UNIVERSIDAD
NACIONAL
DE COLOMBIA

Universidad Nacional de Colombia

Facultad de ciencias

Estadística Bayesiana

Caso de estudio 1: Conteo de Victimas

Autores:

Joan Fernando Lamprea Huertas
jolampreah@unal.edu.co
José Armando Valdés Domínguez
jvaldesd@unal.edu.co

Docente:

Juan Camilo Sosa Martinez

Marzo 2023

Introducción

El Sistema Penal Oral Acusatorio (SPOA) es "una Dirección Nacional creada mediante el decreto 016 de 2014 y reglamentada por la resolución 0-0555 del 02 de abril de 2014 suscrita por el Fiscal General de la Nación concebida con el propósito principal de fortalecer el funcionamiento integral del sistema penal acusatorio y articular a la Fiscalía General de la Nación con las entidades que tengan incidencia en su labor misional"(Fiscalía General de la Nación s.f.).

Entre las diversas funciones que se destacan en este sistema, se considera la clasificación y organización de las entradas de noticias criminales por delito, reguladas en la Ley 906 de 2004 y Ley 1098 de 2006. La Fiscalía General de la Nación, con base estos registros suministrados por el SPOA, recopila la información del Conteo de Víctimas desde hechos ocurridos en 2010, los cuales se presentan a corte del último día del mes anterior. Este conjunto de datos consta de 3651193 registros que muestran el total de víctimas por delito para las cuales se cumple el cruce de las 25 variables relacionadas con la caracterización tanto de las víctimas como de los victimarios (Datos Abiertos 2019).

En este estudio se considera el total de víctimas de delitos sexuales; hombres y mujeres, nacidos en Colombia, menores de edad; cuyo hecho sea activo y realizado, denunciado e ingresado a la Fiscalía en el año 2022; ocurrido en la ciudad de Bogotá D.C. Esto lleva a un total de 352 víctimas, 115 hombres y 237 mujeres, respectivamente.

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

Análisis Bayesiano en 2022

Sea $\boldsymbol{y}_k = (y_{k,1}, \ldots, y_{k,n_k})$ 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

$$y_{k,i} \mid \theta_k \quad \overset{iid}{\sim} \mathsf{Poisson}(\theta_k) \quad i = 1, \ldots, n_k\,, \tag{1}$$

$$\theta_k \sim \mathsf{Gamma}(a_k, b_k) \tag{2}$$

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.
Para ajustar los modelos Gamma-Poisson para las poblaciones se consideran los siguientes hiperparámetros:

$$a_k = b_k = 0{,}01 \qquad \text{con } k = \{1, 2\}$$

donde con $k = 1$ representa la población masculina y $k = 2$ la femenina. Esto nos lleva a considerar la misma distribución previa tanto para hombres como para mujeres.

Haciendo uso del teorema de factorización de Fisher-Neyman (Estadística Bayesiana 2023a) se consideran los estadísticos suficientes, que son aquellos en los que se condensa la información de los respectivos datos, como siguen:

$$s_1 = \sum_{i=1}^{n_1} y_{1,i} = 208 \qquad s_2 = \sum_{i=1}^{n_2} y_{2,i} = 539$$

Estos valores permiten encontrar los hiperparámetros de la distribución posterior dado un modelo Gamma-poisson (Estadística Bayesiana 2023b), tiene la siguiente forma:

$$p(\theta|\boldsymbol{y}_k) \sim Gamma(a_k + s_k, b_k + n_k) \tag{3}$$

Estableciendo $a_{pk} = a_k + s_k$ y $b_{pk} = b_k + n_k$, se puede ver que $a_{p1} = 208{,}01$ y $b_{p1} = 115{,}01$ son los hiperparámetros de la población masculina, y $a_{p2} = 539{,}01$ y $b_{p2} = 237{,}01$ los de la población femenina.
A continuación se muestra la información resumida en la Figura 1 y el Cuadro 1.

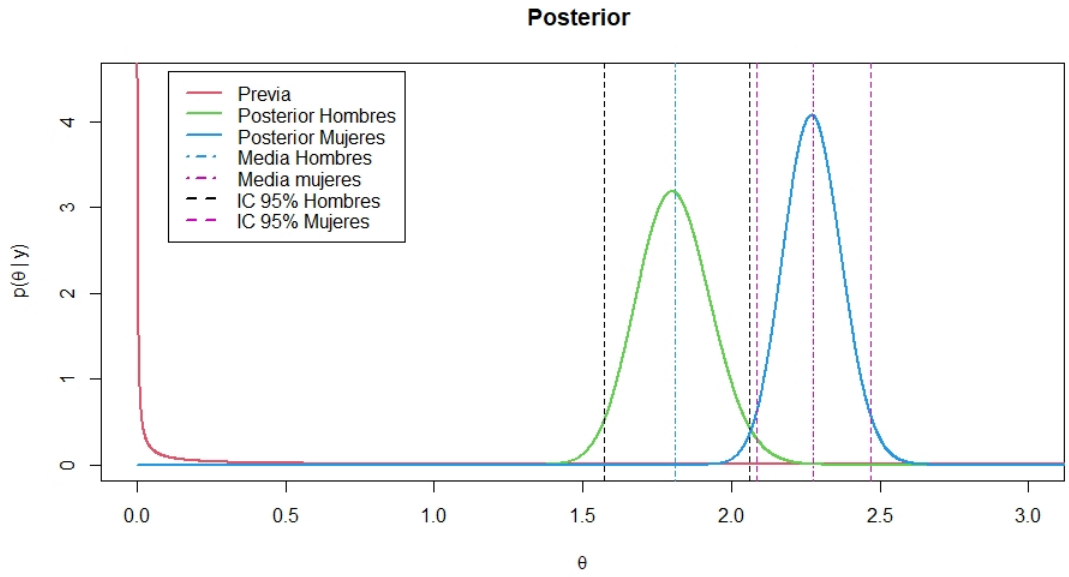

Figura 1: Distribución posterior para poblaciones masculina y femenina junto con la distribución previa.

	Media Posterior	Percentil 2.5 %	Percentil 97.5 %
Hombres	1.809	1.571	**2,063**
Mujeres	2.274	**2,086**	2.470

Cuadro 1: Distribución posterior por género con sus intervalos de credibilidad del 95 %.

Se puede observar en la Figura 1 que la ocurrencia media de víctimas de delitos sexuales es mayor en la población femenina que en la masculina. Además, en el Cuadro 1, se evidencia que los intervalos de credibilidad no llegan a la intersección. Por tanto, se puede asegurar, con una credibilidad del 95 %; que, para el año 2022, la población colombiana femenina, menor de 18 años, en la ciudad de Bogotá, es más propensa a delitos sexuales que la masculina.

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 (máximo 100 palabras). **Nota:** usar métodos de Monte Carlo con una cantidad de muestras adecuada.

   Se define η como la razón entre la diferencia relativa de la ocurrencia media de delitos sexuales entre mujeres y hombre con la ocurrencia media de delitos sexuales en hombres. Por tanto, para interpretar los valores de η se considera la siguiente tabla:
   Empleando un método de Monte Carlo, con 20000 simulaciones, se obtiene la distribución

$\eta > 0$	La ocurrencia de delitos sexuales en mujeres es mayor que en hombres.
$\eta < 0$	La ocurrencia de delitos sexuales en hombres es mayor que en mujeres.
$\eta = 0$	No hay diferencias significativas en la ocurrencia de delitos sexuales entre hombres y mujeres.

Cuadro 2: Interpretación de η

posterior de η, la cual se encuentra resumida en la Figura 2 y Cuadro 3.

	Est. Puntual	Coef. Variación	Percentil 2.5 %	Percentil 97.5 %
η	0.264	0.393	0.077	0.481

Cuadro 3: Estimación puntual de η con su intervalo de credibilidad del 95 %.

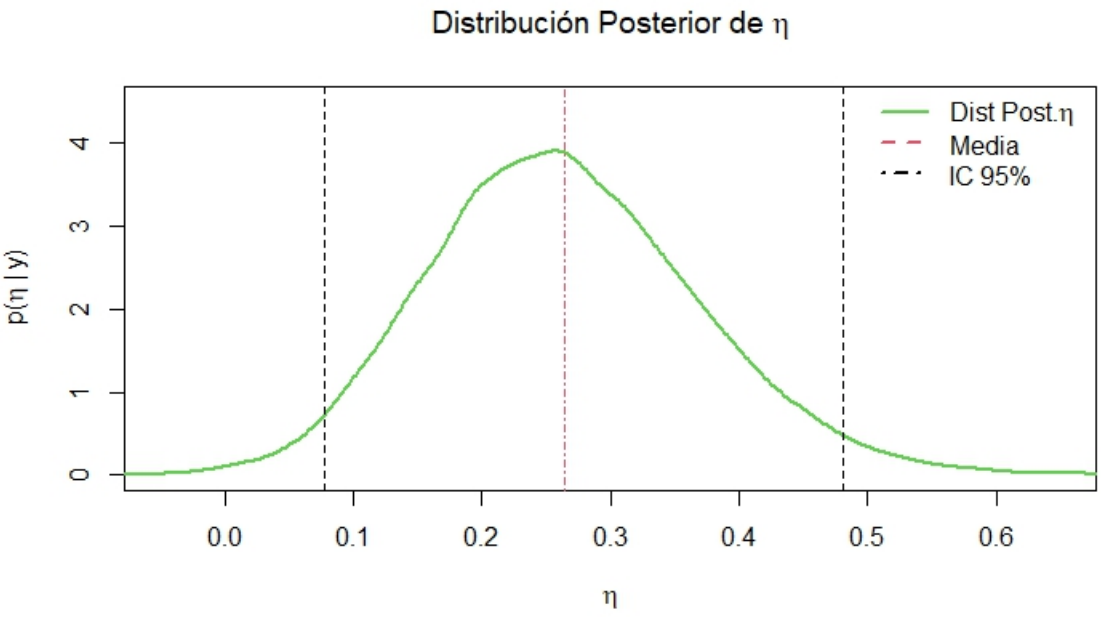

Figura 2: Distribución posterior de η y su media con intervalos de credibilidad del 95 %.

A partir de la información resumida en el Cuadro 2, se pone en evidencia que, con una credibilidad del 95 %, las mujeres sufren un $26{,}4\,\%$ más delitos sexuales que los hombres; sin embargo, el coeficiente de variación para η es del $39{,}3\,\%$, indicando una variabilidad alta en comparación con la media estimada; mostrando que la estimación de η es relativamente inexacta o imprecisa.

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 (máximo 100 palabras). **Nota:** usar un solo panel para la visualización.

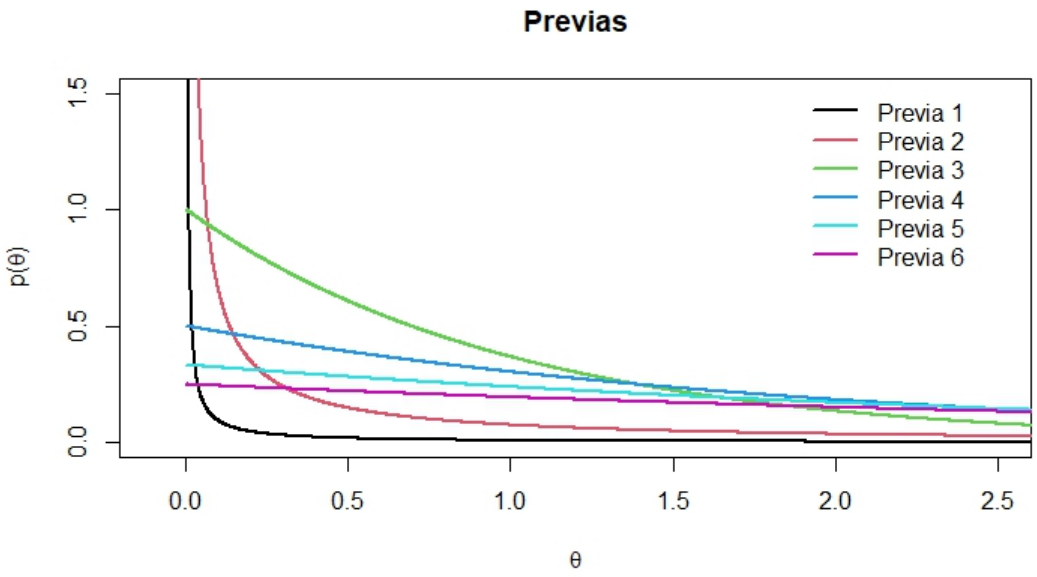

Figura 3: Análisis de Sensitividad de 6 distribuciones previas.

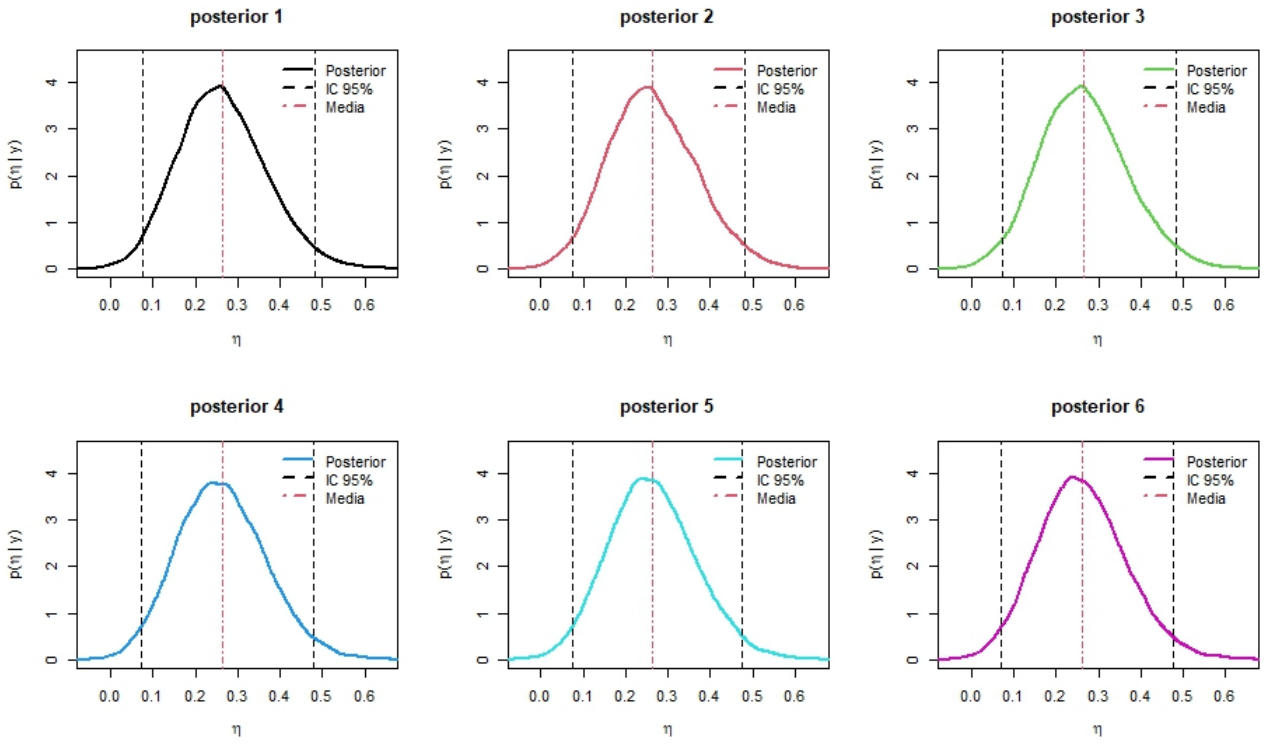

Figura 4: Gráficas de 6 distribuciones previas con diferentes estados de información.

	Previa 1	Previa 2	Previa 3	Previa 4	Previa 5	Previa 6
Media	1	1	1	2	3	4
Coef. Variación	10	3.162	1	1	1	1

Cuadro 4: Media y coeficiente de variación a priori de 6 distribuciones previas.

	Media	Coef. Variación	Percentil 2.5 %	Percentil 97.5 %
Previa 1	0.264	0.393	0.077	0.481
Previa 2	0.264	0.394	0.074	0.481
Previa 3	0.266	0.392	0.073	0.484
Previa 4	0.263	0.395	0.073	0.479
Previa 5	0.263	0.393	0.073	0.475
Previa 6	0.262	0.395	0.072	0.477

Cuadro 5: Distribución posterior asociada a los 6 estados de información externa con sus intervalos de credibilidad del 95 %.

Se puede mencionar que la distribución previa tiene un comportamiento diferente con respecto a la información externa que se tiene de los parámetros. Sin embargo, al analizar la distribución posterior de cada estado de información externo, en la Figura 4 y el Cuadro 5, se puede observar que la distribución no resulta sensible a la información previa que se disponga. (Hoff, P 2009). Por lo que se puede concluir el conocimiento externo no afectan la

distribución posterior ni las estimaciones de los parámetros del modelo.

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 (máximo 100 palabras).

   Nota: calcular los valores p predictivos posteriores.

   A través de la técnica de Monte Carlo se puede evaluar la bondad de ajuste interna de un modelo estadístico, basada en la comparación del valor observado de un estadístico de prueba con su distribución predictiva posterior simulada. Se verifica que el valor ppp (posterior predictive p-value) se encuentre oscilando entre 0.5 para considerar que el modelo tiene una buena capacidad de ajuste interna; de lo contrario, se dice que no la tiene. Para este estudio, se emplean la media y la desviación estándar.

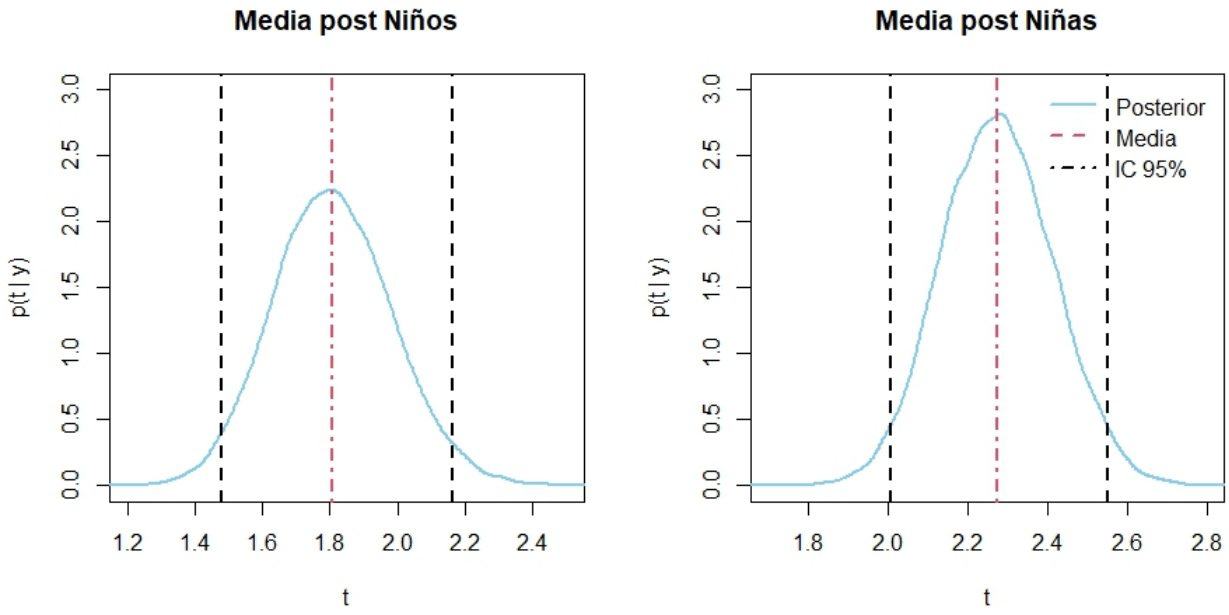

Figura 5: Media observada de distribución posterior de η.

	Media	Coef. Variación	Percentil 2.5 %	Percentil 97.5 %	***ppp***
Masculino	1.809	0.828	1.478	2.165	**0,480**
Femenino	2.274	0.895	2.008	2.553	**0,495**

Cuadro 6: Valor p predictivo posterior para la media con su intervalo de credibilidad del 95 %.

Como se observa en la Figura 5 y Cuadro 6, las medias tienen un valor ppp cercano a 0.5,

entonces se puede decir que el modelo captura adecuadamente el conteo medio de ocurrencia de delitos sexuales en ambas poblaciones.

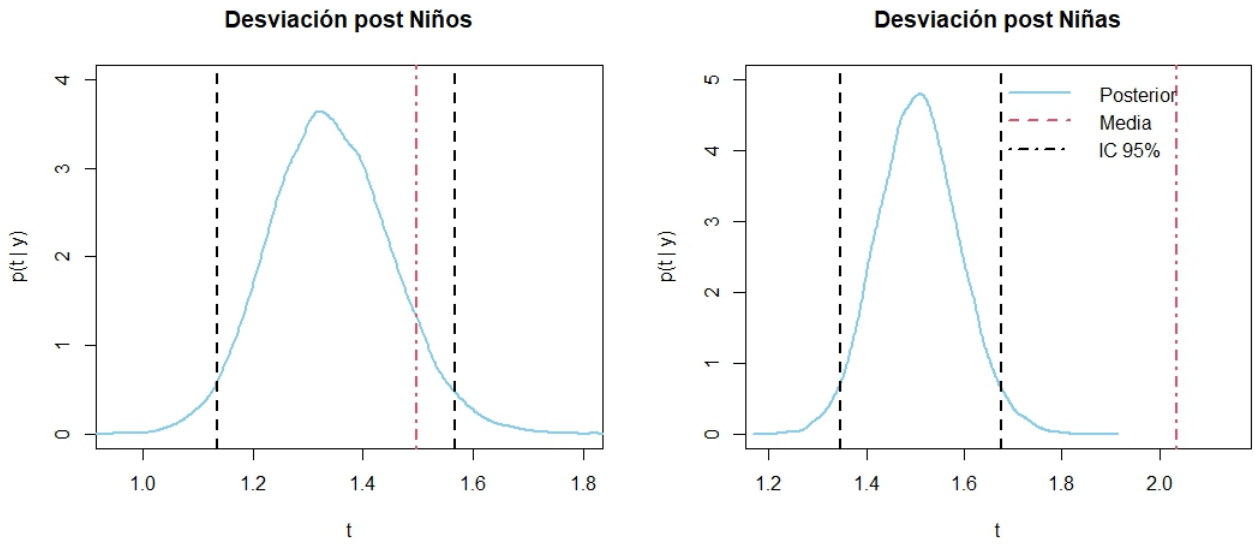

Figura 6: Desviación estándar observada de distribución posterior de η.

	Desviación estándar	Percentil 2.5 %	Percentil 97.5 %	***ppp***
Masculino	1.498	1.135	1.567	**0,082**
Femenino	2.035	1.347	1.676	**0,000**

Cuadro 7: Valor p predictivo posterior para la desviación estándar con su intervalo de credibilidad del 95 %.

Sin embargo, en la Figura 6 y Cuadro 7, se observa que la desviación estándar tiene un valor ppp muy cercano a 0, tanto para la población masculina como la femenina. Por tanto, se determina que el modelo no captura adecuadamente la dispersión de los casos de delitos sexuales en ambas poblaciones; en cambio, las subestima.

Análisis Frecuentista en 2022

1. Repetir el numeral 2. de la PARTE 1 usando Bootstrap paramétrico

   Nota: usar una cantidad de remuestras adecuada.

   Para realizar el bootstrap paramétrico se considera 1900 remuestras (s.n. s.f.) y se parte de la suposición de que los datos provienen de una distribución Poisson, cuyo estimador de máxima verosimilitud es la media muestral. Por tanto, por el teorema de Invarianza del MLE (Arrieta 2021), se calcula el estimador η con base en la media muestral de la población masculina y

femenina. En la Figura 7 y el Cuadro 8, se puede observar el comportamiento de la distribución posterior de η.

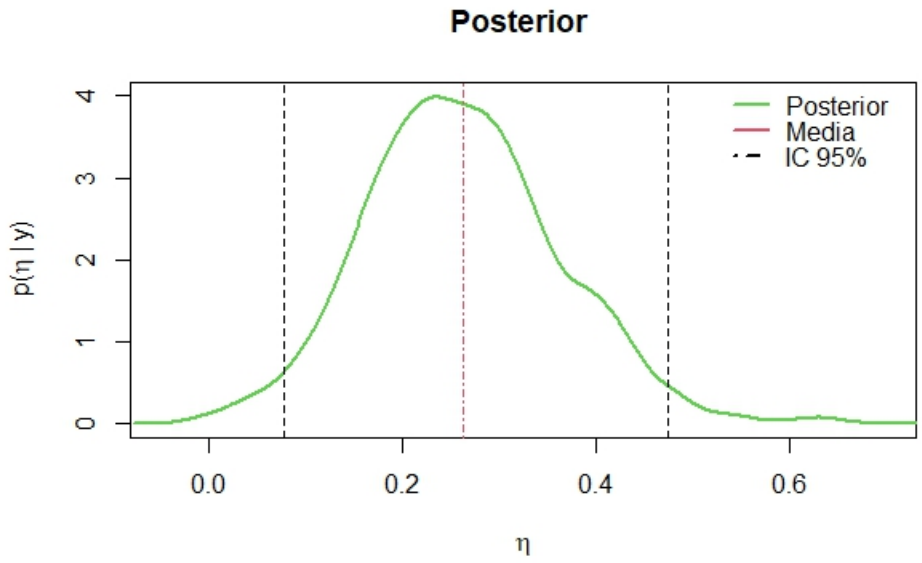

Figura 7: Distribución posterior de η y su valor observado con sus intervalos de confianza del 95 %

	Est. Puntual	Coef. Variación	Percentil 2.5 %	Percentil 97.5 %
η	0.263	0.385	0.078	0.474

Cuadro 8: Estimación puntual con su respectivo intervalo de confianza del 95 %.

Con base en la información resumida en el Cuadro 8, se observa que, con una confianza del 95 %, las mujeres sufren un 26,3 % más delitos sexuales que los hombres (resultado similar al caso bayesiano); sin embargo, el coeficiente de variación para η es del 39,3 %, presentado el mismo inconveniente que en el caso bayesiano.

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

   Escenario 1: $n_1 = n_2 = 10$, $\theta_1 = \bar{y}_1$ y $\theta_2 = \bar{y}_2$

   Escenario 2: $n_1 = n_2 = 20$, $\theta_1 = \bar{y}_1$ y $\theta_2 = \bar{y}_2$

   Escenario 3: $n_1 = n_2 = 50$, $\theta_1 = \bar{y}_1$ y $\theta_2 = \bar{y}_2$

   Escenario 4: $n_1 = n_2 = 100$, $\theta_1 = \bar{y}_1$ y $\theta_2 = \bar{y}_2$

   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 (máximo 100 palabras).

Considerando a $a_k = b_k = 0{,}01$, con $k = 1, 2$, los hiperparámetros del modelo gamma-poisson y 20000 simulaciones para Monte Carlo en el enfoque bayesiano, y 1900 remuestras para el bootstrap paramétrico, para el enfoque frecuentista, se verifica la proporción de intervalos de credibilidad/confianza que contiene al verdadero valor del parámetro η, en los diferentes escenarios a considerar.

	Prop. Bayesiano	Prop. Frecuentista
Escenario 1	100 %	100 %
Escenario 2	94.80 %	94.63 %
Escenario 3	94.82 %	94.64 %
Escenario 4	94.97 %	94.81 %

Cuadro 9: Proporción de intervalos de credibilidad/confianza que contiene al parámetro por cada uno de los escenarios.

Como se puede observar en el Cuadro 9, la proporción de intervalos de credibilidad que contienen al verdadero valor del parámetro no difiere significativamente con respecto a la proporción de intervalos de confianza. Sin embargo, es importante resaltar que, para el primer escenario, el tamaño de las muestras resulta ser bastante pequeño, lo que permite que el tamaño de los intervalos de credibilidad/confianza sean bastante grandes. Además, el costo computacional para hallar la proporción de intervalos de confianza es mayor que la de los intervalos de credibilidad, haciendo más eficiente el modelo bayesiano.

Análisis Mixto en 2012 hasta 2022

1. 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 (máximo 100 palabras).

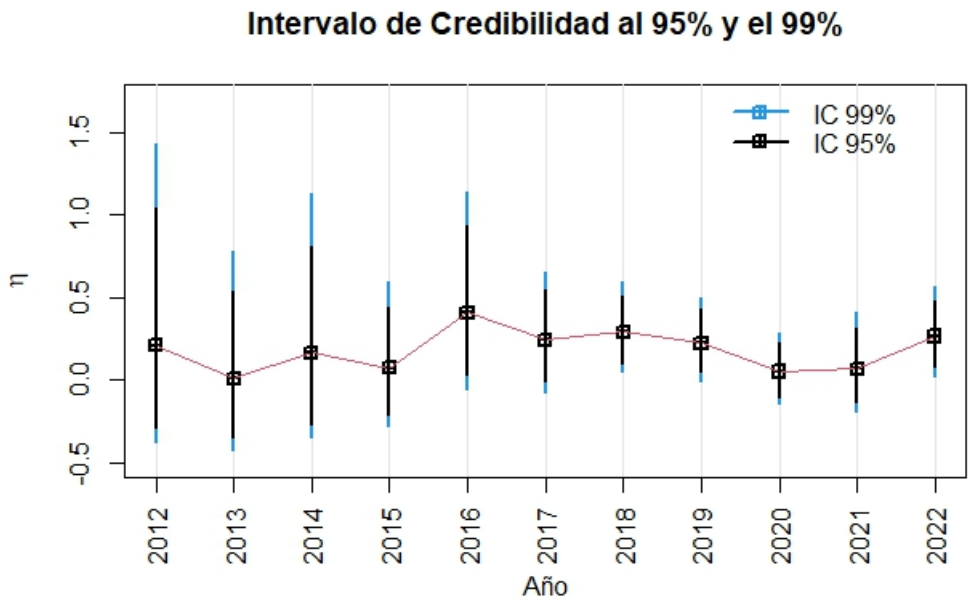

Figura 8: Estimaciones puntuales con sus intervalos de credibilidad del 95 % y 99 %

Año	Tamaño de Muestra		Media	Coef. Variación	IC 95 %		IC 99 %	
	Niños	Niñas			2.5 %	97.5 %	0.5 %	99.5 %
2012	13	64	0.207	1.634	-0.294	1.038	-0.382	1.433
2013	18	85	0.008	26.627	-0.347	0.536	-0.422	0.775
2014	18	66	0.163	1.699	-0.270	0.811	-0.353	1.128
2015	30	94	0.070	2.370	-0.214	0.442	-0.282	0.591
2016	41	132	0.408	0.561	0.028	0.933	-0.060	1.142
2017	77	213	0.240	0.589	-0.011	0.545	-0.076	0.657
2018	116	301	0.291	0.358	0.102	0.511	0.049	0.590
2019	122	295	0.224	0.434	0.045	0.426	-0.004	0.497
2020	111	291	0.050	1.677	-0.104	0.224	-0.145	0.283
2021	75	265	0.067	1.712	-0.135	0.313	-0.189	0.405
2022	115	237	0.264	0.393	0.077	0.481	0.019	0.561

Cuadro 10: Media observada de η para los años 2012 al 2022, con sus respectivos intervalos de credibilidad del 95 % y 99 %

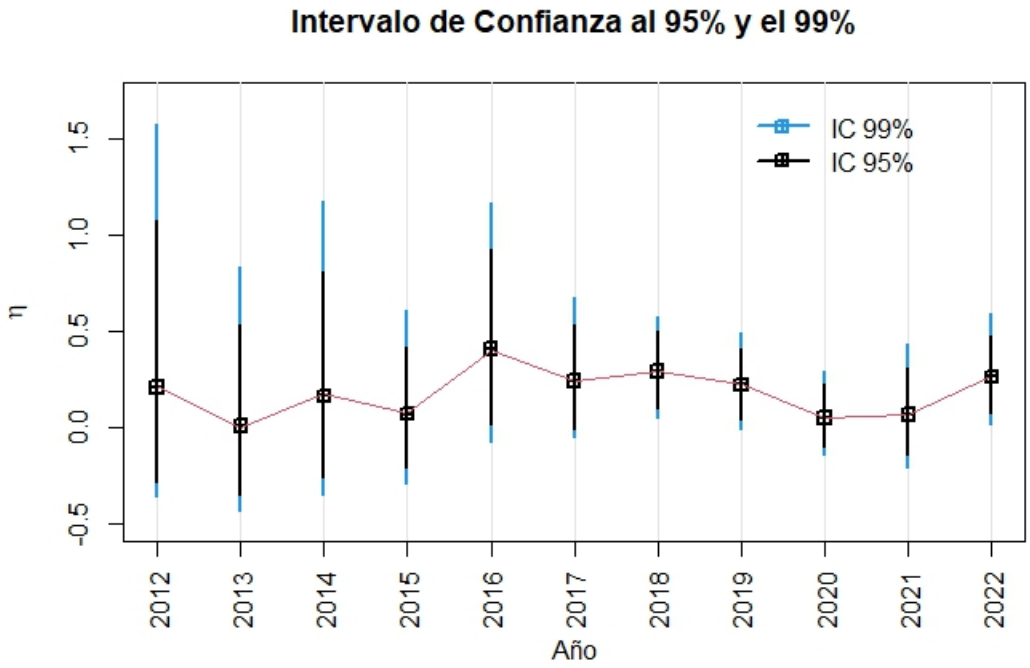

Figura 9: Estimaciones puntuales con sus intervalos de confianza del 95 % y 99 %

Año	Tamaño de Muestra		Media	Coef. Variación	IC 95 %		IC 99 %	
	Niños	Niñas			2.5 %	97.5 %	0.5 %	99.5 %
2012	13	64	0.213	1.654	-0.281	1.074	-0.360	1.573
2013	18	85	0.000	465.995	-0.351	0.532	-0.431	0.836
2014	18	66	0.172	1.647	-0.256	0.809	-0.349	1.173
2015	30	94	0.071	2.340	-0.212	0.417	-0.290	0.604
2016	41	132	0.403	0.576	0.015	0.921	-0.078	1.164
2017	77	213	0.244	0.574	-0.009	0.533	-0.053	0.676
2018	116	301	0.290	0.359	0.098	0.503	0.051	0.577
2019	122	295	0.222	0.434	0.043	0.411	-0.008	0.493
2020	111	291	0.050	1.703	-0.102	0.222	-0.139	0.293
2021	75	265	0.065	1.766	-0.138	0.310	-0.208	0.429
2022	115	237	0.263	0.385	0.078	0.474	0.020	0.589

Cuadro 11: Media observada de η para los años 2012 al 2022, con sus respectivos intervalos de confianza del 95 % y 99 %

Se puede observar que las estimaciones puntuales de obtenidas por Monte Carlo como por Bootstrap no tienen diferencias significativas entre ellas.

Dados los altos valores del coeficiente de variación posterior, indicando una variabilidad alta en comparación con la media estimada; mostrando que la estimación de η es relativamente inexacta o imprecisa, principalmente en el año 2013.

Además, los intervalos de credibilidad/confianza tienden a reducir su tamaño cuando el tiem-

po más se acerca al 2022; más exactamente desde el 2017, que se registran más conteo de ocurrencias de delitos sexuales a víctimas, haciendo que los modelos tanto bayesianos como frecuentistas sean más precisos.

Anexos

Teorema de factorización de Fisher-Neyman

$t(y_1, \ldots, y_n)$ es un estadístico suficiente para θ si y sólo si se pueden encontrar dos funciones no negativas h y g_θ tales que $f(y_1, \ldots, y_n) = h(y_1, \ldots, y_n) g_\theta(t(y_1, \ldots, y_n))$. (Estadística Bayesiana 2023a)

Modelo Gamma-Poisson

Sabemos que los datos se modelan con una dsitribución $poisson$ condicionalmente independiente e igualmente distribuidas así:

$$y_i \mid \theta \quad \overset{iid}{\sim} \mathsf{Poisson}(\theta) \quad i = 1, \ldots, n$$

por tanto la distribución condicional conjunta conocida como distribución muestral es:

$$p(\boldsymbol{y}|\theta) = \prod_{i=1}^{n} \frac{\theta^{y_i} e^{-\theta}}{y_i!} = \frac{\theta^s e^{-n\theta}}{\prod_{i=1}^n y_i!}$$

donde $s = \sum_{i=1}^{n} y_i$ es un estadístico suficiente.

Por otro lado tenemos que $\theta \sim \mathsf{Gamma}(a, b)$ así la distribución previa es un gamma por ende:

$$p(\theta) = \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta}, \quad \theta > 0$$

Por el teorema de bayes tenemos que:

$$\begin{aligned} p(\theta|\boldsymbol{y}) &= \tfrac{p(\boldsymbol{y}|\theta)p(\theta)}{p(\boldsymbol{y})} \\ &\propto p(\boldsymbol{y}|\theta)p(\theta) \end{aligned}$$

Con todo lo anterior la distribución posterior es proporcional a:

$$\begin{aligned} p(\theta|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\theta)p(\theta) \\ &\propto \theta^s e^{-n\theta} \left(\theta^{a-1} e^{-b\theta}\right) \\ &\propto \theta^{(a+s)-1} e^{-(b+n)\theta} \end{aligned}$$

Lo cual cómo vemos la distribución posterior tiene el kernel de un distribución Gamma de parámetros $a + s$ y $b + n$ por lo tanto la distribución Gamma es conjugada para la distribución poisson

y la distribución posterior tiene distribución (Estadística Bayesiana 2023b):

$$\theta|\boldsymbol{y} \sim Gamma(\theta|a+s, b+n)$$

Teorema de factorización de Fisher-Neyman

$t(y_1, \ldots, y_n)$ es un estadístico suficiente para θ si y sólo si se pueden encontrar dos funciones no negativas h y g_θ tales que $f(y_1, \ldots, y_n) = h(y_1, \ldots, y_n) g_\theta(t(y_1, \ldots, y_n))$. (Estadística Bayesiana 2023a)

Teorema: Invarianza del MLE

Si $\hat{\theta}(\boldsymbol{X})$ es el MLE para θ y $\tau(.)$ es un función, el MLE para $\tau(\theta)$ será $\tau\left(\hat{\theta}(\boldsymbol{X})\right)$ (Arrieta 2021)
Para este estudio, se considera la media muestral como estimador de máxima verosimilitud de θ_k, con $k = 1, 2$; es decir, $\hat{\theta}_{1MLE} = \bar{y}_1$ y $\hat{\theta}_{2MLE} = \bar{y}_2$. Se tiene, por definición, que

$$\eta = \frac{\theta_2 - \theta_1}{\theta_1}$$

Entonces, por el teorema de invarianza del estimador de máxima verosimilitud, se tiene que

$$\hat{\eta}_{MLE} = \frac{\hat{\theta}_{2MLE} - \hat{\theta}_{1MLE}}{\hat{\theta}_{1MLE}} = \frac{\frac{\sum y_2}{n} - \frac{\sum y_1}{n}}{\frac{\sum y_1}{n}} = \frac{\bar{y}_2 - \bar{y}_1}{\bar{y}_1}$$

donde $\sum y_1$ y $\sum y_2$ son las sumas de las observaciones en las dos poblaciones.

Referencias

Arrieta, M (2021). *Estimador de Máxima verosimilitud*. Notas de clase adaptadas para la comprensión.

Datos Abiertos (2019). *Conteo de Víctimas*. Recuperado el 12 de marzo de 2023 de. URL: `https://www.datos.gov.co/Justicia-y-Derecho/Conteo-de-V-ctimas/sft7-9im5`.

Estadística Bayesiana (2023a). *Modelo binomial*. Recuperado el 12 de marzo de 2023 de. URL: `https://rpubs.com/jstats1702/932520`.

— (2023b). *Modelo poisson*. Recuperado el 12 de marzo de 2023 de. URL: `https://rpubs.com/jstats1702/933886`.

Fiscalía General de la Nación (s.f.). *Sistema Penal Oral Acusatorio*. Recuperado el 12 de marzo de 2023 de. URL: `https://www.fiscalia.gov.co/colombia/la-entidad/sistema-penal-oral-acusatorio/`.

Hoff, P (2009). "A First Course in Bayesian Statistical Methods". En: Recuperado el 12 de marzo de 2023 de. Springer. Cap. 1, págs. 5-7. URL: `http://metodos.fam.cie.uva.es/~latex/apuntes/apuntes19.pdf`.

s.n. (s.f.). *Chapter 3: Bootstrap*. Recuperado el 12 de marzo de 2023 de. URL: http://www.math.chalmers.se/Stat/Grundutb/CTH/tms150/1112/Boot.pdf.