Índice

Contents

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 alpha 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 (alpha y beta) 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):

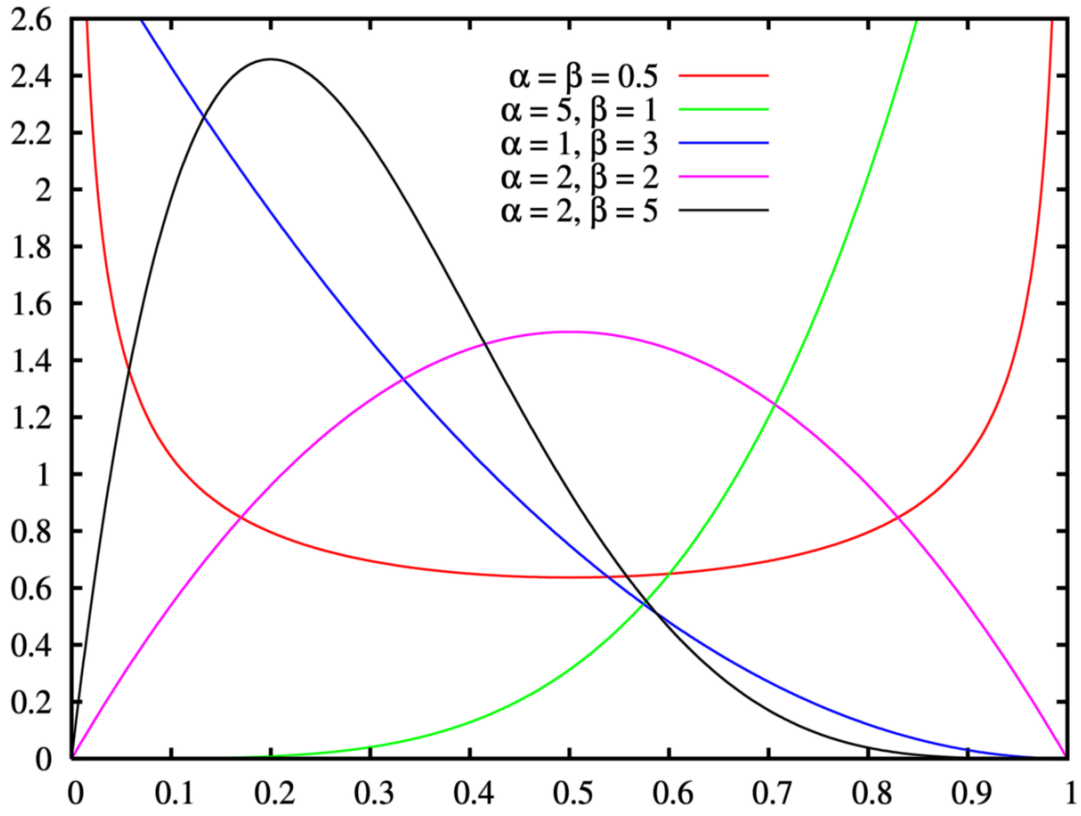
$$B(\alpha, \beta) = B(\beta, \alpha)$$

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

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

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 y Función Generadora de Momentos

$$E[X] = \frac{\alpha}{(\alpha+\beta)}$$

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

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

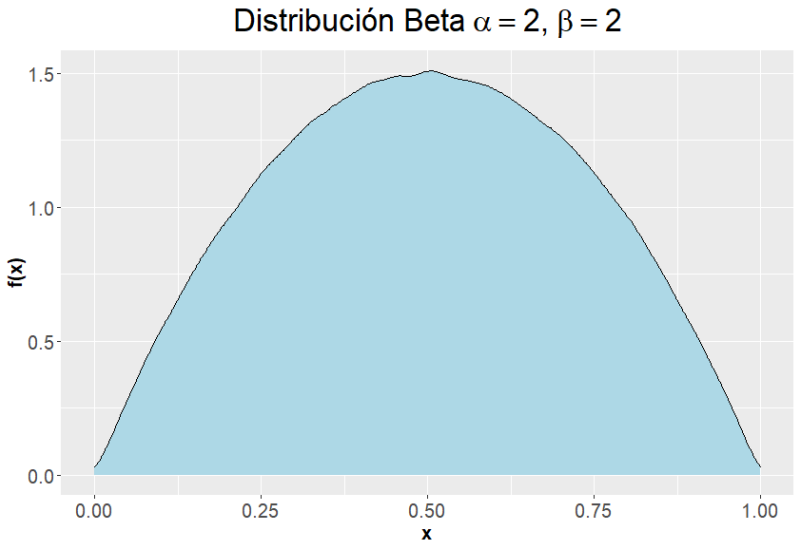
[1]

[1]Nota: Los cálculos para encontrar los momentos y la función generadora están en el Anexo I y II (respectivamente)

3. Creación del Estadístico para estimar la Media Poblacional

3.1 Escogiendo los parámetros de la distribución

En nuestro caso hemos optado por seleccionar una distribución beta con alpha = 2, beta = 2, ya que pensamos que debido a su forma simétrica, seremos capaces de aproximar la media poblacional de manera más precisa que si fuera asimétrica. La forma de una distribución Beta(2,2) es la siguiente:



Los momentos de la distribución Beta(alpha = 2, beta = 2) son los siguientes:

$$E[X] = \mu = 0.5$$
$$Var[X] = \sigma^2 = 0.05$$

3.2 Fórmula del Estadístico

Nuestro estadístico es el promedio de los percentiles 40 y 60. Hemos escogido este estadístico debido a que la distribución por la que hemos optado (distribución beta(2,2)), es bastante simétrica, y nos parece que podemos llegar a estimar la media poblacional con bastante precisión. La fórmula de nuestro Estadístico es:

$$T = \frac{P_{40} + P_{60}}{2}$$

4. Propiedades del Estimador T

En nuestro caso hemos podido demostrar analíticamente la insesgabilidad, la invarianza y la suficiencia de nuestro estimador.

4.1 Insesgabilidad

4.1.1. Analíticamente

$X \sim Beta(\alpha = 2, \beta = 2)$;

$$f(x) = \frac{\Gamma(4)}{\Gamma(2)\Gamma(2)} x(1-x) \qquad 0 \le x \le 1$$

Calculemos los percentiles p_{10} y p_{60}:

$$P(X \le p) = \int_{-\infty}^{p} \frac{\Gamma(4)}{\Gamma(2)\Gamma(2)} x(1-x)\,dx =$$

$$= \frac{\Gamma(4)}{\Gamma(2)\Gamma(2)} \int_{-\infty}^{p} x(1-x)\,dx =$$

$$= \frac{\Gamma(4)}{\Gamma(2)\Gamma(2)} \int_{0}^{p} x - x^2\,dx =$$

$$= \frac{\Gamma(4)}{\Gamma(2)\Gamma(2)} \left(\frac{x^2}{2} - \frac{x^3}{3} \right)_{x=0}^{x=p}$$

$$= \frac{3!}{1} \left(\frac{p^2}{2} - \frac{p^3}{3} \right) =$$

$$= 6 \cdot \left(\frac{3p^2 - 2p^3}{6} \right) = 3p^2 - 2p^3$$

Percentil 40:

$P(X \leq p_{40}) = 0,4$;

$3p_{40}^2 - 2p_{40}^3 = 0,4$ ~~$\Rightarrow p_{40} = 0,4$~~

$3p_{40}^2 - 2p_{40}^3 - 0,4 = 0 \Rightarrow$ $\boxed{p_{40} = 0,4329311}$

Percentil 60:

$P(X \leq p_{60}) = 0,6$

$3p_{60}^2 - 2p_{60}^3 = 0,6$

$3p_{60}^2 - 2p_{60}^3 - 0,6 = 0 \Rightarrow$ $\boxed{p_{60} = 0,5670689}$

- $\boxed{T = \frac{p_{40} + p_{60}}{2}}$
- $\boxed{r = 0,5}$

$E[T] = E\left[\frac{p_{40}+p_{60}}{2}\right] = \frac{1}{2} \cdot E[p_{40}+p_{60}] =$

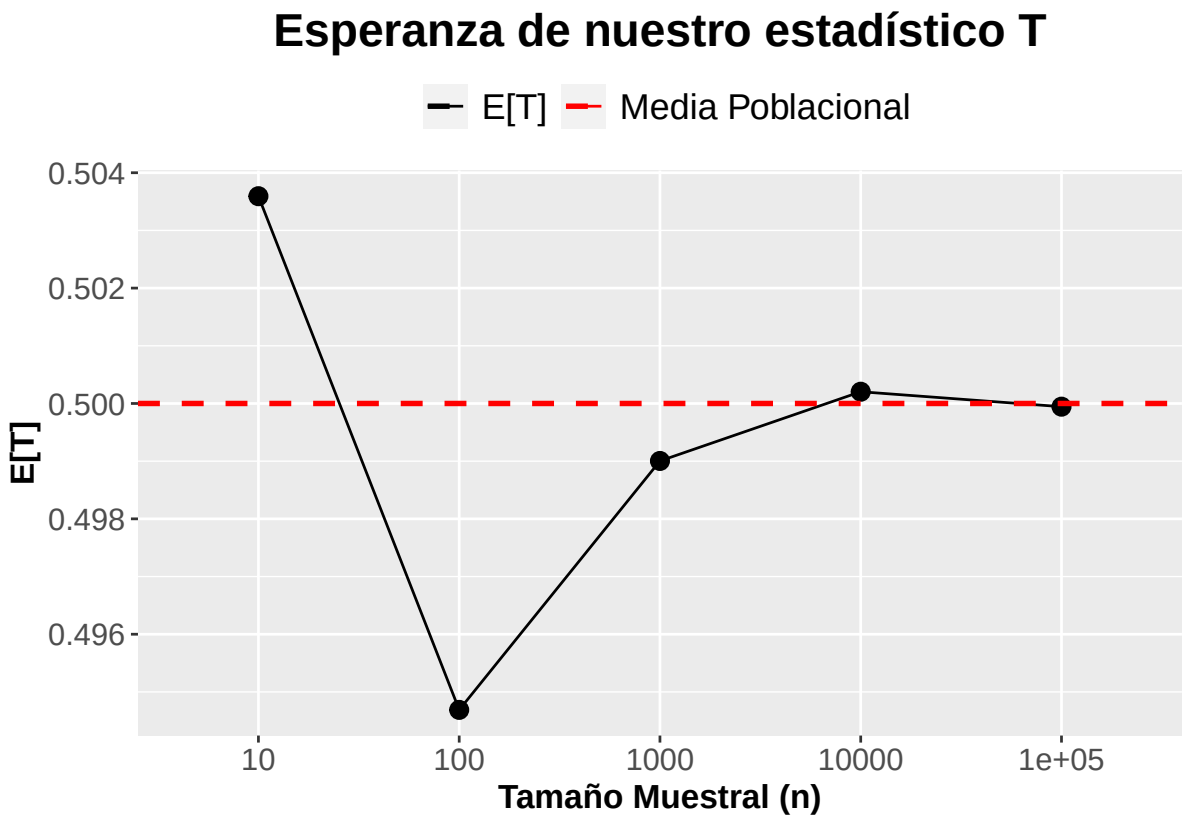
$= \frac{1}{2} \cdot E[0,4329311 + 0,5670689] = 0,5 = r$

$\Rightarrow$ $\boxed{E[T] = 0,5 = r}$ $\Rightarrow$ $\boxed{T \text{ es un estimador insesgado de } r}$

$\boxed{b[T] = E[T] - r = 0}$

4.1.2 Numéricamente

Para evaluar numéricamente la insesgadez hemos generado 40 muestras de distintos tamaños (n = 10, . . . , n = 1000000), entonces hemos calculado nuuestro estadístico y su media (para todos los valores de n). Hemos obtenido el siguiente gráfico.

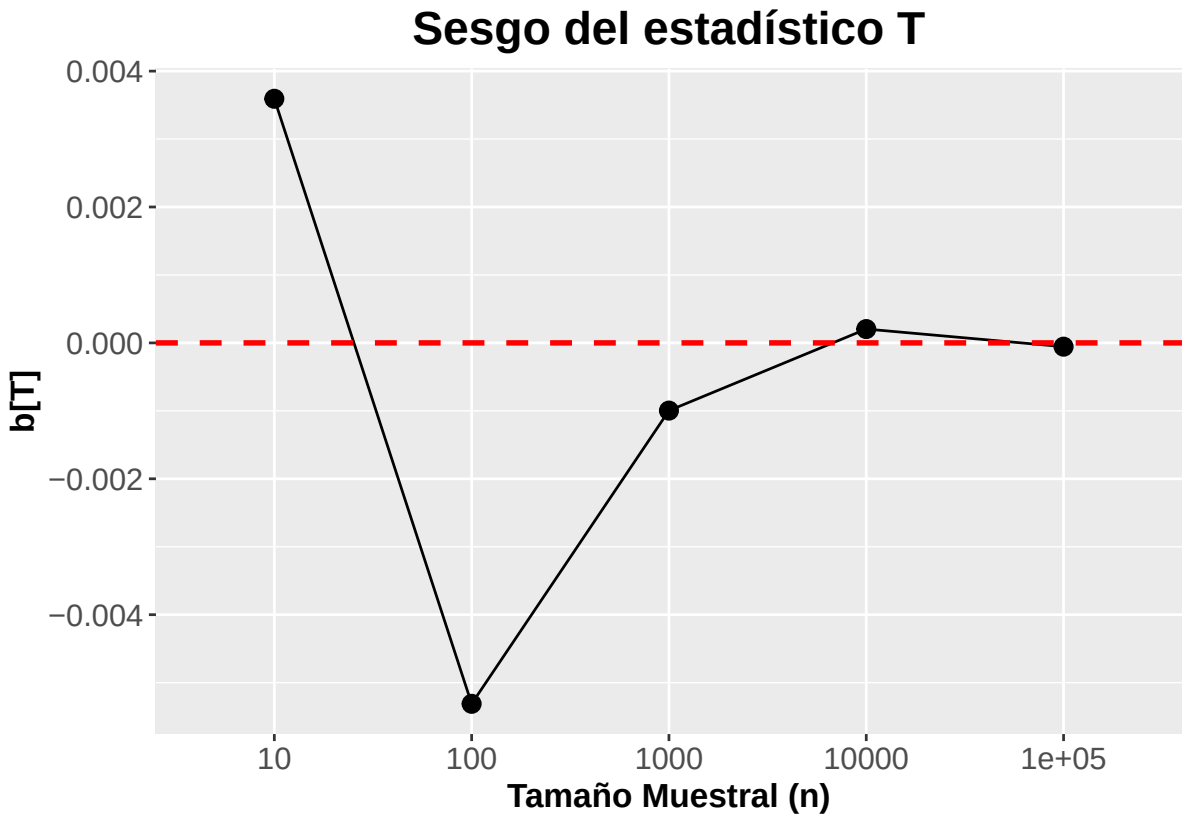


Podemos ver que a medida que aumenta el tamaño muestral (n), la esperanza de nuestro estadístico se va acercando cada vez más a la media poblacional. De aquí podemos suponer que para un n suficientemente grande, la esperanza del estadístico convergerá a la media poblacional.

Definimos el sesgo como:

$$b(T) = E[T] - \mu$$

Hemos obtenido el siguiente gráfico para el sesgo.



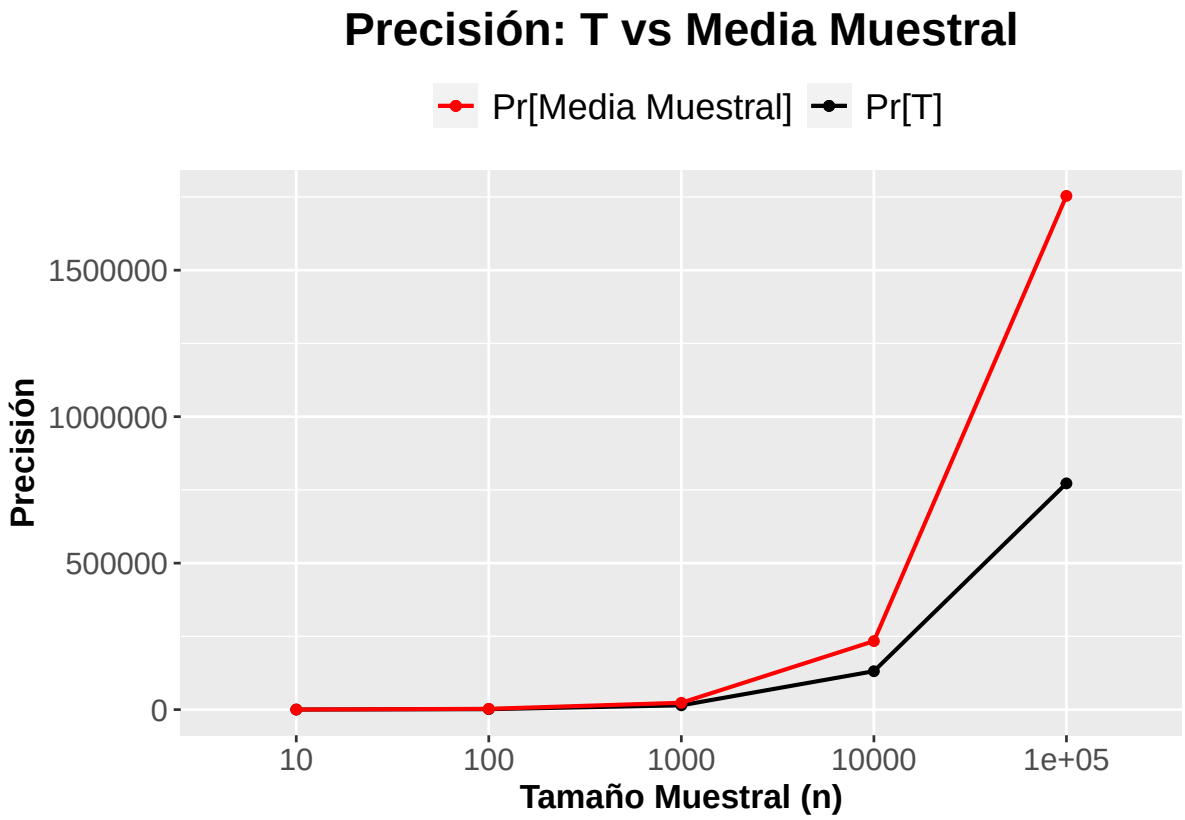
En este caso, podemos observar que el sesgo de nuestro estimador tiende a cero, a medida que aumenta el tamaño muestral. Después de la demostración analítica y de observar los dos gráficos anteriores, podemos concluir que nuestro estimador es **insesgado**.

4.2 Eficiencia / Precisión

La eficiencia o precisión se define como:

$$Pr(T) = \frac{1}{Var[X]}$$

En nuestro caso hemos generado 40 muestras de distintos tamaños (n = 10, . . . , n = 1000000), y hemos calculado la varianza y la precisión de nuestro estadístico para cada uno de los tamaños muestrales. Además hemos decidido comparar la eficiencia de nuestro estimador T y la media muestral. Hemos obtenido el siguiente gráfico para el sesgo.



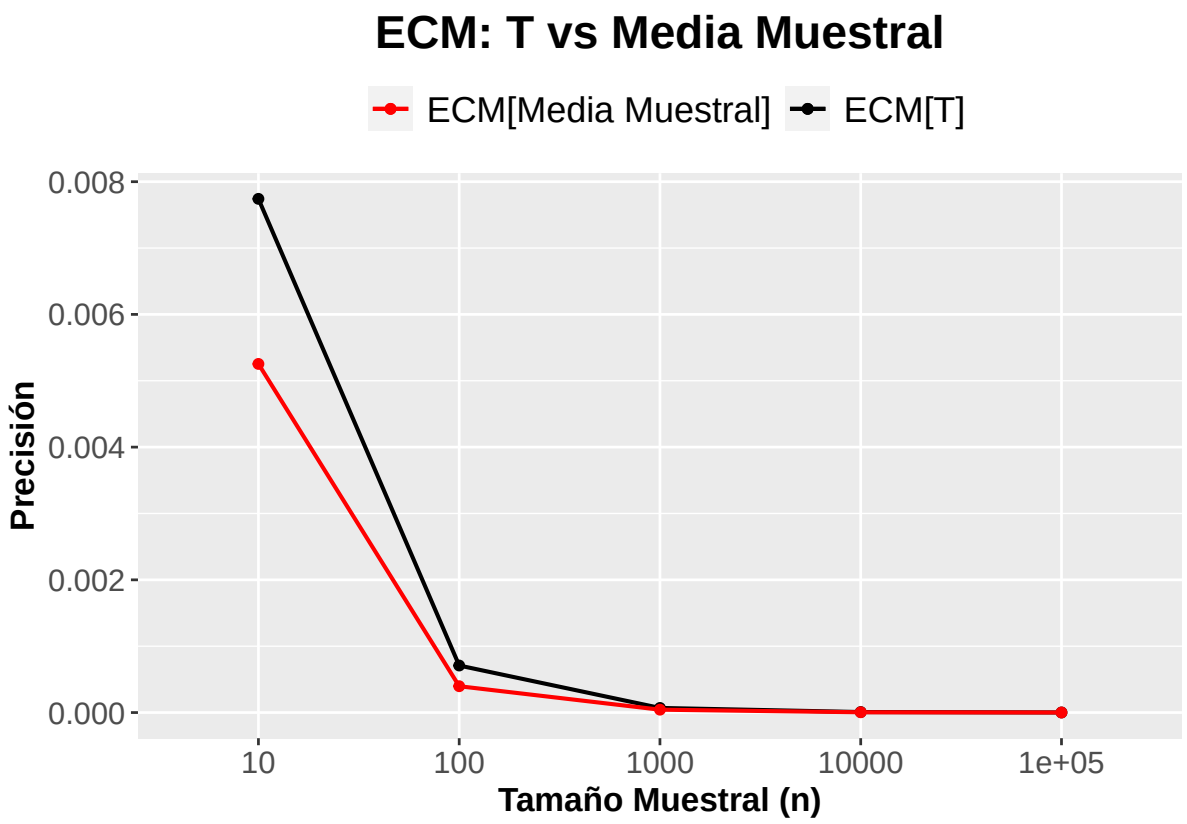
Podemos observar que la media muestral es más precisa estimando la media poblacional que nuestro parámetro. Esto se debe a que la varianza de la media muestral es menor que la de nuestro estadístico, y también a que la media muestral es el estimador máximo verosímil de la media poblacional.

4.3 Error cuadrático Medio (ECM)

El error cuadrático medio de un estimador T se define como el promedio de las desviaciones del estadístico al parámetro que estima. Matemáticamente:

$$ECM[T] = E_{(T-\theta)^2} = Var[X] + b[T]^2$$

En este caso hemos seguido el mismo procedimiento que en las anteriores propiedades, obteniendo el siguiente gráfico:



Como ambos estimadores (la media muestral y T), son insesgados, el ECM de ambos se reduce a su varianza, y como la media muestral tiene menos varianza, tamién tiene menor error cuadrático medio.

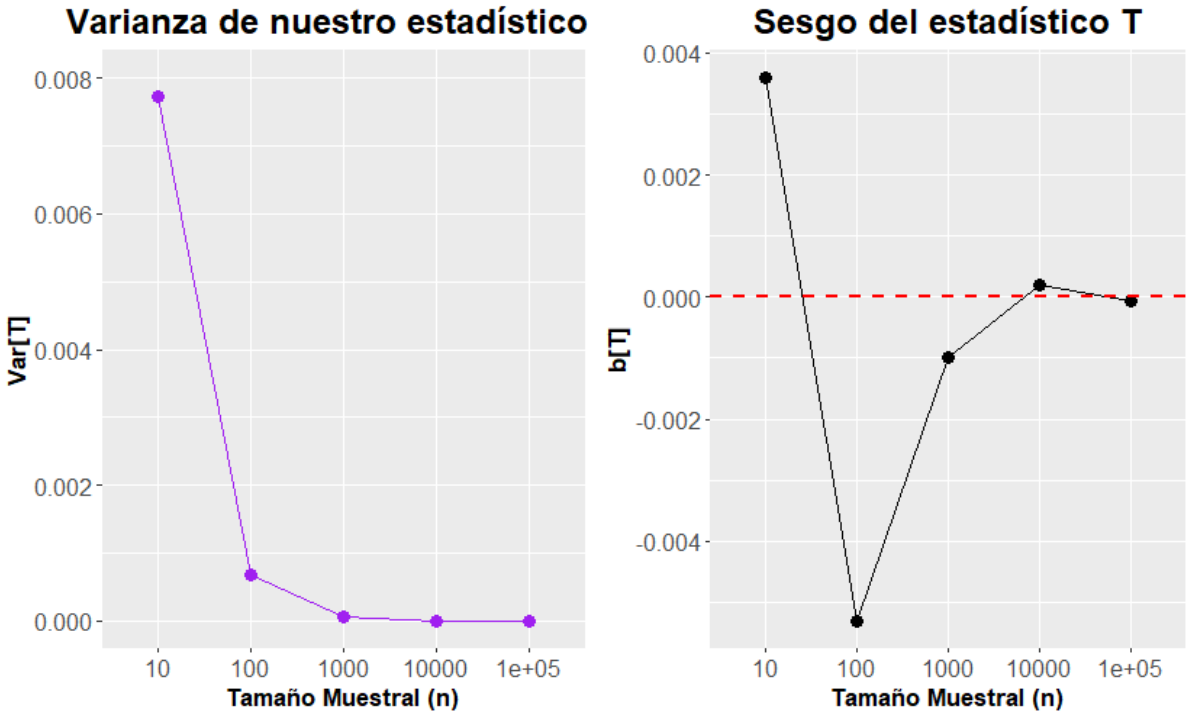
4.4 Consistencia

Un estimador T se dice consistente si como condición suficiente cumple las siguientes propiedades:

$$\lim_{n\to\infty} Var(T) = 0$$

$$\lim_{n\to\infty} b(T) = 0$$

En nuestro caso hemos tenido que probar la consistencia numéricamente, puesto que el cálculo de la varianza de manera analítica es muy complicada. Hemos obtenido los siguientes gráficos:



Podemos observar como al aumentar el tamaño muestral (es decir cuando n tiende a infinito), tanto la varianza de nuestro estadístico y el sesgo tienden a cero. Esto nos permite afirmar con bastante seguridad que nuestro estimador es **consistente** para la media poblacional, es decir, que al aumentar el tamaño muestral se aproxima bien a ella.

4.5 Invarianza

Se dice que un estimador T es invariante por traslaciones (1) o por cambios de escala (2), si se cumple que:

$$1) \quad T(a + X_1, \ldots, a + X_n) = a + T(X_1, \ldots, X_n)$$

$$2) \quad T(bX_1, \ldots, bX_n) = bT(X_1, \ldots, X_n)$$

En nuestro caso, sabemos que los percentiles son **invariantes por traslaciones** ya que al sumar a cada elemento de la muestra una constante el percentil k-ésimo sigue siendo el mismo, y **lo mismo en el caso de los cambios de escala**. No es posible demostrar esto analíticamente, debido a que no disponemos de una expresión escrita del percentil.

4.6 Robustez

Para analizar la robustez hemos generado una distribución contaminada, a partir de nuestra distribución original (Beta(2,2)) y la distribución Uniforme en el intervalo [2, 3]. Para crear la muestra contaminada hemos generado 400 datos aleatorios de una distribución Uniforme[0,1]. En caso de que la probabilidad de esta última distribución fuera menor o igual que 0.9, generabamos un dato de la Beta[2,2], y en caso que p > 0.9 escogíamos un dato de la Uniforme[2,3].

Gráficamente:

$$X_u = U(0, 1)$$

$$X_B = \text{Beta}(\alpha = 2, \beta = 2)$$

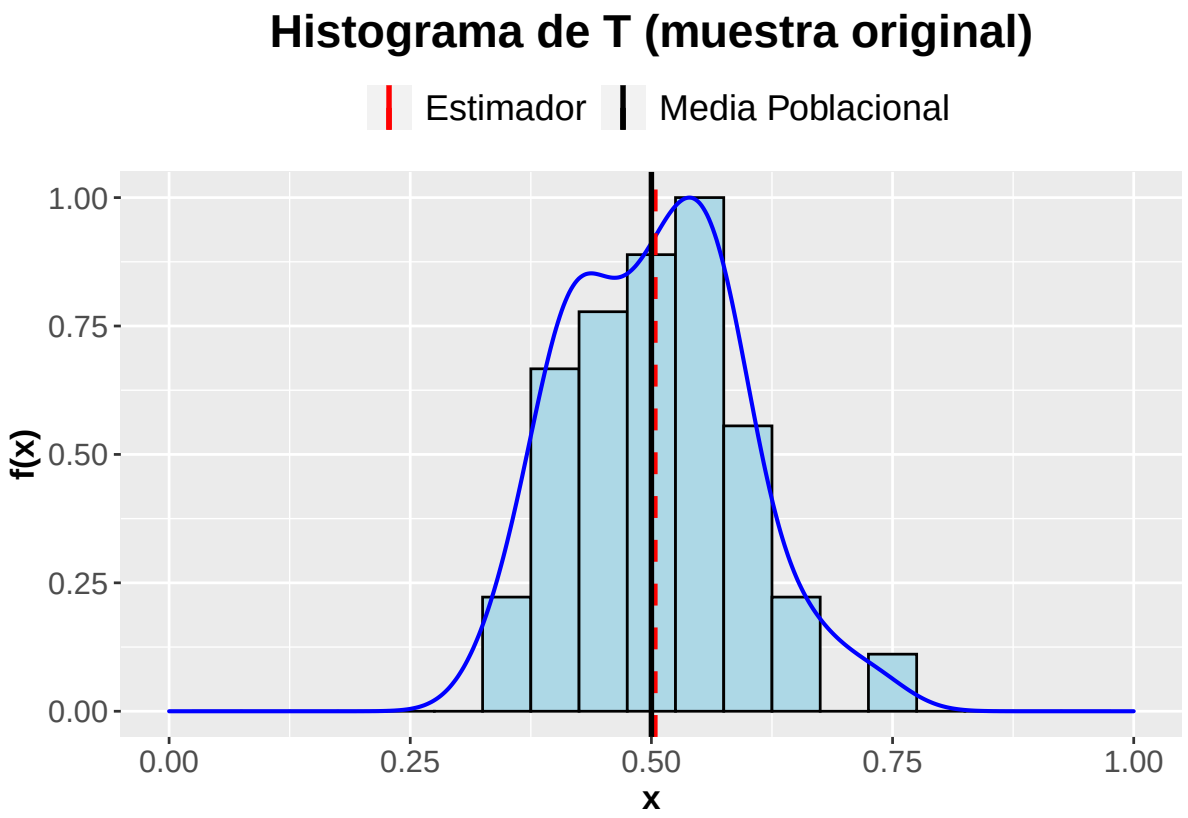
$$X_{u'} = U(2, 3)$$

$$X_c = \begin{cases} X_B & P(X_u) \leq 0.9 \\ X_{u'} & P(X_u) > 0.9 \end{cases}$$

4.6.1 Robustez de nuestro estadístico

4.6.1.1 Muestra sin contaminar

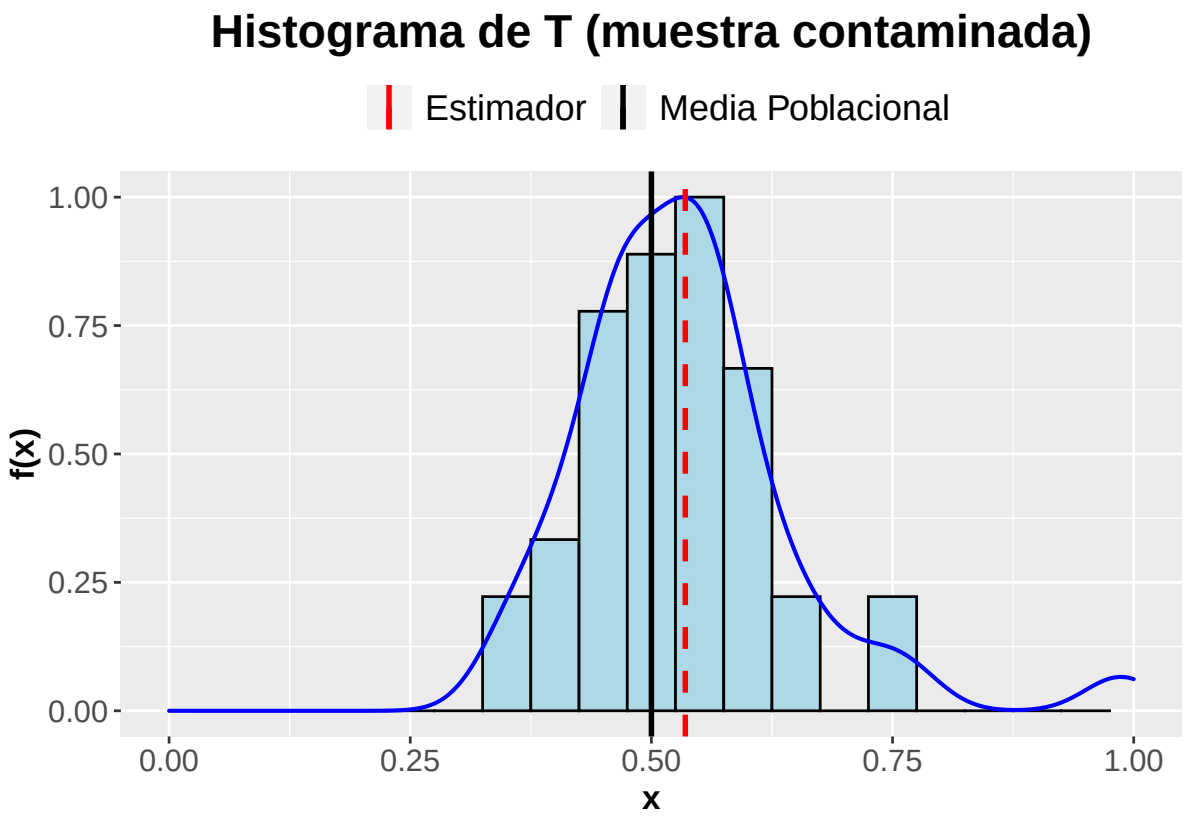
Estadístico						
Medidas de Centralización		***Medidas de Dispersión***			***Medidas de Forma***	
Media	Mediana	SD	IQR	MAD	Curtosis	Asimetría
0.503593	0.5041133	0.08791211	0.1353218	0.09725723	2.656788	0.2748998



Podemos ver que en promedio nuestro estadístico se acerca bastante a la media poblacional (cuyo valor es 0.5) y por otro lado la mediana también está muy centrada, y también podría ser un buen estimador de la media poblacional. En cuanto a las medidas de forma, la curtosis nos indica que se trata de una distribución levemente platicúrtica (su valor es menor que 3), es decir menos apuntada y con colas menos gruesas que la normal y el coeficiente de asimetría nos indica una pequeña asimetría a la derecha, ya que su valor es mayor que 0. Finalmente en cuanto a la dispersión, la distancia entre el cuartil 3 y 1 (IQR) es muy pequeña, por lo que podemos concretar que la mayoría de los datos están concentrados en este rango. Además, la desviación absoluta mediana (MAD) y también la desviación típica (SD) son muy pequeñas, y eso nos indica que hay poca variabilidad en los datos, es decir, estan bastante concentrados alrededor de la media y la mediana.

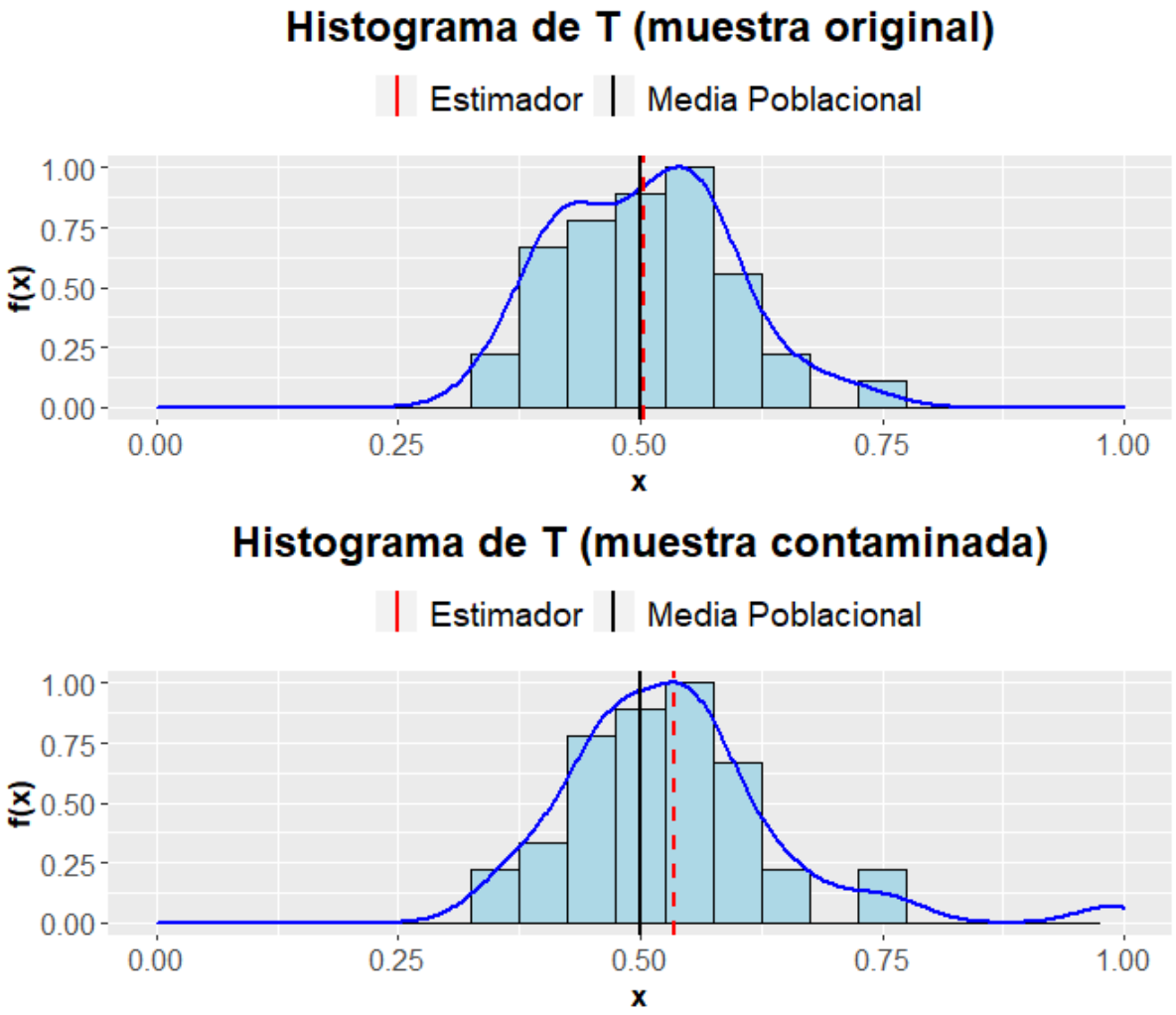
4.6.1.2 Muestra contaminada

Estadístico T						
Medidas de Centralización		***Medidas de Dispersión***			***Medidas de Forma***	
Media	Mediana	SD	IQR	MAD	Curtosis	Asimetría
0.5351557	0.5279007	0.1172313	0.1140922	0.08615515	6.962659	1.455874



Nuestro estadístico para la muestra contaminada se desplaza en promedio relativamente poco de la media poblacional. A la mediana el pasa algo similar siendo también un buen estimador del parámetro. En cuanto a las medidas de dispersión, los datos atípicos no han provocado cambios relevantes en los valores de la SD, IQR y MAD respecto a la muestra sin contaminar. Al contrario, sucede con las medidas de forma que, como indica el coeficiente de asimetría, tiene una leve asimetría a la derecha, aunque mayor que en la muestra sin contaminar. La curtosis indica que la distribución del estimador es leptocúrtica ya que su valor es mayor que 3, por lo que es más apuntada y con colas más gruesas que la normal.

4.6.1.3 Comparando T en la muestra original y contaminada

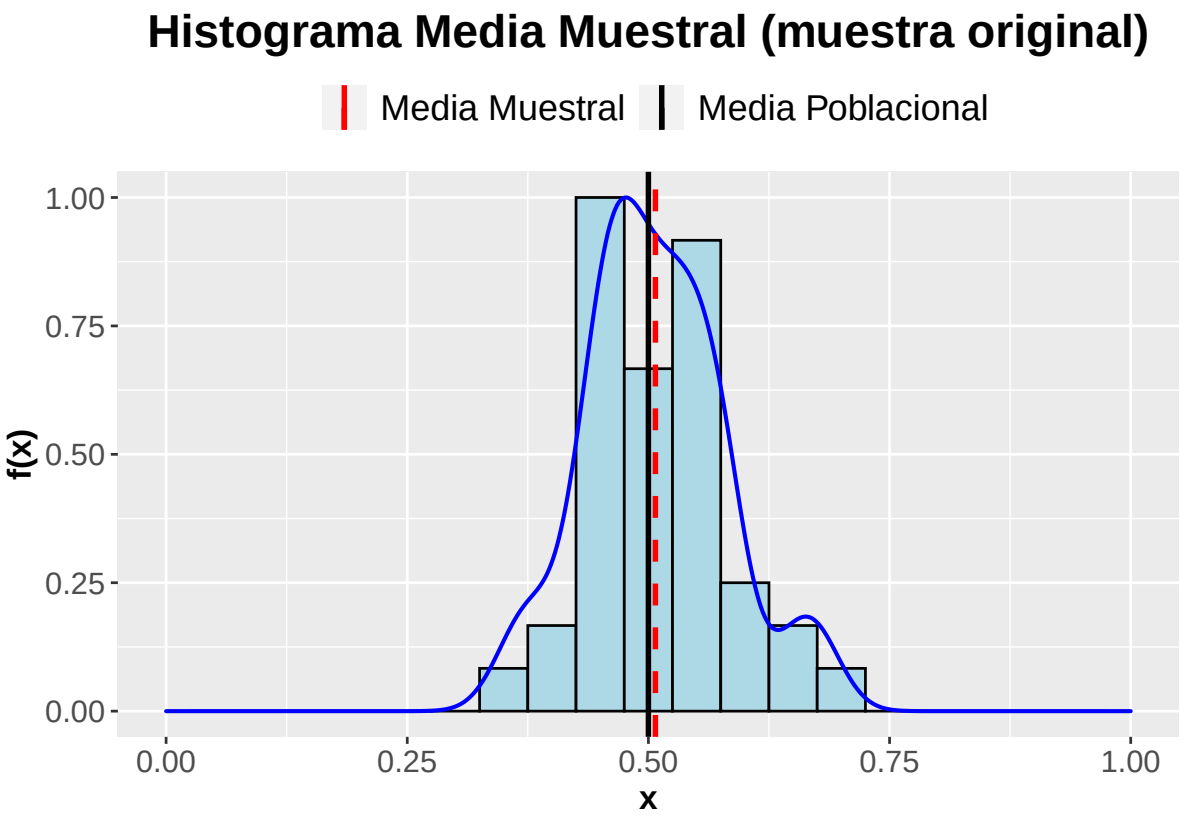


Podemos observar que al añadir datos atípicos, nuestro estimador es bastante robusto, ya que en promedio se desplaza relativamente poco de la media poblacional. Cuando la muestra no contiene datos atípicos, T sobreestima la media poblacional en un 0.7186 %, mientras que al añadir datos lo hace en un 7.03114 % . Eso supone que multiplicamos nuestro sesgo por 10 al añadir datos atípicos. Luego compararemos estos datos con los equivalentes de la media muestral, para saber cuál de los dos estimadores es más robusto.

4.6.2 Robustez de la Media Muestral

4.6.2.1 Muestra sin contaminar

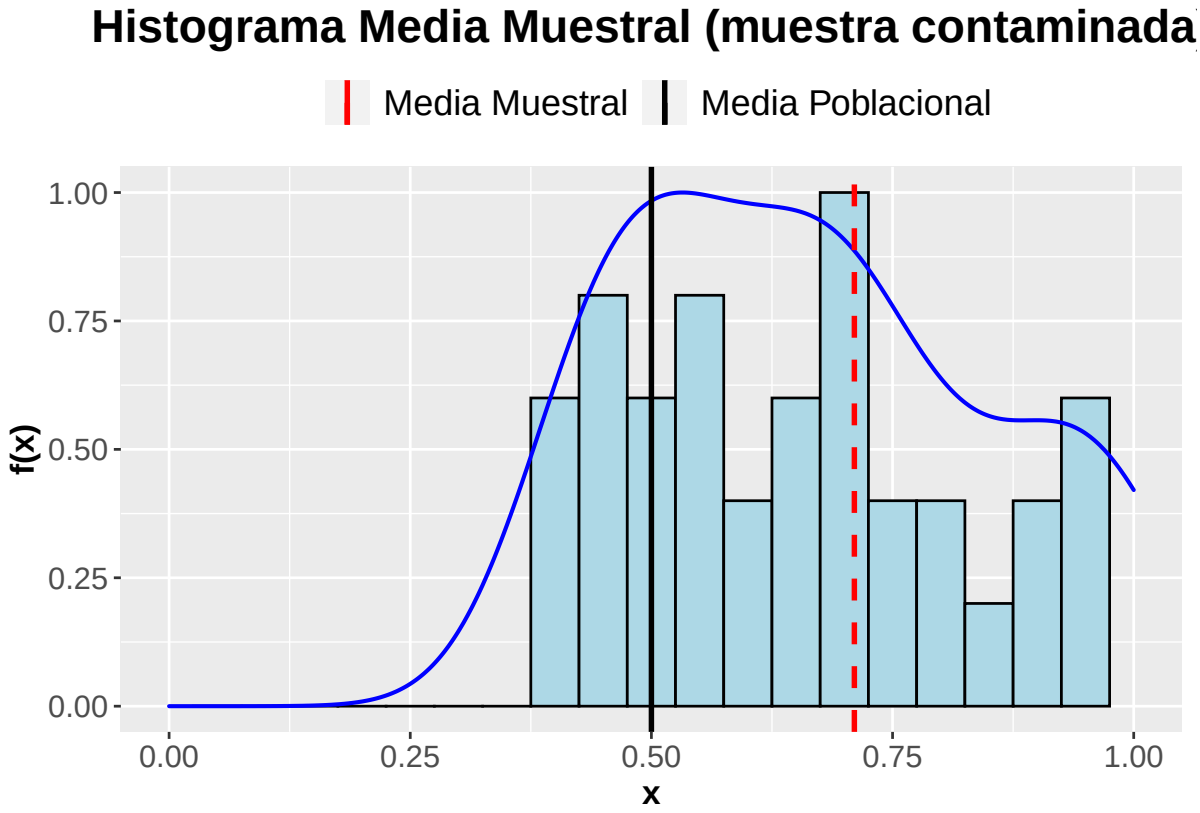
Media Muestral						
Medidas de Centralización		***Medidas de Dispersión***			***Medidas de Forma***	
Media	Mediana	SD	IQR	MAD	Curtosis	Asimetría
0.5072706	0.5012344	0.0721176	0.08862303	0.06388053	3.077518	0.3099279



Con estos datos se confirmaría el que el valor de la esperanza de las medias muestrales es muy parecida a la media teórica. Como indica el coeficiente de curtosis, esta distribución es prácticamente mesocúrtica, es decir se parece mucho a una normal y es muy simétrica pero tiene un poco de asimetría a la derecha, por lo que también la mediana se parece mucho a la media al ser una distribución centrada. En cuanto a las medidas de dispersión, la desviación típica muestral (SD) es muy pequeña, así como el IQR y la MAD, indicando esto que los valores están muy concentrados en torno al centro, que en este caso está cerca de la media.

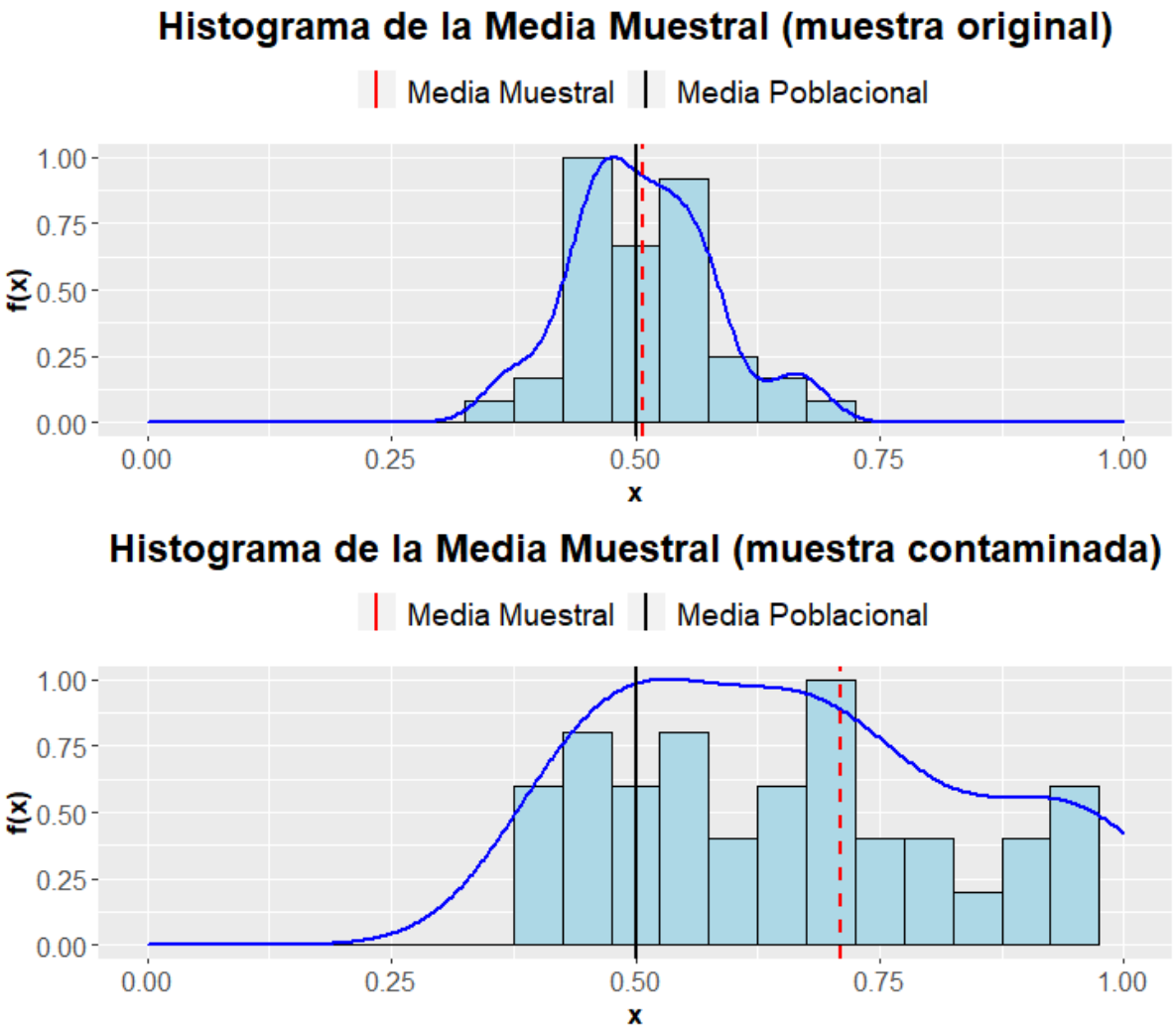
4.6.2.2 Muestra contaminada

Media Muestral						
Medidas de Centralización		***Medidas de Dispersión***			***Medidas de Forma***	
Media	Mediana	SD	IQR	MAD	Curtosis	Asimetría
0.7104073	0.6790156	0.2268685	0.3757027	0.2397981	2.567035	0.5786614



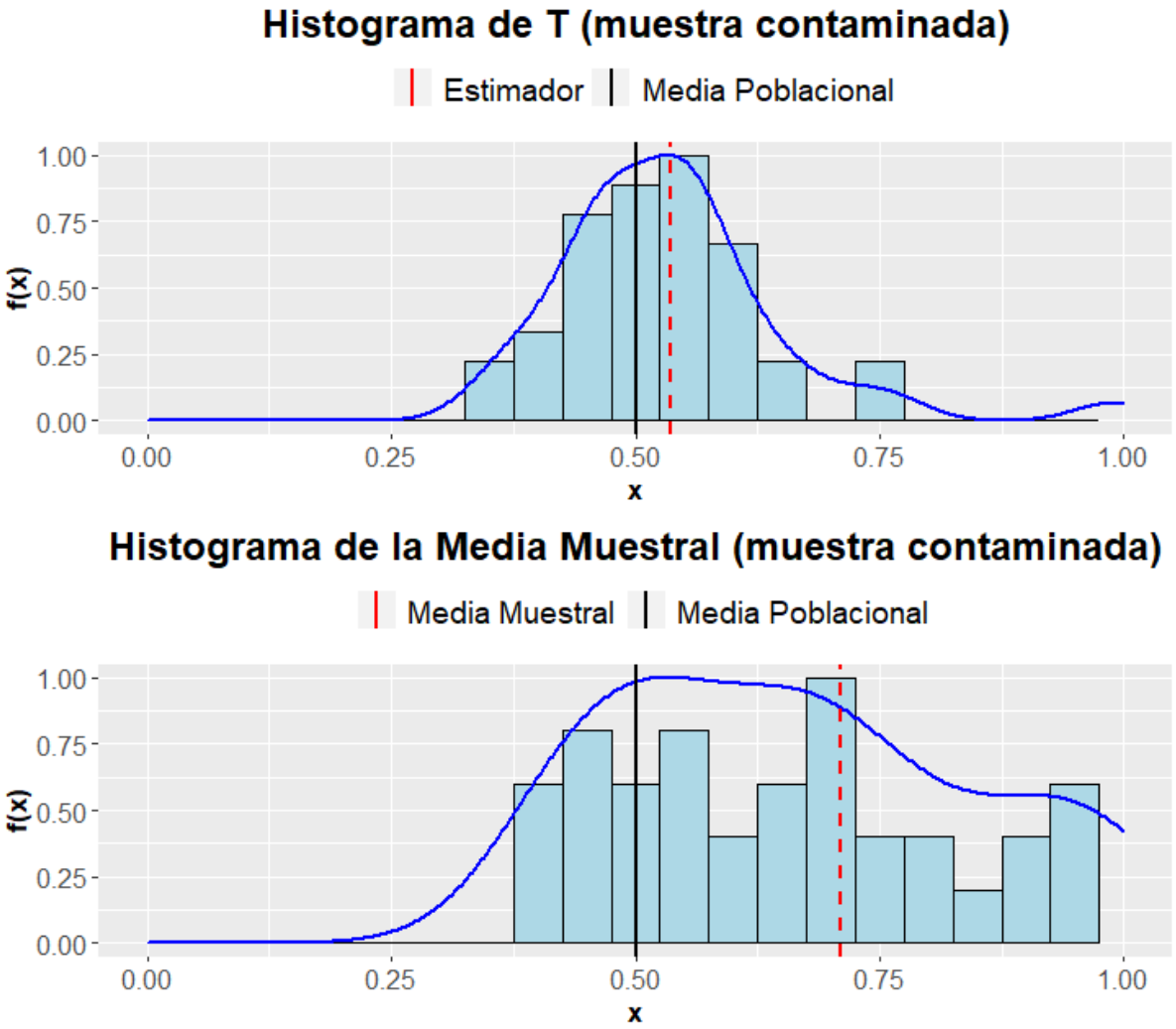
Como podemos observar en el gráfico, la media muestral contaminada es muy diferente a la media sin contaminar, concretamente un 20% mayor, por lo que podemos decir que es bastante asimétrica a la derecha. Como indica el coeficiente de curtosis, esta distribución es platicúrtica, es decir, es menos apuntada y con colas menos gruesas que la normal.Comparando las medidas de dispersión en la muestra contaminada y la original, la desviación típica muestral (SD) es un 15% mayor, así como el IQR y la MAD, que son un 28% y 18% respectivamente.

4.6.2.3 Comparando la Media Muestral en la muestra original y contaminada



Podemos observar que al añadir datos atípicos, nuestro estimador es muy poco robusto, ya que en promedio se desplaza mucho de la media poblacional. Cuando la muestra no contiene datos atípicos, la media muestral sobreestima en promedio la media poblacional en un 1.45412 %, mientras que al añadir datos atípicos lo hace en un 42.08146 % . Esto nos permite concluir que la media muestral no es un estimador robusto de la poblacional.

4.6.2.3 Comparando la robustez de T y la media muestral



Finalmente podemos comparar ambos estimadores (T y la media muestral), observando que **T es mucho más robusto que la media muestral**, ya que se desplaza en promedio mucho menos que esta última.

4.7 Suficiencia