Técnicas de Inferencia Estadística I

Entrega 2: Propiedades de los estimadores

Grado en Estadística y Empresa
Daniel Aramburu, Guillermo Palomo, Jorge Salas
y Marc Pastor

13 de noviembre de 2020

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):

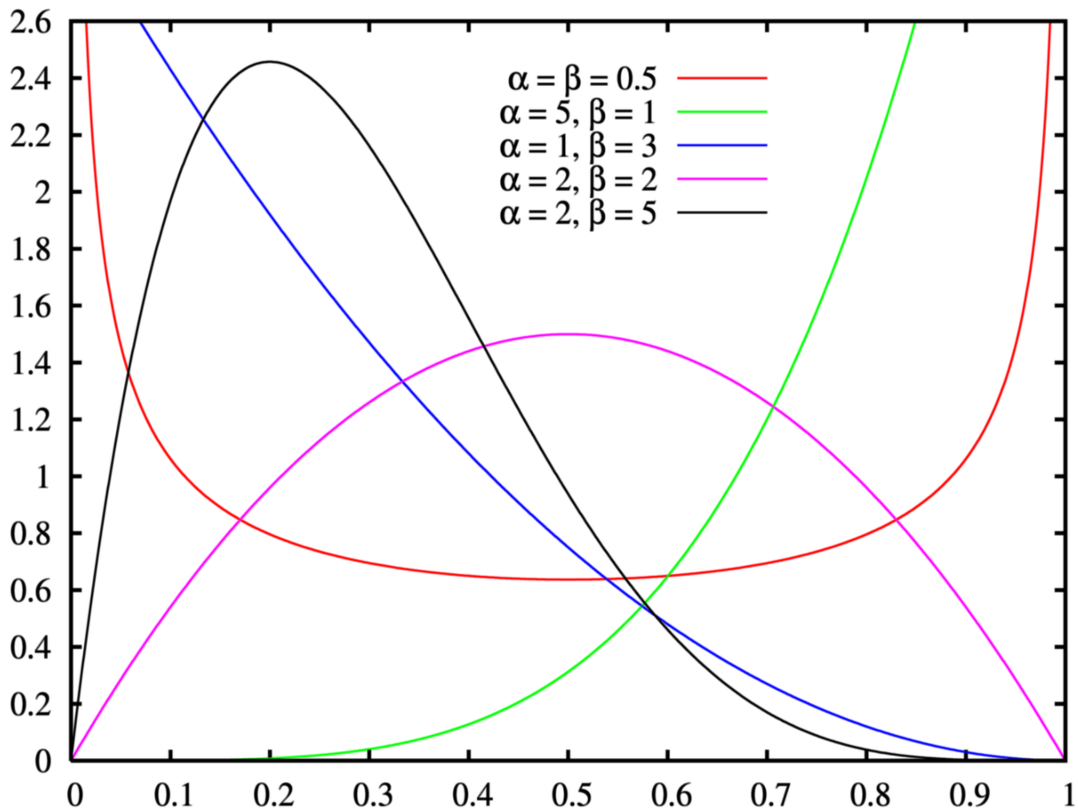
$$B(\alpha, \beta) = B(\beta, \alpha)$$

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

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

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

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

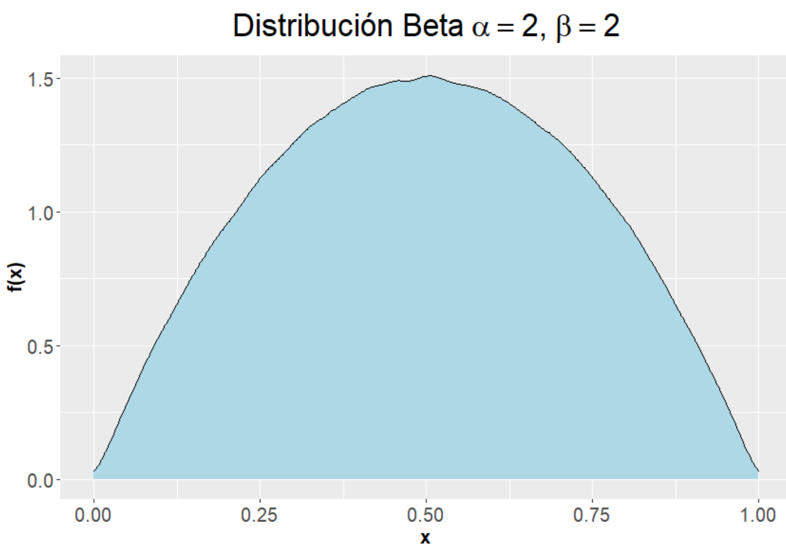
$$Var[X] = \frac{\alpha \beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)}$$

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

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 de T y de la Media Muestral

En este apartado hemos descrito las principales características de nuestro estimador T, y al mismo tiempo lo hemos comparado con la media muestral. Hemos demostrado de manera analítica la insesgadez, la invarianza, y la suficiencia, mientras que el resto de propiedades han sido demostradas numéricamente, a partir de la ley de los grandes números. La ley fuerte de los grandes números asegura que dadas n variables aleatorias independientes e idénticamente distribuidas, con esperanza finita, el promedio muestral de dichas variables aleatorias converge a la media poblacional con probabilidad = 1 cuando el tamaño muestral aumenta infinitamente.

Podemos aplicar esta ley debido a que hemos generado 40 muestras independientes (ya que los valores de cada muestra se generan de manera aleatoria e independiente de los valores de las demás) y de distintos tamaños muestrales (n = 10, n = 100, ... , n = 100000) de la misma distribución Beta(2,2) (muestras idénticamente distribuidas). Para probar las propiedades numéricamente hemos calculado las diferentes medidas (esperanza, sesgo, varianza, ECM, precisión, etc) de los estimadores de la media poblacional para todos los distintos números de muestras de tamaño 10, y con ello hemos obtenido una serie de datos y gráficos, que nos permiten ver que sucede con cada medida cuando el tamaño muestral tiende a infinito.

4.1 Insesgabilidad

4.1.1. Analíticamente

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

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

Calculemos los percentiles P_{40} y P_{60}:

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

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

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

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

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

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

Percentil 40:

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

$3P_{40}^2 - 2P_{40}^3 = 0,4$

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

Percentil 60:

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

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

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

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

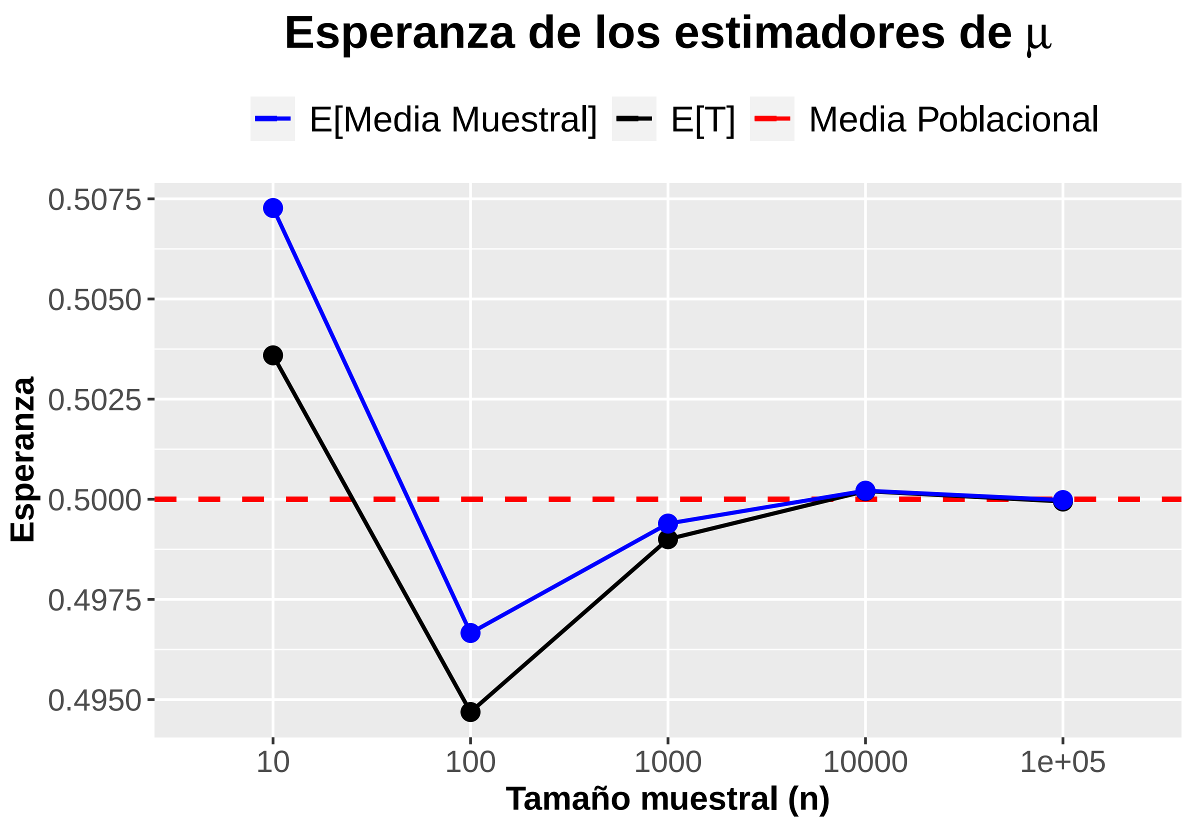
$E[T] = E\left[\frac{P_{40}+P_{60}}{2}\right] = \frac{1}{2} \cdot E[P_{40}+P_{60}] =$

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

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

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

4.1.2 Numéricamente

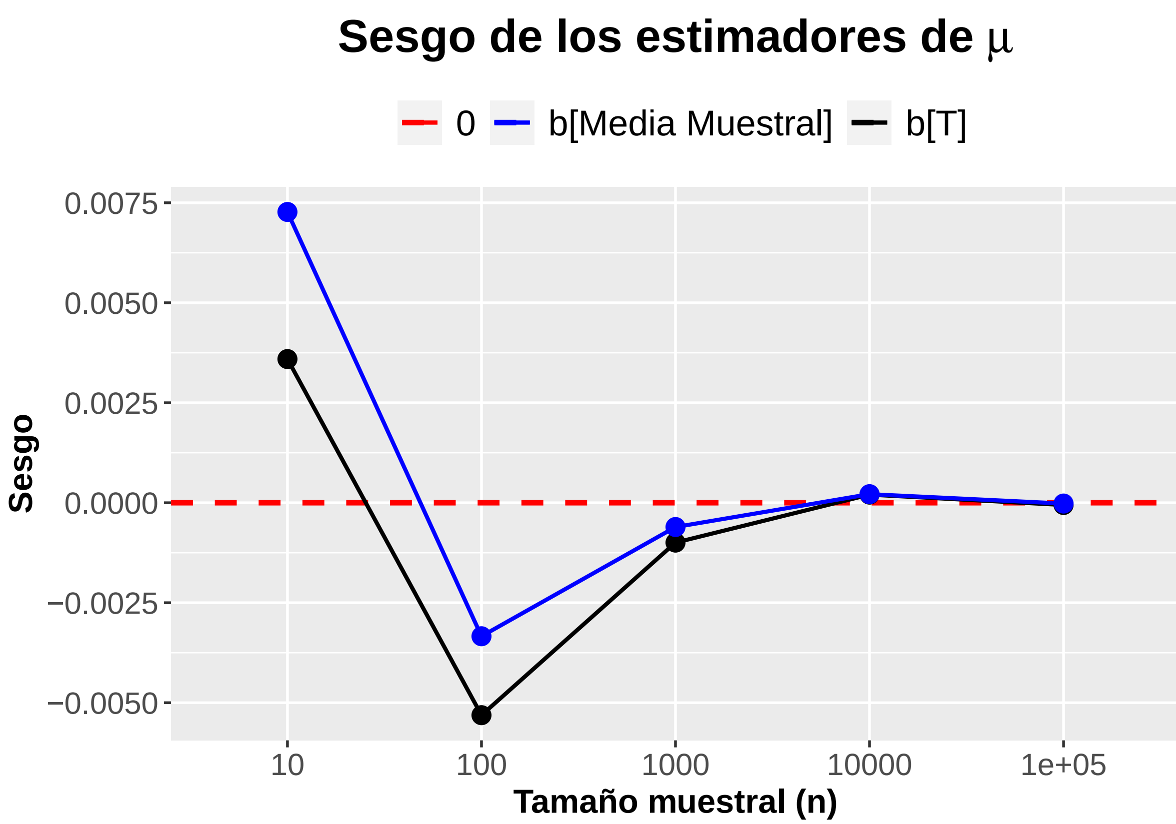


Podemos ver que a medida que aumenta el tamaño muestral, la esperanza de nuestro estadístico (en negro) así como la de la media muestral (en azul), se van acercando cada vez más a la media poblacional (en rojo). Por la ley de los grandes números podemos suponer que para un tamaño muestral suficientemente grande (cuando n tienda a infinito), ambos estimadores serán insesgados para la media poblacional, es decir, en promedio convergerán a la media poblacional, cuyo valor es de 0,5 . Además cabe destacar que sabemos por teoría, que ambos estimadores son insesgados (hemos demostrado solamente la insesgadez de T, ya que la insesgadez de la media muestral respecto a la media poblacional ya la hemos dado en clase)

Definimos el sesgo como:

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

Hemos obtenido el siguiente gráfico para el sesgo.

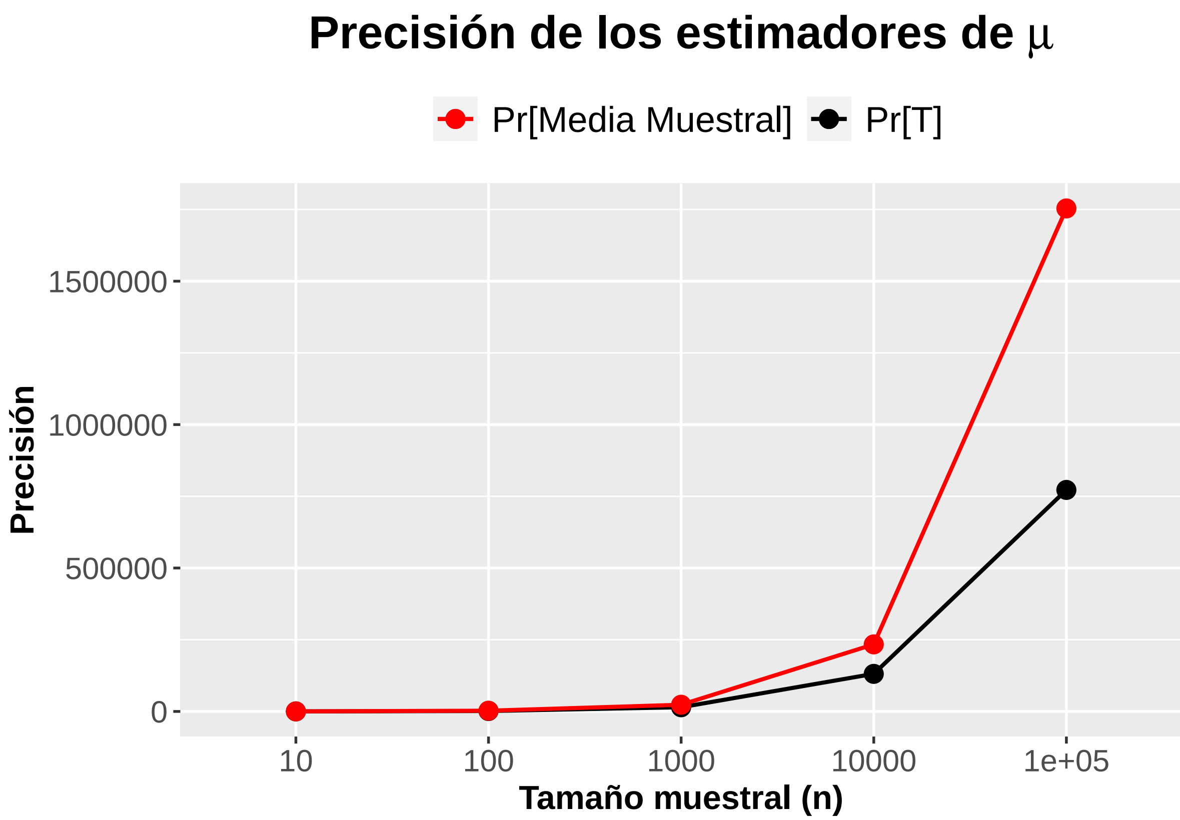


En este caso, podemos observar que el sesgo de ambos estimadores tiende a cero, es decir, que a medida que aumenta el tamaño muestral, la diferencia entre el valor de los estimadores y de la media poblacional converge a 0. Por la ley de los grandes números, podemos suponer que para un tamaño muestral muy grande, el sesgo de ambos estadísticos será 0. Después de la demostración analítica y de observar los dos gráficos anteriores, podemos concluir que **tanto nuestro estadístico como la media muestral son estimadores insesgados de la media poblacional**.

4.2 Eficiencia / Precisión

La eficiencia o precisión se define como:

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



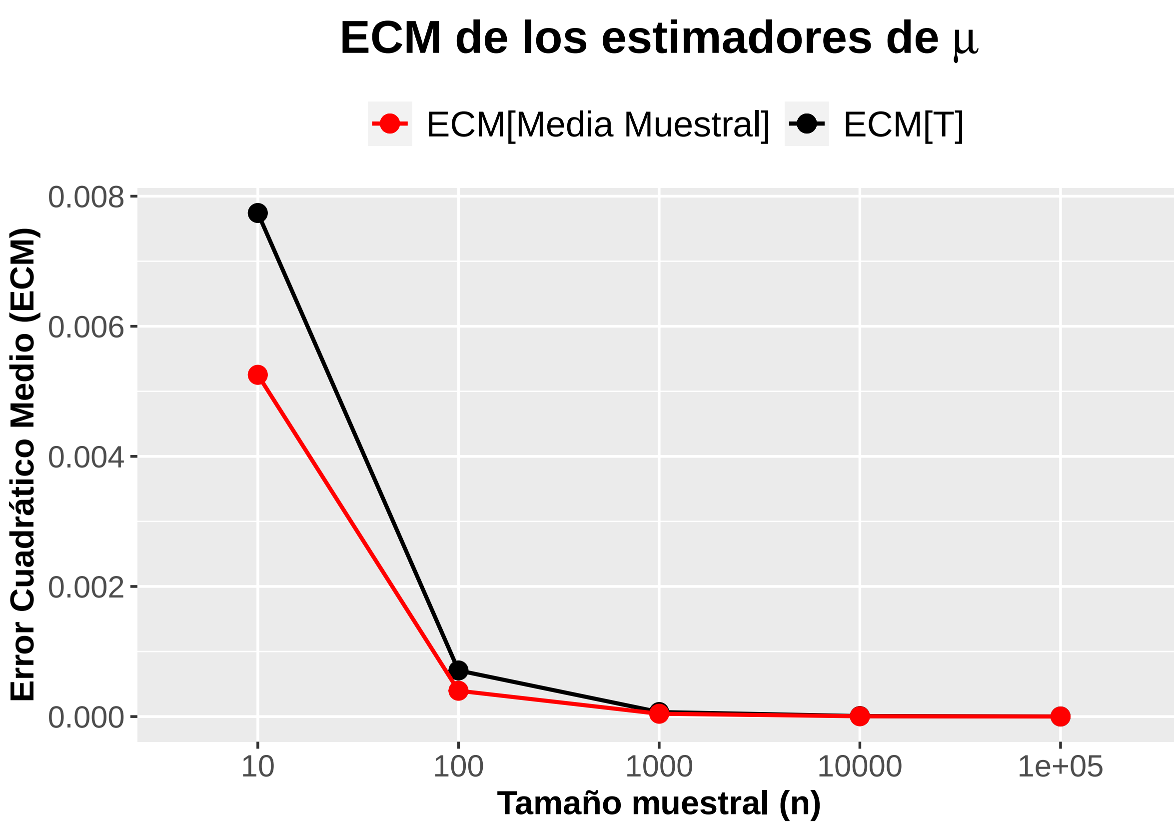
Podemos observar que la media muestral es más precisa estimando la media poblacional que nuestro estadístico T. Esto se debe a que la varianza de la media muestral es menor que la de nuestro estadístico, y también a que la media muestral es el estimador máximo verosímil de la media poblacional.

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.

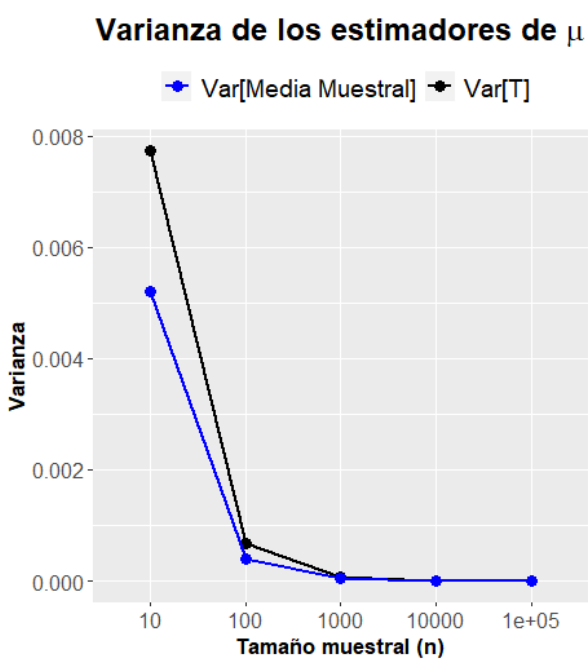
4.4 Consistencia

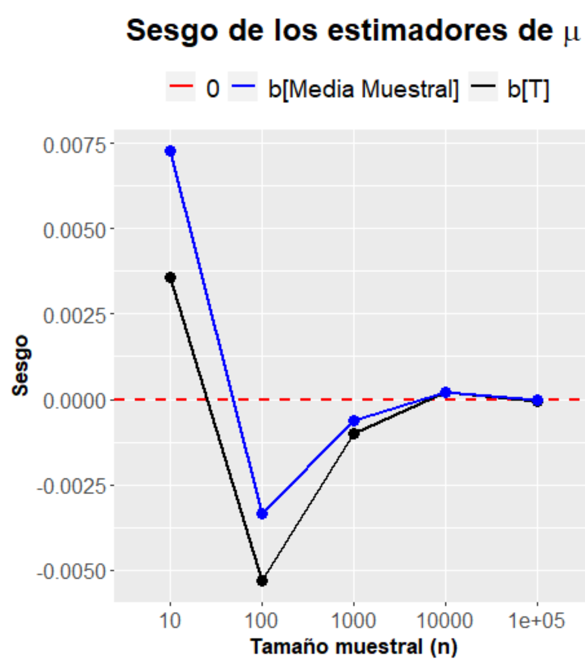
Un estimador T se dice consistente si como condición suficiente cumple las siguientes propiedades:

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

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

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), la varianza y el sesgo de ambos estimadores tienden a cero. Esto nos permite afirmar con bastante seguridad que **la media muestral y T son estimadores consistentes para la media poblacional**, es decir, que al aumentar el tamaño muestral se aproximan bien a ella.

Además sabemos por teoria que la media muestral es consistente para la media poblacional, ya que su varianza es la varianza poblacional dividida entre n. Esto supone que cuando n tiende a infinito la varianza converge a 0, y como el sesgo es 0, es consistente para la media poblacional.

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.

Por otro lado, sabemos por teoria, que **la media muestral es invariante por cambios de escala y por traslaciones**.

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].

Esquema:

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

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

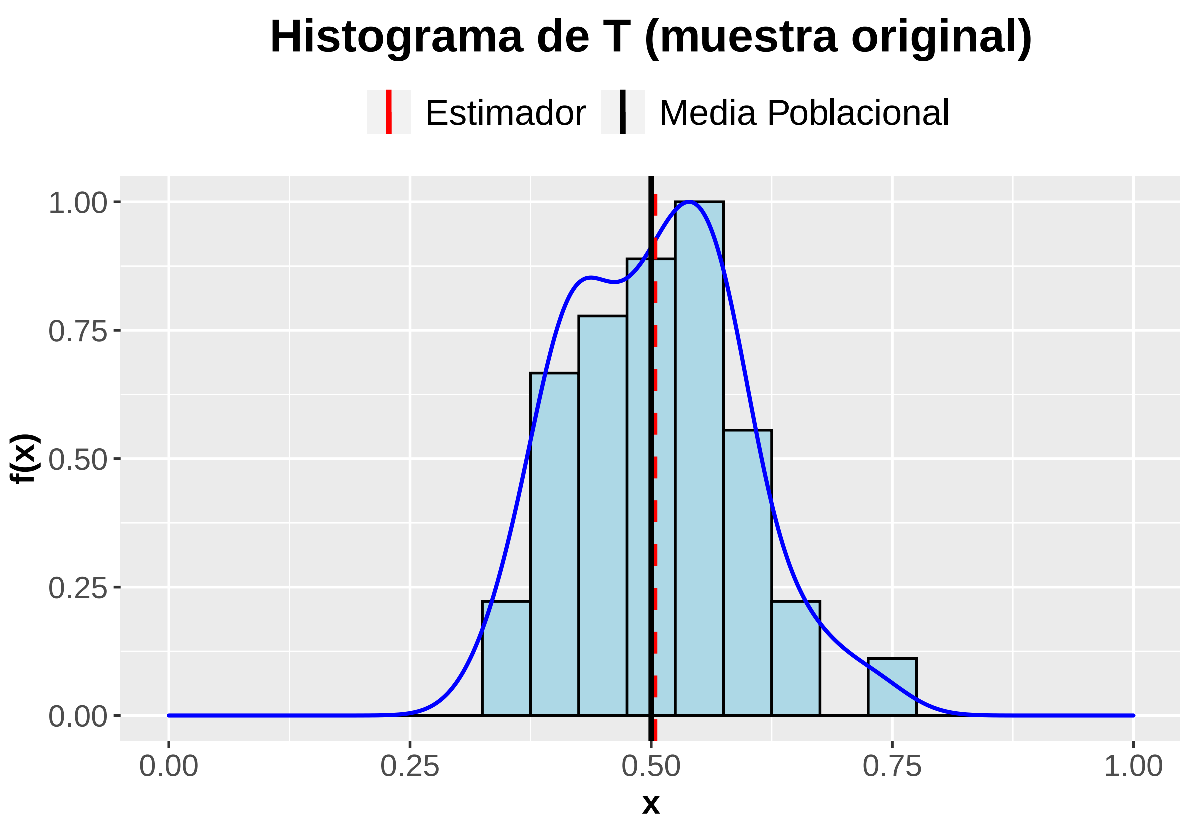
$$X_{u'} = U(2, 3)$$

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

4.6.1 Robustez de nuestro estadístico

4.6.1.1 Muestra sin contaminar

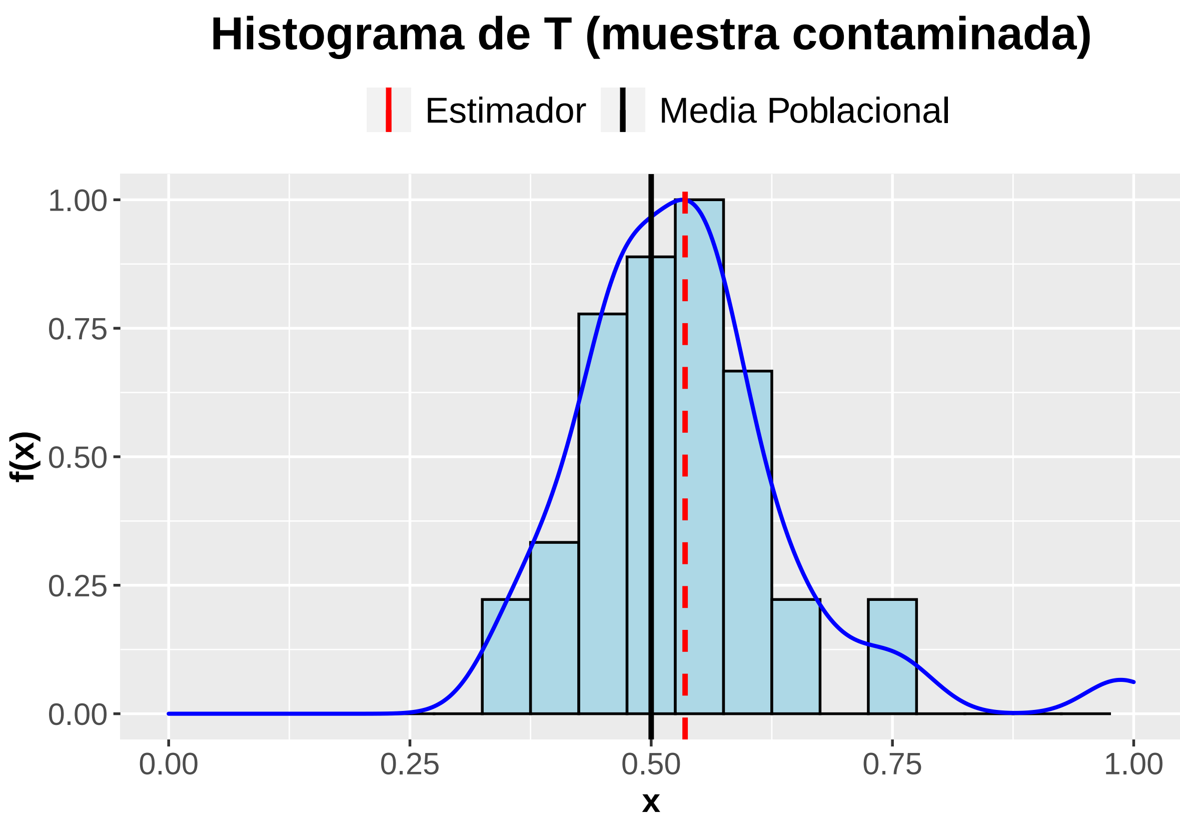
Estadístico						
Medidas de Centralización		*Medidas de Dispersión*			*Medidas de Forma*	
Media	Mediana	SD	IQR	MAD	Curtosis	Asimetría
0.503593	0.5041133	0.08791211	0.1353218	0.09725723	2.656788	0.2748998



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.

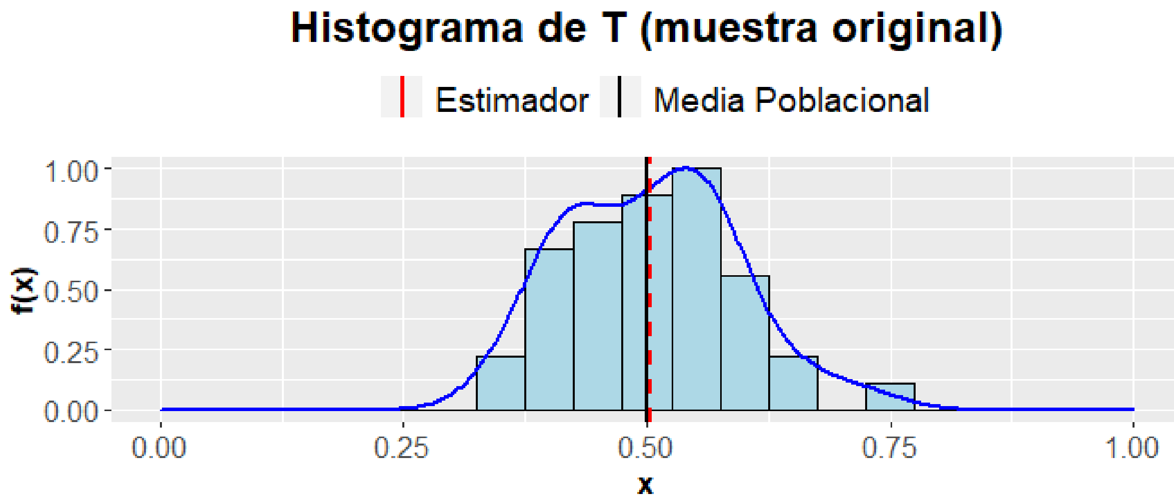
4.6.1.2 Muestra contaminada

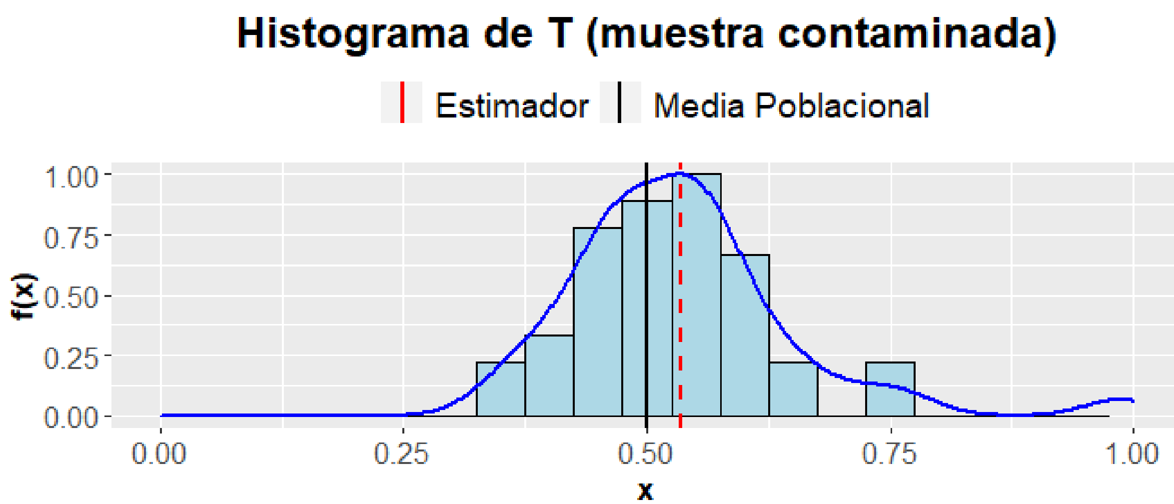
Estadístico T						
Medidas de Centralización		*Medidas de Dispersión*			*Medidas de Forma*	
Media	Mediana	SD	IQR	MAD	Curtosis	Asimetría
0.5351557	0.5279007	0.1172313	0.1140922	0.08615515	6.962659	1.455874



Nuestro estadístico para la muestra contaminada se desplaza en promedio relativamente poco de la media poblacional. A la mediana le 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 (> 1), es asimétrica positiva, siendo su asimetría 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 mucho más apuntada y con colas más pesadas que la normal.

4.6.1.3 Comparando T en la muestra original y contaminada



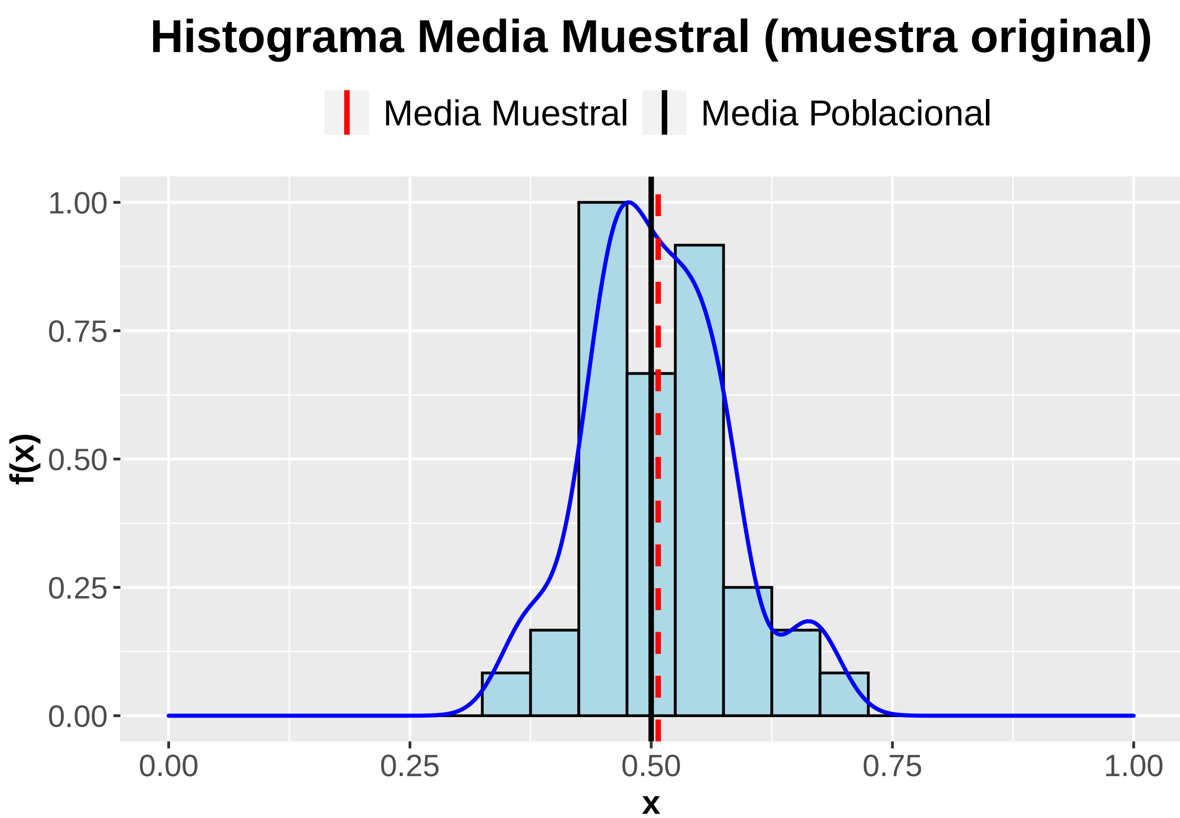


Podemos observar que al añadir datos atípicos, nuestro estimador es bastante robusto, ya que en promedio se desplaza relativamente poco de la media poblacional. Cuando la muestra no contiene datos atípicos, T sobreestima la media poblacional en un 0.7186 %, mientras que al añadir datos atípicos lo hace 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 que el valor de la esperanza de las medias muestrales es muy parecido al valor de 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.

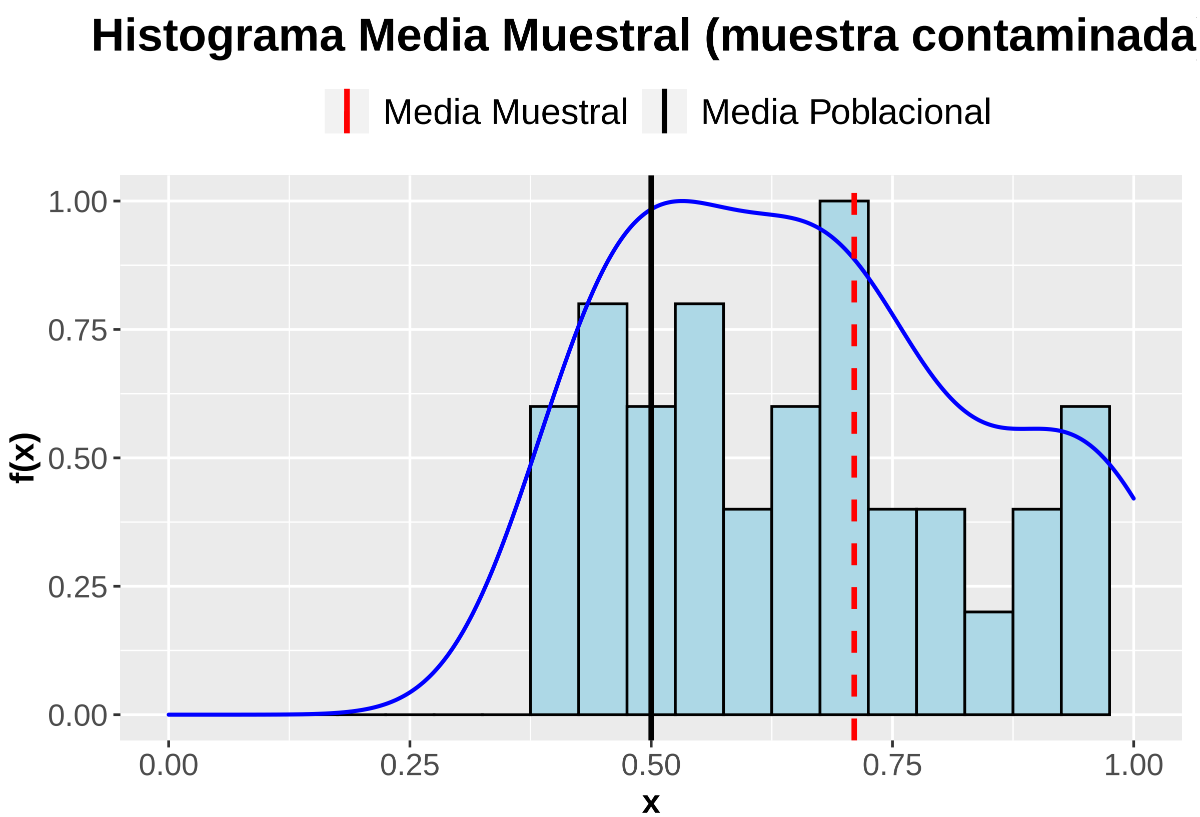
4.6.2.1.1 Varianza de la media muestral

Varianza de la Media Muestral (muestra original)	
Valor Obtenido	Valor Teórico
0.005200948	0.005

Podemos observar que la varianza de la media muestral que hemos obtenido, es muy parecida a la teórica, sobreestimando el valor teórico en tan solo un 9%. Si el tamaño muestral aumentara infinitamente, la varianza muestral convergería a la poblacional.

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



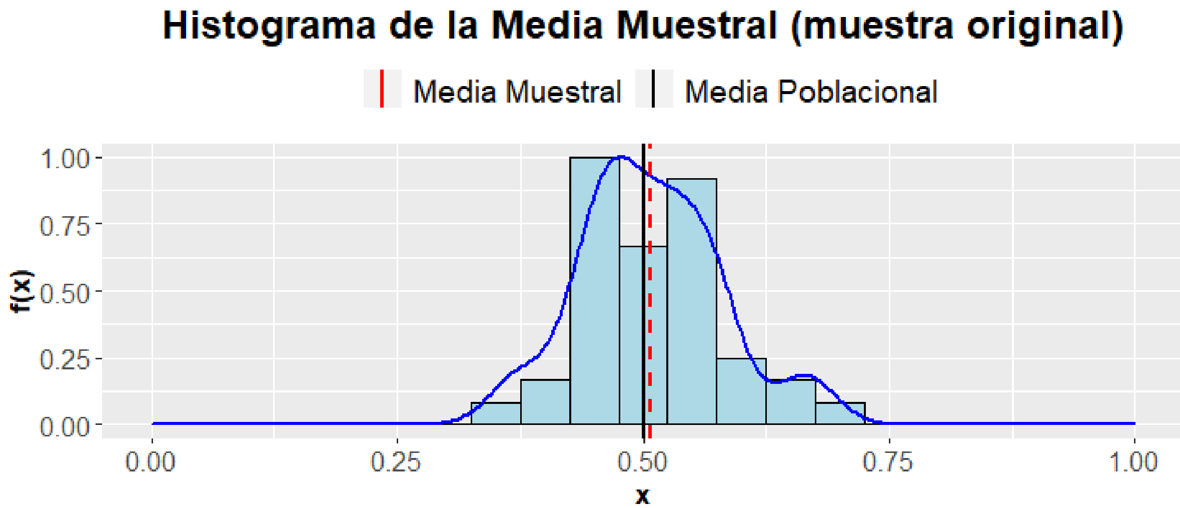
Como podemos observar en el gráfico, la esperanza de la media muestral de la muestra contaminada es muy diferente a la original, concretamente un 20% mayor. Podemos observar que es muy asimétrica a la derecha. El coeficiente de curtosis indica que esta distribución es platicúrtica, es decir, menos apuntada 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.

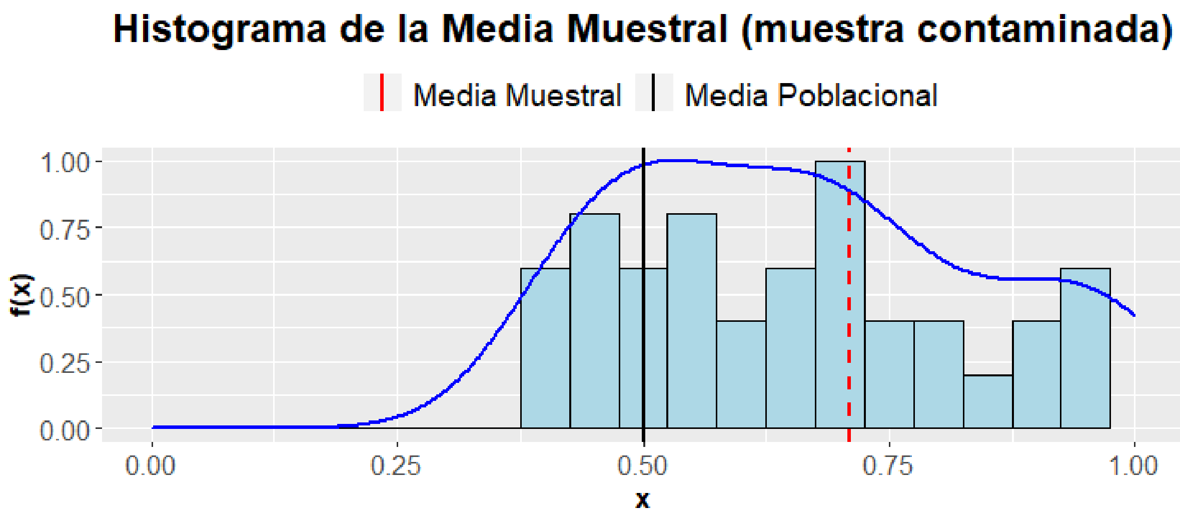
4.6.2.2.1 Varianza de la media muestral

Varianza de la Media Muestral (muestra contaminada)	
Valor Obtenido	Valor Teórico
0.05146932	0.005

Al contaminar los datos, podemos observar como la varianza de la media muestral está muy lejos de su valor teórico, sobreestimándolo en un 900% . Esto es otro indicador de la poca robustez de la media muestral como estimador.

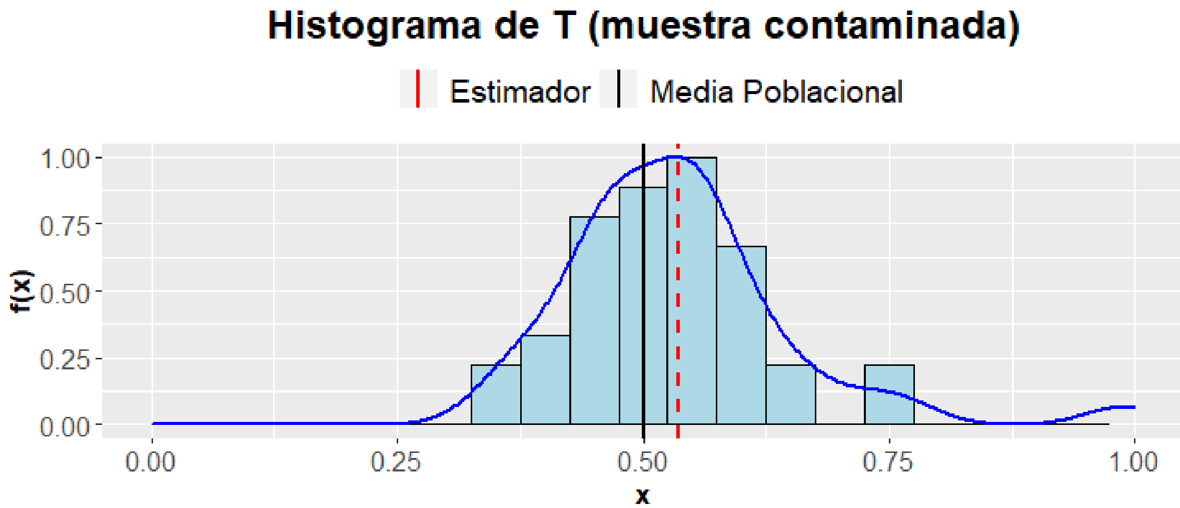
4.6.2.3 Comparando la Media Muestral en la muestra original y contaminada

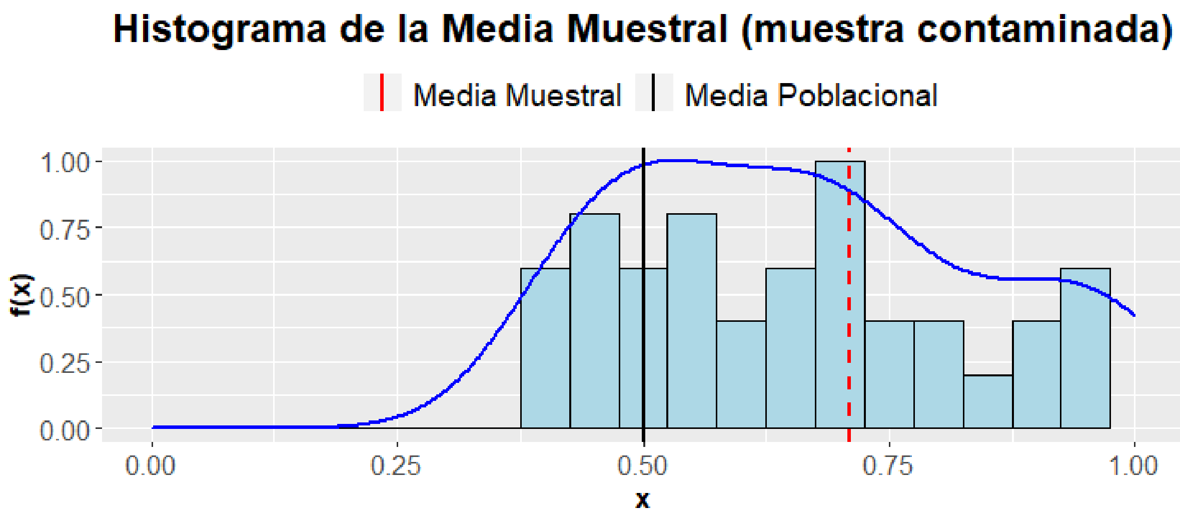




Podemos observar que al añadir datos atípicos, la media muestral es muy poco robusta, ya que en promedio se desplaza mucho de la media poblacional. Cuando la muestra no contiene datos atípicos, la media muestral sobreestima en promedio la media poblacional en un 1.45412 %, mientras que al añadir datos atípicos lo hace en un 42.08146 % . Además, hemos visto que en el caso de la muestra original, la varianza de la media muestral sobreestima la varianza poblacional tan solo en un 4%, mientras que al añadir datos atípicos lo hace en un 900%. 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 un estimador bastante robusto de la media poblacional**, mientras que **la media muestral no es nada robusta**.

4.7 Suficiencia

Suficiencia

- $X \sim \text{Beta}(\alpha, \beta)$

$$\boxed{f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}} \quad 0<x<1$$

Sean $X_1, X_2, \dots, X_n$ m.a.s de X, obtenemos la función de verosimilitud de X:

$$f(x_1,\dots,x_n) \overset{ind}{=} f(x_1)\cdot\ldots\cdot f(x_n) =$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x_1^{\alpha-1}(1-x_1)^{\beta-1} \cdot \ldots \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x_n^{\alpha-1}(1-x_n)^{\beta-1} =$$

$$= \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^n \prod_{i=1}^{n} x_i^{\alpha-1}(1-x_i)^{\beta-1} = L(\alpha,\beta)$$

Recordemos que nuestro estadístico tiene la siguiente expresión:

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

Como podemos ver en la función de verosimilitud $L(\alpha,\beta)$ no aparece T ni la media muestral, y no es posible llevar a cabo transformaciones de $L(\alpha,\beta)$ para incluirlos. Por ello tanto nuestro estimador T, como la media muestral, no son estimadores suficientes de la distribución

5. Bibliografía

- Asimetría estadística. (2020). Recuperado de: https://es.wikipedia.org/w/index.php?title=Asimetr%C3%ADa_estad%C3%ADstica&oldid=125822430
- Beta distribution. (2020). Recuperado de: https://en.wikipedia.org/w/index.php?title=Beta_distribution&oldid=981642297
- Brown, Kyle. (2020). Understanding R: Mathematical Expressions. Recuperado de: https://rpubs.com/kylewbrown/math-expressions
- Características de estimadores. (s. f.). Recuperado de: https://www.uv.es/webgid/Inferencial/42_caractersticas_estimadores.html#:~:text=Diremos%20que%20un%20estimador%20es,muestra%20aproxime%20al%20p
- Commonly used symbols in R Markdown. (s. f.). Recuperado de: https://people.ok.ubc.ca/jpither/modules/Symbols_markdown.html
- Curtosis. (2020). Recuperado de: https://es.wikipedia.org/w/index.php?title=Curtosis&oldid=128327400
- Curtosis - Definición, qué es y concepto. (2017). Recuperado de: https://economipedia.com/definiciones/curtosis.html
- Función gamma. (2020). Recuperado de: https://ast.wikipedia.org/w/index.php?title=Funci%C3%B3n_gamma&oldid=2985655
- Greek letters, symbols, and line breaks inside a ggplot legend label . (s. f.). Recuperado de: https://stackoverflow.com/questions/27690729/greek-letters-symbols-and-line-breaks-inside-a-ggplot-legend-label
- Mathematical Annotation in R. (s. f.). Recuperado de: https://stat.ethz.ch/R-manual/R-devel/library/grDevices/html/plotmath.html
- Medidas de dispersión. (2020). Recuperado de: https://es.wikipedia.org/w/index.php?title=Medidas_de_dispersi%C3%B3n&oldid=128745923
- Molina, Luis. (2009). Apuntes de Latex. Recuperado de: http://metodos.fam.cie.uva.es/~latex/apuntes/apuntes3.pdf
- Moment Generating Function of Beta Distribution. (2018). Recuperado de: https://proofwiki.org/wiki/Moment_Generating_Function_of_Beta_Distribution#:~:text=Let%20X%E2%88%BCBeta(%CE%B1,%2Br)t
- Ramirez, Oscar. (2015). Escribiendo funciones en R. Recuperado de: https://rpubs.com/osoramirez/93049
- Reilly, Archer. (2020). Writing Mathematic Fomulars in Markdown. Recuperado de: http://csrgxtu.github.io/2015/03/20/Writing-Mathematic-Fomulars-in-Markdown/
- Remove space between two graphs in R. (s. f.). Recuperado de: https://stackoverflow.com/questions/22226358/remove-space-between-two-graphs
- Rincón, Luis. (2013). Distribución beta. Recuperado de: https://www.youtube.com/watch?v=cqvEph836gM&ab_channel=LuisRinc%C3%B3n