

ANÁLISIS DE DATOS CENSURADOS PARA INGENIERÍA Y CIENCIAS BIOLÓGICAS

FIDEL ULÍN MONTEJO *

Recibido/Received: 22 Feb 2006; Aceptado/Accepted: 13 Jun 2007

Resumen

El análisis estadístico de tiempos de vida o tiempos de respuesta se ha convertido en un tópico de interés considerable para matemáticos y estadísticos en áreas tales como ingeniería, medicina y ciencias ambientales. Los métodos de máxima verosimilitud han sido una de las herramientas más importantes para resolver problemas desde análisis de tiempos de vida hasta análisis de datos de confiabilidad. Tomando problemas típicos de varias disciplinas y usando esos métodos, se pretende motivar a científicos, ingenieros y estudiantes al análisis de datos censurados. Se muestra un ejemplo sobre un ensayo clínico, el cual fue llevado a cabo para determinar si un tratamiento hormonal es benéfico para las mujeres con cáncer [3]. Por otro lado, un ejemplo sobre análisis de degradación es presentado para estimar los parámetros de un modelo para datos de tamaño de grieta por fatiga para una aleación [5]. Finalmente, se enfatiza en un procedimiento para probar las medianas en un contexto de dos muestras lognormales que contienen dos contaminantes [12]. Se realizan comparaciones entre la metodología lognormal que se presenta aquí y los métodos no paramétricos propuestos por Millard [6]. SPLIDA [10] y SPLUS [11] son usados para implementar la metodología en cada caso.

Palabras clave: Datos censurados, distribución exponencial, distribución lognormal, máxima verosimilitud.

Abstract

The statistical analysis of lifetime or response time data has become a topic of considerable interest to mathematicians and statisticians in areas such as engineering, medicine, and the environmental sciences. Maximum likelihood methods have been one of most important tools to solve problems from analysis of lifetime to reliability

*División Académica de Ciencias Básicas, Universidad Juárez Autónoma de Tabasco, Km.1 Carretera Cunduacán–Jalpa de Méndez. A.P. 24, C.P. 86690. Cunduacán, Tabasco, México. Tel. (+52) 914 3 36 09 28; E-Mail: fidel.ulin@basicas.ujat.mx.

analysis data. Taking typical problems of several areas and using these methods, it intends motivate to scientists, engineers, and students in censored data analysis. An example is presented about a clinical trial, which was conducted to determine whether a hormone treatment benefits women who suffer breast cancer [3]. On the other hand, an example of the degradation analysis is presented to estimate parameters model fatigue crack-size for an alloy [5]. Finally, it emphasizes on a test procedure for comparing medians in a lognormal two-sample context containing two pollutants [12]. Comparisons are made between the lognormal methodology introduced here and the non-parametric methods used by Millard [6]. SPLIDA [10] and SPLUS [11] are used to implement the methodologies.

Keywords: Censored data, exponential distribution, degradation, lognormal distribution, maximum likelihood.

Mathematics Subject Classification: 62N01.

1 Introducción

Una dificultad en el análisis de tiempos de vida, es la posibilidad de que algunos individuos no puedan ser observados hasta tener una recaída, deceso o falla; que al finalizar un experimento de pruebas de vida puede que no todos los componentes hayan fallado; o que la señal producida por un contaminante sea tan pequeña que los instrumentos de medición no puedan registrarla. Tales observaciones incompletas son llamadas datos censurados. La censura es un punto de referencia, en el periodo de observación, que indica hasta donde fue posible tomar una medida de la variable de interés sobre el espécimen en el experimento. Existen tres posibles tipos de censura. Los más comunes son los datos censurados por la derecha; originados por unidades que no fallaron durante el experimento. La censura por intervalos refleja incertidumbre respecto al tiempo exacto en que las unidades fallaron. En los datos censurados por la izquierda, solo se sabe que la falla ocurrió antes de un cierto tiempo. Además, se tienen otros tipos no comunes de censura: censura aleatoria, y hasta tener un número fijo de fallas. En medicina es muy relevante el uso de métodos estadísticos en el estudio de tiempos de vida, en el desarrollo de nuevos fármacos y terapias encaminadas a prolongar la vida de pacientes con enfermedades crónicas [4]. De igual modo, los avances acelerados en tecnología, el desarrollo de productos altamente sofisticados, la intensa competencia global y las crecientes expectativas de clientes han creado nuevas presiones en la manufactura de productos de alta calidad; lo que ha llevado a mejorar la confiabilidad de productos, es decir la calidad sobre el tiempo [5].

Por otro lado, la creación de leyes destinadas a la protección ambiental demanda cada vez más el uso de estadística [7].

El presente trabajo sobre bioestadística, confiabilidad y estadística ambiental, pretende motivar a ingenieros, científicos y estudiantes de ingeniería y ciencias, en el análisis de datos censurados; enfatizando los métodos de máxima verosimilitud. Se muestran algunos problemas típicos tomados de la literatura, concluyendo con un problema típico en contaminación ambiental; usando cómputo y software para ilustrar ideas y conceptos.

2 Metodología

Los métodos de máxima verosimilitud proveen herramientas generales y versátiles para ajustar modelos a los datos. Los métodos pueden ser aplicados a una amplia variedad de modelos paramétricos y no paramétricos con datos censurados y no-censurados.

2.1 Función de verosimilitud

La función de verosimilitud $L(\theta)$ es igual o proporcional a la probabilidad de los datos. Si x_1, x_2, \dots, x_n , es una muestra aleatoria de densidad $f(x; \theta)$, entonces

$$L(\theta) = cP(\text{datos}; \theta) = cf(x_1; \theta)f(x_2; \theta) \dots f(x_n; \theta). \quad (1)$$

2.2 Estimadores de máxima verosimilitud

Sea $L(\theta)$ la función de verosimilitud para las variables aleatorias X_1, X_2, \dots, X_n . Si $\hat{\theta}$ es el valor de θ en Θ el cual maximiza $L(\theta)$, entonces $\hat{\theta}$ es el estimador de máxima verosimilitud (EMV) de θ . Muchas funciones de verosimilitud satisfacen condiciones de regularidad; así el EMV es la solución de la ecuación $\frac{dL(\theta)}{d\theta} = 0$. Sin embargo, $L(\theta)$ y $\ln L(\theta)$ tienen sus máximos en el mismo valor de θ , y algunas veces es más fácil encontrar el máximo de $\ln L(\theta)$.

2.3 Verosimilitud relativa

Las plausibilidades relativas de otros valores de θ pueden ser examinadas comparándolas con $\hat{\theta}$, usando función de verosimilitud relativa:

$$R(\theta) = \frac{L(\theta)}{\sup_{\theta} L(\theta)} = \frac{L(\theta)}{L(\hat{\theta})}. \quad (2)$$

2.4 Prueba de razón de verosimilitud

Es una forma de evaluar si un modelo general ajusta mejor un conjunto de datos que uno restringido. El modelo restringido ajusta los datos casi tan bien como el modelo general si

$$-2 \ln \left[\frac{L(\theta_0)}{L(\hat{\theta})} \right] \sim \chi_{(r-k)}^2. \quad (3)$$

3 Bioestadística

Se realizó un ensayo clínico para determinar si un tratamiento hormonal beneficia a mujeres con cáncer de mama. Cuando una mujer ha tenido una recaída es tratada con irradiación y asignada al grupo de terapia hormonal o al grupo control [3]. Los tiempos hasta que ocurra una segunda recaída se muestran en el Cuadro 1 del Apéndice. Muchas de las mujeres no tuvieron una segunda recaída antes de concluir el estudio, de modo que estos datos fueron censurados.

3.1 Distribución exponencial

Puede suponerse que estos tiempos siguen una distribución exponencial con media θ_H para el grupo de terapia hormonal y θ_C para el grupo control. Esta distribución ha sido ampliamente usada en investigaciones sobre tiempos de reincidencia en enfermedades crónicas [2]. Si se asume que X tiene una distribución exponencial con media θ , entonces su *fdp* es

$$f(x) = \frac{1}{\theta}e^{-x/\theta}, \quad F(x) = 1 - e^{-x/\theta}, \quad x > 0. \quad (4)$$

Ahora, la función de verosimilitud para los datos censurados por la derecha es:

$$L(\theta) = \left[\prod_{i=1}^m \frac{1}{\theta} e^{-x_i/\theta} \right] \prod_{j=1}^{n-m} e^{-T_j/\theta} = \theta^{-m} e^{-s/\theta}, \quad s = \sum_{i=1}^n x_i. \quad (5)$$

De aquí, se obtiene el EMV de θ , y $R(\theta)$:

$$\hat{\theta} = \frac{s}{m}; \quad R(\theta) = (\theta/\hat{\theta})^{-m} e^{m-s/\theta}. \quad (6)$$

3.2 Resultados y discusión

La Figura 1 muestra los EMV y las $R(\theta)$ para cada grupo. Note que $R(\theta_H)$ se desplaza a la izquierda del grupo control, $R(\theta_C)$, manteniendo sus valores de mayor plausibilidad menores a los de $R(\theta_C)$; de donde el EMV para $\theta_H(101.7333)$, es menor que el del grupo control, $\theta_C(123.3)$. Por lo tanto, no existe indicio de que la terapia hormonal aumente el tiempo de recaída.

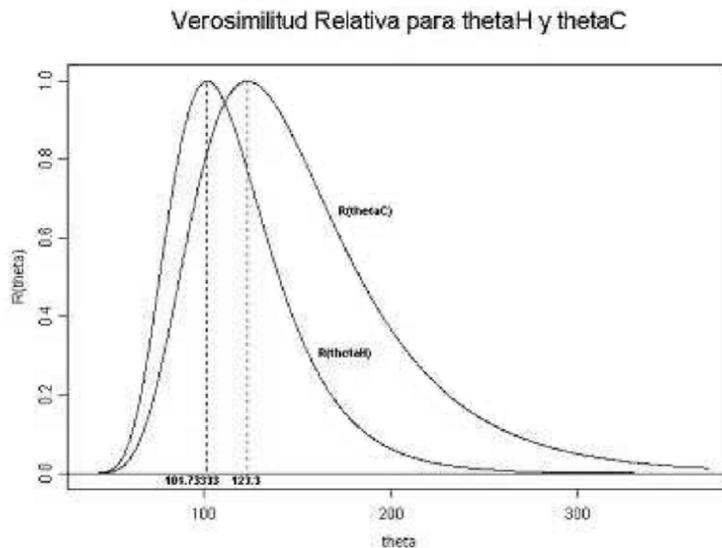


Figura 1: Verosimilitudes relativas y EMV para cada grupo.

4 Confiabilidad

Los sistemas de alta confiabilidad requieren que los componentes individuales tengan una confiabilidad muy alta en periodos largos; de modo que en pruebas de vida, estos componentes no fallan. Así, es difícil evaluar la confiabilidad. Una relación entre las fallas y una medida de degradación hace posible usar modelos de degradación para inferir y predecir tiempos de falla.

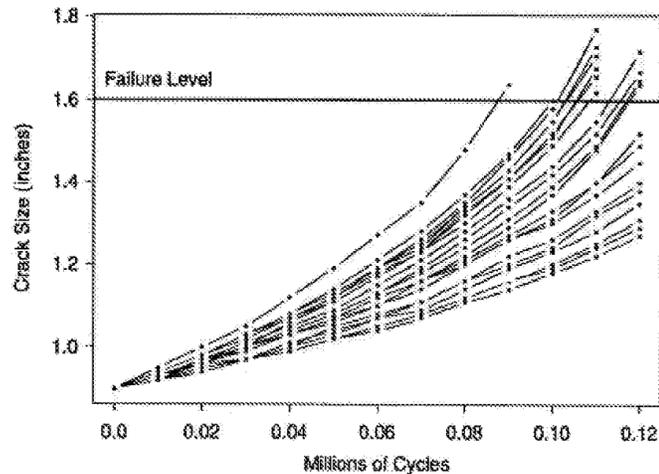


Figura 2: Trayectorias individuales de degradación.

4.1 Datos de tamaño de fractura debido a fatiga

El Cuadro 2 del Apéndice, muestra el tamaño de fractura como función de los ciclos de estrés aplicado a 21 especímenes y la Figura 2 muestra las trayectorias individuales de degradación. Para el análisis, los ingenieros refieren que una fractura de 1.6 plg. es considerada una falla [4]. Se desea modelar la degradación para realizar inferencia y predicciones sobre tiempos de falla.

4.2 Degradación convexa

Sea $a(t)$ el tamaño de fractura al tiempo t , C y m parámetros, entonces una versión del modelo de Paris $\frac{da(t)}{dt} = C [\Delta K(a)]^m$, provee un modelo útil para este tipo de degradación [1].

La trayectoria de degradación real de una unidad particular sobre el tiempo se denota por $a(t), t > 0$. Así, la degradación muestral observada y_{ij} de la i -ésima unidad al tiempo

t_j es:

$$y_{ij} = a(t_{ij}; \beta_{1i}, \dots, \beta_{2i}) + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma_\epsilon).$$

Aunque los valores de los parámetros $\beta_{1i}, \dots, \beta_{2i}$, para unidades individuales, pueden ser importantes en algunas aplicaciones, en confiabilidad se busca inferir sobre proceso poblacional y futuras fallas, de modo que el interés radica en los modelos parámetros poblacionales del proceso, $\theta_\beta = (\mu_\beta, \Sigma_\beta, \sigma_\epsilon)$.

4.3 Estimación de parámetros

Entonces, la función de verosimilitud para los parámetros aleatorios del modelo de degradación pueden ser expresada como

$$L(\theta_\beta, \sigma_\epsilon) = \prod_{i=1}^{21} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left[\prod_{j=1}^{n_i} \frac{1}{\sigma_\epsilon} \phi_{norm}(\zeta_{ij}) \right] f_\beta(\beta_{1i}, \beta_{2i}; \theta_\beta) d\beta_{1i}, d\beta_{2i}. \quad (7)$$

donde $\zeta_{ij} = [y_{ij} - a(t_{ij}, \beta_{1i}, \beta_{2i})]/\sigma_\epsilon$ y $f_\beta(\beta_{1i}, \dots, \beta_{2i}; \theta_\beta)$ es la función de densidad normal multivariada para los parámetros fijos y aleatorios del modelo.

4.4 Resultados y discusión

Aplicando una transformación *log* a la trayectoria de degradación convexa y para los parámetros $\beta_1 = C$ y $\beta_2 = 1 - m$, se calculan la verosimilitud y los parámetros estimados a través de la función nlme de Splus [9]. Se obtiene la distribución bivariada de los parámetros y su *fdp* se muestra en la Figura 3. Los resultados anteriores permiten modelar la trayectoria de degradación global del proceso y gracias a la distribución estimada se hace inferencia y predicciones sobre los tiempos de falla, considerando el umbral de falla de 1.6 plg.

$$\hat{\mu}_\beta = \begin{pmatrix} 5 \cdot 17 \\ 3 \cdot 73 \end{pmatrix}, \quad \hat{\Sigma}_\beta = \begin{pmatrix} 0 \cdot 251 & -0 \cdot 194 \\ -0 \cdot 194 & 0 \cdot 519 \end{pmatrix}, \quad \hat{\sigma}_\epsilon = 0 \cdot 0034.$$

5 Estadística ambiental

En [6] se compararon los niveles medios de zinc y cobre en aguas freáticas en la Zona Aluvial Fan y la Zona Basin-Trough, dos zonas en el Valle San Joaquín, Cal. En el Apéndice, el Cuadro 3 muestra los datos, donde el 20% de ellos son no detectados (censurados). Stoline [11] propone una metodología paramétrica alterna al protocolo de comparación empleado en [6], con el objetivo de tomar cuenta la homogeneidad de la información.

5.1 Protocolo y procedimiento de prueba

Investigadores de las ciencias ambientales han reportado que las concentraciones de varios contaminantes tienen distribuciones lognormal. Esto es, cuando los logaritmos de las concentraciones observadas son graficadas como una distribución de frecuencias, el resultado es una distribución aproximadamente normal o Gaussiana [8].

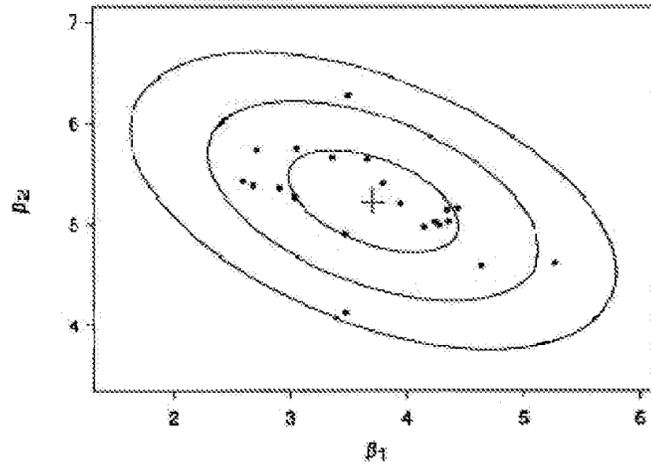


Figura 3: Contorno de la distribución normal multivariada para los parámetros β_1 y β_2 .

$LN(\mu, \sigma)$ denota una variable aleatoria x distribuida lognormalmente con fdp

$$f(x, \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-(\ln x - \mu)^2 / 2\sigma^2}, \quad 0 < x < \infty; \quad -\infty < \mu < \infty, \sigma > 0. \quad (8)$$

Entonces, las dos medianas poblacionales son iguales cuando $H_0 : \mu_1 = \mu_2$ no es rechazada. Sin embargo, la interpretación depende de si los parámetros σ_1 y σ_2 son iguales o no. Se tienen entonces dos casos, el caso homogéneo ($\sigma_1 = \sigma_2$) y el caso heterogéneo ($\sigma_1 \neq \sigma_2$). Así como en el caso normal, las pruebas recomendadas de $H_0 : \mu_1 = \mu_2$, para muestras lognormales dependen de si $\sigma_1 = \sigma_2$ o $\sigma_1 \neq \sigma_2$. Por esta razón se definen cuatro hipótesis:

$$\begin{aligned} H_1 : \mu_1 = \mu_2 = \mu, \quad \sigma_1 = \sigma_2 = \sigma, \quad H_2 : \mu_1 \neq \mu_2, \quad \sigma_1 = \sigma_2 = \sigma, \\ H_3 : \mu_1 = \mu_2 = \mu, \quad \sigma_1 \neq \sigma_2, \quad H_4 : \mu_1 \neq \mu_2, \quad \sigma_1 \neq \sigma_2. \end{aligned} \quad (9)$$

Con lo que se definen cuatro pruebas:

$$\begin{aligned} \text{Prueba 1: } H_1 \text{ vs } H_4, \quad \text{Prueba 2: } H_2 \text{ vs } H_4, \\ \text{Prueba 3: } H_1 \text{ vs } H_2, \quad \text{Prueba 4: } H_3 \text{ vs } H_4. \end{aligned}$$

Usando pruebas asintóticas ji-cuadradas y un algoritmo de probabilidades se calculan los p -valores asociados a cada prueba, p_i . Esos p -valores son usados bajo la estrategia siguiente: La *prueba 1* determina la homogeneidad total de las dos poblaciones lognormales. La *prueba 3* y *4* prueban la igualdad de las dos medianas asumiendo homogeneidad de

asimetría ($\sigma_1 = \sigma_2$) en la *prueba 3* y heterogeneidad de asimetría ($\sigma_1 \neq \sigma_2$) en la *prueba 4*. La *prueba 2* determina la homogeneidad ($\sigma_1 = \sigma_2$) contra la heterogeneidad de asimetría ($\sigma_1 \neq \sigma_2$), y puede ser usada para determinar cual prueba (*prueba 3 o 4*) usar para evaluar la homogeneidad de medianas. Si la *prueba 2* no es significativa, entonces la *prueba 3* es recomendada para la igualdad de mediana; de otra forma, la *prueba 4* es la indicada.

5.2 Derivación del procedimiento

Para llevar a cabo las pruebas asintóticas ji-cuadradas con las pruebas **1 - 4**, es necesario obtener primero los EMV de μ_i y σ_i para cada una de las hipótesis $H_1 - H_4$.

Caso H_1 . La función de verosimilitud para $H_0 : \mu_1 = \mu_2 = \mu, \sigma_1 = \sigma_2 = \sigma$, es

$$L_1(\mu, \sigma) = \prod_{i=1}^2 \left[\prod_{j=1}^{r_i} \frac{1}{\sigma} \phi(z_{ij}) \prod_{j=r_i+1}^{n_i} \Phi(z_{ij}) \right], \quad z_{ij} = (y_{ij} - \mu)/\sigma. \quad (10)$$

$\phi(\cdot)$ y $\Phi(\cdot)$ denotan la *fdp.* y la *fda.* de la $n(0, 1)$, ya que si x es *LN*, entonces $y = \ln(x)$ es $n(\mu, \sigma)$.

Los EMV de μ y σ son obtenidos como soluciones simultáneas a las ecuaciones,

$$\frac{\partial \ln L_1}{\partial \mu} = 0, \quad \frac{\partial \ln L_1}{\partial \sigma} = 0. \quad (11)$$

Caso H_2 . La función de verosimilitud para $H_0 : \mu_1 \neq \mu_2 = \mu, \sigma_1 = \sigma_2 = \sigma$, es

$$L_2(\mu_1, \mu_2, \sigma) = \prod_{i=1}^2 \left[\prod_{j=1}^{r_i} \frac{1}{\sigma} \phi(z_{ij}) \prod_{j=r_i+1}^{n_i} \Phi(z_{ij}) \right], \quad z_{ij} = (y_{ij} - \mu)/\sigma. \quad (12)$$

Los EMV de μ_1, μ_2 y σ son obtenidos como soluciones simultáneas a las ecuaciones,

$$\frac{\partial \ln L_2}{\partial \mu_1} = 0, \quad \frac{\partial \ln L_2}{\partial \mu_2} = 0, \quad \frac{\partial \ln L_2}{\partial \sigma} = 0. \quad (13)$$

Caso H_3 . La función de verosimilitud para $H_0 : \mu_1 = \mu_2 = \mu, \sigma_1 \neq \sigma_2 = \sigma$, es

$$L_3(\mu, \sigma_1, \sigma_2) = \prod_{i=1}^2 \left[\prod_{j=1}^{r_i} \frac{1}{\sigma_i} \phi(z_{ij}) \prod_{j=r_i+1}^{n_i} \Phi(z_{ij}) \right], \quad z_{ij} = (y_{ij} - \mu)/\sigma_i. \quad (14)$$

EMV de μ, σ_1 y σ_2 son obtenidos como soluciones simultáneas a las ecuaciones,

$$\frac{\partial \ln L_3}{\partial \mu} = 0, \quad \frac{\partial \ln L_3}{\partial \sigma_1} = 0, \quad \frac{\partial \ln L_3}{\partial \sigma_2} = 0. \quad (15)$$

Caso H_4 . La función de verosimilitud para $H_0 : \mu_1 \neq \mu_2, \sigma_1 \neq \sigma_2$, es

$$L_4(\mu_1, \mu_2, \sigma_1, \sigma_2) = \prod_{i=1}^2 \left[\prod_{j=1}^{r_i} \frac{1}{\sigma_i} \phi(z_{ij}) \prod_{j=r_i+1}^{n_i} \Phi(z_{ij}) \right], \quad z_{ij} = (y_{ij} - \mu_i)/\sigma_i. \quad (16)$$

Los EMV de μ_1, μ_2, σ_1 y σ_2 son obtenidos como soluciones simultáneas a las cuatro ecuaciones:

$$\frac{\partial \ln L_4}{\partial \mu_i} = 0, \quad \frac{\partial \ln L_4}{\partial \sigma_i} = 0, \quad i = 1, 2. \quad (17)$$

5.3 Pruebas ji-cuadrada asintótica

Las pruebas ji-cuadradas asintóticas, de nivel α , pueden ser descritas como sigue para las pruebas **1 - 4**, respectivamente:

$$\begin{aligned}\mathcal{X}_1^2 &= -2 [\ln L_1(\hat{\mu}, \hat{\sigma}) - \ln L_4(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2)] > \mathcal{X}_{\alpha,2}^2, \\ \mathcal{X}_2^2 &= -2 [\ln L_2(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}) - \ln L_4(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2)] > \mathcal{X}_{\alpha,1}^2, \\ \mathcal{X}_3^2 &= -2 [\ln L_1(\hat{\mu}, \hat{\sigma}) - \ln L_2(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma})] > \mathcal{X}_{\alpha,1}^2, \\ \mathcal{X}_4^2 &= -2 [\ln L_3(\hat{\mu}, \hat{\sigma}_1, \hat{\sigma}_2) - \ln L_4(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2)] > \mathcal{X}_{\alpha,1}^2.\end{aligned}\quad (18)$$

Un programa sencillo de SPLUS (o una rutina de SPLIDA) calcula los EMV, L_i estimada y los p -valores para cada una de las *pruebas*.

5.4 Resultados y discusión

En el Cuadro 4 del apéndice, se muestran los resultados, sin embargo, los casos de el cobre y zinc son discutidos separadamente usando un nivel de significancia de $\alpha = 0.05$.

Cobre. La prueba preliminar (*prueba 2*) para asimetría no es significativa ($p = 0.4844$). De aquí, la *prueba 3* es una prueba apropiada para la igualdad de medianas, la cual no es significativa ($p = 0.6629$). La hipótesis de igualdad de medianas, la cual no es rechazada. Este resultado concuerda con el resultado de la prueba no paramétrica de [5] ($p = 0.32$). Además, la prueba para homogeneidad total de las dos poblaciones lognormales (*prueba 1*) no es significativa ($p = 0.7122$) para el cobre. De aquí, puede concluirse que los percentiles y las medianas de cobre son iguales entre esas dos zonas.

Zinc. La prueba preliminar (*prueba 2*) para la asimetría no es significativa ($p = 0.5034$), y de aquí la prueba 3 es la más apropiada para la igualdad de los niveles medios. Esta prueba no es significativa ($p = 0.1104$), y se concluye que los niveles medios no difieren, aunque esta prueba se aproxima a la significancia estadística de $\alpha = 0.10$. También, la prueba de homogeneidad total (*prueba 1*) indica que las dos distribuciones lognormal no difieren ($p = 0.2236$). Estos resultados difieren de los reportados por Millard [5], donde se implementó un procedimiento no paramétrico, indicando que los niveles medios difieren ($p = 0.02$).

5.5 Conclusiones

En Bioestadística, puede observarse la relevancia de los métodos de verosimilitud para involucrar la información censurada en el experimento. Sin embargo, cuando existen muchos datos censurados, debe considerarse un nuevo experimento para comparar estos resultados obtenidos previamente.

Los modelos determinísticos de degradación en Confiabilidad, ofrecen una herramienta muy útil cuando los métodos de análisis de tiempos de falla no pueden ser aplicados. Los métodos numéricos y computacionales son muy útiles, pues las soluciones son aproximadas.

Por último, en Estadística Ambiental, el modelo lognormal para las dos muestras provee una alternativa a los modelos no paramétricos en la comparación de concentraciones medias de contaminantes ambientales. La relevancia del modelo lognormal estriba en que provee información adicional en la comparación de medianas para las muestras, tanto para la homogeneidad total ($\sigma_1 = \sigma_2$) como heterogeneidad ($\sigma_1 \neq \sigma_2$) en ambas poblaciones, lo cual es importante en la comparación de medianas y concentraciones medias de contaminantes.

Referencias

- [1] Dowling, N.E. (1993) *Mechanical Behavior of Materials*. Prentice Hall, Englewood Cliffs, N.J.
- [2] Feigl, P.; Zelen, M. (1965) "Estimation of exponential survival probabilities with concomitant information", *Biometrics* **21**: 826–838.
- [3] Kalbfleisch, J.G. (1985) *Probability and Statistical Inference*. John Wiley and Sons, New York.
- [4] Meeker, W.Q.; Escobar, L.A. (1998) *Statistical Methods for Reliability Data*. John Wiley and Sons, New York.
- [5] Lawless, J.F. (1986) *Statistical Models and Methods for Life Time Data*. John Wiley and Sons, New York.
- [6] Millard, S.P.; Deverel, S.J. (1988) "Nonparametric statistical methods for comparing two sites on data with multiple nondetect limits", *Water Resources Research* **22**: 2087–2098.
- [7] Millard, S.P.; Neerchal, N.K.M (2001) *Environmental Statistics*. CRC Press, New York.
- [8] Ott, W.R. (1990) "A physical explanation of the lognormality of pollutant concentrations", *Journal of the Air and Waste Management Association* **40**: 1378–1383.
- [9] Pinheiro, J.C.; Bates, D.M. (1995) "Mixed effects model, methods and classes for S and S-PLUS", Department of Statistics, University of Wisconsin.
- [10] SPLIDA (2002) *SPLIDA User's Guide, Ver. 5.9.38*. Iowa State University and Louisiana State University.
- [11] SPLUS (2000) *SPLUS User's Guide, Ver.3.0.*. MathSoft, Inc., Seattle, WA.
- [12] Stoline, M.R. (1993) "Comparison of two medians using a two-sample lognormal model in environmental", *Environmetrics* **4**(3): 323–339.

Apéndice

	Hormona tratada						Control					
Tiempo de recaída	2	4	6	9	9	9	1	4	6	7	13	24
	13	14	18	23	31	32	25	35	35	39		
	33	34	43									
Tiempo de censura	10	14	14	16	17	18	1	1	3	4	5	8
	18	19	20	20	21	21	10	11	13	14	14	15
	23	24	29	29	30	30	17	19	20	22	24	24
	31	31	31	33	35	37	24	25	26	26	26	28
	40	41	42	42	44	46	29	29	32	35	38	39
	48	49	51	53	54	54	40	41	44	45	47	47
	55	56					47	50	50	51		

Cuadro 1: Tiempos de recaída, censurados y no censurados para ambos grupos.

	Millones de Ciclos												
Unidad	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09	.10	.11	.12
1	.90	.95	1.00	1.05	1.12	1.19	1.27	1.35	1.48	1.64			
2	.90	.94	.98	1.03	1.08	1.14	1.21	1.28	1.67	1.47	1.60		
3	.90	.94	.98	1.03	1.08	1.13	1.19	1.26	1.35	1.46	1.58	1.77	
4	.90	.94	.98	1.03	1.07	1.12	1.19	1.25	1.34	1.43	1.55	1.72	
5	.90	.94	.98	1.03	1.07	1.12	1.19	1.24	1.34	1.43	1.55	1.71	
6	.90	.94	.98	1.03	1.07	1.12	1.18	1.23	1.33	1.41	1.51	1.68	
7	.90	.94	.98	1.02	1.07	1.11	1.17	1.23	1.32	1.41	1.52	1.66	
8	.90	.93	.97	1.00	1.06	1.11	1.17	1.23	1.30	1.39	1.49	1.62	
9	.90	.92	.97	1.01	1.05	1.09	1.15	1.21	1.28	1.36	1.44	1.55	1.72
10	.90	.92	.96	1.00	1.04	1.08	1.13	1.19	1.26	1.34	1.42	1.52	1.67
11	.90	.93	.96	1.00	1.04	1.08	1.13	1.18	1.24	1.31	1.39	1.49	1.65
12	.90	.93	.97	1.00	1.03	1.07	1.10	1.16	1.22	1.29	1.37	1.48	1.64
13	.90	.92	.97	.99	1.03	1.06	1.10	1.14	1.20	1.26	1.31	1.40	1.52
14	.90	.93	.96	1.00	1.03	1.07	1.12	1.16	1.20	1.26	1.30	1.37	1.45
15	.90	.92	.96	.99	1.03	1.06	1.10	1.16	1.21	1.27	1.33	1.40	1.49
16	.90	.92	.95	.97	1.00	1.03	1.07	1.11	1.16	1.22	1.26	1.33	1.40
17	.90	.93	.96	.97	1.00	1.05	1.08	1.11	1.16	1.20	1.24	1.32	1.38
18	.90	.92	.94	.97	1.01	1.04	1.07	1.09	1.14	1.19	1.23	1.28	1.35
19	.90	.92	.94	.97	.99	1.02	1.05	1.08	1.12	1.16	1.20	1.25	1.31
20	.90	.92	.94	.97	.99	1.02	1.05	1.08	1.12	1.16	1.19	1.24	1.29
21	.90	.92	.94	.97	.99	1.02	1.04	1.07	1.11	1.14	1.18	1.22	1.27

Cuadro 2: Tamaño de grieta por fatiga como función de ciclos.

Zona Aluvial Fan				Zona Basin-Trough			
Conc. Cobre		Conc. Zinc		Conc. Cobre		Conc. Zinc	
Cobre	Frec.	Zinc	Frec.	Cobre	Frec.	Zinc	Frec.
1ND	4	3ND	1	1ND	2	3ND	1
5ND	8	10ND	15	2ND	2	10ND	3
10ND	3	5	1	5ND	5	3	2
20ND	2	7	1	10ND	4	4	2
1	5	8	1	15ND	1	5	2
2	21	9	1	1	7	6	1
3	6	10	20	2	4	8	1
4	3	11	2	3	8	10	5
5	3	12	1	4	5	11	1
7	3	17	1	5	1	12	2
8	1	18	1	6	2	13	1
9	1	19	1	8	1	14	1
10	1	20	14	9	2	15	1
11	1	23	1	12	1	17	2
12	1	29	1	14	1	20	11
16	1	30	1	15	1	25	1
20	1	33	1	17	1	30	4
	65	40	1	23	1	40	3
		50	1		49	50	2
		620	1			60	2
			67			70	1
						90	1
							50

Cuadro 3: Datos de concentraciones de cobre y zinc.

Descripción	Prueba	p-valores y EMVs			
		Cobre		Zinc	
Homogeneidad Total ($H : \mu_1 = \mu_2, \sigma_1 = \sigma_2$)	1	0.7122	$\hat{\mu} = 1.00$ $\hat{\sigma} = 0.88$	0.2236	$\hat{\mu} = 2.85$ $\hat{\sigma} = 0.85$
Asimetría preliminar ($H : \sigma_1 = \sigma_2$)	2	0.4844	$\hat{\mu}_1 = 0.97$ $\hat{\mu}_2 = 1.05$ $\hat{\sigma} = 0.88$	0.5034	$\hat{\mu}_1 = 2.47$ $\hat{\mu}_2 = 2.72$ $\hat{\sigma} = 0.84$
Igualdad de Medianas (asumiendo $\sigma_1 = \sigma_2$)	3	0.6629	$\hat{\mu}_1 = 1.00$ $\hat{\sigma}_1 = 1.05$ $\hat{\sigma}_2 = 0.88$	0.1104	$\hat{\mu}_1 = 2.57$ $\hat{\sigma}_1 = 0.80$ $\hat{\sigma}_2 = 0.90$
Igualdad de Medianas (asumiendo $\sigma_1 \neq \sigma_2$)	4	0.7500	$\hat{\mu}_1 = 0.97$ $\hat{\mu}_2 = 1.03$ $\hat{\sigma}_1 = 0.84$ $\hat{\sigma}_2 = 0.94$	0.1320	$\hat{\mu}_1 = 2.47$ $\hat{\mu}_2 = 2.72$ $\hat{\sigma}_1 = 0.80$ $\hat{\sigma}_2 = 0.85$
Igualdad de Medianas (Milland & Deverel)	No Paramétrica	0.32		0.02	

Cuadro 4: Resultados de las pruebas lognormales para datos de zinc y cobre.