Parcial 1 Estadística Bayesiana

Alejandro Urrego López
aurrego@unal.edu.co

September 2023

1. Introducción

Verizon es la principal compañía telefónica local (**ILEC**, *incumbent local exchange carrier*) para una gran área del este de Estados Unidos. Como tal, es responsable de brindar servicio de reparación a los clientes de otras empresas telefónicas de la competencia (**CLEC**, *competing local exchange carrier*) en esta región. Verizon está sujeto a multas si los tiempos de reparación (el tiempo que lleva solucionar un problema) para los clientes de alguna CLEC son sustancialmente peores que los de los clientes de Verizon. El conjunto de datos `Verizon.csv` contiene una muestra de los tiempos de reparación de $n_1 = 1664$ clientes de Verizon (ILEC) y $n_2 = 23$ clientes de la competencia (CLEC). De acuerdo con los datos, los tiempos medios de reparación son $\bar{x}_1 = 8.41$ y $\bar{x}_2 = 16.51$ horas para ILEC y CLEC, respectivamente.

El objetivo de este caso es determinar si la diferencia entre los tiempos promedio de reparación es lo suficientemente grande para declararse como significativa, y por tanto, ser tenida en cuenta como evidencia para llevar a cabo una intervensión y multar a Verizon.

2. Modelo

La distribución Exponencial es popular para modelar tiempos dado que este modelo permite producir distribuciones con diferentes tasas de decaimiento y variedades de sesgo (para más información acerca de este modelo probabilístico. Así, considere modelos Exponenciales independientes de la forma

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

para $i = 1, \ldots, n_k$ 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.

3. Preguntas

Sea $\eta = \lambda_1 - \lambda_2$. A continuación, se hace inferencia estadística sobre η con el fin de responder al objetivo propuesto.

4. Análisis Bayesiano

1. 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 (máximo 100 palabras).

Nota: use métodos de Monte Carlo con una cantidad de muestras adecuada.

En primer lugar cabe destacar que la distribución gamma inversa es conjugada para la distribución exponencial. Por lo tanto, $\lambda_1|y_{1,i} \sim \text{GL}(a_k + n_1, b_k + s_{k,1})$, y $\lambda_2|y_{2,i} \sim \text{GL}(a_k + n_2, b_k + s_{k,2})$, con $s_k = \sum_{i=1}^{n} y_{k,i}$ es decir:

$$\lambda_1|y_{1,i} \sim \text{GL}(1667, 14013.92)$$
$$\lambda_2|y_{2,i} \sim \text{GL}(26, 396.71)$$

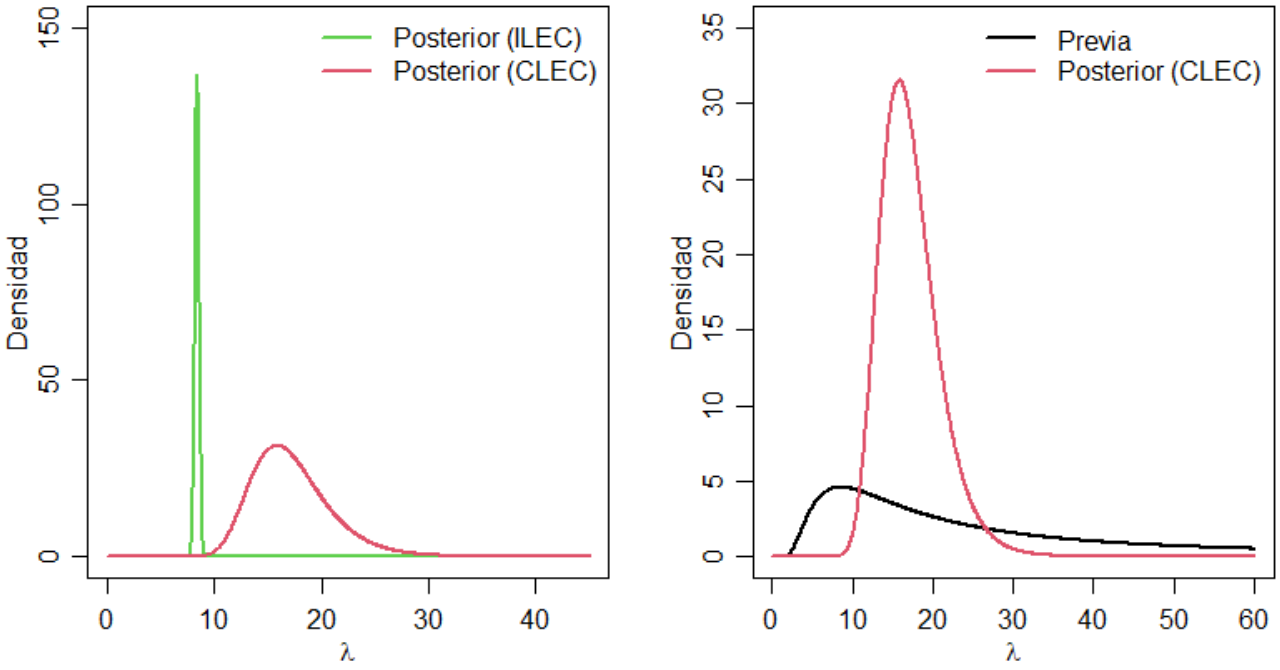

Figura 1: Distribución previa y posterior de λ_k

Grupo	Media Posterior	$P_{2.5}$	$P_{97.5}$
ILEC	8.41	8.02	8.83
CLEC	15.26	10.75	23.36

Tabla 1: Estadísticas de la Distribución posterior de λ_k

Se observa que la información del grupo ILEC tiende a concentrarse en su media debido a que el valor esperado de la distribución posterior es un promedio ponderado entre la previa y su media (Ver Anexo 6.3).

Ahora se utiliza el método de Monte Carlo simulando $B = 10000$ muestras aleatorias de $\lambda_1|y_{1,i}$ y $\lambda_2|y_{2,i}$, y se realiza la inferencia sobre $\eta = \lambda_1 - \lambda_2$.

	Media	Mediana	$P_{2.5}$	$P_{97.5}$	C.V	$P(\eta > 0)$
η	-7.499	-7.074778	-15.227	-2.367	0.437	0

Tabla 2: Estadísticas de η

En conclusión, al observar, la estimación puntual, que el intervalo de credibilidad al 95 % no incluye el cero y que $p(\eta > 0) \approx 0$, se evidencian diferencias significativas entre los tiempos de ILEC y CLEC. Específicamente, el tiempo de reparación para CLEC es mayor que el de ILEC. Por lo tanto, este resultado respalda la necesidad de llevar a cabo una investigación en contra de Verizon.

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

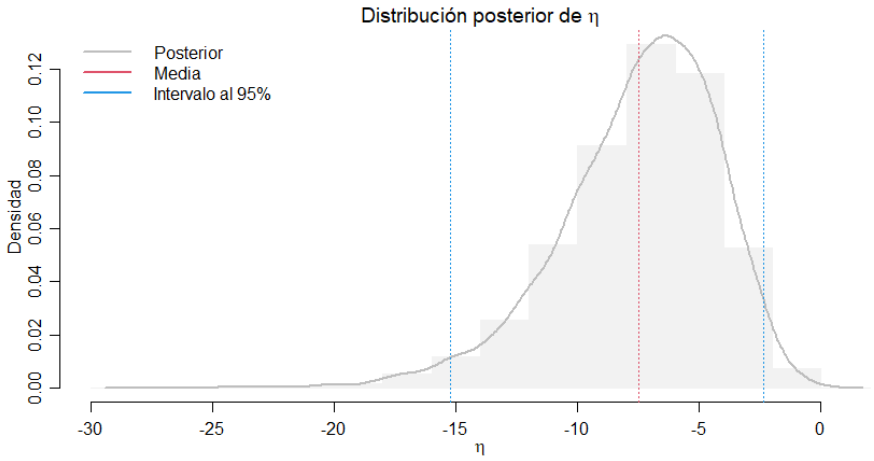

Figura 2: Distribución posterior de η

- 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 repita el numeral anterior. Presente los resultados visual y tabularmente. Interprete los resultados obtenidos (máximo 100 palabras).

a_k	b_1	b_2	Media	Des. Estandar	Coef. Var
3	17	17	0	0	NaN
2	8.5	8.5	0	0	NaN
3	16.8	33	-8.252	9.310	1.128
2	8.4	16.5	-7.854	13.270	1.689

Tabla 3: Estadisticas de η apriori

Se observa que existen dos distribuciones previas muy informativas en las cuales prácticamente toda su masa está concentrada en 0, mientras que las otras dos son poco informativas presentando un alto coeficiente de variación.

a_k	b_1	b_2	Media	CV	$P_{2.5}$	$P_{97.5}$
3	17	17	-7.429	0.430	-14.812	-2.370
2	8.5	8.5	-7.735	0.430	-15.467	-2.469
3	16.8	33	-8.068	0.412	-15.747	-2.807
2	8.4	16.5	-8.068	0.421	-15.959	-2.692

Tabla 4: Estadísticas de η posterior

Es evidente que, a pesar de los estados de información, la inferencia no cambia. Además, dado que el coeficiente de variación y la estimacíon puntual de η posterior es muy parecido en los cuatros escenarios significa que la variabilidad en los resultados es consistente y no cambia significativamente a medida que se ajustan los parámetros, lo que brinda más confianza en la inferencia.

3. 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).

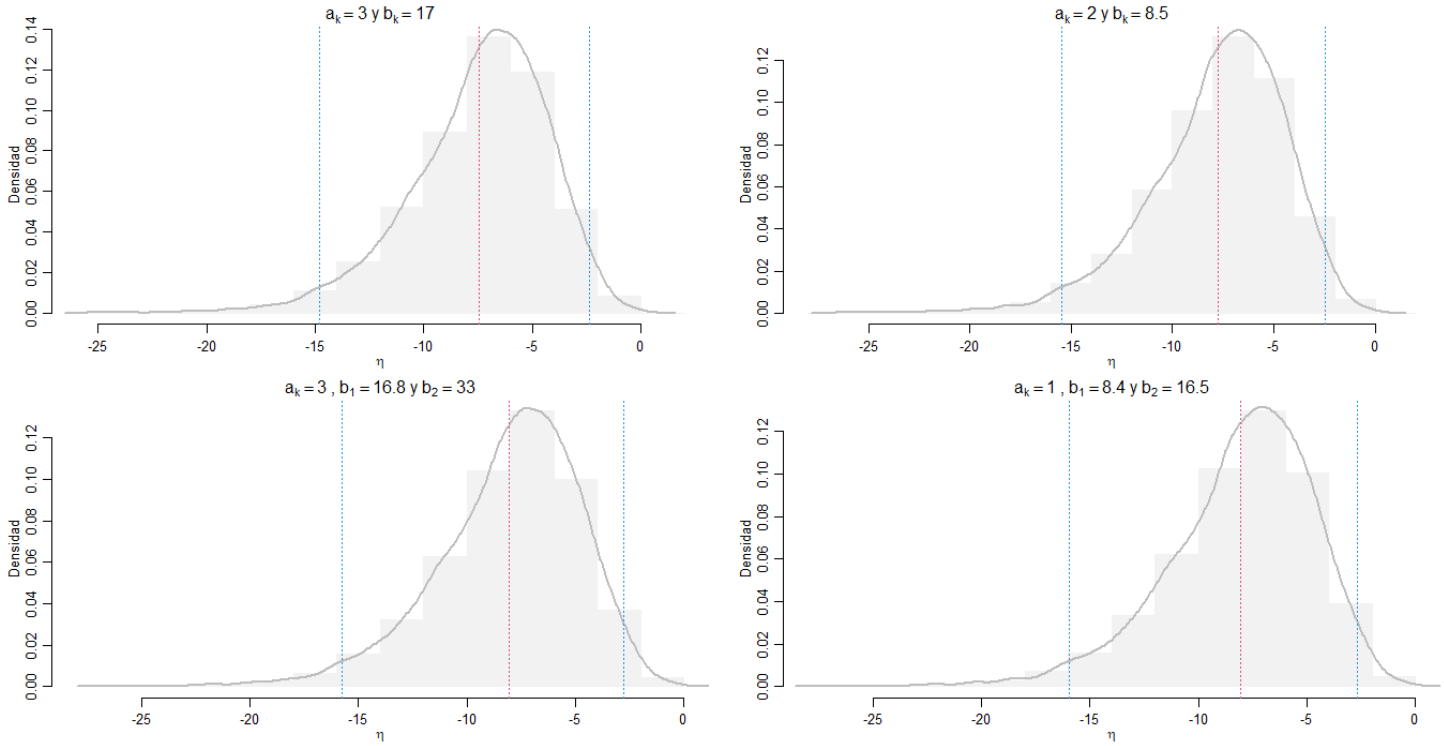

Figura 3: Distribución posterior de η

Nota: Calcule los valores p predictivos posteriores y, en cada grupo, realice la visualización de las distribuciones predictivas de los estadísticos de prueba de manera conjunta (dispersograma con histogramas marginales).

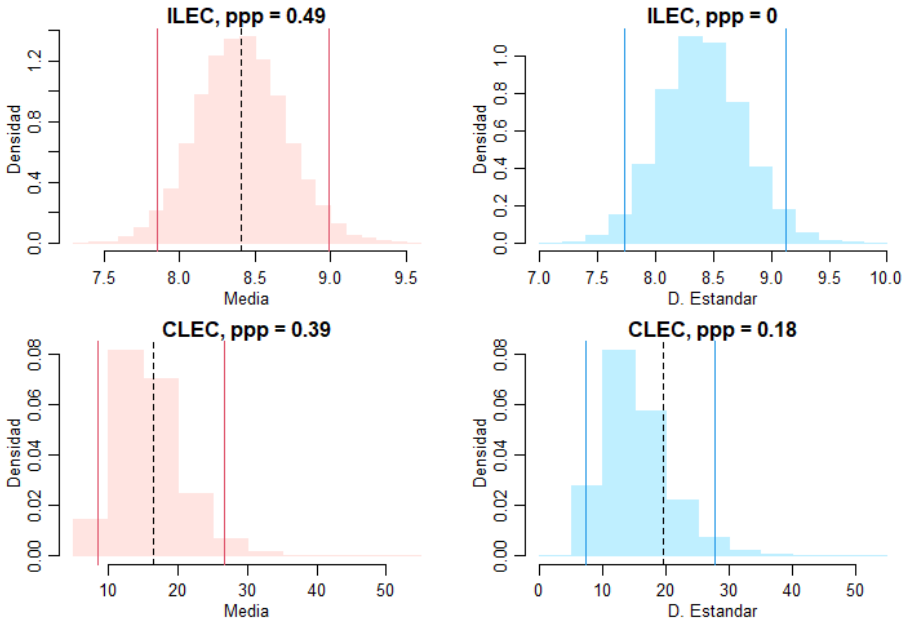

Figura 4: Histogramas Marginales

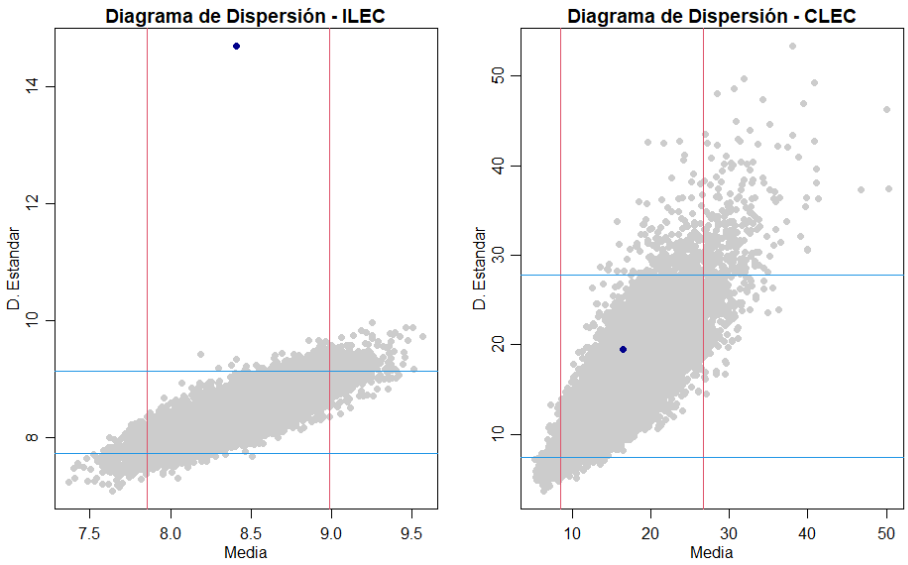

Figura 5: Dispersograma

Esto se puede ver de manera más resumida en un dispersograma con histogramas marginales.

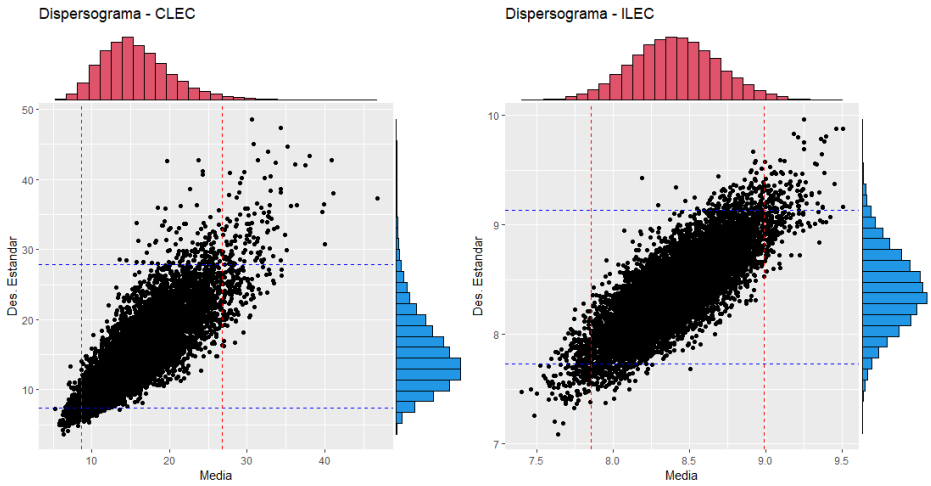

Figura 6: Dispersograma con histogramas marginales

Variable	ILEC	CLEC
Media	0.491	0.387
D. Estandar	0	0.182

Se observa que el modelo no captura adecuadamente el comportamiento de la desviación estándar, pero sí el de la media. En particular, en el caso de ILEC, subestima la desviación estándar, mientras que en CLEC, la desviación estándar no se acerca tanto a 0.5 como se desearía. No obstante, dado que toda la inferencia se basa en la media de ambas distribuciones y que el modelo no capta la desviación estándar, esto no afecta la inferencia presentada anteriormente.

5. PARTE 2: Análisis frecuentista

Repita el punto 1 de la PARTE 1 utilizando la Normalidad Asintótica del MLE (Estimador de Máxima Verosimilitud), Bootstrap paramétrico y Bootstrap no paramétrico. Presente los resultados de manera visual y tabular.Interprete los resultados obtenidos (en un máximo de 100 palabras).

Nota: Según la asintóticamente, se tiene que $\hat{\lambda}_{MLE} \approx N(\lambda, \hat{I}^{-1})$, donde $\hat{\lambda}_{MLE}$ es el MLE de λ, y $\hat{I}$ es la información observada de Fisher.

Nota: Cuando utilice Bootstrap, asegúrese de utilizar una cantidad adecuada de remuestras y el método de los percentiles para calcular los intervalos de confianza.

Para la distribucion exponencial el estimador maximo verosimil λ_k^{MLE} y su respectiva informacion de fisher observada es: (Ver Anexos 6.4).

$$\lambda_k^{\text{MLE}} = \frac{\sum_{i=1}^{n} y_{k,i}}{n} = \bar{y}_k$$
$$\hat{I}(\lambda_k^{\text{MLE}}) = \frac{n_k}{\bar{y}_k^2}$$

Por lo tanto, $\lambda_k^{\text{MLE}} \approx N(\bar{y}_k, \frac{\bar{y}_k^2}{n_k})$.

Ahora al utilizar la normalidad asintótica y la invarianza del MLE se observar que

$$\eta^{\text{MLE}} = \lambda_1^{\text{MLE}} - \lambda_2^{\text{MLE}} \approx N(\bar{y}_1 - \bar{y}_2, \frac{\bar{y}_1^2}{n_1} + \frac{\bar{y}_2^2}{n_2}).$$

Para realizar un Bootstrap no paramétrico, primero se toman n_1 muestras de ILEC y se calcula su media. Luego, se toman n_2 muestras de CLEC y se le resta la media previamente calculada de las muestras de ILEC. Este proceso se repite un total de 10,000 veces. Posteriormente, con las 10,000 estadísticas obtenidas, se hace la inferencia sobre η. De la misma manera se hace boostrap parametrico, pero en este caso generamos la muestras a partir de una distribución exponencial con parámetro λ_k^{MLE}. De esto obtenemos que:

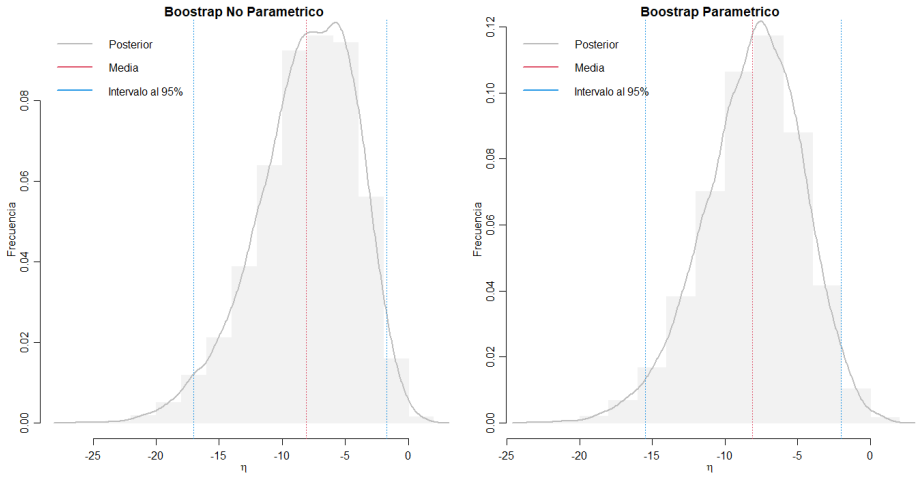

Figura 7: Boostrap

Tabla 5: Resultados de Estimación

Metodo	Media	Coef.Var	$P_{2.5}$	$P_{97.5}$
Bayesiano	-7.499	0.437	-15.227	-2.367
MLE	-8.098	0.426	-14.857	-1.338
No Parametrico	-8.093	0.490	-17.042	-1.681
Parametrico	-8.093	0.424	-15.494	-2.029

La inferencia sigue siendo la misma que en el analisis bayesiano 1.1, lo que significa que la diferencia entre los tiempos promedio de reparación es lo suficientemente grande como para considerarla significativa.Se observa que

tanto las estimaciones, los intervalos de confianza y el coeficiente de variación son muy parecidos. El método no paramétrico es el que tiene un intervalo más grande y un coeficiente de variación mayor.

PARTE 3: Simulación

Simule 100,000 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 $\lambda_2 = \bar{y}_2$.

Donde $\bar{y}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} y_{k,i}$ 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).

Se emplearon los métodos para calcular los intervalos de confianza descritos anteriormente. Además, para llevar a cabo la simulación, se utilizó el software R y sus librerías `foreach` y `doParallel` (ver código adjunto). Los resultados obtenidos son los siguientes:

Tabla 6: Resumen de Resultados para Diferentes Valores de n

n	Bayesiano	MLE	Bootstrap No Paramétrico	Bootstrap Paramétrico
10	0.93362	0.94686	0.89335	0.94097
20	0.93798	0.94726	0.91833	0.94283
50	0.94602	0.94981	0.93692	0.94837
100	0.94775	0.95008	0.94311	0.94869

Se observa que, en este caso particular, el método que capturó de mejor manera la media fue el MLE con su distribución asintótica, seguido por el Bootstrap paramétrico. En tercer lugar, el método bayesiano con los intervalos de credibilidad y, finalmente, el Bootstrap no paramétrico. Es claro que cuando la muestra es lo suficientemente grande, los cuatro métodos tendrán un rendimiento similar.

6. Anexos:

6.1 La distribución Exponencial pertenece a la familia exponencial de densidades uniparamétrica.

Sea $Y \sim \exp(\lambda)$. Así,

$$f_Y(y;\lambda) = \frac{1}{\lambda}\exp\left(-\frac{y}{\lambda}\right) I_{(0,\infty)}(y).$$

Para demostrar que Y pertenece a la familia exponencial de densidades, se verifica si puede expresarse como

$$f_Y(y;\lambda) = a(\lambda)b(y)\exp(c(\lambda)d(y))$$

para todo $y \in \mathbb{R}^n$ y $\lambda \in (0, \infty)$. Es evidente que:

$$\begin{aligned} a(\lambda) &= \frac{1}{\lambda} \\ b(y) &= I_{(0,\infty)}(y) \\ c(\lambda) &= -\frac{1}{\lambda} \\ d(y) &= y \end{aligned}$$

Por lo tanto, se concluye que la distribución Exponencial pertenece a la familia exponencial de densidades uniparamétrica.

6.2 $s_{k,i} = \sum_{i=1}^{n} y_{k,i}$ es un estadístico suficiente para λ_i.

Se observa que la función de verosimilitud es

$$L\left(\lambda|y\right) = \prod_{i=1}^{n} \frac{1}{\lambda} \exp\left(-\frac{1}{\lambda} y_i\right) I_{(0,\infty)}(y_i)$$

Dado que $y_{k,i}|\lambda_k \overset{\text{ind}}{\sim} \text{EXP}(\lambda_k)$, se puede escribir como sigue:

$$L\left(\lambda|y\right) = \frac{1}{\lambda^n} \exp\left(-\frac{1}{\lambda} \sum_{i=1}^{n} y_i\right) \prod_{i=1}^{n} I_{(0,\infty)}(y_i)$$

Finalmente, se observa la función de máxima verosimilitud como

$$L\left(\lambda|y\right) = g(t(y), \lambda) h(y)$$

con

$$\begin{aligned} g(t(y), \lambda) =& \frac{1}{\lambda^n} \exp\left(-\frac{1}{\lambda} \sum_{i=1}^{n} y_i\right) \\ h(y) =& \prod_{i=1}^{n} I_{(0,\infty)}(y_i) \\ t(y) =& \sum_{i=1}^{n} y_i \end{aligned}$$

Por lo tanto, según el Criterio de factorización de Fisher-Neyman, $s_{k,i} = \sum_{i=1}^{n} y_{k,i}$ es un estadístico suficiente para λ. También se observa que Y, al pertenecer a la familia exponencial de densidades, hace que $\sum_{i=1}^{n} d(y)$ sea una estadística suficiente para λ, es decir, $s_{k,i} = \sum_{i=1}^{n} y_{k,i}$.

6.3 Muestre que si $X \sim \text{Gamma}(\alpha, \beta)$, entonces $\frac{1}{X} \sim \text{GI}(\alpha, \beta)$.

Dado que el soporte de X está en $(0, \infty)$, se observa que $h(x) = \frac{1}{x}$ es una función monótona y diferenciable. Por lo tanto, la función de densidad de la variable aleatoria $Y = \frac{1}{X}$ es la siguiente:

$$\begin{aligned} p(y) &= \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{y}\right)^{\alpha-1} e^{-\beta \frac{1}{y}} |h'(y)| \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{y}\right)^{\alpha-1} e^{-\beta \frac{1}{y}} \frac{1}{y^2} \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{y}\right)^{\alpha+1} e^{-\frac{\beta}{y}} \end{aligned}$$

Así, se puede afirmar que $\frac{1}{X} \sim \text{GI}(\alpha, \beta)$.

6.4 Considere el modelo Bayesiano Gamma-Inversa-Exponencial dado por la distribución muestral (1) junto con la distribución previa $\lambda_k \sim \text{GI}(a_k, b_k)$, donde a_k y b_k son los hiperparámetros del modelo:

1. Represente el modelo por medio de un grafo acíclico dirigido (DAG, por sus siglas en inglés).

2. Halle la distribución posterior $p(\lambda_k|y_k)$.

Se tiene que:

$$\begin{aligned}
p(\lambda_k|y_k) &\propto p(y_k|\lambda_k)p(\lambda_k) \\
&\propto \prod_{i=1}^{n} \frac{1}{\lambda_k} \exp\left(-\frac{1}{\lambda_k} y_{k,i}\right) \frac{b_k^{a_k}}{\Gamma(a_k)} \left(\frac{1}{\lambda_k}\right)^{a_k+1} \exp\left(-\frac{b_k}{\lambda_k}\right) \\
&\propto \left(\frac{1}{\lambda_k}\right)^{n} \exp\left(-\frac{1}{\lambda_k} \sum_{i=1}^{n} y_{k,i}\right) \left(\frac{1}{\lambda_k}\right)^{a_k+1} \exp\left(-\frac{b_k}{\lambda_k}\right) \\
&\propto \left(\frac{1}{\lambda_k}\right)^{a_k+n+1} \exp\left(-\frac{1}{\lambda_k}\left(b_k + \sum_{i=1}^{n} y_{k,i}\right)\right)
\end{aligned}$$

Se puede ver que la distribución posterior $p(\lambda_k|y_k)$ es proporcional a un kernel de una distribución gama inversa. Por lo tanto, se deduce que $\lambda_k|y_k \sim \text{GI}(a_k + n, b_k + \sum_{i=1}^{n} y_{k,i})$, y se puede interpretar los hiperparámetros del modelo como a_k siendo el número de experimentos previos y b_k corresponde a la suma de los tiempos en todos los experimentos. Además, se concluye que la distribución gamma inversa es conjugada de la exponencial.

4. Muestre que la media posterior $E(\lambda_k|y_k)$ es un promedio ponderado entre la media previa $E(\lambda_k)$ y la media muestral $\bar{y}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} y_{k,i}$. Dado que se conoce la distribución posterior de λ_k, se observa que:

$$\begin{aligned}
(\lambda_k|y_k) &= \frac{b_k + \sum_{i=1}^{n} y_{k,i}}{a_k + n - 1} \\
&= \frac{a_k - 1}{a_k + n - 1} \frac{b_k}{a_k + n - 1} + \frac{n\left(\frac{1}{n}\sum_{i=1}^{n} y_{k,i}\right)}{a_k + n - 1} \\
&= \frac{a_k - 1}{a_k + n - 1} E(\lambda_k) + \frac{n}{a_k + n - 1} \bar{y}_k \\
&= \omega E(\lambda_k) + (1 - \omega)\bar{y}_k
\end{aligned}$$

Es decir, la esperanza condicional de λ_k es un promedio ponderado entre la media previa $E(\lambda_k)$ y la media muestral $\bar{y}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} y_{k,i}$.

6.5 Halle el estimador de máxima verosimilitud (MLE, por sus siglas en inglés) de λ_k y la información observada (¡no esperada!) de Fisher correspondiente.

Por los preliminares se tiene que:

$$\begin{aligned}
\lambda_k^{\text{MLE}} &= \frac{\sum_{i=1}^{n} y_{k,i}}{n} = \bar{y}_k \\
\hat{I}(\lambda_k^{\text{MLE}}) &= \frac{n_k}{\bar{y}_k^2}
\end{aligned}$$

Por lo tanto $\lambda_k^{\text{MLE}} \approx N(\bar{y}_k, \frac{\bar{y}_k^2}{n_k})$.