

Caso de estudio #1: Tiempo de falla

Sergio Andres Niño Marin, Sergio Andres Gerena Gomez

26/sept/2022

Contents

Caso de estudio	1
Pregunta 1	1
Pregunta 2	2
Punto 3	2
Punto 4	3
Punto 5	4
Punto 6	5
Punto 7	7
Punto 8	8
Punto 9	8
Punto 10	9
Analisis	9
References	9

Caso de estudio

Un investigador del Departamento de Ingeniería Electrónica y Eléctrica de una universidad necesita analizar unos datos sobre los tiempos de falla de un determinado tipo de alambre (Tipo 1). En este problema, el tiempo de falla se define como el número de veces que una máquina podría tensionar el alambre antes de romperse. Los siguientes datos corresponden a \$ n = 14 \$ tiempos de falla de una parte del experimento:

495 541 1461 1555 1603 2201 2750 3468 3516 4319 6622 7728 13159 21194

A partir de este contexto, Su incertidumbre acerca de estos datos antes de que fueran observados es intercambiable. Por lo tanto, resulta apropiado modelar los datos como condicionalmente independientes e idénticamente distribuidos. El modelo más simple para los datos del tiempo de falla involucra la distribución Exponencial:

$$y_i | \lambda \stackrel{\text{iid}}{\sim} \text{Exp}(\lambda), \text{ i.e., } p(y_i | \lambda) = \frac{1}{\lambda} \exp\left(-\frac{y_i}{\lambda}\right) \text{ para } y_i > 0 \text{ y } \lambda > 0, \text{ con } i = 1, \dots, n. \quad (1)$$

Pregunta 1

Muestre que $s = \sum_{i=1}^n y_i$ es un estadístico suficiente para λ .

Solucion:

Aplicando el criterio de factorizacion de Fisher-Neyman se observa que:

$$\begin{aligned}\mathcal{L}(\lambda|y_1, \dots, y_n) &= \prod_{i=1}^n \frac{1}{\lambda} e^{-\frac{y_i}{\lambda}} I_{(0, \infty)}(y_i) \\ &= \underbrace{\frac{1}{\lambda^n} e^{-\sum_{i=1}^n \frac{y_i}{\lambda}}}_{g(T(y), \lambda)} \overbrace{\prod_{i=1}^n I_{(0, \infty)}(y_i)}^{h(y)}\end{aligned}$$

Con esto, vemos que un estadístico suficiente es $s = \sum_{i=1}^n y_i$ dado que este estadístico cumple al poderse factorizar de la expresión de verosimilitud en una función g que solo depende del estadístico suficiente y el parámetro.

Pregunta 2

Se dice que la variable aleatoria X tiene distribución Gamma-Inversa con parámetros $\alpha > 0$ y $\beta > 0$, si la función de densidad de X está dada por:

$$X \sim \text{GI}(\alpha, \beta), \quad \text{i.e.,} \quad p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} \exp\left(-\frac{\beta}{x}\right) \quad \text{para } x > 0.$$

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

Solucion:

Siendo X una V.a absolutamente continua, se usa el teorema de la transformación para determinar la distribución de $Y = \frac{1}{X}$ con $X > 0$ a partir de la distribución conocida X

$$P\left(\frac{1}{X}\right) = f_X\left(\frac{1}{Y}\right) \left| \frac{\delta X}{\delta Y} \right| = \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{Y}\right)^{\alpha-1} \exp\left(-\frac{\beta}{Y}\right) \left| -\frac{1}{Y^2} \right|$$

Resolviendo como se indica a continuación

$$\begin{aligned}P\left(\frac{1}{X}\right) &= \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{Y}\right)^{\alpha-1} \exp\left(-\frac{\beta}{Y}\right) \left(\frac{1}{Y^2}\right) \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{Y}\right)^{\alpha+1} \exp\left(-\frac{\beta}{Y}\right) \\ P\left(\frac{1}{X}\right) &= P(Y) = \frac{\beta^\alpha}{\Gamma(\alpha)} Y^{-(\alpha+1)} \exp\left(-\frac{\beta}{Y}\right) \sim \text{GI}(\alpha, \beta)\end{aligned}$$

Dado que lo obtenido al aplicar el teorema de transformación fue la distribución descrita como Gamma inversa (GI) con los respectivos parámetros, podemos asegurar que la afirmación propuesta en el punto resulta verdadera.

Punto 3

Considere la distribución previa $\lambda \sim \text{GI}(\alpha, \beta)$ junto con la distribución muestral (1). Halle la distribución posterior de λ .

Solucion:

Para esto se emplea el teorema de bayes para encontrar la distribucion posterior de λ a partir de la distribucion muestral y la distribucion previa.

$$\begin{aligned}
 p(\lambda|\vec{y}) &\propto p(\vec{y}|\lambda)p(\lambda) \\
 &\propto \left[\prod_{i=1}^n \frac{1}{\lambda} \exp\left(-\frac{y_i}{\lambda}\right) \right] \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{-(\alpha+1)} \exp\left(-\frac{\beta}{\lambda}\right) \\
 &\propto \lambda^{-n} \exp\left(-\frac{1}{\lambda} \sum_{i=1}^n y_i\right) \lambda^{-(\alpha+1)} \exp\left(-\frac{\beta}{\lambda}\right) \\
 p(\lambda|\vec{y}) &\propto \lambda^{-(\alpha+n+1)} \exp\left(-\frac{(\beta+s)}{\lambda}\right) \sim GI(\alpha+n, \beta+s)
 \end{aligned}$$

Punto 4

Se tiene información externa de otro experimento de acuerdo con el cual la distribución previa de λ debería tener una media $\mu_0 = 4500$ y una desviación estándar $\sigma_0 = 1800$. Haga un gráfico de las distribuciones previa y posterior en el mismo gráfico.

Solucion:

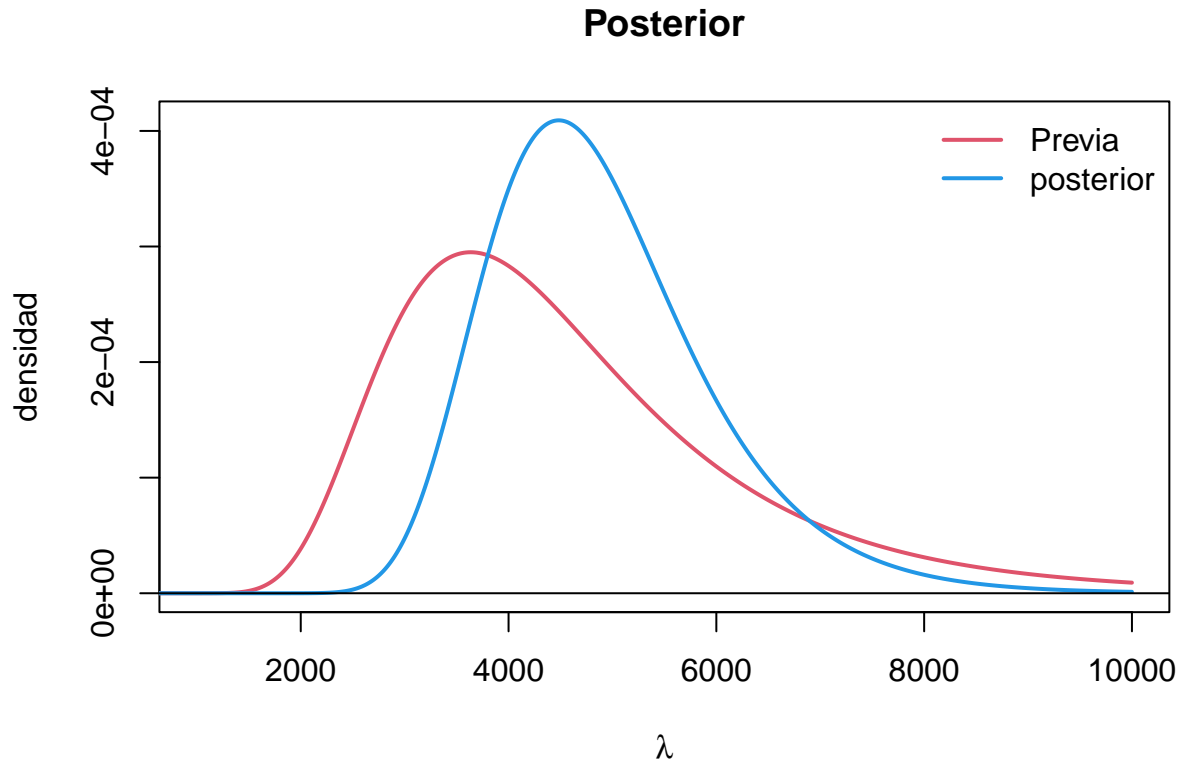
Lo primero es, encontrar los hiperparametros de la distribucion previa tal que satisfaga la media $\mu_0 = 4500$ y la desviacion estandar $\sigma_0 = 1800$, dando el siguiente sistema de ecuaciones:

$$\begin{aligned}
 4500 &= \frac{\beta_0}{\alpha_0 - 1} & \text{para } \alpha_0 > 1 \\
 1800 &= \sqrt{\frac{\beta_0^2}{(\alpha_0 - 1)^2(\alpha_0 - 2)}} & \text{para } \alpha_0 > 2
 \end{aligned}$$

Resolviendo el sistema de ecuaciones, tenemos que $\alpha_0 = 8.25$ y $\beta_0 = 32625$. con esto ya se puede construir la distribucion previa.

Ahora, en base al punto anterior donde se calculo la distribucion posterior para este modelo sigue la forma $GI(\alpha+n, \beta+s)$ donde n es el tamaño de la muestra y s es la suma de todos los y_i . Empleando una suma sobre los 14 terminos que se tiene en los datos, resulta que $n = 14$ y $s = 70612$.

Ahora sigue construir y graficar ambas distribuciones, tanto la previa como la posterior, usando codigo en R para realizar dicha accion, el resultado es el observado a continuacion



Punto 5

Halle el estimador de máxima verosimilitud (MLE, por sus siglas en inglés) de λ .

Solucion:

Para encontrar el estimador de maxima verosimilitud, lo primero es calcular la funcion de verosimilitud del parametro dada la muestra y a dicha funcion se le encuentra el maximo global, pero antes, se usa la parametrizacion de la exponencial donde $\theta = \frac{1}{\lambda}$, luego de encontrar el MLE para dicho parametro, se emplea la invarianza funcional del estimador MLE para encontrar el estimador MLE de λ .

Se empieza encontrando la funcion de verosimilitud del parametro θ de la siguiente manera

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

Para simplificar la notacion y tambien las matematicas en los siguientes pasos , se usa como convencion que $s = \sum_{i=1}^n$ ademas de que en el proceso de maximizacion se emplea la funcion log-verosimilitud $\uparrow(\theta|\vec{y})$ la cual es el logaritmo natural de la anteriormente encontrada.

Siguiendo con la maximización de la función, primero se encuentran los puntos críticos empleando la primera derivada

$$\begin{aligned}\frac{\delta}{\delta\theta}l(\theta|\vec{y}) &= \frac{\delta}{\delta\theta}(n * \ln(\theta) - \theta s) \\ &= \frac{n}{\theta} - s \\ 0 &= \frac{n}{\hat{\theta}} - s \\ \hat{\theta} &= \frac{n}{s}\end{aligned}$$

para evaluar el punto crítico encontrado se aplica el criterio de la segunda derivada, en el tenemos como criterio especial que si la función es siempre negativa en todo el soporte, el punto encontrado es un máximo global.

$$\frac{\delta}{\delta\theta}\left(\frac{n}{\theta} - s\right) = -\frac{n}{\theta^2}$$

Se tiene que la función obtenida es siempre negativa ya que $n > 0$ al igual que $\theta^2 > 0$, por lo que el punto crítico encontrado no solo es un máximo local sino también, un máximo global, al ser así se tiene entonces que $\hat{\theta}_{MLE} = \frac{n}{s}$.

Ahora trabajamos con la invarianza funcional del estimador MLE para encontrar el estimador de máxima verosimilitud de $\lambda = \frac{1}{\theta}$, tenemos entonces que.

$$\begin{aligned}g(\hat{\theta})_{MLE} &= g(\hat{\theta}_{MLE}) \\ \hat{\lambda}_{MLE} &= \frac{1}{\hat{\theta}_{MLE}} = \frac{1}{\frac{n}{s}} \\ \hat{\lambda}_{MLE} &= \frac{s}{n}\end{aligned}$$

Punto 6

Complete la siguiente tabla:

Método	Estimación	CV (%)	Intervalo al 95%
Bayesiano			
Frec. Asintótico			
Frec. Bootstrap			

Table 1: Inferencia Bayesiana y Frecuentista sobre λ .

Para completar la Tabla 1 tenga en cuenta que:

- Bayesiano: Inferencia Bayesiana basada en la distribución posterior.
- Frec. Asintótico: Inferencia frecuentista basada en que $\hat{\lambda}_{MLE} \approx N(\lambda, \hat{I}^{-1})$ siempre que $n \rightarrow \infty$, donde $\hat{\lambda}_{MLE}$ es el MLE de λ y \hat{I} es la información observada de Fisher.
- Frec. Bootstrap: Inferencia frecuentista basada en Bootstrap no paramétrico.

Solución:

Enfoque Bayesiano Para completar la fila mediante el método bayesiano, se parte de la distribución posterior encontrada anteriormente $GI(22.25, 104237)$, Donde para crear una estimación puntual del parámetro

se hace uso de la media de la distribucion posterior $\mu_p = \frac{\beta}{\alpha-1}$. Para el coeficiente de variacion se tiene que $CV\% = \frac{\sigma}{\mu} * 100$ donde σ es la desviacion estandar de la distribucion posterior y μ la media de la misma.

Para el intervalo, que en el panorama estadistico se denomina region de confianza, se plantea un (l, u) tal que $p(l < \lambda < u | \vec{y}) = 0.95$, asi para encontrar los limites del intervalo se emplean los cuantiles de la distribucion posterior tal que $l = \lambda_{0.025}$ y $u = \lambda_{0.975}$ de forma tal que se encierre la probabilidad pedida.

Realizando estos calculos empleando R se llega a los siguientes resultados:

- $\mu_p = 4638.865$
- $CV\% = 23.27\%$
- $IC_{95\%} = (3186, 7382)$

Enfoque frecuentista asintotico Para la parte frecuentista asintotico, se parte de la base que $\hat{\lambda}_{MLE} \approx N(\lambda, \hat{I}^{-1})$. Para la estimacion puntual esto no afecta ya que se emplea el estimador encontrado en el punto 5, que es igual a la media muestral.

Para el CV se plantea primero, gracias a la libreria mle.tools, encontrar la informacion de fisher para el parametro lambda, una vez realizado esto, se tiene que $CV\% = \frac{\sqrt{\frac{1}{I_n}}}{\hat{\lambda}_{MLE}} * 100$.

Por ultimo, en el intervalo de confianza, ya que se tiene la distribucion asintotica, se plantea la variable pivote de forma tal que $Q(\vec{y}, \lambda) = \sqrt{I_n}(\hat{\lambda}_{MLE} - \lambda) \approx N(0, 1)$, de la ecuacion presentada se despeja λ y se acotan los intervalos de la siguiente manera

$$p\left(z_{0.025} < \sqrt{I_n}(\hat{\lambda}_{MLE} - \lambda) < z_{0.975}\right) = 0.95$$

$$p\left(-\frac{z_{0.975}}{\sqrt{I_n}} + \hat{\lambda}_{MLE} < \lambda < -\frac{z_{0.025}}{\sqrt{I_n}} + \hat{\lambda}_{MLE}\right) = 0.95$$

Ya con este intervalo y el codigo en r, se calculan las fronteras del mismo, llegando al resultado.

A continuacion se muestra la lista de los datos pedidos

- $\hat{\lambda}_{MLE} = 5043.71$
- $CV\% = 26.73\%$
- $IC_{95\%} = (2402, 7686)$

0.0.0.1 Enfoque frecuentista bootstrap En este, se generan varios samples con remplazo del mismo tamaño de los datos, esto con el fin de generar un vector de tamaño 10000 con 10000 muestras cradas a partir de samples random de los datos.

Lo que se hace con este vector es sacarle la media a cada conjunto de datos creado, esto debido a que en el paradigmaa frecuentista, el estimador MLE para el parametro era la media muestral, de est manera se obtiene un vector de medias de tamaño 10000, uno por cada muestra generada mediante bootstrap.

Para los datos pedidos en la tabla, se hace inferencia respecto a ese vector de medias, la estimacion puntual del parametro seria la media del vector de medias, para el CV, se toma la varianza del vector de media asi como su media, con la varianza empirica del vector obtenemos la desviacion estandar empirica y siguiendo la formula $CV\% = \frac{\sigma_{empirica}}{\mu_{empirico}} * 100$ se calcula el CV.

Para el intervalo de confianza, se manejan los cuantiles del vector de medias, esto debido a que estamos usando este mismo como una distribucion del verdadero parametro, los cuantiles tomados tal que satisfagan los requerimientos de confianza pedidos son el 0.025 y 0.975.

A continuacion se presenta los valores calculados por este metodo:

- $puntual = 5053.89$
- $CV\% = 29.5\%$
- $IC_{95\%} = (2558.7, 8283.9)$

La tabla completa se puede ver a continuacion

Método	Estimación	CV (%)	Intervalo al 95%
Bayesiano	4638.865	23.27%	(3186,7382)
Frec. Asintótico	5043.71	26.73%	(2402,7686)
Frec. Bootstrap	5053.89	29.5%	(2558.7,8283.9)

Table 2: Inferencia Bayesiana y Frecuentista sobre λ Completa.

Punto 7

Calcule e interprete $\Pr(\lambda < 4000 \mid \mathbf{y})$ y $\Pr(y^* < 4000 \mid \mathbf{y})$, donde y^* es un tiempo de falla futuro.

Solucion:

El primero resulta ser la probabilidad de que lambda sea menor a 4000 una vez se observaron los datos, se sabe que esto se cuantifica con la distribucion posterior. Para el calculo de esta cantidad se emplea la libreria invgamma, donde se encuentra pinvgamma para calcular la probabilidad acumulada hasta cierto valor, la funcion que se envia es la que ya se encontro para la distribucion posterior.

La segunda probabilidad pedida es la de prediccion, esta quiere decir cual es la probabilidad de que un dato no observado (que no es parte de la muestra) sea menor a 4000, como se observa, ambas son cantidades distintas ya que hacen inferencia sobre cantidades distintas, la distribucion posterior hace inferencia sobre el parametro en cambio, la funcion predictiva posterior hace referencia a el valor que puede tomar una unidad no observada, es decir, la probabilidad de que el tiempo de falla de un nuevo alambre sea menor a 4000.

Para calcular la probabilidad acumulada de la distribucion predictiva primero hay que encontrar la misma, para eso se usa la siguiente ecuacion:

$$\begin{aligned}
p(y^* \mid \vec{y}) &= \int_{\Theta} p(y^* \mid \lambda) p(\lambda \mid \vec{y}) d\lambda \\
&= \int_{\Theta} \frac{1}{\lambda} e^{-\frac{y^*}{\lambda}} \frac{(\beta + s)^{\alpha+n}}{\Gamma(\alpha + n)} \lambda^{-(\alpha+n+1)} \exp\left(-\frac{(\beta + s)}{\lambda}\right) d\lambda \\
&= \frac{(\beta + s)^{\alpha+n}}{\Gamma(\alpha + n)} \int_{\Theta} \lambda^{-(\alpha+n+1+1)} e^{\left(-\frac{(\beta+s+y^*)}{\lambda}\right)} d\lambda \\
&= \frac{(\beta + s)^{\alpha+n}}{\Gamma(\alpha + n)} \frac{\Gamma(\alpha + n + 1)}{(\beta + s + y^*)^{(\alpha+n+1)}} \\
&= \frac{(\beta + s)^{(\alpha+n)} (\alpha + n) \Gamma(\alpha + n)}{\Gamma(\alpha + n) (\beta + s + y^*)^{(\alpha+n)} (\beta + s + y^*)} \\
p(y^* \mid \vec{y}) &= \left(\frac{\beta + s}{\beta + s + y^*} \right)^{\alpha+n} \frac{\alpha + n}{\beta + s + y^*}
\end{aligned}$$

Ya con la funcion de densidad de la predictiva posterior, se calcula $\Pr(y^* < 4000 \mid \mathbf{y})$ aplicando la integral sobre esta desde 0 hasta 4000, ya que los valores que puede tomar la variable son los positivos dada la

definición de la misma, esta integral se observa a continuación.

$$\begin{aligned}
\Pr(y^* < 4000 \mid \mathbf{y}) &= \int_0^{4000} \left(\frac{\beta + s}{\beta + s + y^*} \right)^{\alpha+n} \frac{\alpha + n}{\beta + s + y^*} dy^* \\
&= (\beta + s)^{\alpha+n} (\alpha + n) \int_0^{4000} \frac{1}{(\beta + s + y^*)^{\alpha+n+1}} dy^* \\
&= 103237^{22.25} 22.25 \int_0^{4000} \frac{1}{(103237 + y^*)^{23.25}} dy^* \\
&= 103237^{22.25} 22.25 \int_{103237}^{107237} u^{-23.25} du \\
&= 103237^{22.25} 22.25 \left(\frac{-1}{22.25(107237)^{22.25}} + \frac{1}{22.25(103237)^{22.25}} \right) \\
&= - \left(\frac{103237}{107237} \right)^{22.25} + 1 \\
\Pr(y^* < 4000 \mid \mathbf{y}) &= 0.5708
\end{aligned}$$

Empleando el código de R se obtuvo que $\Pr(\lambda < 4000 \mid \mathbf{y}) = 0.2155$, como se mencionaba anteriormente, ambas cantidades no son iguales por lo que son inferencia sobre cosas distintas como se explicaba anteriormente, estos resultados dan soporte de la afirmación mencionada anteriormente.

Punto 8

Pruebe el sistema de hipótesis $H_0 : \lambda = \lambda_0$ frente a $H_1 : \lambda \neq \lambda_0$, con $\lambda_0 = 4000$. Para ello tenga en cuenta que

$$p(\mathbf{y} \mid H_0) = \int_0^\infty \lambda_0^{-n} \exp\left(-\frac{1}{\lambda_0} \sum_{i=1}^n y_i\right) \delta_{\lambda_0}(\lambda) d\lambda$$

y

$$p(\mathbf{y} \mid H_1) = \int_0^\infty \lambda^{-n} \exp\left(-\frac{1}{\lambda} \sum_{i=1}^n y_i\right) \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \lambda^{-(\alpha_0+1)} \exp\left(-\frac{\beta_0}{\lambda}\right) d\lambda$$

donde $\delta_a(x)$ es la función delta de Dirac. Reporte el factor de Bayes B_{10} e interprete los resultados.

Solución:

Punto 9

Experimentación adicional bajo las mismas condiciones con otro tipo de alambre (Tipo 2) produjo los siguientes resultados:

294 569 766 1576 1602 2015 2166 3885 8141 10285

Considerando modelos independientes de la forma $y_{i,k} \mid \lambda_k \stackrel{\text{iid}}{\sim} \text{Exp}(\lambda_k)$ con $\lambda_k \sim \text{GI}(\alpha_0, \beta_0)$, para $i = 1, \dots, n_k$ y $k = 1, 2$, donde $y_{i,k}$ es el tiempo de falla del alambre i de tipo k , y n_k es el número de alambres de tipo k sometidos a experimentación (la distribución previa es la misma para ambos tipos de alambre). Pruebe el sistema de hipótesis $H_0 : \lambda_1 = \lambda_2$ frente a $H_1 : \lambda_1 \neq \lambda_2$. Reporte el factor de Bayes B_{10} e interprete los resultados.

Solucion:

Punto 10

Verifique la idoneidad del modelo para ambos tipos de alambre empleando como estadística de prueba la media del tiempo de falla. Presente sus resultados gráficamente comparando la distribución predictiva posterior con el valor observado correspondiente. Así mismo, reporte el valor p predictivo posterior en cada caso.

Solucion:

Analisis

References

<!-- placeholder for References in toc --!>