Universidad Nacional de Colombia

Facultad de Ciencias

Departamento de Estadística

Estadística Bayesiana

Autores

Juan David Duitama Correa

jduitama@unal.edu.co

Daniel Hoyos Mateus

dhoyosm@unal.edu.co

Docente

Juan Camilo Sosa Martínez

Bogotá, marzo de 2023

Caso de estudio 1

El objetivo de este caso de estudio es modelar el conteo total de víctimas en Bogotá D. C. entre 2012 y 2022 para establecer si existen diferencias significativas por sexo respecto a delitos sexuales en menores de edad.

PARTE 1: Análisis Bayesiano en 2022

Sea $Y_k = (y_{k,1}, \ldots, y_{k,n_k})$ el vector de observaciones correspondientes al conteo total de víctimas asociados con la población k, con $k = 1$ (hombres) y $k = 2$ (mujeres). Considere modelos Gamma-Poisson de la forma

$$\begin{aligned} y_{k,i} \mid \theta_k &\mathsf{Poisson}(\theta_k)\,, \quad i = 1, \ldots, n_k\,, \\ \theta_k &\sim \mathsf{Gamma}(a_k, b_k) \end{aligned}$$

donde a_k y b_k son hiperparámetros, para $k = 1, 2$.

1. **Ajustar los modelos Gamma-Poisson de manera independiente con $a_k = b_k = 0{,}01$, para $k = 1, 2$. Hacer una visualización donde se presenten simultáneamente las distribuciones posteriores y las distribuciones previas correspondientes.**

   A continuación se presenta un resumen de la información de la víctimas de delitos sexuales en la ciudad de Bogotá del año 2022.

Año 2022	Cantidad total de víctimas (Suma)	Registros(n)
Femeninas	539	237
Masculinas	208	115
TOTAL	747	352

   La familia de distribuciones Gamma es conjugada para la distribución Poisson, por lo tanto, las distribuciones de probabilidad posteriores para las variables aleatorias θ_1 y θ_2 son las siguientes:

$$\begin{aligned} \theta_1 \mid Y &\sim \Gamma(a_1 + s_1, b_1 + n_1) \\ \theta_2 \mid Y &\sim \Gamma(a_2 + s_2, b_2 + n_2) \end{aligned}$$

   donde a_1, b_1, a_2, b_2 son los hiperparámetros de las distribuciones previas, s_1, s_2 representan los estadísticos suficientes de las distribucines muestrales, en este caso la suma de las observaciones, y finalmente, n_1 y n_2 representan el tamaño de la muestra de cada población.

   Dado que todos los hiperparámetros de las distribuciones previas son iguales a 0.01, es decir, ambas poblaciones poseen la misma distribución previa; junto al resumen presentado en la tabla anterior se tiene que las distribuciones posteriores para cada variable aleatoria es la siguiente.

$$\begin{aligned} \theta_1 \mid Y &\sim \Gamma(208{,}01, 115{,}01) \\ \theta_2 \mid Y &\sim \Gamma(539{,}01, 237{,}01) \end{aligned}$$

   La comparación entre la ditribución previa y las distribuciones posteriores se encuentra en el siguiente gráfico.

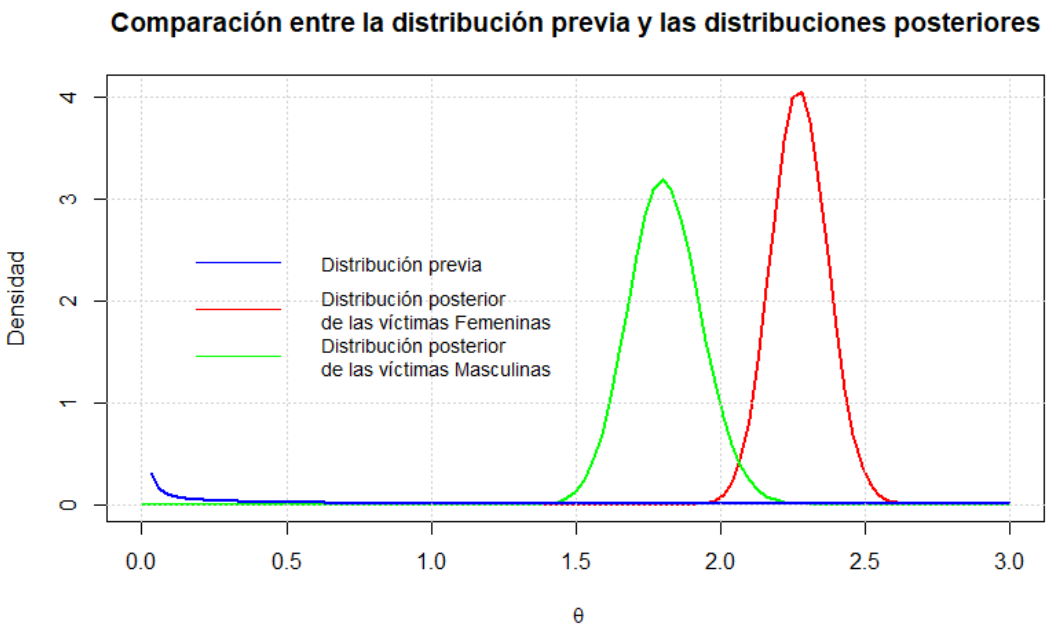


Figura 1: Distribuciones Gamma previa y posteriores de las víctimas de violencia sexual en 2022.

La distribución posterior de las víctimas femeninas es más apuntada que la posterior de las víctimas masculinas, además de que la probabilidad se encuentra concentrada en un rango menor, 2 y 2.5 aproximadamente, a comparación de la masculina que se encuentra entre 1.5 y 2.2 aproximadamente. Esto indica que la frecuencia media con la que se presentan víctimas de violencia sexual en víctimas femeninas es mayor que la de masculinas, esto puede ser observado más claramente utilizando estimaciones puntuales, en este caso, los valores esperados de cada distribución. Teniendo una estimación de 1.81 víctimas para el caso masculino y 2.27 para el femenino.

2. **Sea $\eta = (\theta_2 - \theta_1)/\theta_1$. Obtener la distribución posterior de η. Reportar la media, el coeficiente de variación, un intervalo de credibilidad al 95 %. Presentar los resultados visual y tabularmente. Interpretar los resultados obtenidos.**

   Para obtener la distribución posterior de η se generaron 100000 muestras de cada distribución posterior de θ_1 y de θ_2, para así operarlas como se indica. Esta variable aleatoria representa la diferencia entre las frecuencias medias de víctimas femeninas y masculinas, ponderada por la frecuencia masculina.

Estadística	Valor
Media	0.264
Coeficiente de variación	0.4
Intervalo de credibilidad 95 %	[0.075 , 0.480]

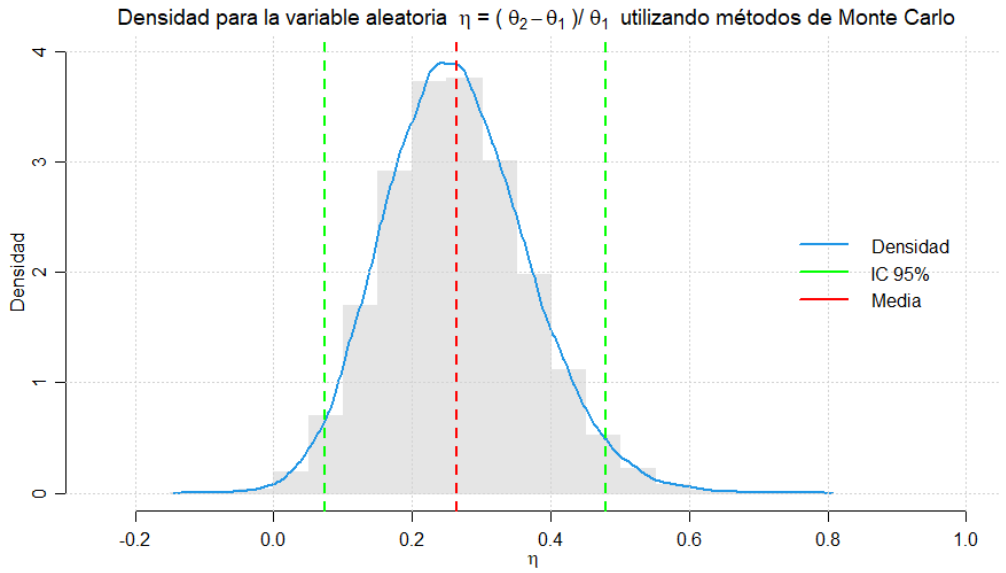


Figura 2: Densidad posterior y estadísticas de η en el año 2022.

Del gráfico se concluye que la probabilidad de que η sea mayor a cero es prácticamente 1, puesto que los η que fueron negativos representaron solo el 0.2 % de la probabilidad, esto indica que casi siempre la frecuencia media de la cantidad de víctimas femeninas es superior al de las víctimas masculinas. Se obtuvo también que la media de esta distribución es 0.264 y un coeficiente de variación de 0.4, lo cual indica que hay bastante variabilidad en los valores que puede tomar η.

3. **Llevar a cabo un análisis de sensitividad. Para ello, considere los siguientes estados de información externos al conjunto de datos:**

   - **Distr. Previa 1:** $a_k = b_k = 0{,}01$**, para** $k = 1, 2$**.**
   - **Distr. Previa 2:** $a_k = b_k = 0{,}10$**, para** $k = 1, 2$**.**
   - **Distr. Previa 3:** $a_k = b_k = 1{,}00$**, para** $k = 1, 2$**.**
   - **Distr. Previa 4:** $a_k = 1{,}00$ **y** $b_k = 1/2$**, para** $k = 1, 2$**.**
   - **Distr. Previa 5:** $a_k = 1{,}00$ **y** $b_k = 1/3$**, para** $k = 1, 2$**.**
   - **Distr. Previa 6:** $a_k = 1{,}00$ **y** $b_k = 1/4$**, para** $k = 1, 2$**.**

   En cada caso calcular la media y el coeficiente de variación a priori, y repetir el númeral anterior. Presentar los resultados visual y tabularmente. Interpretar los resultados obtenidos.

   Las densidades de las distribuciones previas y posteriores se encuentran en el siguiente gráfico.

Hiperparámetros Distribución Previa	Media a Priori	CV a priori	Media de η	CV Posterior	IC 95 %
$a_k = 0{,}01; b_k = 0{,}01$	1	10	0.263	0.4	[0.075,0.480]
$a_k = 0{,}1; b_k = 0{,}1$	1	3.162	0.263	0.4	[0.075,0.480]
$a_k = 1; b_k = 1$	1	1	0.264	0.4	[0.074,0.482]
$a_k = 1; b_k = 0{,}5$	2	1	0.265	0.4	[0.076,0.483]
$a_k = 1; b_k = 0{,}333$	3	1	0.262	0.4	[0.075,0.478]
$a_k = 1; b_k = 0{,}25$	4	1	0.261	0.4	[0.072,0.478]

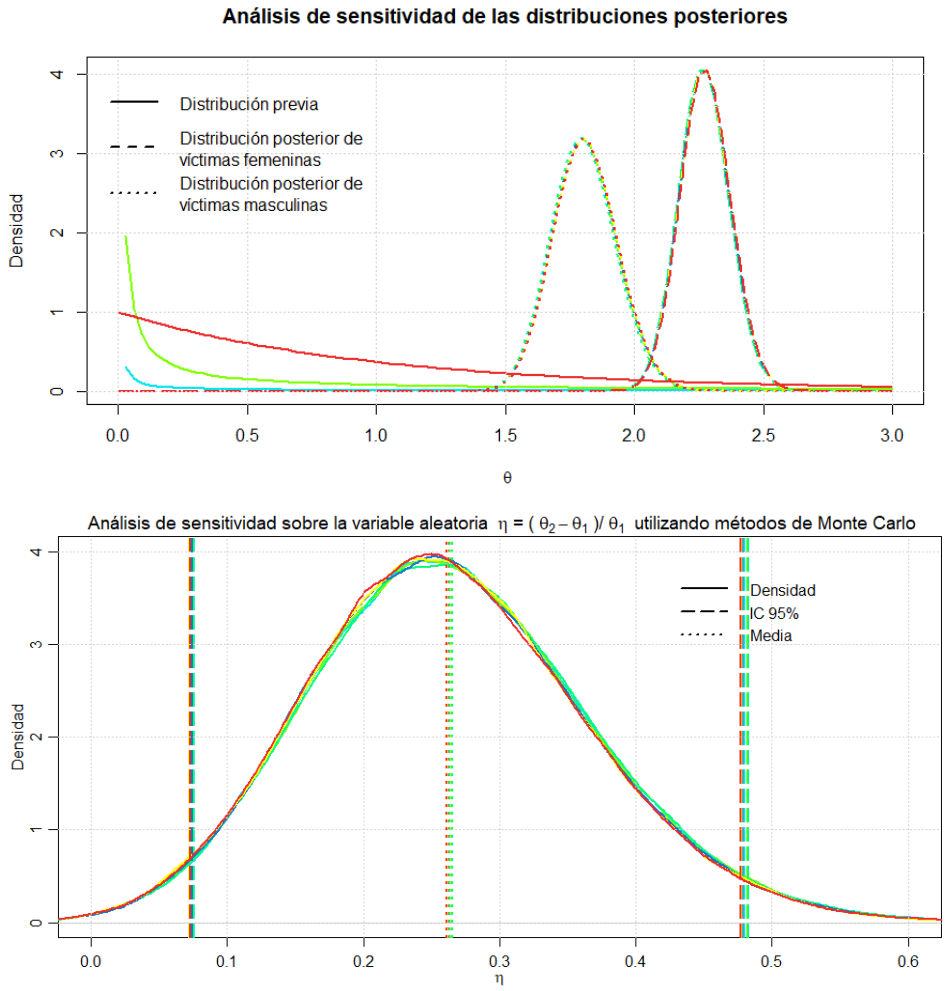


Figura 3: Densidades previas y posteriores de θ_1, θ_2 y η en el año 2022.

Tanto en la tabla como en las gráficas se aprecia el poco impacto de las distribuciones previas propuestas sobre las distribuciones posteriores de θ_1 y θ_2 como la de la variable η, las diferencias en las previas radica en el grado de incertidumbre, puesto que algunas tienen coeficientes de variación mayores, sin embargo, el reflejo de estas en las posteriores es nulo, ya que sus coeficientes de variación son prácticamente iguales, lo mismo sucede con la media posterior y los intervalos de credibilidad. Por lo tanto se puede concluir que toda la información es brindada por las muestras.

4. **En cada población, evaluar la bondad de ajuste del modelo propuesto utilizando como estadísticos de prueba la media y la desviación estándar. Presentar los resultados visual y tabularmente. Interpretar los resultados obtenidos.**

Población	Media observada	Media Modelo	ppp	Desv. observada	Desv. modelo	ppp
Femenina	2.274	2.274	0.486	2.034	1.507	0
Masculina	1.808	1.809	0.482	1.497	1.341	0.086

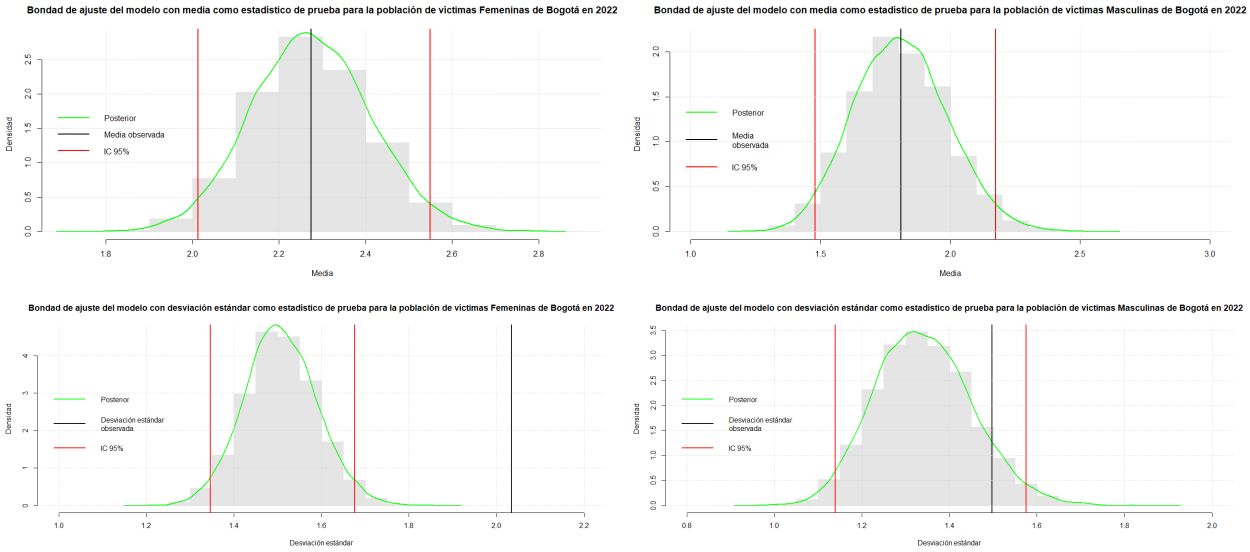


Figura 4: Bondad de ajuste para el modelo del año 2022 utilizando como estadísticos de prueba la media y la desviación estándar.

Ambos modelos se ajustan bien a la realidad usando como estadística de prueba la media, esto confirmado con el valor PPP(Posterior Predictive P-value), ya que tienen valores cercanos a 0.5, sin embargo, utilizando como estadística de prueba la desviación estándar se tienen resultados desafortunados, puesto que en ambos casos el valor observado de la desviación estándar encuentra alejado de las partes de la distribución con mayor masa de probabilidad, nuevamente confirmado con el valor PPP, en ambos casos cercano a 0.

PARTE 2: Análisis frecuentista en 2022

1. **Repetir el númeral 2. de la PARTE 1 usando *Bootstrap* paramétrico**

   Para encontrar la distribución aproximada de $\eta = \frac{\theta_2 - \theta_1}{\theta_1}$ se generaron 2000 remuestreos de cada población de víctimas masculinas y femeninas, en cada remuestreo se calculó una estimación de η dada por $\hat{\eta} = \frac{\hat{\theta}_{2MLE} - \hat{\theta}_{1MLE}}{\hat{\theta}_{1MLE}}$, donde $\hat{\theta}_{kMLE}$ representa el estimador máximo verosímil de la distribución muestral, en este caso, el estimador máximo verosímil para una distribución Poisson es la media muestral.

   A continuación se presenta un gráfico de los resultados obtenidos.

Estadística	Valor
Media	0.258
Coeficiente de variación	0.461
Intervalo de confianza 95 %	[0.042 , 0.523]

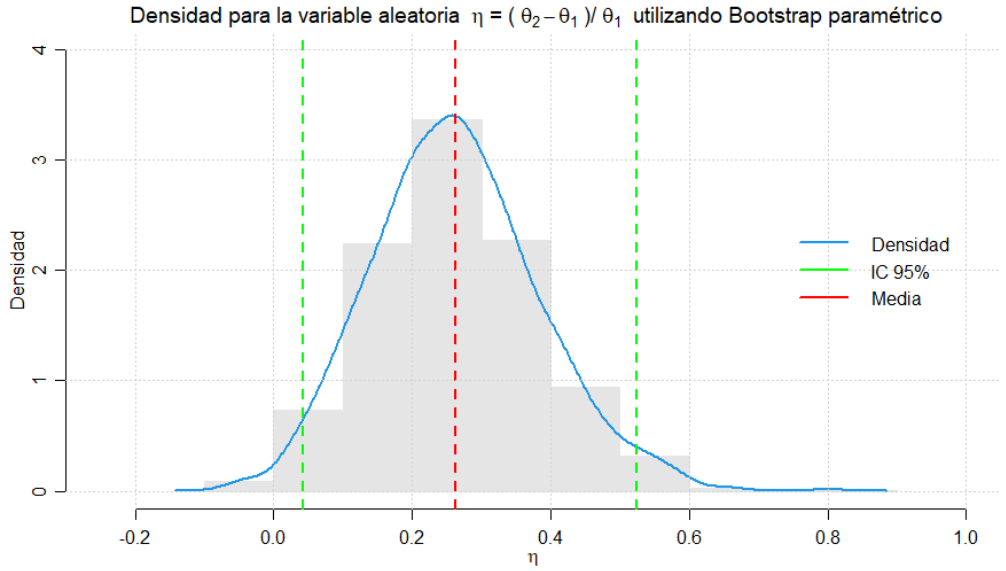


Figura 5: Densidad de la variable η del año 2022.

En la imagen podemos observar, que la media de la variable aleatoria η está alrededor de 0.25, y, al igual que en el modelamiento bayesiano, la mayoría de la masa de probabilidad se encuentra en los η mayores a 0. El modelamiento frecuentista arrojó una densidad menos apuntada que la bayesiana, esto reflejado en los intervalos de confianza y credibilidad, ya que el intervalo frecuentista es más amplio que el bayesiano. Así, nuevamente se afirma que la frecuencia media de cantidad de víctimas femeninas de violencia sexual en el año 2022 es mayor que la de víctimas masculinas.

2. Simular 100,000 muestras aleatorias de poblaciones Poisson bajo los siguientes escenarios:

- Escenario 1: $n_1 = 10$, $n_2 = 10$, $\theta_1 = \bar{y}_1$, y $\theta_2 = \bar{y}_2$.
- Escenario 2: $n_1 = 20$, $n_2 = 20$, $\theta_1 = \bar{y}_1$, y $\theta_2 = \bar{y}_2$.
- Escenario 3: $n_1 = 50$, $n_2 = 50$, $\theta_1 = \bar{y}_1$, y $\theta_2 = \bar{y}_2$.
- Escenario 4: $n_1 = 100$, $n_2 = 100$, $\theta_1 = \bar{y}_1$, y $\theta_2 = \bar{y}_2$.

donde $\bar{y}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} y_{k,i}$ es la media muestral observada de la población k, para $k = 1, 2$. En cada escenario el valor verdadero de η es $\eta = (\bar{y}_2 - \bar{y}_1)/\bar{y}_1$.

Usando cada muestra, ajustar el modelo de manera tanto Bayesiana (usando la primera configuración previa) como frecuentista (usando *Bootstrap* paramétrico), y en cada caso calcular la proporción de veces que el intervalo de credibilidad/confianza al 95 % contiene el valor verdadero de η. Reportar los resultados tabularmente. Interpretar los resultados obtenidos.

El valor del parámetro η es 0.257, a continuación se presentan los porcentajes obtenidos después de realizar las respectivas simulaciones.

Escenario	Tamaño pobalción	Modelo Bayesiano	Modelo Frecuentista
1	10	94.44 %	91.68 %
2	20	94.61 %	93.4 %
3	50	94.75 %	94.35 %
4	100	94.8 %	94.5 %

En los resultados podemos observar que el porcentaje de veces que el intervalo de credibilidad del modelo bayesiano contuvo al parámetro η estuvo siempre por encima del 94 % independientemente del tamaño de la población. Estos

porcentajes en el modelo frecuentista fueron aumentando en cuanto el tamaño de la población también aumenta, en todos los escenarios los porcentajes fueron menores que los del modelo bayesiano. Este ejercicio de simulación remarca la importancia de los tamaños de las muestras que requieren los modelos frecuentistas, algo que no sucede en los modelos bayesianos.

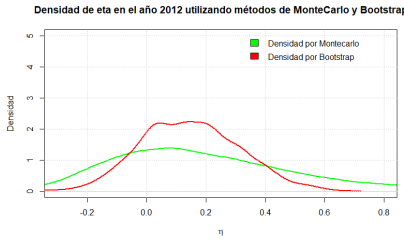
PARTE 3: Análisis Bayesiano y frecuentista en 2012-2022

Para cada año de 2012 a 2022 (inclusive), ajustar el modelo de manera tanto Bayesiana (usando la primera configuración previa) como frecuentista (usando *Bootstrap* paramétrico), y obtener tanto una estimación puntual como intervalos de credibilidad/confianza al 95 % y 99 % para η. Presentar los resultados visualmente. Interpretar los resultados obtenidos.
A continuación se presentan las gráficas de los años 2012 al 2022 donde aparecen las desnsidades utilizando cada uno de los métodos, en cada imagen se pueden observar los dos comportamientos juntos, para permitir hacr una comparación visual.

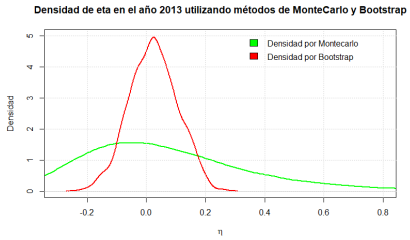
Estadísticos de η a través de los años

Año	Media(Bayes)	Media(Boots.)	IC 95 %(Bayes)	IC 95 %(Boots.)	IC 99 %(Bayes)	IC 99 %(Boots.)
2012	0.209	0.147	[-0.29,1.01]	[-0.15,0.46]	[-0.39,1.45]	[-0.24,0.56]
2013	0.074	0.024	[-0.36,0.77]	[-0.14,0.18]	[-0.45,1.12]	[-0.18,0.2]
2014	0.162	0.12	[-0.27,0.81]	[-0.09,0.39]	[-0.35,1.12]	[-0.17,0.47]
2015	0.070	0.07	[-0.21,0.44]	[-0.27,0.5]	[-0.28,0.59]	[-0.35,0.66]
2016	0.407	0.38	[0.02,0.92]	[0.15,0.63]	[-0.07,1.15]	[0.09,0.7]
2017	0.241	0.23	[-0.01,0.54]	[0.07,0.41]	[-0.07,0.66]	[0.01,0.46]
2018	0.291	0.29	[0.1,0.51]	[0.09,0.49]	[0.05,0.59]	[0.04,0.59]
2019	0.223	0.22	[0.04,0.42]	[0.01,0.47]	[-0.002,0.5]	[-0.03,0.58]
2020	0.049	0.05	[-0.1,0.22]	[-0.12,0.25]	[-0.14,0.29]	[-0.17,0.33]
2021	0.066	0.06	[-0.14,0.31]	[-0.09,0.22]	[-0.2,0.404]	[-0.14,0.28]
2022	0.264	0.26	[0.07,0.48]	[0.04,0.52]	[0.02,0.56]	[-0.02,0.61]

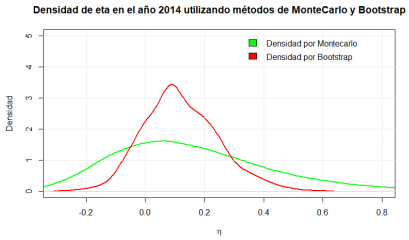
En general, se obtuvieron resultados similares comparando medias e intervalos de credibilidad y confianza, no obstante, en los primeros tres años estudiados (2012, 2013 y 2014) se observan diferencias en estas estimaciones, esto se refleja en las densidades resultantes de cada procedimiento; las realizadas por Bootstrap son más apuntadas a comparación de las realizadas por el método bayesiano, las cuales son platicurticas. En los siguientes años los resultados del estudio de η fueron similares, reflejado en las densidades resultantes, especialmente en los años posteriores a 2018.



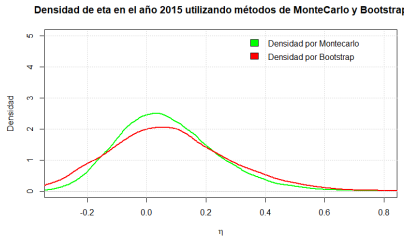
(a) Densidad de η en el año 2012.



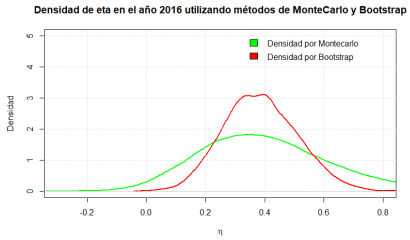
(b) Densidad de η en el año 2013.



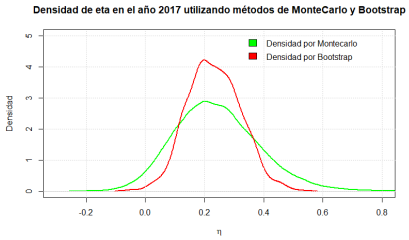
(c) Densidad de η en el año 2014.



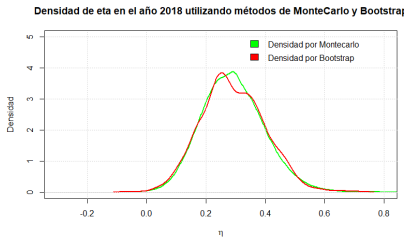
(d) Densidad de η en el año 2015.



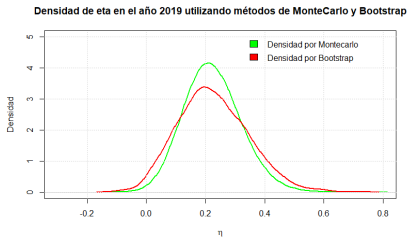
(e) Densidad de η en el año 2016.



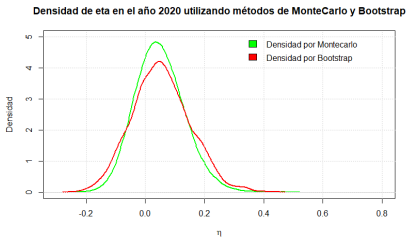
(f) Densidad de η en el año 2017.



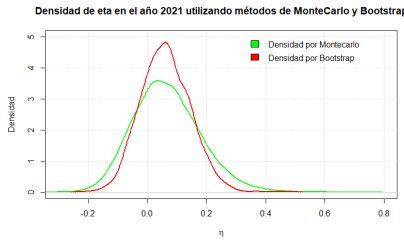
(g) Densidad de η en el año 2018.



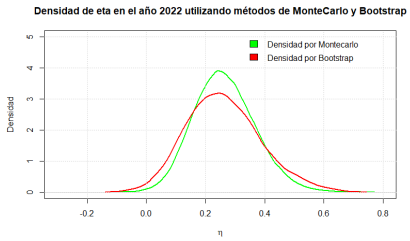
(h) Densidad de η en el año 2019.



(i) Densidad de η en el año 2020.



(j) Densidad de η en el año 2021.



(k) Densidad de η en el año 2022.

Anexo

1. **Familia Conjugada Gamma-Poisson**

   - **Distribución muestral:**
     Sea $\mathbf{y} = (y_1, y_2, ..., y_n)$ una muestra aleatoria de tal forma que $y_i|\theta \overset{iid}{\sim} Poisson(\theta)$, entonces la distribución de $\mathbf{y}$ es la siguiente

$$\begin{aligned} p(\mathbf{y}|\theta) &= \prod_{i=1}^{n} \frac{\theta^{y_i} e^{-\theta}}{y_i!}; \theta > 0 \\ &= \frac{\theta^{\sum_{i=1}^{n} y_i} e^{-n\theta}}{\prod_{i=1}^{n} y_i!} \\ &\propto \theta^{\sum_{i=1}^{n} y_i} e^{-n\theta} \end{aligned}$$

   - **Distribución previa:**
     Se propone la distribución Gamma para θ, es decir, $\theta \sim Gamma(\alpha, \beta)$ con α y β positivos, entonces

$$\begin{aligned} p(\theta) &= \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}; \theta > 0 \\ &\propto \theta^{\alpha-1} e^{-\beta\theta} \end{aligned}$$

   - **Distribución posterior:**
     Finalmente se tiene que la distribución posterior de θ es la siguiente

$$\begin{aligned} p(\theta|\mathbf{y}) &= p(\mathbf{y}|\theta)p(\theta) \\ &\propto \theta^{\sum_{i=1}^{n} y_i} e^{-n\theta} \theta^{\alpha-1} e^{-\beta\theta} \\ &\propto \theta^{\sum_{i=1}^{n} y_i + \alpha - 1} e^{-(n+\beta)\theta} \end{aligned}$$

     Así entonces $\theta|\mathbf{y} \sim Gamma(\sum_{i=1}^{n} y_i + \alpha, n + \beta)$

2. **Estimador Máximo Verosímil de una distribución Poisson**
   El estimador máximo verosímil de una una muestra aleatoria es aquel que maximiza la función de verosimilitud o log-verosimilitud. Para el caso de una muestra aleatoria Poisson se tiene que la función de verosimilitud es la siguiente.

$$\begin{aligned} L(\theta|\mathbf{y}) &= \prod_{i=1}^{n} \frac{\theta^{y_i} e^{-\theta}}{y_i!} \\ &= \frac{\theta^{\sum_{i=1}^{n} y_i} e^{-n\theta}}{\prod_{i=1}^{n} y_i!} \end{aligned}$$

   Para facilitar la maximización se encuentra la log-verosimilitud

$$
\begin{aligned}
ln(L(\theta|\mathbf{y})) &= ln(\frac{\theta^{\sum_{i=1}^{n} y_i} e^{-n\theta}}{\prod_{i=1}^{n} y_i!}) \\
&= ln(\theta^{\sum_{i=1}^{n} y_i}) + ln(e^{-n\theta}) - ln(\prod_{i=1}^{n} y_i!) \\
&= \sum_{i=1}^{n} y_i ln(\theta) - n\theta - \sum_{i=1}^{n} ln(y_i!)
\end{aligned}
$$

Derivando respecto a θ

$$
\frac{dln(L(\theta|\mathbf{y}))}{d\theta}(\hat{\theta}) = \frac{\sum_{i=1}^{n} y_i}{\hat{\theta}} - n - 0
$$

Después de igualar a 0 se tiene que $\hat{\theta} = \frac{\sum_{i=1}^{n} y_i}{n} = \overline{y}$.

Finalmente $\overline{y}$ es un máximo puesto que

$$
\begin{aligned}
\frac{d^2 ln(L(\theta|\mathbf{y}))}{d\theta^2}(\overline{y}) &= -\frac{\sum_{i=1}^{n} y_i}{\overline{y}^2} \\
&= -\frac{n}{\overline{y}}
\end{aligned}
$$

Cantidad siempre negativa, ya que y_i tiene recorrido en los naturales. Así, el estimador máximo verosímil de una muestra aleatoria Poisson es la media muestral.

$$
\hat{\theta}_{MLE} = \overline{y}
$$

3. **Invarianza del estimador máximo verosímil**
   Si $\hat{\theta}$ es el estimador máximo verosímil para θ y $\tau()$ es una función uno a uno, el estimador máximo verosímil de $\tau(\theta)$ es $\tau(\hat{\theta})$.

   Demostración disponible en: Mayorga Álvarez, J. (2004). Inferencia estadística. Universidad Nacional de Colombia. Página 126.