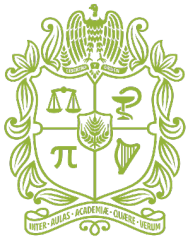

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Facultad de Ciencias
Departamento de Estadística

Caso de Estudio I:
Tiempos de reparación
Verizon

Estadística Bayesiana
Docente:
Juan Camilo Sosa Martinez

Autores

Michel Mendivenson Barragan
mbarraganz@unal.edu.co
Gerardo Sebastian Gil Sanchez
ggil@unal.edu.co

Septiembre 2023

La base de datos `Verizon.csv` está conformada por 1687 observaciones de tiempos de reparación de la empresa telefónica Verizon. Estos registros se dividen en dos grupos:

- ILEC: Servicios de reparación ofrecidos a clientes de Verizon (1664 registros).
- CLEC: Servicios de reparación ofrecidos a clientes de empresas externas (23 registros).

Uno de los principales objetivos del informe es decidir si existe una diferencia significativa entre la media de los tiempos de reparación del grupo ILEC y el tiempo medio de reparación del grupo CLEC en beneficio del primero. Para esto considere modelos exponenciales independientes de la forma

$$y_{k,i}|\lambda_k \sim^{iid} \text{Exp}(\lambda_k) \quad \Longleftrightarrow \quad p(y_{k,i}|\lambda_k) = \frac{1}{\lambda_k}\exp\left(-\frac{y_{k,i}}{\lambda_k}\right), \quad y_{k,i} > 0, \quad \lambda_k > 0$$

para i = 1,...,n_k y k = 1,2 (1: ILEC, 2: CLEC), donde $y_{k,i}$ es el tiempo de reparación (en horas) del individuo i en el grupo k, n_k es el tamaño de la muestra del grupo k, y finalmente, $y_k = (y_{k,1}, \ldots, y_{k,n_k})$ es el vector columna de observaciones correspondiente y considere $\eta = \lambda_1 - \lambda_2$.

Parte 1. Análisis Bayesiano.

Distribución posterior de η.

Ajuste los modelos Gamma-Inversa-Exponencial con $a_k = 3$ y $b_k = 17$ en cada grupo. A partir de las distribuciones posteriores obtenga la distribución posterior de η. Reporte la media, el coeficiente de variación y un intervalo de credibilidad al 95 % para η. Presente los resultados visual y tabularmente. Interprete los resultados obtenidos.

Si queremos muestrear de la población η tenemos que conocer la distribución posterior para cada una de las poblaciones (ILEC y CLEC) primero para poder aplicar el método Monte Carlo. Usando el teorema de Bayes obtenemos lo siguiente:

$$\begin{aligned} p(\lambda \,|\, \mathbf{y}) &\propto p(\mathbf{y} \,|\, \lambda)p(\mathbf{y}) \\ &= \prod_{i=1}^{n} p(y_i \,|\, \lambda)p(\lambda) \end{aligned}$$

Puesto que $\lambda \sim GI(\alpha, \beta)$ se tiene que $p(\lambda) = \dfrac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{-(\alpha+1)}\, exp\left(-\dfrac{\beta}{\lambda}\right)$

$$= \prod_{i=1}^{n} \frac{1}{\lambda}\, exp\left(-\frac{y_i}{\lambda}\right) \cdot \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{-(\alpha+1)}\, exp\left(-\frac{\beta}{\lambda}\right)$$

$$\propto \lambda^{-n}\, exp\left(-\frac{\beta+s}{\lambda}\right)\cdot\lambda^{-(\alpha+1)}exp\left(-\frac{\beta}{\lambda}\right)$$
$$= \lambda^{-(\alpha+n-1)}\, exp\left(-\frac{\beta+s}{\lambda}\right)$$

Y esto último, es el núcleo de una distribución Gamma Inversa de parámetros $\alpha+n$ y $\beta+s$ siendo $s=\sum_{i=1}^{n} y_i$. Es decir, la distribución posterior de λ dado un vector de datos $\mathbf{y}=(y_1,y_2,\cdots,y_n)$ es $\lambda\,|\,\mathbf{y}\sim GI(\alpha+n,\beta+s)$ o lo que es lo mismo la gamma inversa es una distribución previa conjugada para la distribución exponencial.
Con este resultado podemos muestrear las distribuciones posteriores de λ_1 y λ_2 por separado y calcular estimaciones de η usando el método de Monte Carlo con un tamaño de 50000 muestras obteniendo los siguientes resultados:

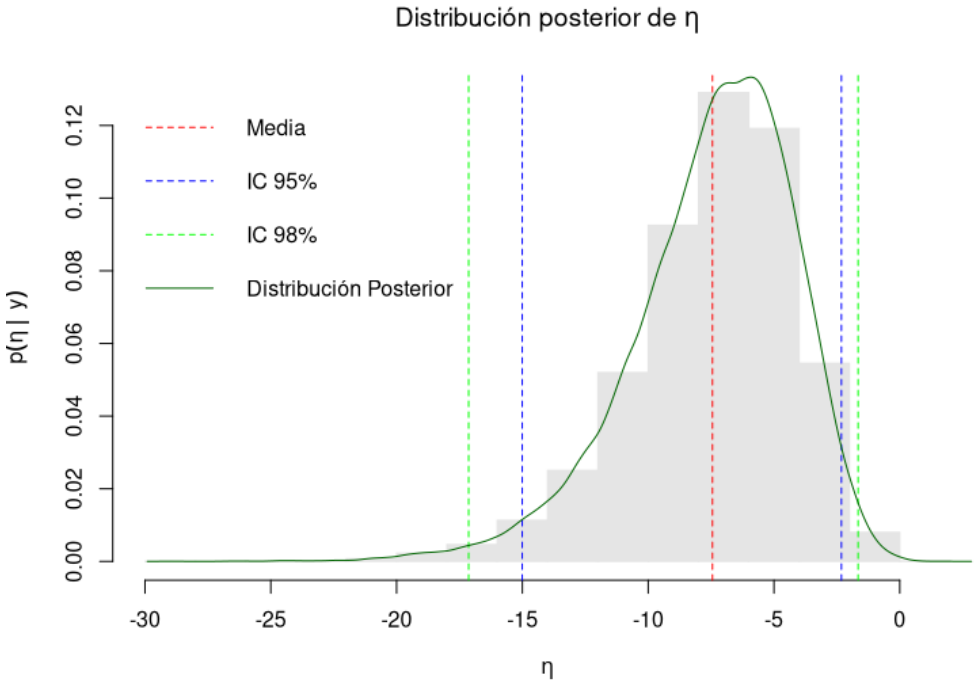

Figura 1: Distribución posterior de η basada en el método de Montecarlo

Y en forma tabular consignados aquí:

Media	IC 95 %	CV
-7.445	(-15.01, -2.32)	43.7 %

Tabla 1: Estimaciones sobre η

Como $\eta=\lambda_1-\lambda_2$ tendremos que $\lambda_1<\lambda_2$ si $\eta<0$. Con la información presentada anteriormente, podemos afirmar que con probabilidad 95 % η se encuentra en el intervalo de $-15{,}01$ a $-2{,}32$, más aún podemos afirmar que con apróximadamente 99.96 % de probabilidad η es

menor que cero o lo que es lo mismo $\lambda_1 < \lambda_2$. Parece existir una diferencia significativa en los tiempos de reparación a favor de los clientes de Verizon. También debemos tener en cuenta que la estimación tiene una variabilidad alta pues su coeficiente de variación es de aproximadamente 43.7 %.

Análisis de sensitividad (Respecto a η)

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

- Distribución previa 1: $a_k = 3$ y $b_k = 17$, para $k = 1, 2$.
- Distribución previa 2: $a_k = 2$ y $b_k = 8.5$, para $k = 1, 2$.
- Distribución previa 3: $a_k = 3$ y $b_1 = 16.8$ y $b_2 = 33$, para $k = 1, 2$.
- Distribución previa 4: $a_k = 2$ y $b_1 = 8.4$ y $b_2 = 16.5$, para $k = 1, 2$.

En cada caso calcule la media y el coeficiente de variación a priori, y repetir el núméral anterior. Presente los resultados visual y tabularmente. Interprete los resultados obtenidos (máximo 100 palabras).

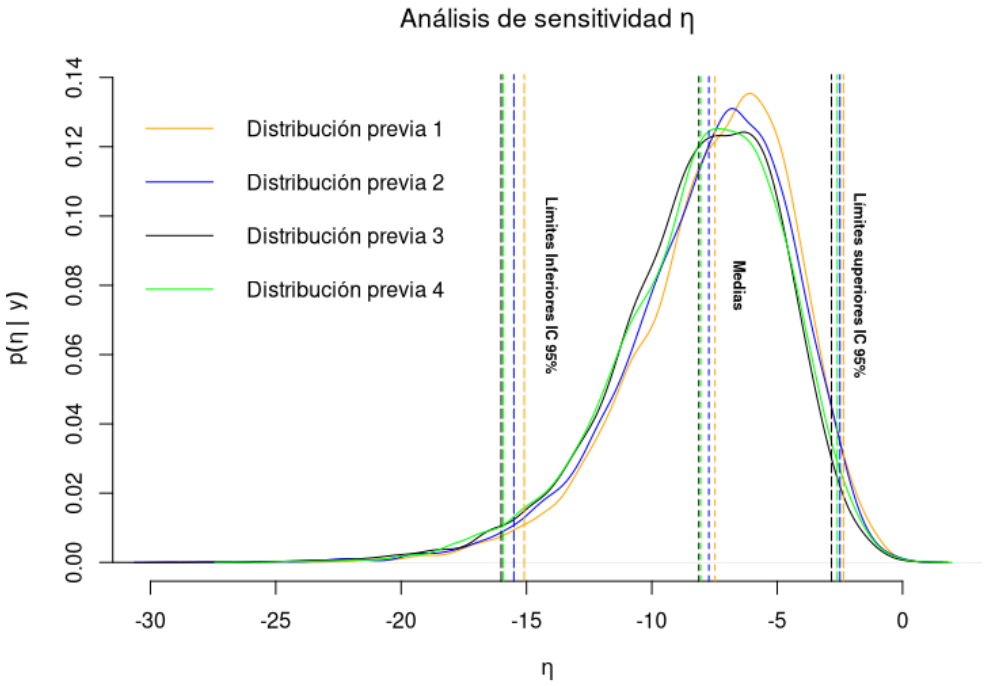

Figura 2: Análisis de sensitividad η

Dist. previa	Media	IC 95 %	CV	Media a priori	CV a priori
1	-7.482	(-15.089, -2.354)	43.5 %	0.0	–
2	-7.726	(-15.503, -2.496)	43.5 %	0.0	–
3	-8.124	(-16.010, -2.836)	41.6 %	-8.1	228.582 %
4	-8.053	(-15.921, -2.604)	42.1 %	-8.1	–

Tabla 2: Resultados análisis de sensitividad sobre η

Para calcular los coeficientes de variación y medias a priori de η se toma la diferencia de las medias de los parámetros λ_1 y λ_2 como la media de η y la varianza de η como la suma de las varianzas de λ_1 y λ_2. Para el caso de las distribuciones previas 1 y 2 como a_k y b_k son iguales para los dos grupos la media de η será cero y los coeficientes de variación no podrán ser definidos mientras que en el caso de la distribución previa cuatro el problema está en que dada una gamma inversa de parámetros α y β su media será dada por $\frac{\beta}{\alpha-1}$ y su varianza $\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$ por lo cual la varianza para la gamma inversa no estará definida para $\alpha = 2$ y por ende no estará definida la varianza de η. Puede ser prudente revisar las estimaciones para los coeficientes de variación a priori (Y las distribuciones previas) de η mediante el método Monte Carlo con 25000 muestras de la distribución previa de λ_1 y de λ_2 obtenemos las siguientes distribuciones previas para η:

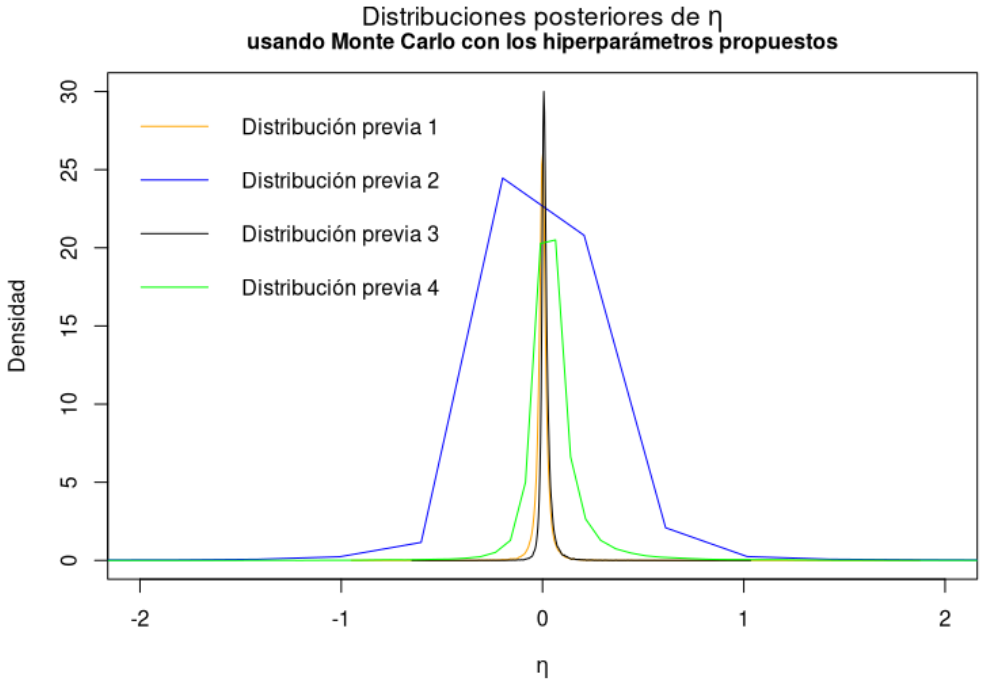

Figura 3: Distribución previa de η

Y las siguientes estimaciones de los coeficientes de variación:

Dist. previa	1	2	3	4
Coef. Variación a priori	24255.944 %	83154.455 %	224.5733 %	435.5787 %

Tabla 3: Coeficientes de variación previos de η estimados mediante Monte Carlo

Con los datos obtenidos podemos decir que el modelo no parece tener demasiada sensibilidad frente a la distribución previa pues la estimación puntual de η (Media), el intervalo de credibilidad y el coeficiente de variación no cambian de forma considerable. Además se debe notar que todas las distribuciones previas pese a tener distintos hiperparámetros no son informativas pues sus coeficientes de variación son altos.

Bondad de ajuste para las poblaciones por separado

En cada población, evalúe la bondad de ajuste del modelo propuesto utilizando la distribución previa 1, utilizando como estadísticos de prueba la media y la desviación estándar. Presente los resultados visual y tabularmente. Interprete los resultados obtenidos (máximo 100 palabras).

Para este análisis de la bondad de ajuste se generaron 50000 valores de la distribución posterior de λ_1 y λ_2 y por cada uno de estos valores se generó una muestra de 1664 y 23 elementos respectivamente. Obteniendo los siguientes resultados:

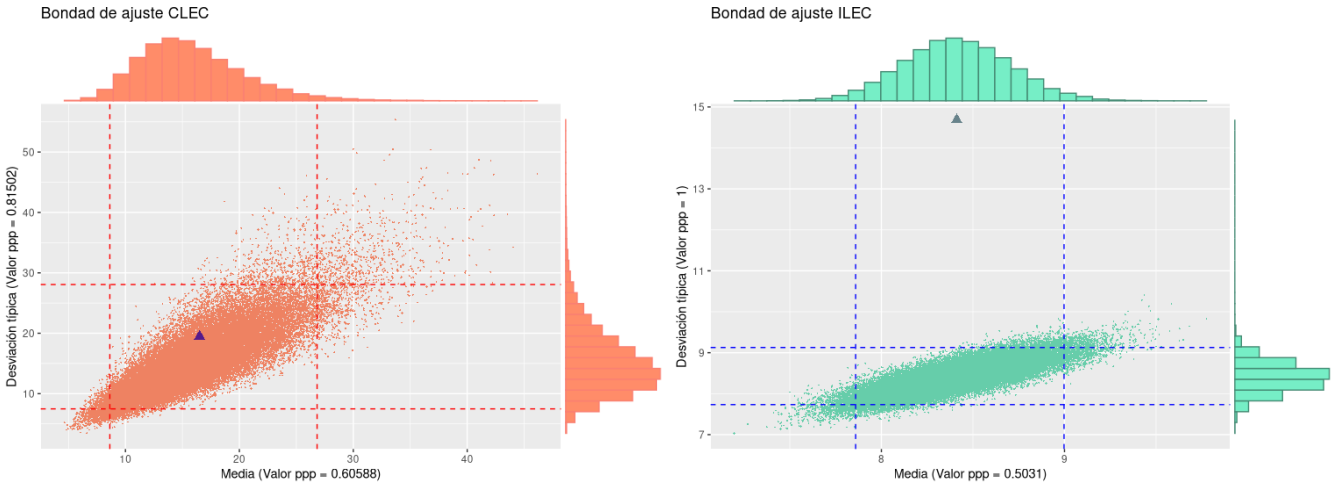

Figura 4: Dispersogramas CLEC e ILEC

	ILEC	CLEC
Media	0.5031	0.60588
Desv. estándar	1.0000	0.81502

Tabla 4: Valores ppp

Los valores ppp para cada caso nos están mostrando que nuestros modelos capturan bien los valores de las medias de ambas poblaciones. Sin embargo, no lo hacen tan bien para las desviaciones estándar lo cual supondría un problema si pretendierámos estimar esta estádistica de la población, pero como nuestro parámetro de intéres es la media deberíamos ser capaces de hacer estimaciones buenas con el modelo propuesto. Eso sí, cabe mencionar que el modelo parece ajustarse mejor en la población de clientes de Verizon, pero esto pudiera deberse a la cantidad de observaciones que se tienen.

Parte 2. Análisis frecuentista

Repita el númeral 1. de la PARTE 1 usando la Normalidad asintótica del MLE, *Bootstrap* paramétrico y *Bootstrap* no paramétrico. Presente los resultados visual y tabularmente. Interprete los resultados obtenidos (máximo 100 palabras).

Antes de empezar con este análisis, veamos cuál es el MLE de una distribución exponencial definiendo su función de verosimilitud como:

$$\begin{aligned}\mathscr{L}(\lambda) &= \prod_{i=1}^{n} \frac{1}{\lambda} e^{-\frac{x_i}{\lambda}} \\ &= \frac{1}{\lambda^n} exp\left(-\frac{1}{\lambda}\sum_{i=1}^{n} x_i\right) \\ &= \frac{1}{\lambda^n} exp\left(-\frac{n}{\lambda}\,\bar{x}\right)\end{aligned}$$

Y usando la log-verosimilitud obtenemos:

$$l(\lambda) = ln(\mathscr{L}(\lambda)) = -n\,ln(\lambda) - \frac{1}{\lambda}\,n\,\bar{x}$$

Y para encontrar el máximo derivamos e igualamos a cero:

$$\begin{aligned}\frac{d}{d\lambda} l(\lambda) = -\frac{n}{\lambda} + \frac{1}{\lambda^2}\,n\bar{x} &= 0 \\ -\frac{1}{\lambda} + \frac{\bar{x}}{\lambda^2} &= 0 \\ \frac{-\lambda + \bar{x}}{\lambda^2} &= 0 \\ -\lambda + \bar{x} = 0 \lambda &= \bar{x}\end{aligned}$$

Es decir, un punto crítico para la función de máxima verosimilitud es $\lambda = \bar{x}$. Usando el criterio de la segunda derivada tenemos que

$$\frac{d^2}{d^2\lambda} l(\lambda) = \frac{n}{\lambda^2} - \frac{2}{\lambda^3} n\bar{x}$$

Al reemplazar λ por $\bar{x}$ obtenemos $\frac{1-2n}{\bar{x}}$ que es una expresión menor que cero para $n \geq 1$. Es decir, $\hat{\lambda}_{MLE} = \bar{x}$.

Análisis por Normalidad Asintótica del MLE

Como ya sabemos que el estimador por MLE de λ para la parametrización dada inicialmente es la media muestral y sabemos que $\hat{\lambda}_{MLE} \sim N(\lambda, \mathbf{I}^{-1}(\lambda))$ lo que hacemos es definir a η

como la resta de dos normales (Con media la resta de las medias y varianza la suma de las varianzas) y evaluar los intervalos de confianza, estimadores puntuales, etc. sobre esta distribución:

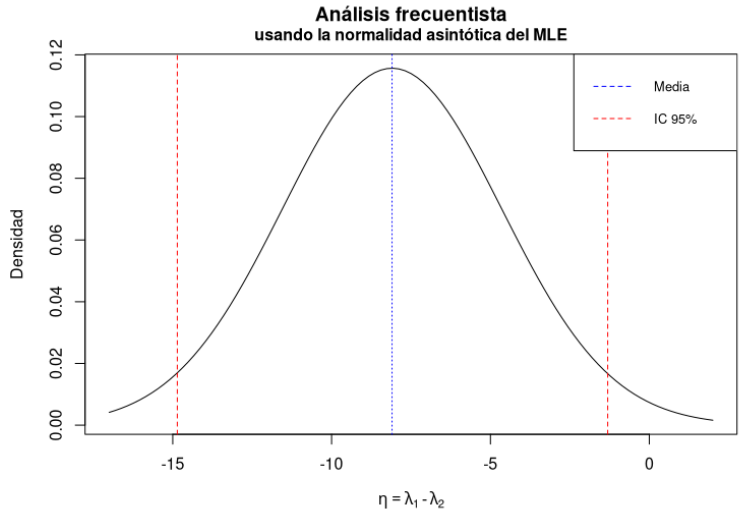

Figura 5: Estimación de η usando la normalidad asintótica del MLE

Media	CV	IC (95 %)
-8.1	42.59 %	(-14.86, -1.34)

Tabla 5: Resultados estimaciones sobre η por normalidad asintótica

Análisis por Boostrap No Paramétrico

Para el bootstrap no paramétrico se generaron 50000 remuestras de cada uno de los grupos de tamaño 1000 y se restan las medias de las remuestras para generar los valores de η. Los resultados se presentan a continuación:

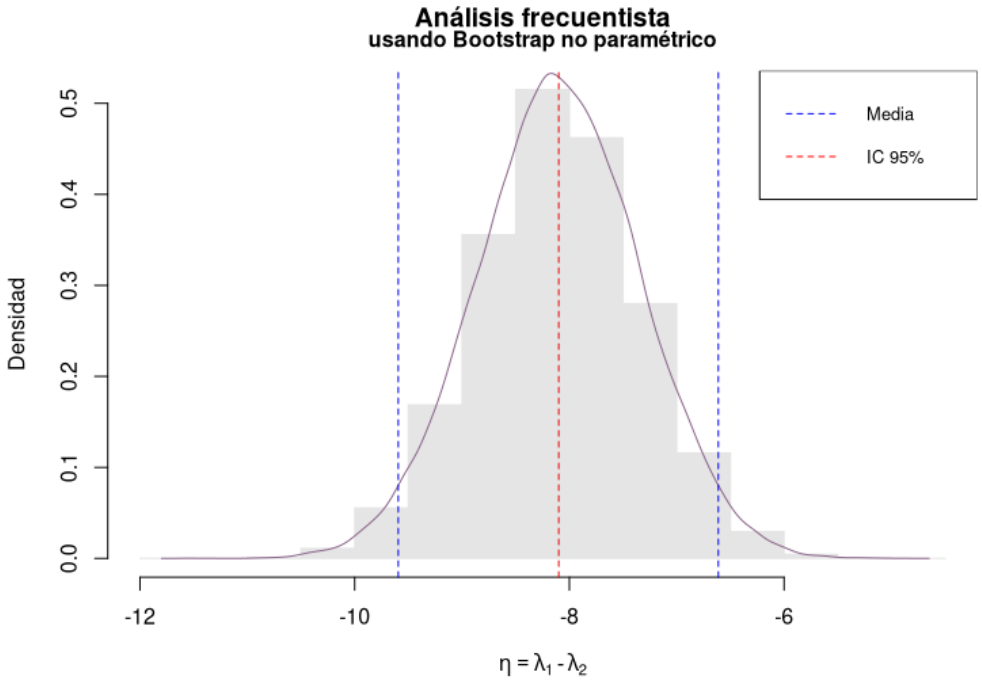

Figura 6: Estimación de η usando Bootstrap no paramétrico

Media	CV	IC 95 %
-8.1	9.39 %	(-9.59, -6.61)

Tabla 6: Resultados estimaciones sobre η por Bootstrap paramétrico

Análisis por Bootstrap Paramétrico

Nuevamente, como ya conocemos el estimador MLE del parámetro λ para una distribución exponencial, con este tipo de bootstrap lo que hacemos es simular muestras separadas para cada grupo siguiendo la distribución $Exp(\hat{\lambda}_{i_{MLE}})$ con $i = 1, 2$ y luego restar las muestras generadas para generar a η. Así obtenemos los siguientes resultados:

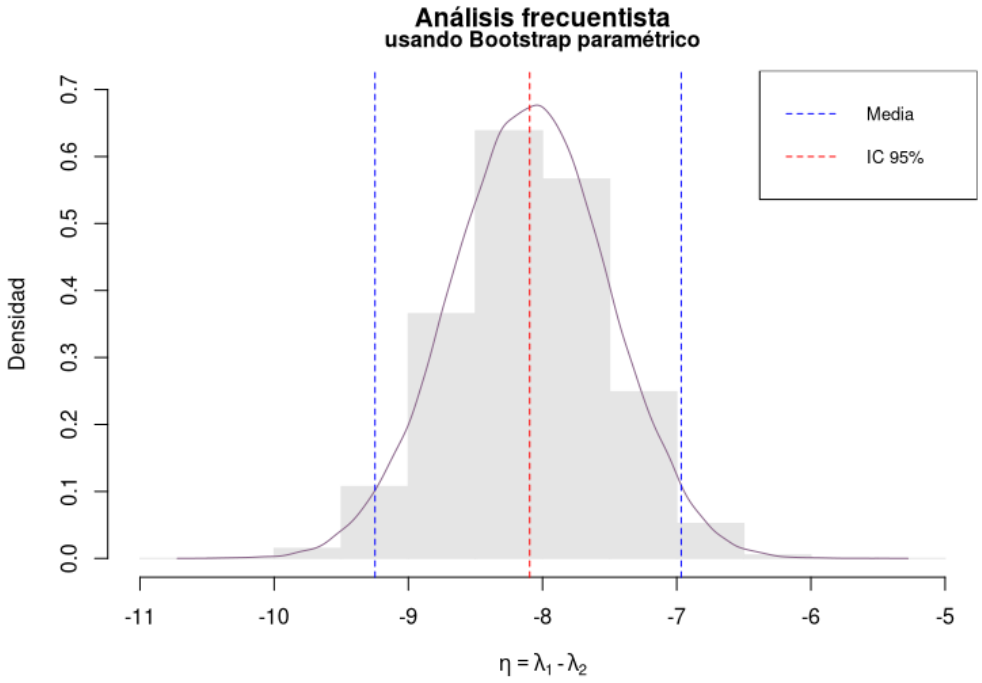

Figura 7: Estimación de η usando Bootstrap paramétrico

Media	CV	IC 95 %
-8.1	7.21 %	(-9.25, -6.97)

Tabla 7: Resultados estimaciones sobre η por Bootstrap paramétrico

Si bien es cierto que ninguno de los resultados encontrados por medio de análisis frecuentista contradice a los resultados del análisis Bayesiano también es cierto que para este caso la vía de la normalidad asintótica del MLE y el análisis bayesiano parecen mostrar estimadores con mayor variablidad (CV más altos) mientras que los bootstrap generan estimaciones con mucha menos variabilidad.

Parte 3. Simulación

Simule 100000 muestras aleatorias de poblaciones Exponenciales bajo los siguientes escenarios:

- Escenario 1: $n_1 = 10$, $n_2 = 10$, $\lambda_1 = \bar{y_1}$, y $\lambda_2 = \bar{y_2}$.
- Escenario 2: $n_1 = 20$, $n_2 = 20$, $\lambda_1 = \bar{y_1}$, y $\lambda_2 = \bar{y_2}$.
- Escenario 3: $n_1 = 50$, $n_2 = 50$, $\lambda_1 = \bar{y_1}$, y $\lambda_2 = \bar{y_2}$.
- Escenario 4: $n_1 = 100$, $n_2 = 100$, $\lambda_1 = \bar{y_1}$, y $\lambda_2 = \bar{y_2}$.

esdonde $\bar{y_k} = \frac{1}{n_k}\sum_{i=1}^{n_k} y_{k,i}$ es la media muestral observada del grupo k. Observe que el valor verdadero de η en cada caso es $\eta = \lambda_1 - \lambda_2 = \bar{y_1} - \bar{y_2}$. Usando cada muestra, ajuste el modelo de manera tanto Bayesiana (usando la distribución previa 1) como frecuentista (usando la Normalidad asintótica, Bootstrap paramétrico, Bootstrap no paramétrico), y en cada caso calcule la proporción de veces que el intervalo de credibilidad/confianza al 95 % contiene el valor verdadero de η. Reporte los resultados tabularmente. Interprete los resultados obtenidos (máximo 100 palabras).

Escenario	**Bayesiano**	**Frecuentista**		
		Asintótico	**Bootstrap P.**	**Bootstrap no P.**
1	92,93 %	94,15 %	92,89 %	100 %
2	92,69 %	94,81 %	94,26 %	100 %
3	94,69 %	95,06 %	93,89 %	100 %
4	94,32 %	95,08 %	94,91 %	100 %

Tabla 8: Proporción de η en los Intervalos de credibilidad/confianza al 95 %

La tabla muestra cómo las tasas de éxito de estimar si η está dentro del intervalo de confianza varían según el tamaño de la muestra y el enfoque estadístico utilizado. En general, a medida que aumenta el tamaño de la muestra, las tasas de éxito tienden a ser más altas, lo que es consistente con la idea de que las muestras más grandes proporcionan estimaciones de intervalo de credibilidad/confianza más precisas. Además, el enfoque bayesiano y los métodos bootstrap (tanto paramétricos como no paramétricos) parecen funcionar bastante bien en este contexto.