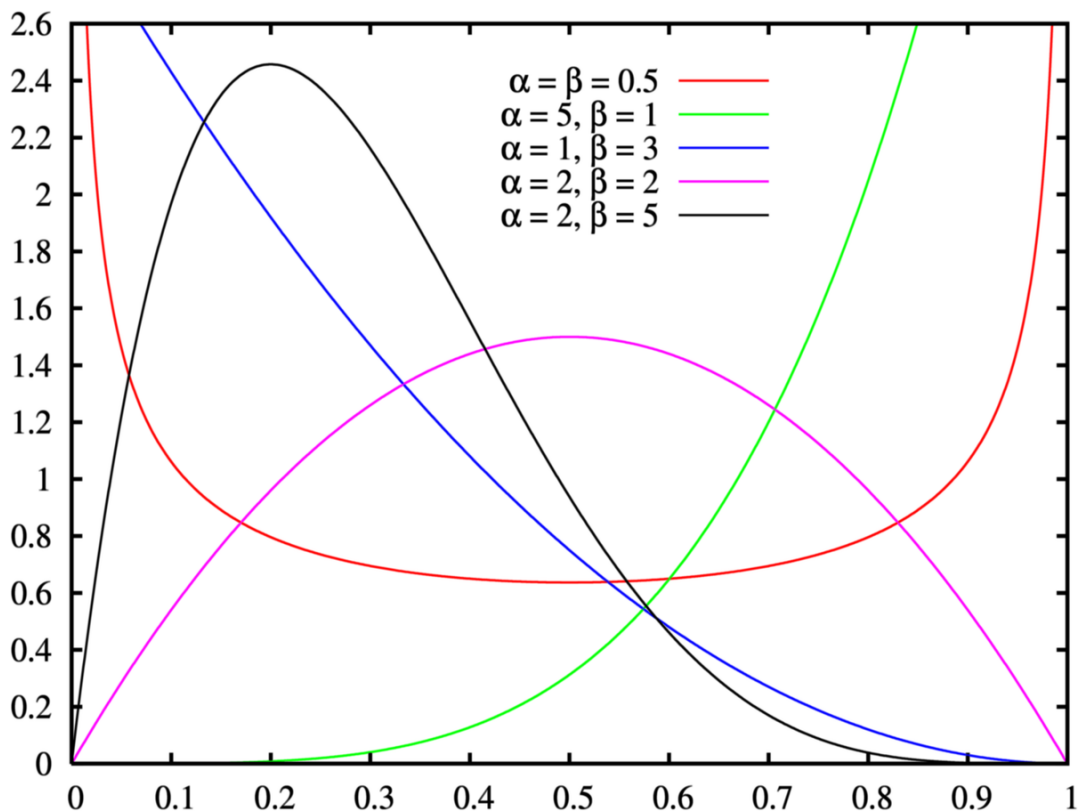
$$B(\alpha, \beta) = B(\beta, \alpha)$$

$$B(\alpha, 1) = \frac{1}{\alpha}$$

$$B(\alpha + 1, \beta) = \frac{\alpha}{(\alpha + \beta)} B(\alpha, \beta)$$

2.0 Forma de la distribución Beta

Esta distribución adopta formas muy diversas, en función de los valores de sus parámetros, por ello se utiliza mucho para modelar datos de manera empírica (es decir, observando la forma de la distribución de los datos, se intenta aproximar por la distribución beta con la forma más parecida)



2.1 Momentos

$$E[X] = \frac{\alpha}{(\alpha + \beta)}$$
$$\text{Var}[X] = \frac{\alpha \beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)}$$

2.2 Función Generadora de Momentos

$$M_x(t) = \sum_{k=0}^{\infty} \frac{B(\alpha + k, \beta)}{B(\alpha, \beta)} \frac{t^k}{k!}$$

3. Creación del Estadístico para estimar la Media Poblacional

En primer lugar hemos creado la función *muestreo*. Esta función es la base de toda la práctica, ya que nos permite crear muestras aleatorias, calcular nuestro Estadístico, y la media muestral en cada una de esas 40 muestras, etc.

```
library(tidyverse)
library(e1071)
library(moments)
library(gt)
library(ggpubr)
library(ggplotify)
library(grid)

muestreo <- function(alpha, beta){
  lista <- list()
  tabla_muestras <- matrix(ncol = 40, nrow = 10) %>% as.data.frame()
  lista[[6]] <- matrix(ncol = 40, nrow = 1) %>% as.data.frame()
  media_poblacional <- alpha / (alpha + beta)
  varianza_poblacional <- alpha * beta / (((alpha + beta) ^ 2) * (alpha + beta + 1))
  for(j in 1:40){
    set.seed(j)
    tabla_muestras[, j] <- rbeta(10, shape1 = alpha, shape2 = beta)
    colnames(tabla_muestras)[j] <- as.numeric(gsub("V", "",
                                                    colnames(tabla_muestras)[j]))
    colnames(lista[[6]])[j] <- as.numeric(gsub("V", "",
                                                    colnames(tabla_muestras)[j]))
    lista[[6]][, j] <- 0.5*(quantile(tabla_muestras[, j],
                                    probs = c(0.6)) +
                          quantile(tabla_muestras[, j],
                                    probs = c(0.4)))
  }
  lista[[1]] <- tabla_muestras
  lista[[2]] <- media_poblacional
  lista[[3]] <- varianza_poblacional
  lista[[4]] <- data.frame(Muestra = names(colMeans(tabla_muestras)),
                          mediamuestral = unname(colMeans(tabla_muestras)))
  lista[[5]] <- data.frame(Estimador = "Media Muestral",
                          Media = mean(lista[[4]][, 2]),
                          Mediana = median(lista[[4]][, 2]),
                          SD = sd(lista[[4]][, 2]),
                          IQR = IQR(lista[[4]][, 2]),
                          MAD = mad(lista[[4]][, 2]),
                          Curtosis = moments::kurtosis(lista[[4]][, 2]),
                          Asimetría = e1071::skewness(lista[[4]][, 2]))
  lista[[6]] <- gather(as.data.frame(lista[[6]]))
  colnames(lista[[6]]) <- c("Muestra", "Estadístico")
  lista[[7]] <- data.frame(Estimador = "Estadístico",
                          Media = mean(lista[[6]][, 2]),
                          Mediana = median(lista[[6]][, 2]),
                          SD = sd(lista[[6]][, 2]),
                          IQR = IQR(lista[[6]][, 2]),
                          MAD = mad(lista[[6]][, 2]),
                          Curtosis = moments::kurtosis(lista[[6]][, 2]),
```

```

                                Asimetría = e1071::skewness(lista[[6]][, 2]))
lista[[8]] <- rbind(lista[[5]], lista[[7]])
names(lista) <- c("tabla_muestras", "media_poblacional",
                  "varianza_poblacional", "media_muestral",
                  "medidas_media_muestral", "estadistico",
                  "medidas_estadistico", "tabla_comparacion")

  lista
}

resultados <- muestreo(alpha = 2, beta = 2)

datosgrafico <- inner_join(resultados$media_muestral , resultados$estadistico)

```

4. Evaluación de nuestro Estadístico

4.1 Tabla de Medidas

Estadístico						
Medidas de Centralización		Medidas de Dispersión			Medidas de Forma	
Media	Mediana	SD	IQR	MAD	Curtosis	Asimetría
0.503593	0.5041133	0.08791211	0.1353218	0.09725723	2.656788	0.2748998

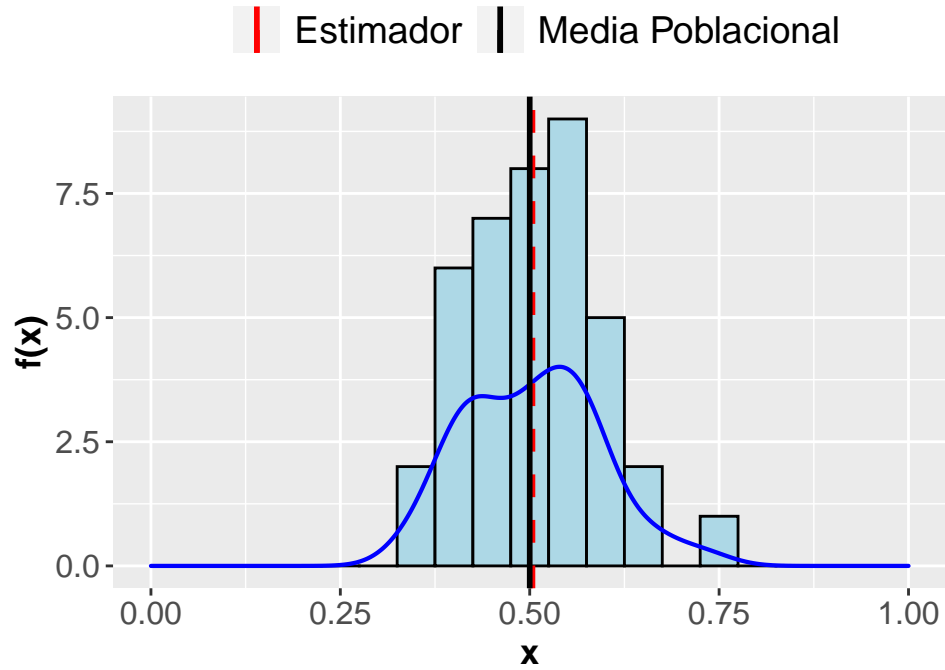
4.2 Histogramas

```
grafico_estadistico_histogram_density <- ggplot(data = datosgrafico, aes(x = Estadístico)) +  
  geom_histogram(fill = "lightblue", binwidth = 0.05, color = "black") +  
  geom_density(color = "blue", size = 0.75) +  
  ggtitle("Histograma de nuestro Estadístico") +  
  geom_vline(aes(xintercept = resultados[[7]]$Media, col = "Estimador"), linetype = "dashed", size = 1) +  
  geom_vline(aes(xintercept = resultados$media_poblacional, col = "Media Poblacional"), size = 1) +  
  scale_color_manual(name = "", values = c("Media Poblacional" = "black", Estimador = "red")) +  
  xlab("x") + ylab("f(x)") +  
  theme(legend.position = "top",  
        legend.text = element_text(size = 14),  
        axis.text.x = element_text(size = 12),  
        axis.text.y = element_text(size = 12),  
        axis.title = element_text(size = 13, face = "bold"),  
        plot.title = element_text(size = 18, face = 'bold', hjust = 0.5)) + xlim(0, 1)
```

Este es el histograma resultante:

grafico_estadistico_histogram_density

Histograma de nuestro Estadístico



5. Evaluación de la Media Muestral

5.1 Tabla de Medidas

Media Muestral						
Medidas de Centralización		Medidas de Dispersión			Medidas de Forma	
Media	Mediana	SD	IQR	MAD	Curtosis	Asimetría
0.5072706	0.5012344	0.0721176	0.08862303	0.06388053	3.077518	0.3099279

Comparación con la varianza poblacional

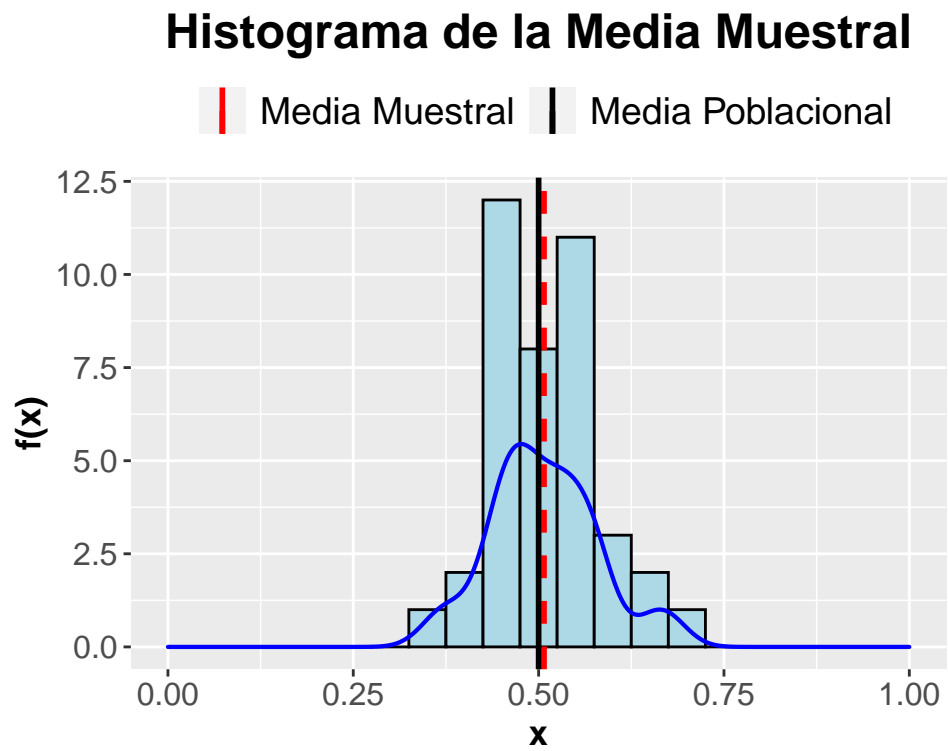
Varianza de la Media Muestral	
Valor Obtenido	Valor Teórico
0.005200948	0.005

5.2 Histogramas

```
grafico_mediamuestral_histogram_density <- ggplot(data = datosgrafico, aes(x = mediamuestral)) +  
  geom_histogram(fill = "lightblue", color = "black", binwidth = 0.05) +  
  geom_density(size = 0.75, col = "blue") +  
  ggtitle("Histograma de la Media Muestral") +  
  geom_vline(aes(xintercept = resultados[[5]]$Media, col = "Media Muestral"), linetype = "dashed") +  
  geom_vline(aes(xintercept = resultados$media_poblacional, col = "Media Poblacional"), size = 1) +  
  scale_color_manual(name = "", values = c("Media Poblacional" = "black", "Media Muestral" = "red")) +  
  xlab("x") + ylab("f(x)") +  
  theme(legend.position = "top",  
        legend.text = element_text(size = 14),  
        axis.text.x = element_text(size = 12),  
        axis.text.y = element_text(size = 12),  
        axis.title = element_text(size = 13, face = "bold"),  
        plot.title = element_text(size = 18, face = 'bold', hjust = 0.5)) + xlim(0, 1)
```


Este es el gráfico resultante.

grafico_mediamuestral_histogram_density



6. Comparando nuestro estadístico con la media muestral

6.1 Tabla de Medidas

Comparativa de Estimadores							
Estimador	Medidas de Centralización		Medidas de Dispersión			Medidas de Forma	
	Media	Mediana	SD	IQR	MAD	Curtosis	Asimetría
Media Muestral	0.5072706	0.5012344	0.07211760	0.08862303	0.06388053	3.077518	0.3099279
Estadístico	0.5035930	0.5041133	0.08791211	0.13532182	0.09725723	2.656788	0.2748998

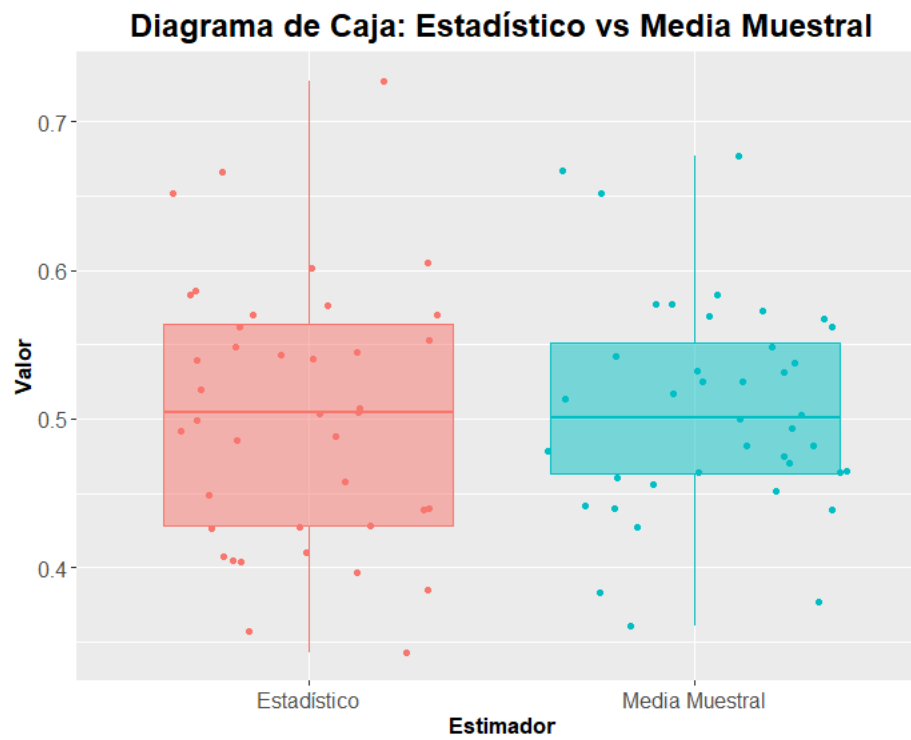
6.2 Gráficos Comparativos

```
datos_grafico_comparativo <- gather(datosgrafico[, -1])
colnames(datos_grafico_comparativo) <- c("Estimador", "Valor")
datos_grafico_comparativo[, 1] <- gsub("mediamuestral", "Media Muestral",
                                       datos_grafico_comparativo[, 1])
cutoff <- data.frame( yintercept = resultados$media_poblacional,
                     cutoff = factor(50) )

grafico_comparativo_boxplot <- ggplot(data = datos_grafico_comparativo,
                                     aes(x = Estimador, y = Valor,
                                         fill = Estimador,
                                         color = Estimador)) +

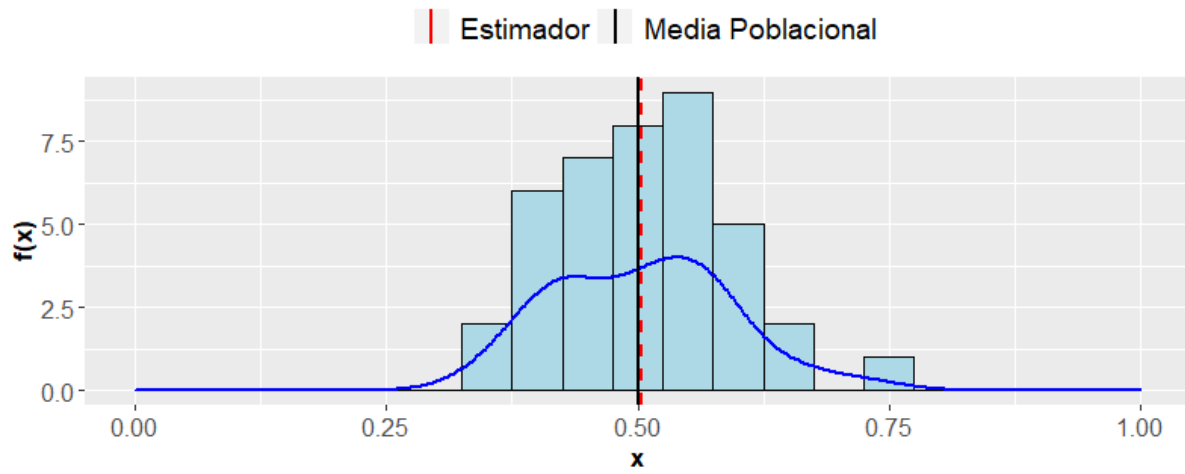
  geom_boxplot(alpha = 0.5) +
  geom_jitter() +
  ggtitle("Diagrama de Caja: Estadístico vs Media Muestral") +
  theme(legend.position = "none",
        legend.text = element_text(size = 14),
        axis.text.x = element_text(size = 12),
        axis.text.y = element_text(size = 12),
        axis.title = element_text(size = 13, face = "bold"),
        plot.title = element_text(size = 18, face = 'bold',
                                   hjust = 0.5))
```

Ahora observamos los gráficos que hemos creado:



Aquí tenim un altre gràfic

Histograma de nuestro Estadístico



Histograma de la Media Muestral

