Universidad Nacional de Colombia

Facultad de ciencias
Departamento de estadística

Caso de Estudio I

Estadística Bayesiana
Prof. Juan Camilo Sosa

Autores

Mateo Santiago Cardona Ayala - Estadística - mcardonaay@unal.edu.co
Santiago Arias Garzón - Estadística - saarias@unal.edu.co

Bogotá, 15 de Marzo de 2023

PARTE 1: 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

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

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.

   Nota: usar un solo panel para la visualización.

   Solución:
   Sean $y_1, \ldots, y_n$ una secuencia de variables de conteo intercambiables tal que su distribución muestral está dada por

   $$\begin{aligned} y_i \mid \theta &\overset{\text{iid}}{\sim} \mathsf{Poisson}(\theta)\,, \quad i = 1, \ldots, n_k\,, \\ \theta &\sim \mathsf{Gamma}(a, b) \end{aligned}$$

   Para algunos a, b reales. Ahora veamos la distribución conjunta de los y_i.

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

   Donde $s = \sum_{i=1}^{n} y_i$, el cual gracias al criterio de factorización de Fisher-Neyman podemos ver que es un estadístico suficiente a partir de la expresión anterior podemos expresar la distribución conjunta de las y_i como un producto de una función de la muestra $\left(\frac{1}{\prod_{i=1} y_i!}\right)$ y una función de s y θ ($\theta^s e^{-n\theta}$), gracias a lo cual podemos usar s para resumir la muestra. Además dado que las y_i son independientes e identicamente distribuidas condicionalmente entonces

   $$s \mid \theta \sim \mathsf{Poisson}(\theta)$$

   Entonces;

   $$p(\theta \mid s) = \int_0^{\infty} p(s \mid \theta) \cdot p(\theta)\; d\theta$$

$$\begin{aligned} &\propto \int_0^\infty \theta^s e^{-n\theta} \cdot \theta^{a-1} e^{-b\theta}\ d\theta \\ &\propto \int_0^\infty \theta^{a+s-1} e^{-(b+n)\theta}\ d\theta \end{aligned}$$

El cual es el núcleo de una distribución Gamma de parámetros $a+s$ y $b+n$ por lo que podemos concluir que

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

Ahora procedemos a obtener la distribución predictiva posterior del modelo;

$$\begin{aligned} p(y^* \mid s) &= \int_0^{infty} p(y^* \mid \theta) \cdot p(\theta \mid s)\ d\theta \\ &= \int_0^\infty \frac{\theta^{y^*} e^{-\theta}}{y^*!} \cdot \frac{(b+n)^{a+s}}{\Gamma(a+s)} \theta^{a+s} e^{-(b+n)\theta}\ d\theta \\ &= \frac{1}{y^*!} \cdot \frac{(b+n)^{a+s}}{\Gamma(a+s)} \int_0^\infty \theta^{y^*} e^{-\theta} \cdot \theta^{a+s} e^{-(b+n)\theta}\ d\theta \\ &= \frac{1}{y^*!} \cdot \frac{(b+n)^{a+s}}{\Gamma(a+s)} \int_0^\infty \theta^{a+s+y^*-1} e^{-(b+n+1)\theta}\ d\theta \\ &= \frac{1}{y^*!} \cdot \frac{(b+n)^{a+s}}{\Gamma(a+s)} \cdot \frac{\Gamma(a+s+y^*)}{(b+n+1)^{a+s+y^*)}} \\ &= \frac{\Gamma(a+s+y^*)}{\Gamma(a+s)\Gamma(y^*+1)} \cdot \left(\frac{b+n}{b+n+1}\right)^{a+s} \cdot \left(\frac{1}{b+n+1}\right)^{y^*} \end{aligned}$$

Lo cual es la función de masa de probabilidad de una variable aleatoria binomial negativa de parámetros $a+s$ y $b+n$, por consiguiente tenemos que

$$y^* \mid s \sim \mathsf{BN}(a+s, b+n)$$

Con lo cuál podemos proceder a hacer el ajuste del modelo Gamma-Poisson al conteo total de victimas por género. De la muesta obtenida en la base de datos obtenemos

Tabla 1: Estadísticas obtenidas de la muestra

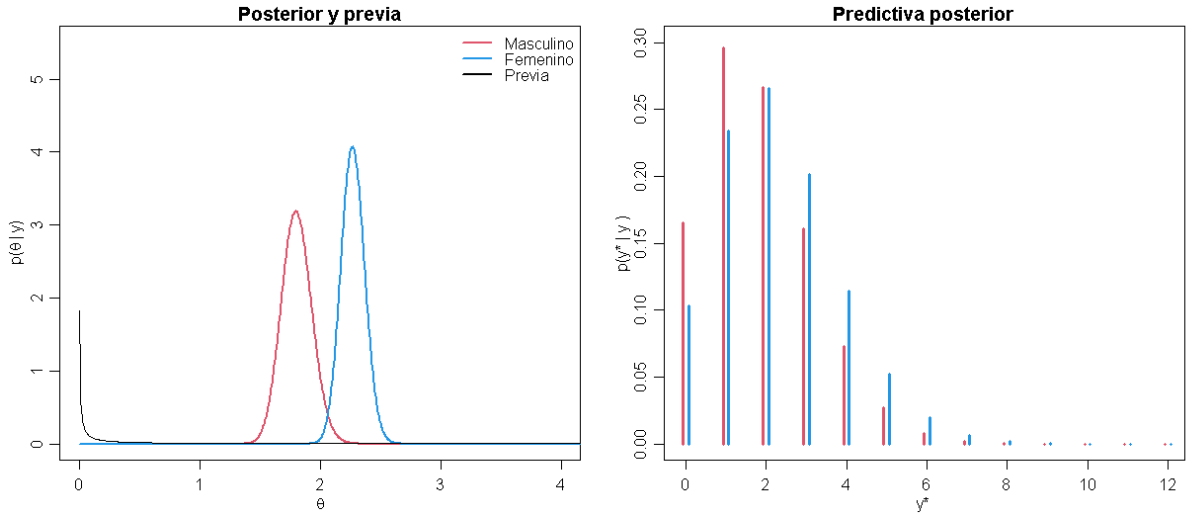
Población	s_k	n_k	σ_k
1	208	115	1.498
2	539	237	2.035

Al calcular la relación media-varianza para ambos grupos vemos que amabs poblaciones presentan sobredispersión considerable por lo que un mejor modelo sería usando la binomial negativa como previa. Por lo cual se tienen las distribuciones posteriores

$$\theta_1 \mid s_1 \sim \mathsf{Gamma}(208{,}01; 115{,}01)$$

$$\theta_2 \mid s_2 \sim \mathsf{Gamma}(539{,}01; 237{,}01)$$

Figura 1: Distribuciones posteriores del modelo



Y las distribuciones predicitivas posteriores

$$y_1^* \mid s_1 \sim \mathsf{BN}(208{,}01; 115{,}01)$$

$$y_2^* \mid s_2 \sim \mathsf{BN}(539{,}01; 237{,}01)$$

Ahora podemos visualizar estas distribuciones posteriores

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.

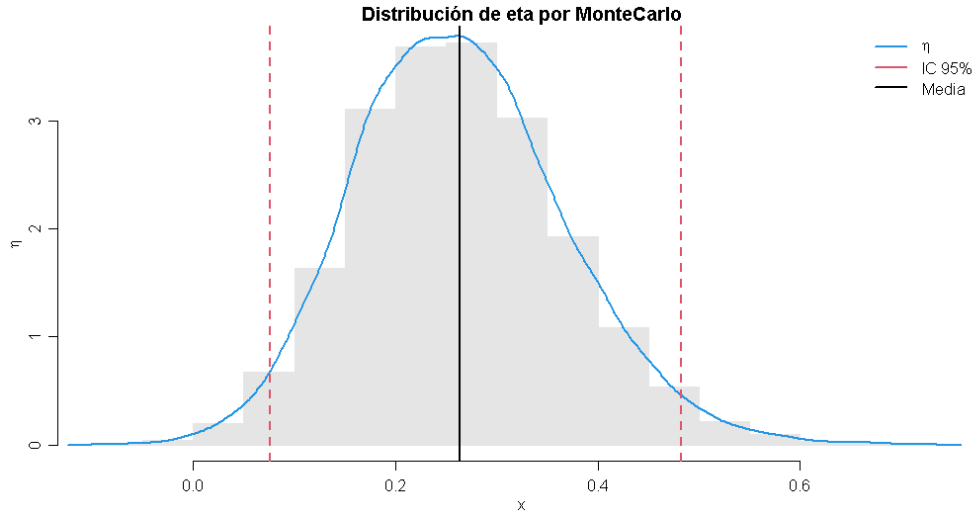
   Solución:
   Se usaron 10000 muestras en la aplicación del método de Monte Carlo, con las cuales se obtuvieron los siguientes resultados

Tabla 2: Media, Coeficiente de variación e intervalo de credibilidad de η por Monte carlo

	Valor
Media	0.264
Coeficiente de Variación	0.394
Límite Inferior al 95 %	0.076
Límite Superior al 95 %	0.482

Además de la siguiente distribuciión empírica estimada Se estima que las mujeres fueron victimas de violencia sexual a menores de edad en Bogotá D.C durante el año 2022 un 26.4 % más que los hombres. Se puede afirmar con una credibilidad del 95 % que las mujeres menores de edad

Figura 2: Distribución empírica de η por Monte Carlo



de Bogotá fueron victimas de violencia sexual más frecuentemente que los hombres menores de edad de esta misma ciudad, sin embargo con una diferencia porcentual no mayor al 48.2 %. Sin embargo la variación de esta estimación es alta por lo que se debe tener cuidado con cualquier medida tomada al respecto.

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.

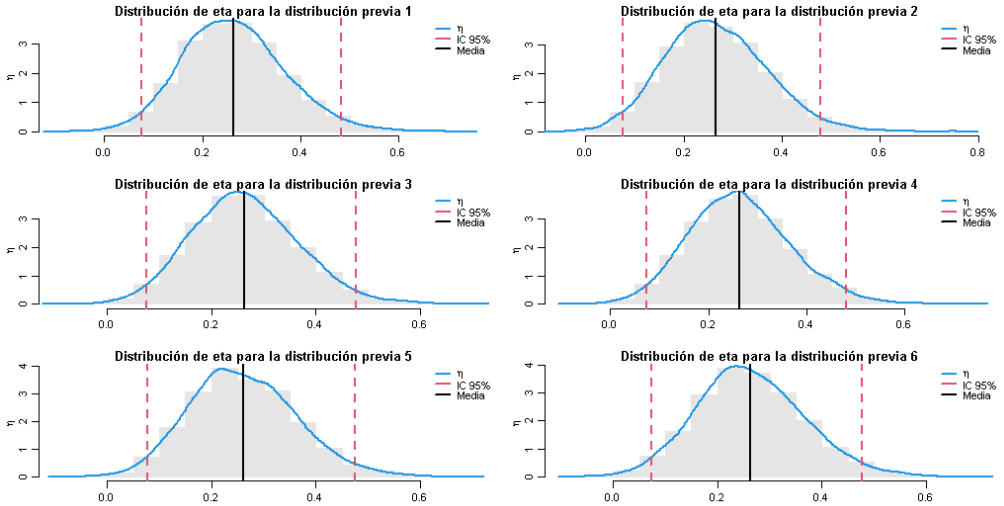
   Solución:
   Se usaron 10000 muestras para cada caso en la aplicación del método de Monte Carlo del cual se obtuvieron los siguientes resultados

   Además de las siguientes visualizaciones de las distribuciones empíricas estimadas. Podemos apreciar en las tres primeras distribuciones previas que a pesar de tener la misma media apriori son muy diferentes en cuanto a variación y aún así no hay cambios significativos en la estimación

Tabla 3: Media apriori y aposteriori, CV apriori y aposteriori e intervalo de credibiliad para η por cada distribución previa

Dist. Previa	**Media**	**CV**	**LI 95 %**	**LS 95 %**	**Media apriori**	**CV apriori**
Distr. Previa 1	0.264	0.394	0.076	0.482	1	10
Distr. Previa 2	0.265	0.390	0.076	0.479	1	3.162
Distr. Previa 3	0.264	0.390	0.076	0.477	1	1
Distr. Previa 4	0.264	0.393	0.074	0.481	2	1
Distr. Previa 5	0.260	0.392	0.077	0.476	3	1
Distr. Previa 6	0.263	0.391	0.074	0.476	4	1

Figura 3: Distribuciones empíricas de η por cada previa



de η lo cual podemos notarlo en que el coeficiente de varianza calculado para las tres previas es muy parecido, además de que gráficamente las tres distribuciones tambien son bastante similares. Por otro lado para las últimas tres distribuciones tenemos el caso contrario (mismo coeficiente de variación apriori y medias apriori distintas) y vemos que incluso así la estimación de η no se ve afectada por estas diferencias de forma significativa. Por lo que podemos concluir que el modelo Gamma-Poisson es muy poco sensible al cambio de los hiperparámetros cuando se tienen previas poco informativas.

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).

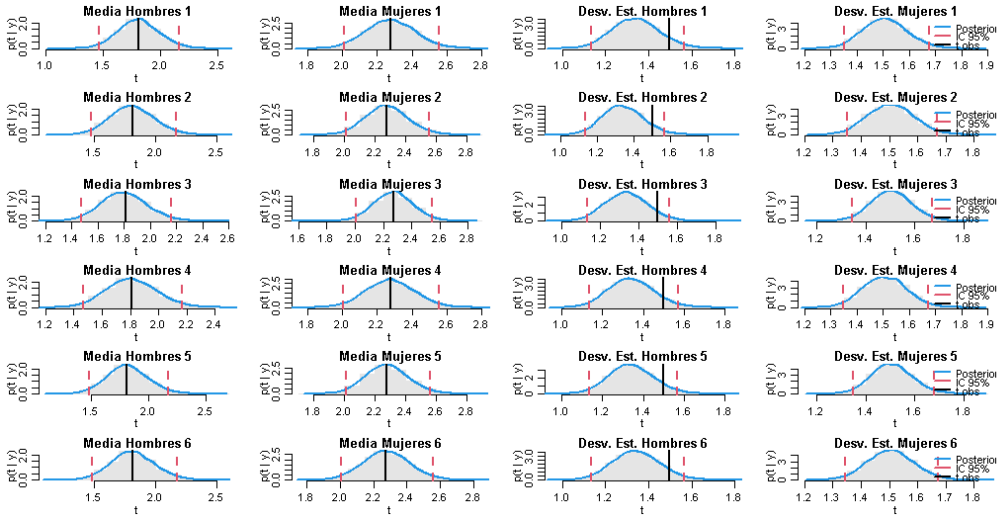
   Nota: calcular los valores p predictivos posteriores. **Solución:**
   Se usaron 10000 muestras para calcular las distribuciones empiricas de los estadísticos de prueba de donde se obtuvieron los siguientes valores p predictivos posteriores

Tabla 4: Tabla de Valores p predictivos posteriores de los estadísticos de prueba

Dist. Previa	Media Hombres	Media Mujeres	Desv. Est. Hombres	Desv. Est. Mujeres
Distr. Previa 1	0.486	0.496	0.081	0
Distr. Previa 2	0.477	0.491	0.081	0
Distr. Previa 3	0.458	0.484	0.075	0
Distr. Previa 4	0.479	0.491	0.081	0
Distr. Previa 5	0.488	0.496	0.079	0
Distr. Previa 6	0.490	0.494	0.082	0

Además de las siguientes visualizaciones para las distribuciones empiricas de los estadísticos de prueba para las poblaciones 1 y 2.
Lo más destacable por su ausencia es la desviación estandar observada para las mujeres la cual

Figura 4: Distribuciones empíricas de los estadísticos de prueba por grupo y por cadad previa



para todas las distribuciones previas al ajustar el modelo no se encuentra dentro del intervalo de credibiilidad, caso similar al de la desviación estandar de los hombres la cual si se encuentra dentro de los intervalos de credibilidad pero con valores p predicitivos posteriores bastante alejados de 0.5. Esto se puede deber a lo mencionado en el punto 1 de la sobredispersión de los datos observados. Mientras que para la media vemos que la bondad de ajuste es bastante alta con valores ppp muy cercanos a 0.5.

PARTE 2: Análisis frecuentista en 2022

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

   Nota: usar una cantidad de remuestras adecuada.

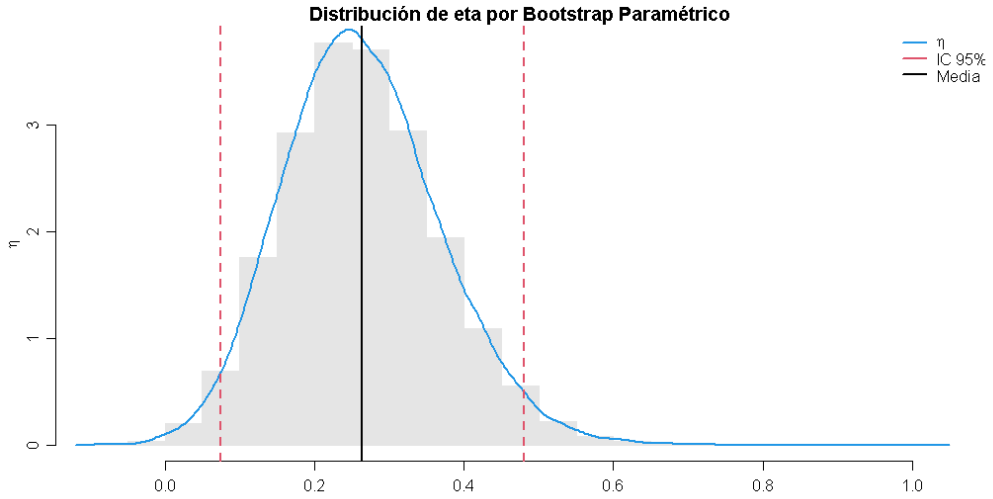
   Solución: Se usaron 50000 remuestras en el Bootstrap de donde se obtuvieron los siguientes resultados

Tabla 5: Media, Coeficiente de variación e intervalo de credibilidad de η por Bootstrap

	Valor
Media	0.263
Coeficiente de Variación	0.395
Límite Inferior al 95 %	0.075
Límite Superior al 95 %	0.481

Además de la visualización de la distribución empírica de η calculada mediante Bootstrap Com-

Figura 5: Distribución de η por Bootstrap paramétrico



parando los resultados obtenidos con los vistos en el punto 2 de la primera parte podemos notar

que son bastante parecidos y de donde podemos obtener las mismas conclusiones, ademas de ver que ambos enfoques estadísticos producen resultados similares.

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 (máximo 100 palabras).

Se usaron 2000 remuestras para el Bootstrap y 2000 muestras para Monte Carlo de donde se obtuvieron los siguientes resultados

Ambos intervalor contienen el valor verdadero de η aproximadamente 95 % de las veces en todos

Tabla 6: Cobertura de los intervalos de confianza/credibilidad

Escenario	**Tamaño de muestra**	**Cobertura Int. Credibilidad**	**Cobertura Int. Confianza**
1	10	0.945	0.946
2	20	0.947	0.948
3	50	0.948	0.949
4	100	0.949	0.950

los casos, lo que es importante mencionar es que a medida que aumenta el tamaño de muestra más se acerca, sin embargo incluso con solo muestras de tamaño 10 se tiene una cobertura muy cercana a la deseada, también es notable lo cercanas que son las coberturas entre el intervalo de credibilidad y el de confianza para cada tamaño de muestra, lo que nos vuelve a mostrar que ambos enfoques producen resultados similares.

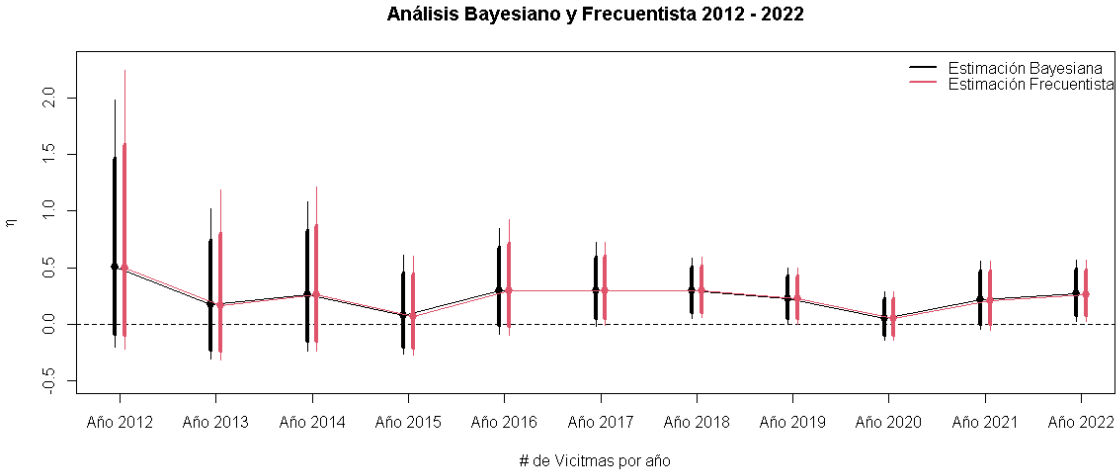
PARTE 3: Análisis Bayesiano y frecuentista en 2012-2022

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 (máximo 100 palabras).

Nota: usar un solo panel para la visualización.

Solución:

Figura 6: Análisis bayensiano y frecuentista en 2012 - 2022



Se estima que entre el 2012 y el 2022 las mujeres han sido más victimas de violencia sexual en menores que los hombres en la ciudad de Bogotá, sin embargo en años como el 2015 y el 2020 la diferencia es menor, incluso del 2012 al 2015 no hay evidencia estadística para rechazar que sean iguales los promedios de victimas para ambas poblaciones, igualmente en en el 2020 el cual posiblemente sea por la intervención de la pandemia y el confinamiento ya que en los años posteriores la media general parece regresar a la tendencia adquirida desde el 2016 hasta el 2019. Cabe mencionar que ambas estimaciones son bastantes similares aunque el estimador frecuentista es más sensible a los bajos tamaños de muestra de algunos años como caso más notable el año 2012.