**Primer Trabajo Estadística II**

Inferencia Estadística

**Integrantes:**

1. Juan David Mena Gamboa
2. Miguel Angel Bolaño López
3. Heyner David Marquez Garnica

* **Fuente de los datos (url):**

<https://microdatos.dane.gov.co/index.php/catalog/819/get-microdata>

* **Descripción de los datos seleccionados:**

Cuantitativas:P3094S3: Cuánto ahorro por cultivar.P3087S1: Valor mensual por prácticas o pasantías.P3095S3: Valor ahorra por criar animales.CualitativasP3101: ¿Fue a reuniones familiares durante las últimas 4 semanas? sí o no.

* **Variables seleccionadas:**

|  |  |  |
| --- | --- | --- |
| **Variables Cualitativas** | | |
| **Nombre Variable** | | **Categorías o Niveles** |
| **1** | P3101 | Sí y No |

|  |  |  |
| --- | --- | --- |
| **Variables Cuantitativas** | | |
| **Nombre Variable** | | **Unidad de medición** |
| **1** | P3094S3 | COP |
| **2** | P3087S1 | COP |
| **3** | P3095S3 | COP |

**Desarrollo**

**Variable 1**

* ***Análisis Descriptivo***

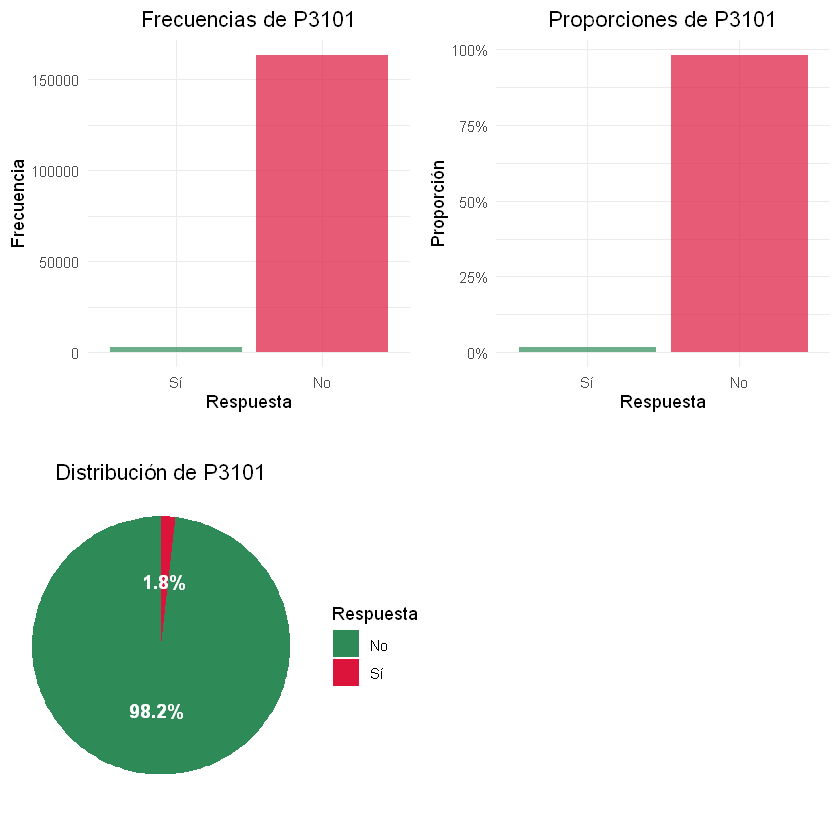
*Descriptivos Básicos*

|  |  |  |
| --- | --- | --- |
|  | Medida | Valor |
| 1 | Registros | 166341 |
| 2 | Número de personas que dijeron “Sí” | 2918 |
| 3 | Número de personas que dijeron “No” | 163423 |
| 4 | Porcentaje de “Sí” | 1.75% |
| 5 | Porcentaje de “No” | 98.25% |
| 6 |  |  |
| 7 |  |  |
| 8 |  |  |
| 9 |  |  |
| 10 |  |  |

Comentarios:

Los resultados descriptivos muestran que, de un total de 166,341 registros, solo 2,918 personas (1.75%) respondieron afirmativamente, mientras que la gran mayoría (163,423 personas, 98.25%) respondió negativamente. Esto refleja un marcado desequilibrio en las respuestas, donde la opción “No” domina de manera abrumadora sobre el “Sí”. Por lo tanto, cualquier análisis posterior debe considerar esta desproporción en la distribución, ya que puede influir en la interpretación estadística y en la validez de los modelos que se construyan con estos datos.

*Análisis Gráfico*



Comentarios:

* ***Cálculo de los estimadores***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Estimador** | **Variable** | **Estimadores Puntuales** | | **Estimadores por Intervalo** | |
| **Analogía** | **Máxima Verosimilitud** | **Límite Inferior** | **Límite Superior** |
| (Media o Proporción) | 1. Si | 0.0175 | 0.0175 | 0.0169 | 0.0182 |
| (Media o Proporción) | 2. No | 0.9825 | 0.9825 | 0.9818 | 0.9831 |
| Comentario | | La gran mayoría respondió “No”, y solo una fracción muy pequeña respondió “Sí” | La estimación por máxima verosimilitud coincide con la proporción muestral, por lo que se obtienen los mismos valores: **0.0175 para “Sí”** y **0.9825 para “No”**. | Para la proporción **“Sí”**, en el peor escenario, la proporción real sería cercana al **1.69%**. Para la proporción **“No”**, asegura que al menos un **98.18%** de la población respondió “No” | Para **“Sí”**, la proporción real no superaría el **1.82%**.  Para **“No”**, como máximo un **98.31%** de la población respondió “No”. |

* ***Evaluación del estimador:***

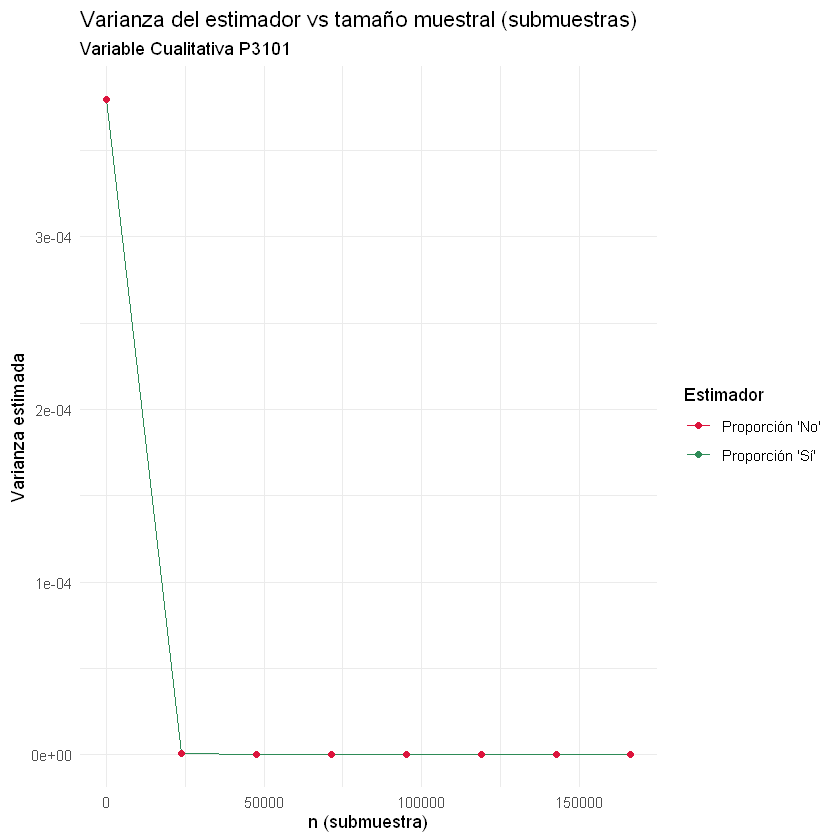
*Insesgamiento*

|  |  |  |
| --- | --- | --- |
| **Media** | **Mediana** | **Sesgo** |
| 0.0175 | 0.0175 | -5.7e-05 |

Comentario:

La proporción estimada (1.75%) es estable, sin sesgos relevantes, y tanto la media como la mediana reflejan adecuadamente el valor central de la distribución.

*Consistencia*



Comentario:

A medida que aumenta el tamaño muestral, la varianza de los estimadores de las proporciones disminuye rápidamente hasta estabilizarse cerca de cero. Esto indica que con muestras grandes las estimaciones son muy precisas y prácticamente no presentan variabilidad.

*Eficiencia*

|  |  |
| --- | --- |
| **Medida** | **Valor** |
| Media | 0 |
| Mediana | 0 |

Comentario:

Para variables cualitativas binarias, ambos estimadores (proporción 'Sí' y 'No')

tienen la misma eficiencia teórica, ya que son complementarios (p + q = 1).

Sintaxis empleada con esta variable:

--- Sintaxis clave usada en este análisis ---

1) Limpieza categórica: clean\_categorical(...)

2) Descriptivos: table(), prop.table() para frecuencias y proporciones

3) Gráficas: geom\_bar(), coord\_polar() para barras y pie charts

4) IC proporciones: binom.confint() (exacto y asintótico)

5) Bootstrap: boot() para IC y bias de proporciones

6) Consistencia: submuestreo comparando var(proporción) vs n

7) Eficiencia: comparar var\_bootstrap(proporción)

**Variable 2**

* ***Análisis Descriptivo***

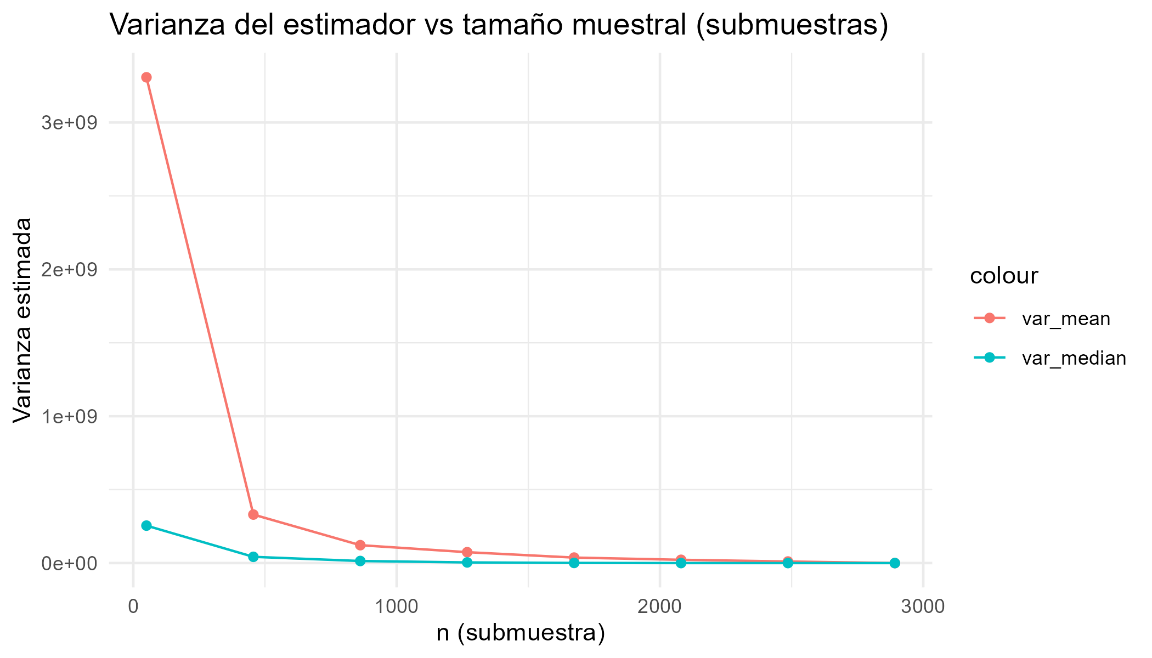
*Descriptivos Básicos*

|  |  |  |
| --- | --- | --- |
|  | Medida | Valor |
| 1 | Registros | 166341 |
| 2 | Valores faltantes | 166005 |
| 3 | Mínimo | 80000 |
| 4 | Primer cuartil | 750000 |
| 5 | Mediana | 975000 |
| 6 | Media | 998708 |
| 7 | Tercer cuartil | 1300000 |
| 8 | Máximo | 9000000 |
| 9 | Desviación Estándar | 621977 |
| 10 | Rango intercuartílico | 550000 |

Comentarios:

La variable **P3094S3** presenta 166,341 observaciones, de las cuales solo 2,892 son válidas debido al alto nivel de no respuesta. Los valores muestran gran dispersión, con un mínimo de 1,000 pesos y un máximo de 20 millones, evidenciando la presencia de outliers. La media (126,938 pesos) resulta más del doble de la mediana (60,000 pesos), lo que indica una distribución sesgada a la derecha. El rango intercuartílico (30,000–150,000) refleja concentración en valores moderados, mientras que la elevada desviación estándar confirma la alta variabilidad. En conclusión, la **mediana** es un mejor indicador de la tendencia central que la media en este conjunto de datos.

*Análisis Gráfico*



Comentarios:

Los gráficos muestran que la variable P3094S3 (COP) presenta una alta varianza inicial, debido a la presencia de valores extremos. A medida que aumenta el número de observaciones, tanto la media como la varianza se estabilizan, indicando mayor consistencia en los datos. Además, se observa una diferencia clara entre los grupos “SI” y “NO”, donde el grupo “SI” reporta ingresos considerablemente más altos. Esto sugiere una distribución asimétrica y la necesidad de considerar medidas como la mediana para representar el ingreso típico.

* ***Cálculo de los estimadores***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Estimador** | **Estimadores Puntuales** | | **Estimadores por Intervalo** | |
| **Analogía** | **Máxima Verosimilitud** | **Límite Inferior** | **Límite Superior** |
| (Media o Proporción) | 126,937 | 126,937.6 | 111,985.4 | 141,889.7 |
| Comentario | La media estimada es 126,937.6 pesos, representando el ingreso promedio en la muestra. Este valor está influenciado por ingresos altos que elevan el promedio. | La mediana es 60,000 pesos, indicando que la mitad de los individuos gana menos y la otra mitad más, lo que sugiere una distribución sesgada hacia ingresos bajos. | El estimador de media por máxima verosimilitud coincide con el estimador por analogía (126,937.6 pesos), lo que valida la estabilidad de la estimación para el ingreso promedio. | La mediana estimada por máxima verosimilitud es 60,000 pesos, igual a la estimación por analogía, reforzando la representatividad del ingreso típico en la muestra. |

* ***Evaluación del estimador:***

*Insesgamiento*

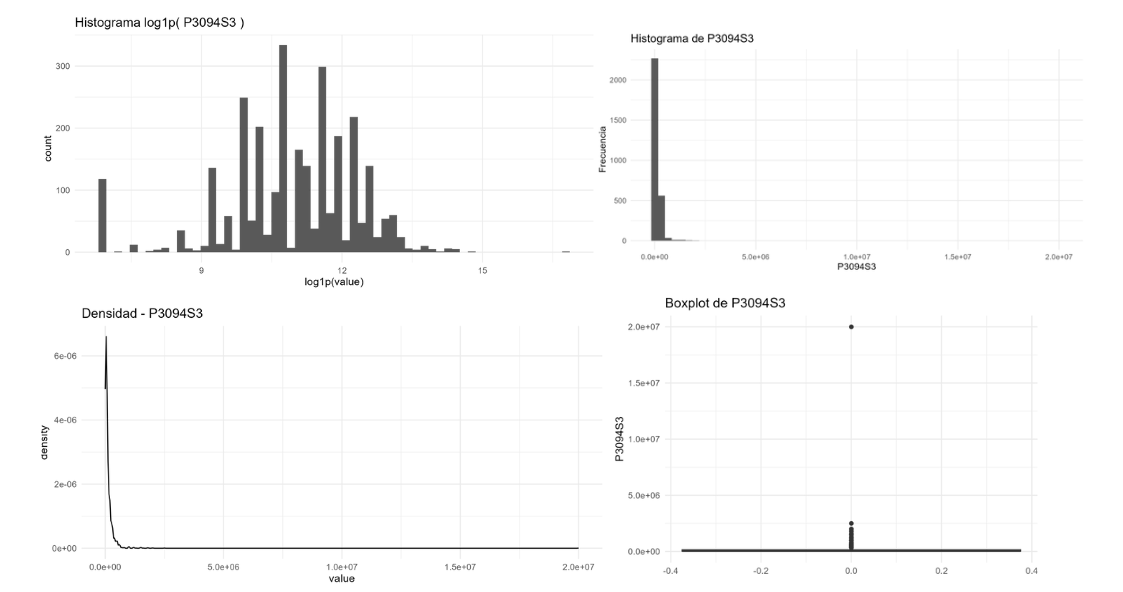
|  |  |  |
| --- | --- | --- |
| **Media** | **Mediana** | **Sesgo** |
| -166.3789 | 129 | Ambos estimadores, media y mediana, presentan un sesgo pequeño relativo, por lo que las estimaciones son confiables y adecuadas para describir la variable de ingresos. |

Comentario:

Los análisis de sesgo para los estimadores de media y mediana indican que ambos presentan un sesgo pequeño y relativo, lo que sugiere que las estimaciones obtenidas son confiables y representativas de la distribución real de los ingresos. Este bajo nivel de sesgo respalda el uso de estos estadísticos para describir adecuadamente la variable P3094S3 en el análisis de la encuesta.

*Consistencia*

Ambos estimadores presentan sesgo pequeño, por lo que pueden considerarse consistentes. Sin embargo, la media muestra un intervalo de confianza más amplio y mayor variabilidad, reflejando sensibilidad a los valores extremos. En contraste, la mediana exhibe intervalos muy concentrados y gran estabilidad frente al remuestreo, lo que confirma su mayor consistencia y robustez en esta distribución de ingresos.



Comentario:

La variable P3094S3 presenta una distribución altamente asimétrica a la derecha, con ingresos mayoritariamente bajos y pocos valores atípicos muy elevados (superiores a 20 millones). El boxplot y el gráfico de densidad muestran una fuerte concentración cerca de cero, mientras que el histograma confirma la escasa representación de los ingresos altos. Tras aplicar la transformación logarítmica (log1p), la distribución se aproxima a la normal y permite identificar con mayor claridad la concentración principal entre 8.000 y 160.000 pesos. En este contexto, se recomienda utilizar medidas robustas como la mediana y los percentiles, dado que la media se ve afectada por los valores extremos.

*Eficiencia*

|  |  |
| --- | --- |
| **Medida** | **Valor** |
| Media | 58,275.694 |
| Mediana | 2,852.046 |

Comentario:

La varianza de la media resulta mucho mayor que la de la mediana, lo que indica que la **mediana es un estimador más eficiente** en esta muestra de la variable P3094S3. En otras palabras, la mediana presenta menor variabilidad frente al re-muestreo bootstrap y ofrece una descripción más estable del ingreso típico, mientras que la media se ve fuertemente afectada por los valores extremos (outliers). Esto confirma que, en distribuciones sesgadas como los ingresos, la mediana es preferible a la media como medida de tendencia central.

Sintaxis empleada con esta variable:

1) Limpieza money: clean\_numeric2(...)

2) Descriptivos: tabla\_vertical (N, N missing, Min, Q1, Median, Mean, Q3, Max, SD, IQR)

3) Gráficas guardadas en: analisis\_P3087S1\_check/plots

4) IC mean (t): t.test(x)$conf.int (o calculado manualmente)

5) Bootstrap bias y CI: uso package 'boot' (boot(...), boot.ci(...))

6) Consistencia: submuestreo comparando var(estimator) vs n

7) Eficiencia: comparar var\_bootstrap(estimator)

**Variable 3**

* ***Análisis Descriptivo***

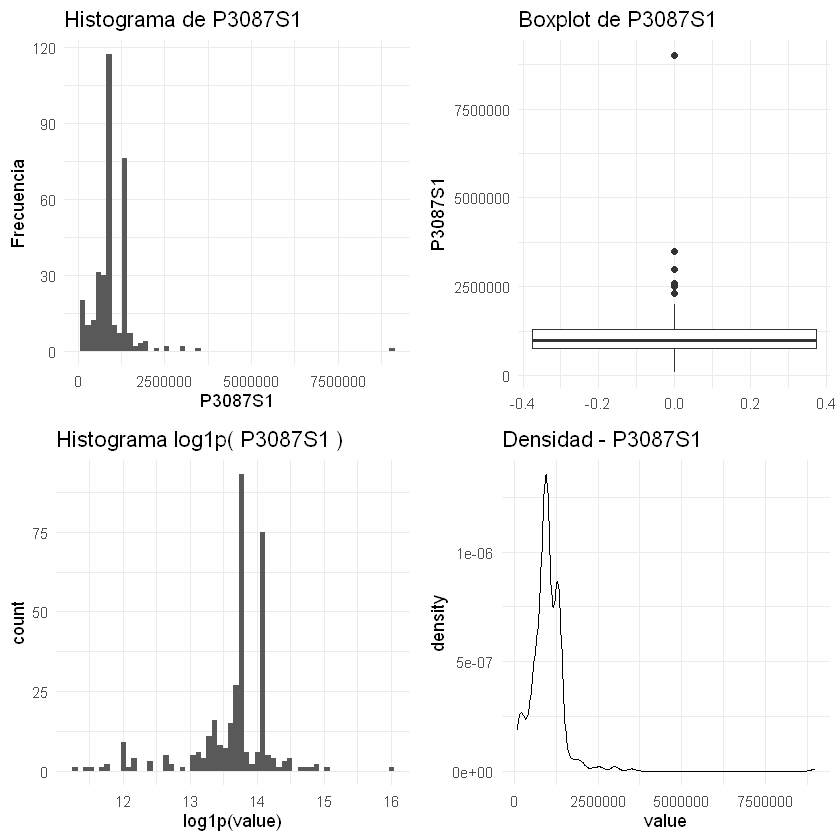
*Descriptivos Básicos*

|  |  |  |
| --- | --- | --- |
|  | Medida | Valor |
| 1 | Registros | 166341 |
| 2 | Valores faltantes | 166005 |
| 3 | Mínimo | 80000 |
| 4 | Primer cuartil | 750000 |
| 5 | Mediana | 975000 |
| 6 | Media | 998708 |
| 7 | Tercer cuartil | 1300000 |
| 8 | Máximo | 9000000 |
| 9 | Desviación Estándar | 621977 |
| 10 | Rango intercuartílico | 550000 |

Comentarios:

El conjunto de datos presenta una alta proporción de valores faltantes (más del 99%), y entre los datos válidos se observa gran dispersión y valores atípicos elevados que influyen en la media. La mediana es una mejor representación del valor típico, ya que no se ve tan afectada por los extremos.

*Análisis Gráfico*



Comentarios:

Las visualizaciones muestran que la variable presenta una distribución altamente asimétrica a la derecha, con varios valores atípicos extremos evidenciados en el boxplot. El histograma y la curva de densidad confirman una concentración significativa de observaciones en valores bajos, mientras que la transformación logarítmica (log1p) reduce la dispersión, revelando una distribución más cercana a la normalidad.

* ***Cálculo de los estimadores***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Estimador** | **Estimadores Puntuales** | | **Estimadores por Intervalo** | |
| **Analogía** | **Máxima Verosimilitud** | **Límite Inferior** | **Límite Superior** |
| (Media o Proporción) | 998,708.3 | 998,708.3 | 931,962.4 | 1,065,454 |
| Comentario | La media es el valor promedio de los datos, representando el centro de la distribución. | El estimador de máxima verosimilitud para la media coincide con la media muestral, ya que bajo normalidad ambos son equivalentes. | El límite inferior del IC al 95% indica que, con alta confianza, la media poblacional no es menor que este valor | El límite superior del IC al 95% establece que, con alta confianza, la media poblacional no excede este valor. |

* ***Evaluación del estimador:***

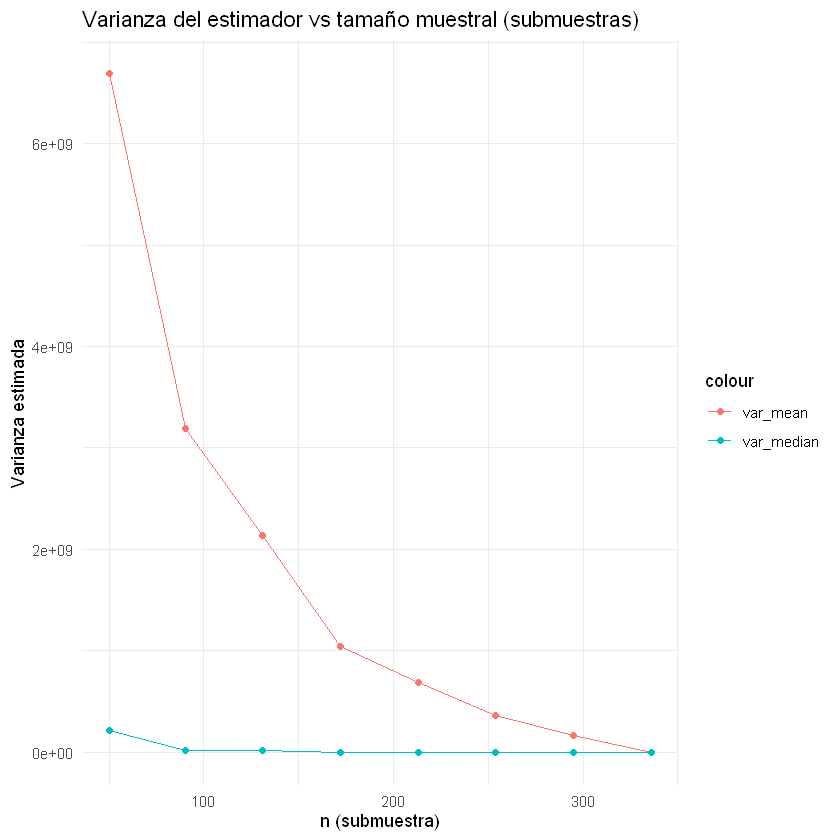
*Insesgamiento*

|  |  |  |
| --- | --- | --- |
| **Media** | **Mediana** | **Sesgo** |
| 996,995.5 | 975,000 | Media: sesgo ≈ +333.15  Mediana: sesgo ≈ -282.5 |

Comentario:

Tanto la media como la mediana muestran **sesgos pequeños y poco significativos** respecto a sus verdaderos valores poblacionales, ambos estimadores son **casi insesgados**, lo que respalda su fiabilidad en el análisis.

*Consistencia*



Comentario:

Se observa una disminución notable de la varianza de ambos estimadores a medida que aumenta el tamaño muestral, lo que confirma que muestras más grandes proporcionan estimaciones más estables. Sin embargo, la varianza de la media es significativamente mayor en tamaños de muestra pequeños y desciende de forma más lenta, lo que evidencia su sensibilidad a valores extremos presentes en los datos, estos resultados sugieren que, para este conjunto de datos con alta dispersión, la mediana es un estimador más confiable y eficiente que la media.

*Eficiencia*

|  |  |
| --- | --- |
| **Medida** | **Valor** |
| Media | 1120377604 |
| Mediana | 0.001749369 |

Comentario:

La mediana es más eficiente en esta muestra.

Sintaxis empleada con esta variable:

1) Limpieza money: clean\_numeric2(...)

2) Descriptivos: tabla\_vertical (N, N missing, Min, Q1, Median, Mean, Q3, Max, SD, IQR)

3) Gráficas guardadas en: analisis\_P3087S1\_check/plots

4) IC mean (t): t.test(x)$conf.int (o calculado manualmente)

5) Bootstrap bias y CI: uso package 'boot' (boot(...), boot.ci(...))

6) Consistencia: submuestreo comparando var(estimator) vs n

7) Eficiencia: comparar var\_bootstrap(estimator)

**Variable 4**

* ***Análisis Descriptivo***

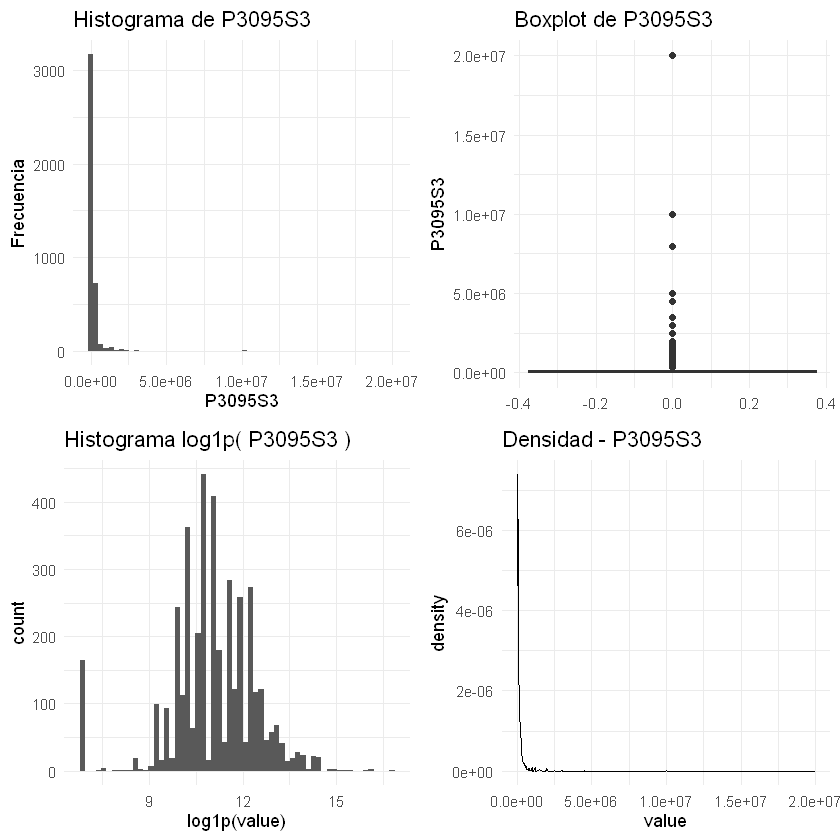
*Descriptivos Básicos*

|  |  |  |
| --- | --- | --- |
|  | Medida | Valor |
| 1 | Registros | 166341 |
| 2 | Valores faltantes | 162261 |
| 3 | Mínimo | 1000 |
| 4 | Primer cuartil | 30000 |
| 5 | Mediana | 60000 |
| 6 | Media | 159471 |
| 7 | Tercer cuartil | 150000 |
| 8 | Máximo | 20000000 |
| 9 | Desviación Estándar | 512519 |
| 10 | Rango intercuartílico | 120000 |

Comentarios:

El conjunto de datos contiene 166,341 registros, de los cuales 162,261 presentan valores faltantes, evidenciando una disponibilidad limitada de información. Los valores observados oscilan entre 1,000 y 2,000,000, con una media de 159,471 y mediana de 60,000, lo que refleja una distribución asimétrica positiva influenciada por valores extremos. El rango intercuartílico (120,000) indica una alta dispersión incluso en los datos centrales, mientras que la desviación estándar (512,519) confirma una marcada variabilidad. Se sugiere aplicar técnicas robustas o transformaciones para su análisis.

*Análisis Gráfico*



Comentarios:

La variable P3095S3 presenta una distribución altamente sesgada a la derecha con numerosos valores atípicos. Tras aplicar la transformación logarítmica (log1p), los datos muestran una distribución más equilibrada y adecuada para análisis estadísticos y modelado.

* ***Cálculo de los estimadores***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Estimador** | **Estimadores Puntuales** | | **Estimadores por Intervalo** | |
| **Analogía** | **Máxima Verosimilitud** | **Límite Inferior** | **Límite Superior** |
| (Media o Proporción) | 159,471.4 | |  | | --- | |  |  |  | | --- | | IC 95% t: [143,740.4, 175,202.4] | | 143,740.4 | |  | | --- | |  |  |  | | --- | | 175,202.4 | |
| Comentario | Muestra una amplia dispersión, lo que refleja variabilidad. | Con un 95% de confianza, la media poblacional se encuentra entre los límites estimados. | Es poco probable que la media real sea menor que este valor. | Valor máximo estimado de la media, reforzando la amplitud del rango y la posible variabilidad de los datos. |

* ***Evaluación del estimador:***

*Insesgamiento*

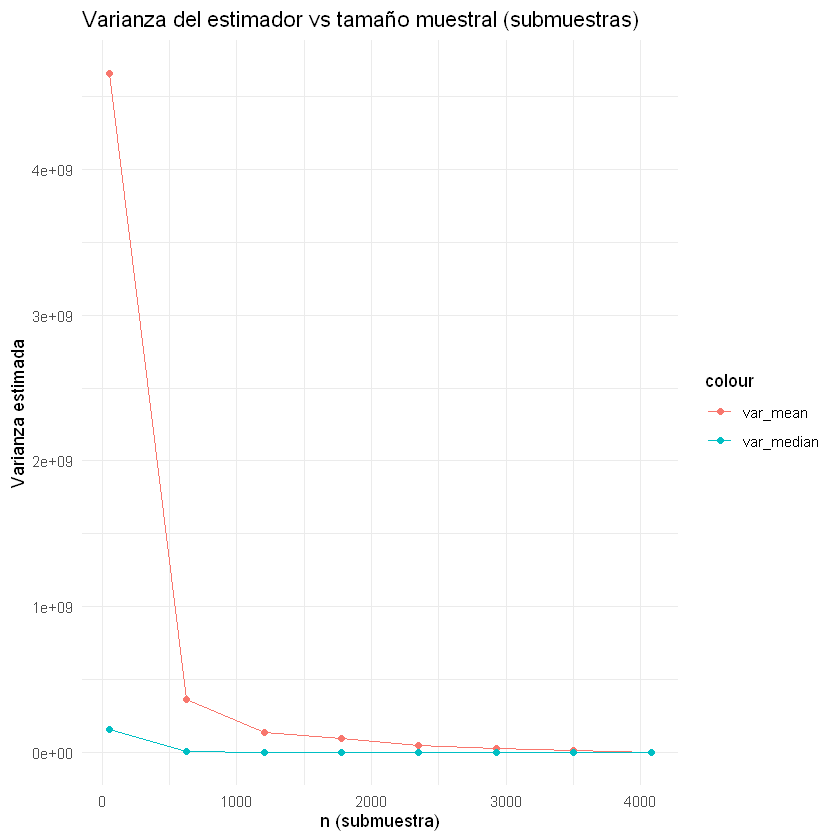
|  |  |  |
| --- | --- | --- |
| **Media** | **Mediana** | **Sesgo** |
| 159172.9 | 60000 | |  | | --- | |  |  |  | | --- | | Media: 166.16  Mediana: 0 | |

Comentario:

La media muestra sesgo pequeño relativo.

La mediana muestra sesgo pequeño relativo.

*Consistencia*



Comentario:

La varianza del estimador (mean o median) disminuye al crecer n, evidenciando consistencia.

*Eficiencia*

|  |  |
| --- | --- |
| **Medida** | **Valor** |
| Media | 64150092 |
| Mediana | 0 |

Comentario:

La mediana es más eficiente en esta muestra.

Sintaxis empleada con esta variable:

Coloque aquí la sintaxis empleada para procesar esta variable

1) Limpieza money: clean\_numeric2(...)

2) Descriptivos: tabla\_vertical (N, N missing, Min, Q1, Median, Mean, Q3, Max, SD, IQR)

3) Gráficas guardadas en: analisis\_P3095S3\_check/plots

4) IC mean (t): t.test(x)$conf.int (o calculado manualmente)

5) Bootstrap bias y CI: uso package 'boot' (boot(...), boot.ci(...))

6) Consistencia: submuestreo comparando var(estimator) vs n

7) Eficiencia: comparar var\_bootstrap(estimator)