Facultad de Ciencias
Departamento de Estadística

Sergio Andrés Díaz Vera
Samuel Ruiz Martínez

UNIVERSIDAD NACIONAL DE COLOMBIA

Estadística Bayesiana: Caso de Estudio 1

septiembre 2022

Tiempo de falla

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:

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

Preguntas

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

   Desarrollo :

   Se tiene que $y_i \mid \lambda$ se distribuye exponencial de parámetro λ es decir

   $y_i \mid \lambda \overset{\text{iid}}{\sim} \mathsf{Exp}(\lambda)$, i.e., $p(y_i \mid \lambda) = \frac{1}{\lambda} \exp\left(-\frac{y_i}{\lambda}\right)$ para $y_i > 0$ y $\lambda > 0$, con $i = 1, \ldots, n$.

   entonces definimos $s = \sum_{i=1}^{n} y_i$,veamos que es estadístico suficiente para λ si se cumple el criterio de factorización de Fisher-Neyman

   $$p(y_1, y_2, ..., y_n \mid \lambda) = g(T(y_1, y_2, ..., y_n) \mid \lambda) * h(y_1, y_2, ..., y_n)$$

   entonces :

   $$\begin{aligned} p(y_1, y_2, ..., y_n \mid \lambda) =& \frac{1}{\lambda^n} e^{\frac{-1}{\lambda}(\sum_{i=1}^{n} y_i)} \\ =& \underbrace{(\frac{1}{\lambda^n} e^{\frac{-1}{\lambda}(s)})}_{g(T(y_1,y_2,...,y_n)|\lambda)} * \overbrace{(1)}^{h(y_1,y_2,...,y_n)} \end{aligned}$$

   Puesto que es posible lograr la factorización de Fisher-Neyman entonces s es un estadístico suficiente para λ.

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:

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

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

   Desarrollo:

   Utilizaremos el teorema de la transformacion, suponga $Y = 1/X$, $X \sim Gamma(\alpha, \beta)$.

   $$\begin{aligned} f_Y(y) =& f_X(1/y) \left(\frac{d}{dy} y^{-1}\right) \\ =& \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{-\alpha+1} exp(-1/\beta y) y^{-2} \\ =& \frac{(1/\beta)^\alpha}{\Gamma(\alpha)} y^{-\alpha-1} exp(-(1/\beta)/y) \end{aligned}$$

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

   Desarrollo:
   Aplicando el teorema de Bayes se tiene que

$$\begin{aligned} p(\lambda \mid \boldsymbol{y}) \propto p(\lambda)p(\boldsymbol{y} \mid \lambda) =& \prod_{i=1}^{n} p(y_i \mid \lambda)p(\lambda) \\ =& \prod_{i=1}^{n} \frac{1}{\lambda}\exp\left(-\frac{y_i}{\lambda}\right) \times \frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}exp\left(-\frac{\beta}{\alpha}\right) \\ \propto& \lambda^{-n}\exp\left(-\frac{s}{\lambda}\right) \times \lambda^{-(\alpha+1)}\exp\left(-\frac{\beta}{\lambda}\right) \\ =& \lambda^{-(\alpha+n+1)}\exp\left(-\frac{\beta+s}{\lambda}\right) \end{aligned}$$

   Donde $s = \sum_{i=1}^{n} y_i$ y $\boldsymbol{y} = (y_1, \ldots, y_n)$. Dado que $p(\lambda \mid \boldsymbol{y})$ es proporcional a el núcleo de una distribución gamma inversa de parámetros $\alpha + n$ y $\beta + s$, se concluye que la distribución posterior $\lambda \mid \boldsymbol{y} \sim GI(\alpha + n, \beta + s)$

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.

   Desarrollo :

   Es necesario observar el sistema formado por la media y la varianza de la distribución gamma Inversa de la siguiente manera :

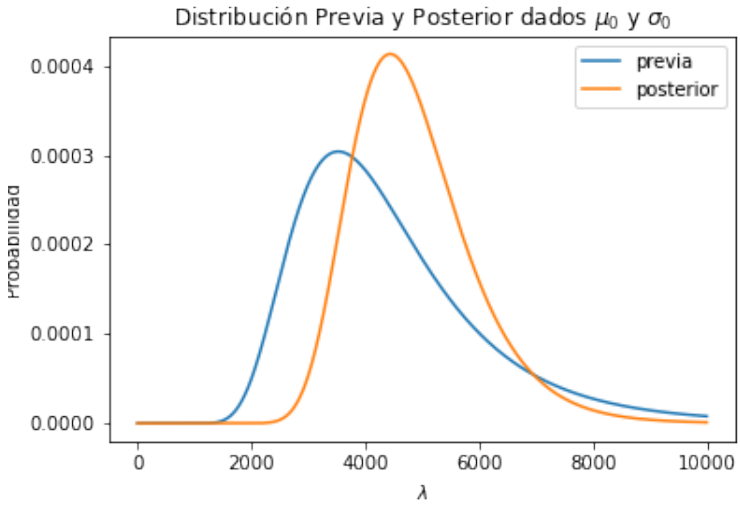
$$\begin{cases} \mu_0 = \frac{\beta}{\alpha-1} & = 4500 \longrightarrow \beta = 4500(\alpha - 1) \quad (1) \\ \sigma_0 = \sqrt{\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}} & = 1800 \quad (2) \end{cases}$$

   Resolviendo el sistema y usando la ecuación (1) en la ecuación (2)

$$\begin{aligned} 1800 =& \sqrt{\frac{(4500(\alpha-1)^2)}{(\alpha-1)^2(\alpha-2)}} \\ (\alpha-2)1800^2 =& 4500^2 \\ \alpha = 33/4 \longrightarrow \beta =& 4500(33/4 - 1) = 32625 \end{aligned}$$

Entonces $n = 14$ y $s = 70612$, luego ,la distribución previa será $\lambda \sim GI(\frac{33}{4}, 32625)$ y la distribución posterior será $\lambda \mid \boldsymbol{y} \sim GI(\frac{33}{4} + 14, 32625 + 70612) = (22{,}25, 103237)$

La forma visual de estas distribuciones es :



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

Desarrollo :

Suponemos una muestra de tamaño n de una sucesión de variables aleatorias $x_1, \ldots, x_n$, un vector de parámetros $\lambda = (\lambda_1, \ldots, \lambda_k)$ que serán las proporciones de cada realización x_i dentro de la muestra y el vector aleatorio del conteo de cada realización en la muestra $\mathbf{n}= (n_1, \ldots, n_k)$, la función de densidad conjunta de las realizaciones será la función de verosimilitud y la denotaremos como

$$L(\lambda; x_1, \ldots, x_n) = \prod_{i=1}^{n} f_X(x_i, \lambda)$$

El estimador MLE de λ será $\hat{\lambda}$ que maximiza la función de verosimilitud. Usando las herramientas de calculo diferencial para maximizar debemos derivar e igualar a cero y para ayudarnos a reducir los cómputos maximizaremos la log-verosimilitud $(\ell(\lambda; x_1, \ldots, x_n))$ dado que el máximo de esta se alcanza en el mismo punto que $L(\lambda; X)$. Dado que $\lambda_k = 1 - \sum_{i=1}^{k-1} \frac{n_j}{n_k} \lambda_k$ implica que λ_k es una combinación lineal de los demás, entonces existen $k-1$ parámetros independientes entre si. Por lo tanto la función de log-verosimilitud es

$$\ell(\lambda; X) = ln(L(\lambda_1, \ldots, \lambda_{k-1}; X) = \sum_{i=1}^{n} n_i ln(\lambda_j)$$

Por lo tanto para $j = 1, 2, \ldots, k-1$, derivando e igualando a cero tenemos que

$$\frac{n_j}{\lambda_j} - \frac{n_k}{\lambda_k} = 0 \rightarrow \lambda_j = \frac{n_j}{n_k}\lambda_k \tag{1}$$

Ahora observe que

$$\lambda_k = 1 - \sum_{j=1}^{k-1} \frac{n_j}{n_k}\lambda_k \rightarrow \lambda_k = \frac{n_k}{n}$$

Por lo tanto, reemplazando λ_k en (1) tenemos que $\lambda_j = n_j/n$. Entonces el punto crítico es

$$\lambda_{MLE} = (n_1/n, \ldots, n_k/n)$$

6. Complete la siguiente tabla:

   Realizando lo métodos sugeridos se tiene que :

Método	Estimación	CV (%)	Intervalo al 95 %
Bayesiano	4862.243	22.2 %	(3196.974, 7356.66)
Frec. Asintótico	5043.714	110 %	(-5849.46, 15936.89)
Frec. Bootstrap	5039.758	29 %	(2516.577, 8189.007)

Cuadro 1: Inferencia Bayesiana y Frecuentista sobre λ.

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

Desarrollo :

Recordando que la distribución a posterior es $\lambda \mid \boldsymbol{y} \sim GI(22{,}25, 103237)$ y la distribución no es conocida ,utilizamos la solución de la distribución posterior y de la predictiva posterior a través de los métodos de Montecarlo Entonces

$$\mathsf{Pr}(\lambda < 4000 \mid \boldsymbol{y}) = 0{,}2151$$

la probabilidad de que λ el tiempo de falla medio sea inferior a 4000 es de 21,51 %y

$$\mathsf{Pr}(y^* < 4000 \mid \boldsymbol{y}) = 0{,}5708$$

la probabilidad de que un nuevo tiempo de falla que se incorpore a los datos sea menor a 4000 es del 57,08 %

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

$$p(\boldsymbol{y} \mid H_0) = \int_0^\infty \lambda_0^{-n} \exp\left(-\tfrac{1}{\lambda_0} \textstyle\sum_{i=1}^n y_i\right) \delta_{\lambda_0}(\lambda)\, \mathrm{d}\lambda$$

y

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

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

Desarrollo:

Veamos que para reportar el factor de Bayes del sistema de hipótesis

$$H_0 : \lambda = \lambda_0$$
$$H_1 : \lambda \neq \lambda_0$$

con $\lambda_0 = 4000$, es necesario $B_{10} = \frac{p(\boldsymbol{y}|H_1)}{p(\boldsymbol{y}|H_0)}$ entonces calculamos:

$$\begin{aligned} p(\boldsymbol{y} \mid H_0) = &\int_0^\infty \lambda_0^{-n} \exp\left(-\tfrac{1}{\lambda_0} \textstyle\sum_{i=1}^n y_i\right) \delta_{\lambda_0}(\lambda)\, \mathrm{d}\lambda \\ =&\lambda_0^{-n} \exp\left(-\tfrac{1}{\lambda_0} \textstyle\sum_{i=1}^n y_i\right) \int_0^\infty \delta_{\lambda_0}(\lambda)\, \mathrm{d}\lambda \\ =&\lambda_0^{-n} \exp\left(-\tfrac{1}{\lambda_0} \textstyle\sum_{i=1}^n y_i\right) I_{(0,\infty)}(\lambda) \end{aligned}$$

y

$$\begin{aligned} p(\boldsymbol{y} \mid H_1) = &\int_0^\infty \lambda^{-n} \exp\left(-\tfrac{1}{\lambda} \textstyle\sum_{i=1}^n y_i\right) \dfrac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \lambda^{-(\alpha_0+1)} \exp\left(-\dfrac{\beta_0}{\lambda}\right) \mathrm{d}\lambda \\ =&\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \int_0^\infty \lambda^{-(\alpha_0+n+1)} \exp\left(-\frac{(\beta_0+s)}{\lambda}\right) d\lambda \\ =&\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \frac{\Gamma(\alpha_0+n)}{(\beta_0+s)^{(\alpha_0+n)}} \end{aligned}$$

con $s = \sum_{i=1}^{n} y_i = 70612$, $\alpha_0 = 33/4$,$n = 14$, $\beta_0 = 32625$ y $\lambda_0 = 4000$ entonces tenemos que el factor de bayes B_{10} es :

$$\begin{aligned} B_{10} =& \frac{p(\boldsymbol{y} \mid H_1)}{p(\boldsymbol{y} \mid H_0)} \\ =& \frac{\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \frac{\Gamma(\alpha_0+n)}{(\beta_0+s)^{(\alpha_0+n)}}}{\lambda_0^{-n} \exp\left(-\frac{1}{\lambda_0} s\right)} \end{aligned}$$

Usando el logaritmo natural para hallar el factor

$$\begin{aligned} B_{10} =& \exp\left((-n \log \lambda_0 - \frac{s}{\lambda_0}) - ([\alpha_0 \log \beta_0 - \log \Gamma(\alpha_0)] + [\log \Gamma(\alpha_0 + n) - (\alpha_0 + n) \log(\beta_0 + s)]) \right) \\ =& 0{,}78239 \end{aligned}$$

Siguiendo a Kass (1995), B_{10} nos indica que no vale mas que una mención acerca de que no hay una evidencia en contra de H_0 es decir no hay evidencia para rechazar que $\lambda = 4000$

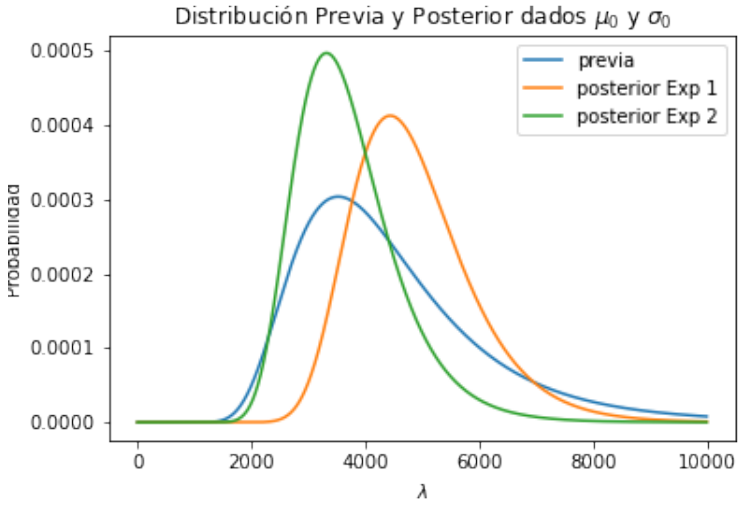
9. Experimentación adicional bajo las mismas condiciones con otro tipo de alambre (Tipo 2) produjo los siguientes resultados:

   294 569 766 1576 1602 2015 2166 3885 8141 10285

   Considerando modelos independientes de la forma $y_{i,k} \mid \lambda_k \overset{\text{iid}}{\sim} \mathsf{Exp}(\lambda_k)$ con $\lambda_k \sim \mathsf{GI}(\alpha_0, \beta_0)$, para $i = 1, \ldots, n_k$ 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.

   Desarrollo:

   Observamos que el alambre de Tipo 2, posee una distribución posterior para λ Gamma Inversa de parámetros $\alpha + n_2$ y $\beta_0 + s_2$, es decir,$\lambda_2 \mid y_{i,2} \sim GI(\alpha + n_2, \beta_0 + s_2)$ con $\alpha_0 = 33/4, \beta_0 = 32625, n_2 = 10, s_2 = 31299$ comparando los modelos



Probaremos el sistema de hipótesis $H_0 : \lambda_1 = \lambda_2$ frente $H_1 : \lambda_1 \neq \lambda_2$, en este caso se tiene que para la simulación:

$$\begin{aligned} p(\boldsymbol{y} \mid H_0) =& p(y_1, y_2 \mid \lambda_1 = \lambda_2) \\ =& p(y_1, y_2 \mid \lambda) \\ =& p(y_1 \mid \lambda) p(y_2 \mid \lambda) \\ =& \lambda^{-n_1} \exp\left(-\frac{s_1}{\lambda}\right) \lambda^{-n_2} \exp\left(-\frac{s_2}{\lambda}\right) \\ =& \lambda^{-(n_1+n_2)} \exp\left(-\frac{(s_1+s_2)}{\lambda}\right) \\ =& 1{,}577\text{x}10^{-98} \end{aligned}$$

$$\begin{aligned} p(\boldsymbol{y} \mid H_1) =& p(y_1, y_2 \mid \lambda_1 \neq \lambda_2) \\ =& p(y_1, y_2 \mid \lambda) \\ =& p(y_1 \mid \lambda_1) p(y_2 \mid \lambda_2) \\ =& \lambda_1^{-n_1} \exp\left(-\frac{s_1}{\lambda_1}\right) \lambda_2^{-n_2} \exp\left(-\frac{s_2}{\lambda_2}\right) \\ =& 1{,}86\text{x}10^{-98} \end{aligned}$$

Así tenemos que el factor de bayes será :

$$B_{10} = \frac{p(\boldsymbol{y} \mid H_0)}{p(\boldsymbol{y} \mid H_1)} = \ln p(\boldsymbol{y} \mid H_0) - \ln p(\boldsymbol{y} \mid H_1) = 1{,}18$$

Siguiendo a Kass (1995), B10 nos indica que no vale mas que una mención acerca de que no hay una evidencia en contra de H_0 es decir no hay evidencia para rechazar que $\lambda_1 = \lambda_2$ es decir los experimentos provienen de poblaciones de igual parámetro λ

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.

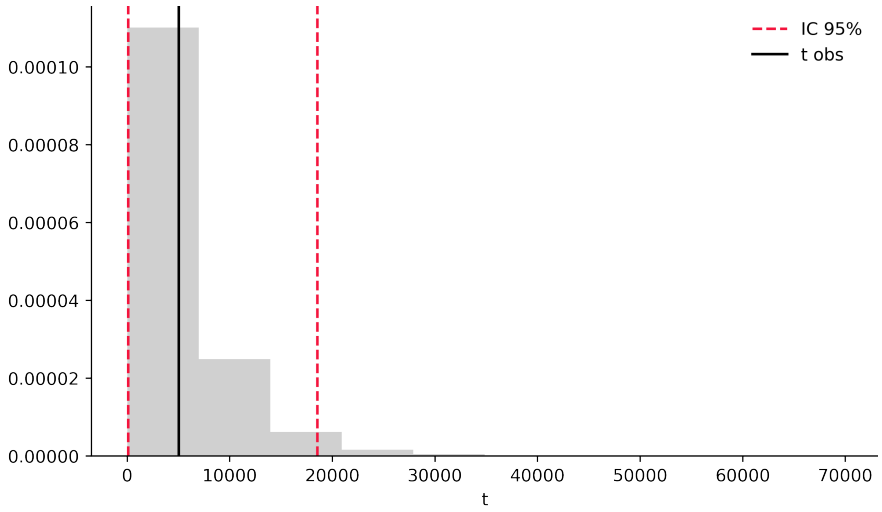


Figura 1: valor observado vs distribución predictiva posterior para los datos del primer alambre

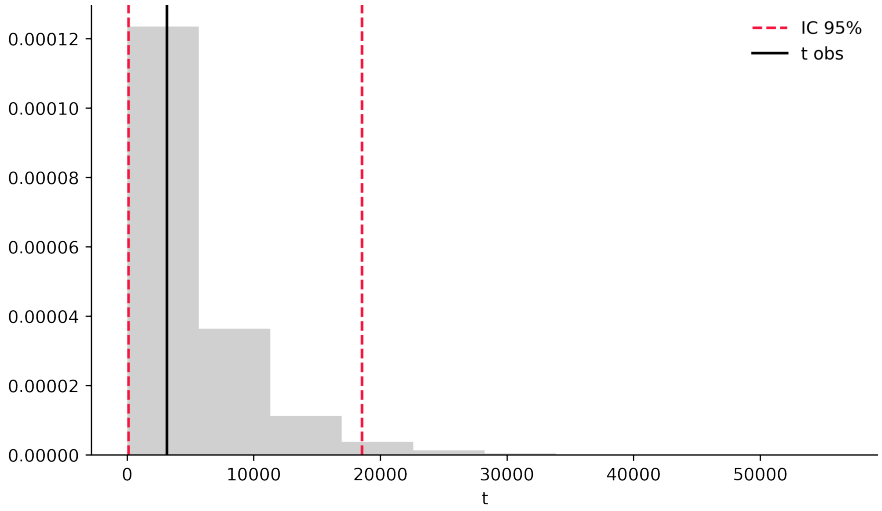


Figura 2: valor observado vs distribución predictiva posterior para los datos del segundo alambre

Los valores ppp para cada uno de los conjuntos de datos fueron 0,34492 y 0,51274 respectivamente.