



COLEGIO DE POSTGRADUADOS

INSTITUCIÓN DE ENSEÑANZA E INVESTIGACIÓN EN CIENCIAS AGRÍCOLAS

CAMPUS MONTECILLO

POSTGRADO DE SOCIOECONOMÍA, ESTADÍSTICA E INFORMÁTICA

ESTADÍSTICA

INFERENCIA SOBRE DATOS NO DETECTADOS DE POBLACIONES LOGNORMALES

FIDEL ULÍN MONTEJO

T E S I S

PRESENTADA COMO REQUISITO PARCIAL
PARA OBTENER EL GRADO DE

DOCTOR EN CIENCIAS

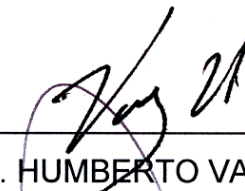
MONTECILLO, TEXCOCO, EDO. DE MÉXICO

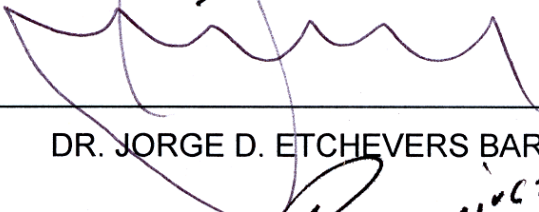
2008

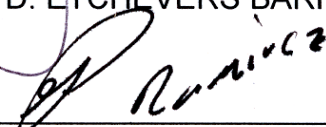
La presente tesis, titulada: **Inferencia sobre datos no detectados de poblaciones lognormales**, realizada por el alumno: **Fidel Ulín Montejo**, bajo la dirección del Consejo Particular indicado, ha sido aprobada por el mismo y aceptada como requisito parcial para obtener el grado de:


DOCTOR EN CIENCIAS
SOCIOECONOMÍA, ESTADÍSTICA E INFORMÁTICA
ESTADÍSTICA

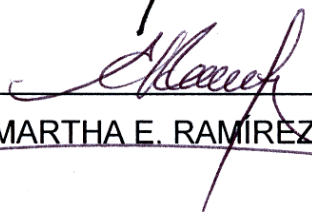
CONSEJO PARTICULAR

CONSEJERO: 
DR. HUMBERTO VAQUERA HUERTA

ASESOR: 
DR. JORGE D. ETCHEVERS BARRA

ASESOR: 
DR. GUSTAVO RAMÍREZ VALVERDE

ASESOR: 
DR. GABRIEL A. RODRÍGUEZ YAM

ASESOR: 
DRA. MARTHA E. RAMÍREZ GUZMÁN

Montecillo, Texcoco, México, 3 de marzo de 2008

INFERENCIA SOBRE DATOS NO DETECTADOS DE POBLACIONES LOGNORMALES

Fidel Ulín Montejo

Colegio de Postgraduados, 2008

Frecuentemente, la descripción y comparación de poblaciones con concentraciones de contaminantes se realiza empleando métodos no paramétricos. En la práctica, si las muestras contienen datos no detectados, éstos se omiten o sustituyen por una fracción del límite de detección (LD). Respecto a datos ambientales que contienen datos no detectados, los organismos de regulación ambiental requieren que los riesgos sean caracterizados en términos de la concentración media del contaminante. Este trabajo aborda el problema de descripción, inferencia y comparación de concentraciones medias de poblaciones lognormales mediante una prueba estadística basada en modelos de regresión lineal con variables indicadoras, empleando un enfoque paramétrico. El algoritmo Expectation-Maximization (EM), la verosimilitud y el método de Wald son empleados para el manejo de datos no detectados, la estimación y la obtención de regiones de confianza. También se presenta el enfoque con modelos mixtos para comparar medias poblacionales con covariables aleatorias. Se realizó un estudio de simulación para conocer la potencia de la prueba paramétrica propuesta al comparar muestras lognormales y no lognormales, confrontándola además, con la prueba no paramétrica logrank. El método resultó simple y versátil para dos poblaciones y puede extenderse a tres o más, ya que su implementación e interpretación no es difícil. La potencia de la prueba propuesta fue mejor que la prueba logrank al comparar muestras lognormales y mostró un desempeño aceptable para muestras exponenciales y de Gumbel.

Palabras Clave: Algoritmo EM, distribución lognormal, máxima verosimilitud, método de Wald, modelo lineal mixto, datos no detectados.

INFERENCE ON NONDETECTS DATA OF LOGNORMAL POPULATIONS

Fidel Ulín Montejo

Colegio de Postgraduados, 2008

Frequently, the summarization and comparison of populations with pollutant concentrations is done using non-parametric methods. In practice, if the samples contain nondetects data, these are omitted or substituted for a fraction of the detection limit. With regard to environmental data which contain nondetects, environmental regulation organisms require that risks be characterized in terms of the pollutant mean concentration. From a parametric approach, this study takes on the problem of summarization, inference, and comparison of mean concentrations of lognormal populations using a linear regression model with indicator variables. The EM algorithm, total likelihood, and Wald's method are used to handle nondetects data, estimation, and detection of confidence regions. We also show an approach with mixed-effects models to compare population means with random covariables. A simulation study is done to know the power of the proposed parametric test when comparing lognormal and non-lognormal samples, besides it is also compared with logrank non-parametric test. The method proved to be simple and adaptable for two populations and can be extended to three or more, since its implementation and interpretation are not difficult. The power of the proposed test was better than the logrank test when comparing lognormal samples, and showed an acceptable performance for exponential and Gumbel samples.

Key words: EM algorithm, lognormal distribution, maximum likelihood, Wald's method, linear mixed-effects model, nondetects data.

DEDICATORIA

AL GRAN ARQUITECTO DEL UNIVERSO

Dios, por darme las herramientas con las que trabajo día a día para ser un hombre libre y de buenas costumbres.

A MI ESPOSA

Rosa Ma. Salinas Hernández, por la fortaleza que me dan sus palabras y la grandeza que me inspira su amor.

A LA FAMILIA ULIN MONTEJO

Mis abuelos, mis padres, mis hermanos y mis sobrinos, por el cariño, amor, respeto, orgullo y admiración que le profeso a mi linaje.

AGRADECIMIENTOS

Al **Dr. Humberto Vaquera Huerta**, con admiración y respeto, por la confianza y oportunidad que me ha brindado para desarrollar y concluir exitosamente este trabajo de investigación.

A **mis profesores, asesores y sinodales** del Colegio de Postgraduados, por compartir conmigo sus conocimientos, experiencia y sabiduría.

Al **Dr.: Sergio Pérez Elizalde**, por sus valiosas sugerencias para este trabajo, y por la amistad y fraternidad que nos une.

A **Emma, Lucy, Gris, Carmen**, ... por su extraordinaria amistad e invaluable ayuda en todo momento.

A todo el **personal, técnicos y trabajadores** del Edificio de Estadísticas del Colegio de Postgraduados, por su aprecio y apoyo.

A las autoridades de la **Dirección de Investigación y Posgrado**, y de la **División Académica de Ciencias Básicas** de la **UJAT**, por darme las facilidades y respaldo para iniciar, desarrollar y concluir este proyecto doctoral que armoniza con la superación académica que impulsan para el engrandecimiento de nuestra Universidad.

Al **PROMEP-UJAT**, por el otorgamiento de recursos y apoyos económicos, para la realización de estudios de doctorado.

Al **COMECYT**, por el apoyo complementario otorgado en la etapa final de este proyecto de investigación.

A **Paulino Pérez Rodríguez y Rene Castro Montoya**, por su amistad, apoyo y compañerismo desde esta etapa de mi vida.

A **mis compañeros de maestría y doctorado** en Estadística, por su camaradería y por ser el semillero de nuestros gremio.

A **mis excompañeros y amigos**, Antonio Murillo Salas y Víctor I. López Ríos, por alentarnos y compartir juntos el inicio de nuestros estudios de doctorado.

A **mis cofraternos en Texcoco**, en especial a Napoleón Ágreda Aragón, por el cariño, protección y luz con que fui cobijado durante mi estancia en este punto geométrico.

CONTENIDO

1. Introducción	1
2. Revisión de Literatura	4
2.1 Datos No Detectados	4
2.2 Análisis de Datos No Detectados	7
2.3 El Modelo Lognormal	8
2.4 Distribuciones de Localización-Escala	13
2.5 Gráficos de Probabilidad para la Distribución Lognormal	14
2.6 Máxima Verosimilitud	15
2.7 Verosimilitud para Datos Lognormales No Detectados	17
2.8 Prueba de Razón de Verosimilitud	17
2.9 El Algoritmo EM (Expectation-Maximization)	17
2.10 Prueba No Paramétrica Logrank	19
3. Objetivos	21
3.1 Objetivo General	21
3.2 Objetivos Particulares	21
4. Metodología Desarrollada	22
4.1 Aplicación del Algoritmo EM en la Familia Exponencial	22
4.2 Obtención de la Matriz de Información Mediante el Algoritmo EM	23
4.3 Construcción de Intervalos y Regiones de Confianza Aproximados	24
4.3.1 Matriz De Varianzas-Covarianzas de los Parámetros	26
4.3.2 Intervalos de Confianza para Funciones de μ y σ	27
4.4 Aplicación de los Modelos de Regresión con Variables Explicatorias	27
4.5 Comparación Mediante Modelo de Regresión con Variables Indicadoras	28
4.6 Aplicación de los Modelos de Efectos-Mixtos	29
4.6.1. Definición del Modelo Lineal de Efectos Mixtos	30
4.6.2. Estimación por Máxima Verosimilitud para Modelos Lineales de Efectos Mixtos	31

5. Estudio de Simulación y Funciones de Potencia	33
5.1 Prueba Propuesta y Logrank Comparando Poblacionales Lognormales	34
5.2 Prueba Propuesta Comparando Muestras Exponenciales	37
5.3 Prueba Propuesta Comparando Muestras Gumbel	38
5.4 Resultados y Discusión del Estudio de Potencia	39
6. Aplicación de la Metodología en Datos Ambientales	40
6.1. Inferencia Sobre Una Muestra Aleatoria Lognormal	40
6.1.1 Gráfico de Probabilidad Lognormal	41
6.1.2 Optimización Directa de la Función de Verosimilitud	41
6.1.3 Optimización de la Función de Verosimilitud Vía Algoritmo EM	43
6.2 Inferencia Sobre Una Muestra Lognormal Con Una Covariable Continua	44
6.2.1 Optimización Directa de la Función de Verosimilitud	46
6.2.2 Optimización de la Función de Verosimilitud Vía EM	47
6.3 Comparación de Dos Poblaciones Lognormales Independientes	48
6.3.1 Modelo 1. Homogeneidad del Parámetro de Escala σ	49
6.3.2 Modelo 2. Heterogeneidad del Parámetro de Escala σ	51
6.3.3 Selección del Mejor Modelo	52
6.4 Comparación y Análisis de Tendencia en un Proceso de Remediación	52
6.5 Comparación de Dos Muestras con Datos Agrupados	56
7. Conclusiones	59
8. Referencias	61
Anexo A. Algunos Resultados de Teoría Estadística	66
A.1 Difusión del Error Estadístico - El Método Delta	66
A.2 Matriz de Información de Fisher	68
A.3 Condiciones de Regularidad para Distribuciones de Localización-Escala	70
A.4 Convergencia en Distribución	71
A.5 Distribución Asintótica de los EMV	73
Anexo B. Programas en R	76
B.1 Estudio de Simulación y Funciones de Potencia	76
B.2 Ejemplos de Aplicación de los Métodos	83
Anexo C. Artículos de Investigación	100

LISTA DE CUADROS

Cuadro 1. Contribución a la función de verosimilitud para datos censurados.	16
Cuadro 2. Concentraciones de arsénico en Manoa Stream, Kanawai Field .	40
Cuadro 3. Estimaciones de MV para los parámetros de los datos de arsénico.	44
Cuadro 4. Dosis de radiación gamma $\mu\text{Sv}\times 100$.	45
Cuadro 5. Estimación de MV para los parámetros del modelo lognormal para las dosis.	47
Cuadro 6. Datos de Concentraciones de Cobre para las zonas del Valle San Joaquín.	48
Cuadro 7. Estimación de MV para parámetros en los datos de Cobre (σ común).	50
Cuadro 8. EMV para los parámetros en los datos de cobre con heterogeneidad de σ .	51
Cuadro 9. Niveles de contaminantes antes y después del proceso de remediación.	53
Cuadro 10. EMV para los parámetros de los datos del proceso remediación (σ común).	55
Cuadro 11. Concentraciones de tolueno para dos muestras en un periodo de 5 meses.	57
Cuadro 12. Estimaciones de MV para los parámetros de concentraciones de tolueno.	57

LISTA DE FIGURAS

Figura 1. Distribuciones lognormal para distintos valores de μ y σ .	10
Figura 2. Grafico de probabilidad lognormal con líneas rectas para $\exp(\mu)=50,500$; $\sigma=1,2$.	14
Figura 3. Intervalos de incertidumbre y contribuciones a la verosimilitud para la censura.	15
Figura 4. Esquema general del Algoritmo EM.	18
Figura 5. Algoritmo EM en la familia exponencial.	22
Figura 6. Prueba propuesta y logrank comparando muestras lognormal de tamaño 15.	34
Figura 7. Prueba propuesta y logrank comparando muestras lognormal de tamaño 30.	35
Figura 8. Prueba propuesta y logrank comparando muestras lognormal de tamaño 100.	36
Figura 9. Gráficos de funciones de densidad para variables exponenciales.	37
Figura 10. Prueba propuesta comparando poblaciones exponenciales.	37
Figura 11. Gráficos de funciones de densidad para variables Gumbel.	38
Figura 12. Prueba propuesta comparando poblaciones Gumbel.	38
Figura 13. Gráfico de probabilidad lognormal para los datos del Cuadro 2.	41
Figura 14. Gráficos de dispersión para las dosis y log-dosis de los datos del Cuadro 4.	45
Figura 15. Gráficos de probabilidad lognormal para los datos del Cuadro 6.	49
Figura 16. Gráficos de dispersión de contaminantes antes y después de la remediación.	54

1. INTRODUCCIÓN

Algunos de los problemas relevantes que requieren el uso de la estadística ambiental son: monitoreo de la calidad del aire y agua, monitoreo de la calidad de mantos freáticos cercanos a depósitos de desechos tóxicos, aseguramiento de zonas contaminadas en proceso de remediación, medición de nivel de riesgo de zonas potencialmente contaminadas, aseguramiento de riesgos en lugares de trabajo, etc. (Millard y Neerchal, 2001; Frome y Watekins, 2004). El campo de la estadística ambiental es relativamente reciente y emplea métodos desarrollados en otras disciplinas para resolver problemas concernientes al medio ambiente.

La evaluación objetiva del riesgo es importante, debido a que grandes concentraciones de contaminantes frecuentemente tienen efectos adversos a la salud, sin embargo, pequeñas concentraciones pueden no ser inocuas. Con la finalidad de conocer el nivel de riesgo de estos contaminantes, se pueden identificar y medir éstos en niveles extraordinariamente bajos. Algunas veces la cantidad presente de un contaminante es tan pequeña para los instrumentos de medición que no pueden ser cuantificadas ni discriminadas (Helsel, 2005). Cuando no es posible cuantificar las concentraciones en una muestra se reportan los datos como *no detectados* (Lambert *et al.* 1991). Tales observaciones incompletas son llamadas *datos censurados por la izquierda*. La censura es un punto de referencia que indica hasta donde fue posible tomar una medida de la variable de interés sobre el espécimen. En este caso, el punto de referencia se denomina *límite de detección* (LD). En el mismo sentido, en monitoreos de exposición a contaminantes en centros de trabajo, las mediciones generalmente son valores positivos; sin embargo, pueden tenerse igualmente datos no detectados. Aquí, las estrategias para asegurar información referente a exposiciones a contaminantes en centros laborales están enfocadas a determinar el nivel medio de exposición, a fin de tomar alguna decisión que evite un riesgo o accidente (Frome y Watkins, 2004).

Investigadores en diferentes ciencias ambientales han reportado que las mediciones de concentraciones de varias sustancias tienen distribuciones de frecuencias de tipo

lognormal o aproximadas a ésta. Es decir, cuando los logaritmos de las concentraciones observadas son graficadas como una distribución de frecuencias, la distribución resultante es aproximadamente normal, sobre la mayor parte del rango observado. Este supuesto se justifica por el hecho de que en muchas situaciones las concentraciones de contaminantes son el resultado de muchas diluciones ligeras (Ott, 1995). Algunos ejemplos incluyen material radiactivo en suelos, contaminantes en aire, calidad del aire, residuos de metales en ríos y en tejidos biológicos. En estos casos la normalidad se asume después de la log-transformación de los datos y entonces las técnicas basadas en la teoría de la normal son usadas para obtener inferencias acerca de la cantidad de interés. Sin embargo, como se discutió en El-Shaarawi y Viveros (1997), frecuentemente los resultados necesitan ser reportados como mediciones obtenidas, más que en la escala log-transformada. Por ejemplo, la US Environmental Protection Agency (EPA) requiere que los riesgos sean caracterizados en términos de la concentración media del contaminante. Este requerimiento es parcialmente responsable del énfasis en la concentración media y su inferencia.

Desde una perspectiva diferente, dado que las concentraciones de un contaminante son físicamente aditivas en un sentido útil, éstas pertenecen a la clase de variables extensivas descritas por Cox y Snell (1981). A pesar de la forma de la distribución, la concentración media del contaminante tiene una interpretación física, lo que no puede decirse respecto a la log-concentración o alguna otra transformación no lineal.

Los datos ambientales pueden presentar características que conducen a problemas de inferencia muy interesantes. La meta es obtener aseveraciones cuantitativas respecto a los parámetros desconocidos, así como regiones de confianza aproximadas usando toda la información paramétrica contenida en la muestra. Un estimador de máxima verosimilitud (EMV) y su varianza son suficientes para especificar o reproducir la función de verosimilitud completa (Díaz-Francés y Sprott, 2000). Las regiones de confianza pueden construirse con el método de Wald, conocido también como método de aproximación normal de la log-verosimilitud (Meeker y Escobar, 1998).

El objetivo de este trabajo es caracterizar, hacer inferencia y proponer una prueba paramétrica para comparar concentraciones medias de contaminantes de muestras lognormales que contienen datos no detectados y covariables. Para esto se emplea una reparametrización en los parámetros del modelo lognormal a través de un modelo de regresión con variables indicadoras (dummies) para indicar cada una de las muestras que se comparan. Los métodos desarrollados se centran en un criterio de comparación mediante regiones de confianza aproximados, enfocándose en la concentración media de contaminantes.

Para ilustrar el procedimiento, primero se caracterizó una población lognormal a través de sus parámetros, su concentración media e intervalos de confianza; del mismo modo, otra población fue analizada considerando la covariable tiempo. Posteriormente se compararon las concentraciones medias de dos poblaciones independientes. También, se presentó la resolución de un problema de comparación, con la covariable tiempo, considerando homogeneidad y heterogeneidad del parámetro σ ; aquí mismo, se presentó una forma de seleccionar el mejor modelo a través de una prueba de razón de verosimilitud. Finalmente, con un enfoque de modelos mixtos, el procedimiento se extiende a muestras poblacionales con covariables continuas y aleatorias, debidas a periodos de monitoreo, diseños de muestreo o datos agrupados. Por otro lado, se realizó un estudio de potencia, para ilustrar las ventajas de la prueba paramétrica propuesta, al compararse con la prueba no paramétrica logrank (Lee y Wang, 2003), típica en la comparación de muestras censuradas en análisis de supervivencia.

En todos los ejemplos de aplicación, se utilizó una modificación del algoritmo EM para ajustar los modelos, considerando la información contenida en los datos no detectados y los errores lognormales, obteniéndose las varianzas de los estimadores y las regiones aproximadas del 95% de confianza. El procedimiento de inferencia y la prueba Paramétrica propuesta fueron implementados en R (Development Core Team, 2006) y se desarrollaron ejemplos de aplicación con datos de estudios en ciencias ambientales, ilustrándose la versatilidad y ventajas de este procedimiento paramétrico.

2. REVISIÓN DE LITERATURA

2.1. Datos No Detectados

Las mediciones cuyos resultados sólo se sabe que son mayores o menores a un límite son llamados datos censurados (Lambert *et al.* 1991). Este tipo de mediciones han sido una parte integral de las ciencias sociales, económicas, médicas y la estadística industrial. Existen procedimientos desarrollados recientemente que permiten incorporar los datos censurados en el cálculo de estadísticas descriptivas, regresión y pruebas de hipótesis. Pero esos procedimientos son raramente usados en estudios ambientales, donde los datos censurados son frecuentemente aquellos valores que se encuentran por debajo de los límites de detección. Esos niveles bajos de contaminantes (tales como residuos de metales o compuestos orgánicos) son conocidos con imprecisión y llamados “menores que” o “no detectados”. Debido a que bajos niveles son usualmente presentados a la izquierda de un gráfico, los no detectados son etiquetados como “censurados por la izquierda”, con valores en algún lugar a la izquierda del límite de detección (LD), pero sin conocer la concentración exacta.

Los datos no detectados son un tema de interés en varias disciplinas y subdisciplinas dentro de las ciencias ambientales, son encontrados en estudios de calidad del aire (Rao *et al.*, 1991), estudios marinos (Huybrechts *et al.*, 2002), análisis de aguas residuales (Kroll y Stedinger, 1996), lluvia ácida (Ahn, 1998), pruebas en bombas de pozos (Wen, 1994), exposición humana a toxinas (Perkins *et al.*, 1990), SIDA y otros estudios de virus (Lynn, 2001), contaminantes en cuerpos de animales diversos (Harris *et al.*, 2003; Hobbs *et al.*, 2003), sedimentos químicos (Clarke, 1998), astronomía (Isobe *et al.*, 1986), química de rocas y minerales (Miesch, 1967), y en estudios de calidad de agua en ríos y mantos freáticos (Helsel y Gilliom, 1986). De esta manera, el tema de cómo mejorar el análisis estadístico de datos censurados por la izquierda es concerniente a muchos científicos en diversas áreas del conocimiento.

En estudios médicos e industriales son mas frecuentes los datos censurados por la derecha, donde sólo se conoce que las observaciones fueron más grandes que un valor establecido (Lee y Wang, 2003). Un ejemplo es el registro del tiempo que la gente vive después de recibir un tratamiento médico. Dos tratamientos pueden ser comparados para medir su efectividad, por ejemplo la mediana del tiempo de sobrevivencia después del tratamiento; esto ocurre en diferentes tiempos para diferentes pacientes, por lo que el tratamiento preferido es el que resulta con mayor expectativa de vida. Para pacientes que aun viven después que el experimento finaliza, los tiempos de supervivencia no son conocidos exactamente, denominándose valores “mayor que” o datos censurados por la derecha.

Pocos estudios en ciencias ambientales han recomendado, o realmente usado, técnicas para datos censurados adoptadas de otras disciplinas. Quizá el primero pudo haber sido Miesch (1967), quien promovió el uso de estimación por máxima verosimilitud para calcular valores medios de datos minerales. Millard y Deverel (1988) demostraron que las pruebas de rangos y pruebas de comparación de dos grupos para datos censurados, podrían ser usadas para estudios de calidad del agua. Helsel (1990) se enfocó en el estudio de técnicas de análisis de supervivencia para pruebas de hipótesis y regresión. Por otro lado, Slyman y otros (1994) usaron máxima verosimilitud para regresión de datos censurados de concentraciones. Asimismo, She (1997) utilizó estimadores Kaplan-Meier para obtener estadísticos descriptivos y caracterizar datos de calidad de agua.

Un antecedente importante es una guía para el uso de técnicas para datos censurados por la izquierda en estudios ambientales, publicado por Akritas (1994) en el Handbook of Statistics Vol 12. Sin embargo, ni los ejemplos presentados y ni el Handbook han cambiado las prácticas estándar, las cuales están basadas en la omisión o sustitución de los datos no detectados por una fracción de los LD. De modo que ninguno de los métodos sugeridos en esta publicación se usa actualmente en ciencias ambientales.

Por lo general, cuando se tienen datos no detectados, dos métodos demasiado simplistas son usados. El primero consiste en no considerar los datos no detectados, lo cual produce resultados sesgados al eliminar los valores más bajos entre las observaciones. En este caso las pruebas o estadísticos que resultan no se aplican realmente para el total de datos muestreados, sino sólo a una parte de ellos, los de la región más alta de la distribución. Esta práctica es muy frecuente cuando el interés radica únicamente en la detección y en las mediciones de las observaciones detectadas (Kolpin et al., 2002; Lundgren y Lopes, 1999). Sin embargo, al comparar entre grupos de datos y entre estudios, los resultados dependerán de los límites de detección en vigor al tiempo en que fueron analizados. Estos resultados pueden diferir erróneamente debido a que las estimaciones son realizadas usando sólo las observaciones detectadas, las cuales estarían alejadas del contexto general de la distribución de probabilidad. Esta práctica hace vulnerable a los autores respecto a sus aseveraciones debido a los sesgos inducidos en el análisis (Till, 2003).

El segundo método comúnmente usado para manejar datos censurados es asignar fracciones arbitrarias de los LD a cada observación censurada (sustitución o fabricación). En numerosos estudios realizados a lo largo de los años, desde Nehls y Akland (1973), a Buckley y otros (1997), y Hobbs y otros (2003); se ha registrado la sustitución para datos no detectados por un medio del LD. Manuales de procedimientos de al menos tres agencias federales en USA, han recomendado también esta práctica (EPA, 1998; Army, 1998; Navy, 1999). Sin embargo, la sustitución puede conducir a indicios no presentes en los datos originales u ocultan una señal que realmente existe.

En un estudio usando simulación para comparar diferentes métodos estadísticos descriptivos de cálculo para datos censurados, Singh y Nocerino (2002) reportaron que la sustitución resulta en estimadores sesgados. El trabajo fue realizado para el caso más simple de datos con sólo un límite de detección, al considerar casos con límites de detección múltiples los errores por sustitución se incrementaron.

En este sentido, el manejo inadecuado de los datos censurados es evidencia de un escepticismo general respecto a la información contenida y obtenida a partir de ese tipo de observaciones. En realidad, los datos censurados contienen una gran cantidad de información que puede ser extraída usando métodos eficientes; esta información es tan valiosa como la obtenida de los valores conocidos. Procedimientos eficientes para datos censurados combinan los valores superiores a los límites de detección con la información contenida en la proporción de los datos debajo de estos mismos límites.

2.2. Análisis de Datos No Detectados

Acorde a lo anterior, Helsel (2005) discutió tres enfoques para extraer información de los conjuntos de datos que incluyen datos no detectados: sustitución, estimación por máxima verosimilitud y métodos no paramétricos. El primer enfoque, que se refiere a la sustitución, no es recomendable y solo se discutirá brevemente. Los otros dos métodos son enfoques viables para el análisis de datos censurados y son usados como métodos estándar en otras disciplinas, además de las ciencias ambientales. A continuación se presenta una descripción general de cada enfoque:

El primer enfoque consiste en la aplicación de métodos de sustitución, en éstos se calculan estadísticas para datos censurados fabricando valores como función de los límites de detección registrados. Estadísticas descriptivas, pruebas de hipótesis y modelos de regresión se calculan usando esos números fabricados, junto con valores detectados superiores al límite de detección. Los métodos de sustitución han sido usados ampliamente, pero no tienen bases teóricas. Varios estudios han mostrado que el método de sustitución no funciona bien, cuando se comparan con los otros dos tipos de enfoques mencionados anteriormente (Gleit, 1985; Helsel y Cohn, 1988; Singh y Nocerino, 2002). La sustitución es usada porque es fácil de hacer, sin embargo, la fabricación de valores no es justificable dado que la elección de dichos valores es arbitraria y los resultados estadísticos difieren dependiendo de los valores elegidos.

El segundo método consiste en la estimación por máxima verosimilitud (MV), ésta ha sido usada muy poco en estudios ambientales. Los estudios realizados por Owen y DeRouen (1980) para calidad del aire y Miesch (1967) para geoquímica son dos de los primeros ejemplos en la literatura. La estimación por MV utiliza tres piezas de información fundamentales: a) datos mayores a los límites de detección, b) proporción de datos por debajo de los límites de detección, y c) una distribución de probabilidad asumida. En este caso se supone que todos los datos, detectados y no detectados, siguen la misma distribución, por ejemplo la lognormal. Los parámetros son estimados de tal forma que se tenga el mejor ajuste a la distribución con los valores detectados y la proporción de los no detectados. La consideración crucial para la estimación por MV consiste en qué tan bueno es el ajuste de los datos al modelo asumido.

El tercer enfoque se refiere al uso de métodos no paramétricos; llamados así, debido a que no se necesita estimar parámetros de alguna distribución asumida. En vez de esto, se usa la posición relativa (rangos) de los datos, una reflexión de los cuantiles de los datos. Debido a que esos métodos no requieren de un supuesto respecto a alguna distribución de los datos, son llamados algunas veces como métodos de distribución libre. Los métodos no paramétricos son especialmente usados para datos censurados debido a que utilizan eficientemente la información disponible. Se sabe que los datos no detectados son menores que sus límites de detección, así que tendrán rangos menores. Estos métodos no requieren estimadores de parámetros o de medidas de centralidad o dispersión, pero si requieren de un orden relativo respecto a las distancias entre los datos detectados y no detectados.

2.3. El Modelo Lognormal

Estudios en ciencias ambientales han reportado que las concentraciones de varias sustancias tienen frecuentemente distribuciones de tipo lognormal, o aproximadas a ésta. Es decir, los *logaritmos* de las concentraciones observadas se ajustan a un modelo aproximadamente normal, o Gaussiano, sobre la mayor parte del rango observado. Ejemplos de este tipo incluyen material radiactivo en suelos, contaminantes

en aire, calidad del aire acondicionado, residuos de metales en ríos, metales en tejidos biológicos (Ott, 1995). De esta manera, la distribución lognormal es usada para modelar varios tipos de datos de contaminación ambiental como datos de calidad del aire, de contaminantes radiactivos, metales pesados en peces y otras concentraciones de residuos radiactivos en tejidos humanos (Gilbert, 1987).

Pueden definirse distribuciones lognormales de dos, tres y cuatro parámetros. La función de densidad de la distribución *lognormal de dos parámetros* esta dada por

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(\log x - \mu)^2\right] \quad x > 0, -\infty < \mu < \infty, \sigma > 0, \quad (2.1)$$

donde μ y σ^2 , los parámetros de la distribución lognormal, son la media y la varianza de la variable aleatoria $Y = \log X$. La mediana, media y varianza para una población lognormal $LN(\mu, \sigma)$ son:

$$\begin{aligned} \text{Mediana: } M &= \exp(\mu) \\ \text{Media: } E &= \exp\left(\mu + \sigma^2/2\right) \\ \text{Varianza: } V &= \exp(\sigma^2)\left[\exp(\sigma^2) - 1\right] \exp(2\mu) \end{aligned} \quad (2.2)$$

Observe que la mediana depende solo de μ , en contraste la media y la varianza dependen de ambos parámetros. Algunos autores refieren la media geométrica $\exp(\mu)$ y la desviación estándar geométrica $\exp(\sigma)$, como los parámetros de la distribución lognormal. Por otro lado, el α -ésimo cuantil para la distribución lognormal, se define por

$$x_\alpha = \exp\left(\mu + \sigma^2 z_{1-\alpha}\right), \quad (2.3)$$

donde $z_{1-\alpha}$ es el cuantil $1-\alpha$ de la distribución normal estándar, $N(0,1)$.

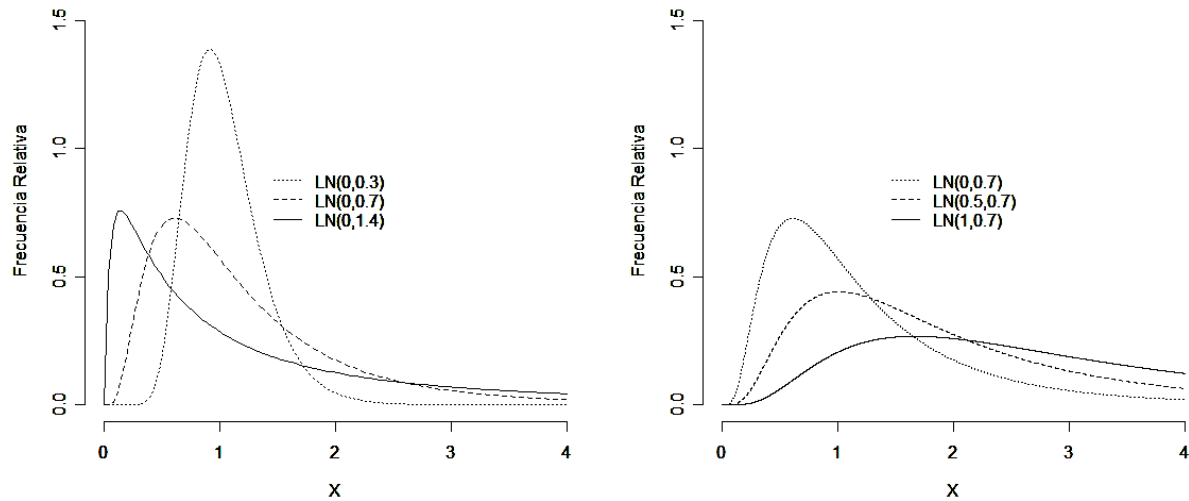


Figura 1. Distribuciones lognormal para distintos valores μ y σ .

Algunas distribuciones lognormales de dos parámetros se ilustran en la Figura 1. La distribución es descrita con detalles por Aitchinson y Brown (1969) y Johnson y Kotz (1970), quienes muestran varios métodos para estimar los parámetros μ y σ . Se ha demostrado que el método por máxima verosimilitud es preferido para obtener:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n \log x_i, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (\log x_i - \hat{\mu})^2$$

La función de densidad de la distribución *lognormal de tres parámetros* está dada por

$$f(x) = \frac{1}{(x-\tau)\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2} [\log(x-\tau) - \mu]^2\right\}, \quad \begin{array}{l} x > \tau, -\infty < \mu < \infty, \\ \sigma > 0, -\infty < \tau < \infty. \end{array} \quad (2.4)$$

Si se comparan 2.1 y 2.4, se observa que $x - \tau$ tiene una distribución lognormal de dos parámetros. El tercer parámetro τ , que puede ser positivo o negativo, simplemente recorre la distribución lognormal de dos parámetros a la derecha o izquierda sin cambiar su forma. Dado que las concentraciones de contaminantes no pueden ser negativas, una lognormal de tres parámetros con τ negativa parece inapropiada. Sin embargo, mediciones negativas pueden ocurrir debido a mediciones erróneas cuando

las concentraciones verdaderas son muy cercanas a cero. Por otro lado, la distribución lognormal de cuatro parámetros está acotada superior e inferiormente sobre los valores posibles de la variable. Las formas sesgadas a la izquierda o derecha también son posibles y han sido aplicadas a datos de calidad del aire por Aitchinson y Brown (1969).

La génesis de la distribución lognormal se debe a Galton (1879), quien demostró que si

X_1, X_2, \dots, X_n son variables aleatorias independientes positivas y $T_n = \prod_{j=1}^n X_j$, entonces,

$$\log T_n = \sum_{j=1}^n \log X_j.$$

De modo que si las variables aleatorias $\log X_j$, son tales que puede aplicarse un tipo de límite central, entonces la distribución estandarizada de $\log T_n$ tiende a una distribución normal estándar $N(0,1)$ cuando n tiende a infinito. Así, la distribución límite de T_n sería lognormal.

Ahora, el momento r -ésimo de X alrededor del cero es

$$\begin{aligned} \mu_r' &= E[X^r] = E[\exp(r \log X)] \\ &= \exp\left(r\mu + \frac{1}{2}r^2\sigma^2\right) \end{aligned} \quad (2.5)$$

μ_r' se incrementa rápidamente con r . Heyde (1963) demostró que la sucesión $\{\mu_r'\}$ no pertenece solamente a la distribución lognormal; por lo que ésta no puede definirse por sus momentos. Entonces, para la distribución lognormal el valor esperado es

$$\mu_1' = \exp\left(\mu + \frac{1}{2}\sigma^2\right) \quad (2.6)$$

y los momentos centrales de ordenes inferiores son

$$\begin{aligned}\mu_2 &= e^{2\mu} e^{\sigma^2} (e^{\sigma^2} - 1) \\ &= w(w-1)e^{2\mu}\end{aligned}\tag{2.7}$$

con lo que

$$\begin{aligned}\sigma(X) &= e^\mu \sqrt{w(w-1)} \\ \mu_3 &= w^{\frac{3}{2}} (w-1)^2 (w+2) e^{3\mu}\end{aligned}\tag{2.8}$$

donde $w = \exp(\sigma^2)$. Wartmann (1956) ha dado la fórmula general

$$\mu_r = \frac{\mu_2^{\frac{1}{2}r}}{(w-1)^{\frac{1}{2}r}} \sum_{j=0}^r (-1)^j \binom{r}{j} w^{\frac{1}{2}(r-j)(r-j-1)}.$$

El coeficiente de variación es $(w-1)^{\frac{1}{2}}$ y tampoco depende de μ . La distribución de X es unimodal, y la moda esta dada por

$$\text{moda}(X) = \exp(\mu - \sigma^2).\tag{2.9}$$

Es claro que el valor X alfa tal que $\Pr[X \leq X_\alpha] = \alpha$ está relacionado al correspondiente cuantil. En particular la mediana de X (correspondiente a $\alpha = 1/2$) es

$$\text{mediana}(X) = e^\mu.\tag{2.10}$$

Ahora, comparando (2.6), (2.9) y (2.10) se tiene que

$$\begin{aligned}E[X] &> \text{mediana}(X) > \text{moda}(X), \\ \frac{\text{moda}(X)}{E[X]} &= e^{-3\sigma^2/2} = \left[\frac{\text{mediana}(X)}{E[X]} \right]^3.\end{aligned}\tag{2.11}$$

Cuando σ tiende a cero, la distribución lognormal tiende a la distribución $N(0,1)$.

La distribución lognormal de dos parámetros es una representación más realista para las distribuciones de atributos como peso, altura, densidad, que la distribución normal. Esas cantidades no pueden tomar valores negativos, pero la distribución normal asigna una probabilidad positiva a tales eventos, mientras que la distribución lognormal de dos parámetros no lo hace. Además, tomando σ suficientemente pequeña, es posible construir una distribución lognormal muy parecida a alguna distribución normal. De aquí, si una distribución normal es aparentemente adecuada, podría remplazarse por una distribución lognormal apropiada.

2.4. Distribuciones de Localización-Escala

Una variable aleatoria Y pertenece a la familia de distribuciones de localización-escala si su función de distribución puede ser expresada como

$$\Pr(y \leq Y) = F(y; \mu, \sigma) = \Phi\left(\frac{y - \mu}{\sigma}\right), \quad (2.12)$$

donde Φ no depende de ningún parámetro desconocido. Se dice que $-\infty < \mu < \infty$ es un parámetro de localización y que $\sigma > 0$ es un parámetro de escala. La sustitución muestra que Φ es la función de distribución de Y cuando $\mu = 0$ y $\sigma = 1$. También, Φ es la función de distribución de $(y - \mu)/\sigma$. Las distribuciones de localización-escala son importantes por que muchas de las distribuciones ampliamente usadas pertenecen a esta familia o están muy relacionadas con ellas. Estas distribuciones incluyen la distribución exponencial, la normal, la logística y de valores extremos.

Análogamente, una variable aleatoria Y pertenece a la familia de distribuciones de log-localización-escala si $Y = \log(y)$ es un miembro de la familia de distribuciones de localización-escala. Las distribuciones Weibull, lognormal y loglogística son las más importantes de esta familia.

2.5. Gráficos de Probabilidad para la Distribución Lognormal

Los gráficos de probabilidad son usados para observar la adecuación de un modelo de probabilidad y obtener estimaciones gráficas de los parámetros del modelo. El gráfico de $\{t \text{ vs } F(t)\}$ puede ser linealizado mediante una transformación de $F(t)$ y t tal que la relación entre las variables transformadas sea lineal. La escala de probabilidad resultante es no lineal y es llamada la “escala de probabilidad”. La escala de datos es usualmente una escala log-lineal, dependiendo de la distribución y el tipo de gráfico.

La función cuantil para $F(t)$ permite linealizar la función de distribución. Para la distribución lognormal, la función cuantil es $t_q = \exp[\mu + \Phi^{-1}(q)\sigma]$, donde $\Phi^{-1}(q)$ es el q -ésimo cuantil de la distribución normal estándar, $N(0,1)$. Esto conduce a

$$\log t_q = \mu + \Phi^{-1}(q)\sigma . \quad (2.13)$$

De donde $\{\log(t_q) \text{ vs } \Phi^{-1}(q)\}$ se grafica como una línea recta. El parámetro $t_{.5} = \exp(\mu)$ se lee desde la escala Y al punto donde la función de distribución interseca $\Phi^{-1}(q) = 0$.

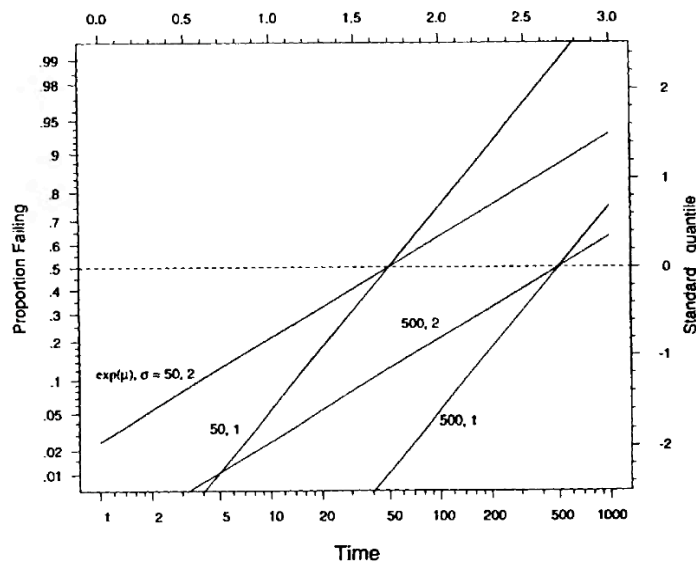


Figura 2. Grafico de probabilidad lognormal con líneas rectas para $\exp(\mu) = 50,500$; $\sigma=1,2$.

Fuente: Meeker and Escobar (1998), Probability Plotting.

La pendiente de esta línea recta contra la escala de cuantiles es $1/\sigma$. Toda distribución lognormal se grafica como una recta con pendiente positiva, Figura 1.5.

2.6 Máxima Verosimilitud

La idea general de la inferencia por máxima verosimilitud es ajustar modelos a los datos, considerando combinaciones de parámetros y modelos para las que la probabilidad de los datos sea grande. Los métodos pueden ser aplicados con una gran variedad de modelos paramétricos con datos censurados. También es posible ajustar modelos con variables explicatorias (es decir, análisis de regresión). La teoría desarrollada, garantiza que estos métodos son estadísticamente eficientes (es decir, proporcionan los estimadores más exactos). Esas propiedades son aproximadas con tamaños de muestras pequeñas y varios estudios han mostrado que estos métodos tienen igual desempeño que otros disponibles (Meeker & Escobar 1998).

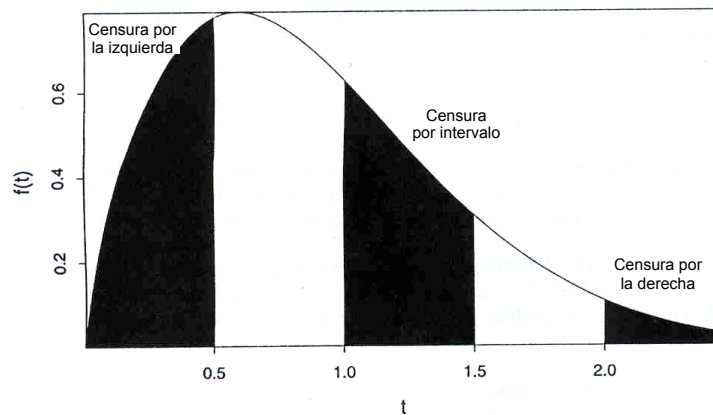


Figura 3. Intervalos de incertidumbre y contribuciones a la verosimilitud para la censura.

Fuente: Meeker and Escobar (1998), Models, Censoring, and Likelihood for Failure-Time Data

La función de verosimilitud puede verse con la probabilidad conjunta de los datos. La Figura 2 ilustra los intervalos de incertidumbre para los ejemplos de datos censurados por la izquierda (no detectados), por intervalo y por la derecha. Las contribuciones a la verosimilitud, para cada caso, mostrada en el Cuadro 1 pueden verse como la probabilidad del evento en el correspondiente intervalo de incertidumbre.

Cuadro 1. Contribución a la función de verosimilitud para datos censurados

Censura	Rango	Verosimilitud
observaciones censuradas por intervalo	$t_{i-1} < T < t_i$	$F(t_i) - F(t_{i-1})$
observaciones censuradas por la izquierda	$T \leq t_i$	$F(t_i)$
observaciones censuradas por la derecha	$T > t_i$	$1 - F(t_i)$

Ahora, dados un conjunto de datos y modelo especificado, la verosimilitud es una función de los parámetros desconocidos del modelo. La forma de la función de verosimilitud dependerá del modelo asumido, los datos disponibles (por ejemplo, censurados) y el problema de estudio. Entonces, para una muestra censurada por la izquierda de n observaciones independientes, la verosimilitud es

$$L(\theta; \text{datos}) = C \prod_{i=1}^n L_i(\theta; \text{datos}_i) = C \prod_{i=1}^n [f(t_i; \theta)]^{\delta_i} [F(t_i; \theta)]^{1-\delta_i}. \quad (2.14)$$

$L_i(\theta; \text{datos}_i)$ es la probabilidad de la observación i , datos_i son los datos para la observación i . Las funciones f, F son la densidad y distribución del modelo probabilístico para la variable aleatoria Y , respectivamente. El vector θ , de dimensión r , contiene a los parámetros a estimar; $\delta_i = 1$ para una observación exacta y $\delta_i = 0$ para una observación censurada por la izquierda; C es una constante que no depende de θ .

El valor de θ que maximiza $L(\theta)$ provee un estimador de máxima verosimilitud (EMV) y se denota por $\hat{\theta}$. En la práctica, se maximiza el logaritmo natural de $L(\theta)$, $\log L(\theta)$, ya que esta expresión ofrece más versatilidad analítica y sus máximos coinciden con el de $L(\theta)$. La maximización de la logverosimilitud se obtiene de sus derivadas parciales con respecto al vector θ . Esto es, $\hat{\theta}$ surge de la solución simultánea a las ecuaciones,

$$\frac{\partial L(\theta)}{\partial \theta_1} = 0, \quad \frac{\partial L(\theta)}{\partial \theta_2} = 0, \quad \dots, \quad \frac{\partial L(\theta)}{\partial \theta_r} = 0. \quad (2.15)$$

2.7. Verosimilitud para Datos Lognormales No Detectados

La verosimilitud para una muestra aleatoria lognormal t_1, \dots, t_n de una variable aleatoria $T > 0$, conteniendo observaciones exactas y no detectados, se escribe como

$$L(\mu, \sigma) = \prod_{i=1}^n \left[\frac{1}{\sigma t_i} \phi\left(\frac{\log t_i - \mu}{\sigma}\right) \right]^{\delta_i} \left[\Phi\left(\frac{\log t_i - \mu}{\sigma}\right) \right]^{1-\delta_i}, \quad (2.16)$$

Φ y ϕ denotan las funciones de densidad y de distribución de una variable aleatoria $N(0,1)$. Pueden omitirse los términos $1/t_i$, ya que no dependen de los parámetros, sin tener algún efecto en la localización de los estimadores de máxima verosimilitud.

2.8. Prueba de Razón de Verosimilitud

Una prueba de significancia de razón de verosimilitud evalúa si un modelo general ajusta mejor los datos que uno restringido (Anexo A.5). Entonces, un modelo restringido ajusta los datos casi tan bien como el modelo general si asintóticamente,

$$-2 \log \left[L_1(\theta_0) / L_2(\hat{\theta}) \right] \sim \chi_v^2, \quad (2.17)$$

donde $L_2(\hat{\theta})$ es la máxima verosimilitud del modelo general en $\hat{\theta}$; $L_1(\theta_0)$ es la máxima verosimilitud del modelo restringido en θ_0 ; χ_v^2 es una variable aleatoria χ^2 con v grados de libertad, obteniéndose de la diferencia entre la dimensión de $\hat{\theta}$ y la dimensión de θ_0 .

2.9. El Algoritmo EM (Expectation-Maximization)

En muchas situaciones la función de verosimilitud es muy compleja y se dificulta su análisis, el problema radica en la presencia de datos faltantes, datos censurados, valores perdidos o variables latentes. Si esa información no observada fuera conocida,

la verosimilitud tendría una forma más tratable. Aquí es donde puede resultar útil el *principio de datos aumentados* que, según Tanner (1994), consiste en aumentar los datos observados con datos latentes, de forma que la distribución final aumentada sea simple para calcular la distribución final observada (Bermúdez, 2004).

El algoritmo EM utiliza el principio de datos aumentados para obtener el EMV, o la moda de la distribución final. Con este algoritmo se consigue formalizar y justificar la idea de que dada una primera estimación de los parámetros, se pueden predecir los datos faltantes, se reestiman los parámetros, continuando iterativamente hasta la convergencia. EM es un proceso iterativo para obtener el EMV de θ , que parte de una primera aproximación $\theta^{(0)}$, (Dempster et al., 1977; Flury y Zoppè, 2001); y la interacción i -ésima ($i = 1, 2, \dots$) consta de dos etapas:

Etapla E (Esperanza). Calcular $Q(\theta|\theta^{(i-1)}) = E(\log f(x|\theta)|y, \theta^{(i-1)})$.

Etapla M (Maximización). Obtener $\theta^{(i)}$ que maximiza en θ a $Q(\theta|\theta^{(i-1)})$.

El flujo del algoritmo EM se muestra en la Figura 2.3. Como en todo proceso iterativo es preciso especificar una regla de parada, basada en una distancia entre $\theta^{(i)}$ y $\theta^{(i-1)}$ o, mejor entre $Q(\theta^{(i-1)}|\theta^{(i-1)})$ y $Q(\theta^{(i)}|\theta^{(i-1)})$, que defina las condiciones de convergencia.

Las dificultades pueden presentarse en el cálculo de $E(\log f(x|\theta)|y, \theta^{(i-1)})$, que incluso puede no existir. También puede resultar complejo el cálculo del máximo de $Q(\theta^{(i)}|\theta^{(i-1)})$. Por fortuna, casi siempre esos problemas se resuelven sin gran dificultad.

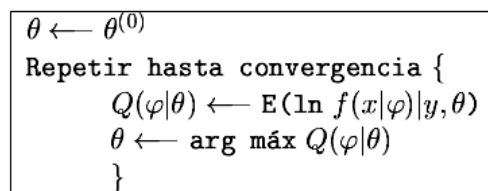


Figura 4. Esquema general del Algoritmo EM

2.10. Prueba No Paramétrica Logrank

El problema de comparar distribuciones de supervivencia surge comúnmente en investigaciones biomédicas; por ejemplo, cuando se desea comparar la capacidad de dos o más tratamientos para prolongar el tiempo de vida o conservar la salud. Los tiempos de supervivencia de los diferentes grupos varían y pueden ser ilustradas con gráficos de funciones de supervivencia estimadas; sin embargo, es solo una idea burda de las diferencias entre distribuciones y no revelan si las diferencias son significativas. De modo que una prueba estadística es necesaria. La prueba no paramétrica logrank es una de las más utilizadas para datos con observaciones censuradas.

En este sentido, supóngase que existen n_1 y n_2 pacientes quienes reciben los tratamientos 1 y 2, respectivamente. Sean $x_1 < \dots < x_{r_1}$ los r_1 tiempos exactos de vida y $x_{r_1+1}^+ < \dots < x_{n_1}^+$ los $n_1 - r_1$ tiempos de vida censuradas en el grupo 1. En el grupo 2, sean $y_1 < \dots < y_{r_2}$ las r_2 tiempos de vida exactos y $y_{r_2+1}^+ < \dots < y_{n_2}^+$ los $n_2 - r_2$ tiempos de vida censuradas. Esto es, al final del estudio $n_1 - r_1$ pacientes del grupo 1 y $n_2 - r_2$ pacientes del grupo 2 aún viven. Suponiendo que las observaciones en el grupo 1 son muestras de una distribución con función de supervivencia $S_1(t)$ y las observaciones en el grupo 2 son muestras de una distribución $S_2(t)$. Entonces el juego de hipótesis a considerar es

$$H_0 : S_1(t) = S_2(t), \text{ contra las alternativas } \begin{cases} H_1 : S_1(t) > S_2(t), \\ H_1 : S_1(t) < S_2(t), \\ H_1 : S_1(t) \neq S_2(t). \end{cases}$$

Cuando no existen observaciones censuradas, pruebas no paramétricas estándar pueden ser usadas para comparar distribuciones de supervivencias. Por ejemplo, la prueba de Wilcoxon, de Mann-Whitney, la prueba de signo, etc.

Sean $t_{(1)} < \dots < t_{(k)}$ los distintos tiempos a la falla en los dos grupos juntos y m_i el número de tiempos a la falla iguales a t_i , o la multiplicidad de t_i , tal que $\sum_{i=1}^k m_{(i)} = r_1 + r_2$. La

generalización de la prueba de Savage (1956), frecuentemente referida como la *prueba logrank* (Peto y Peto, 1972), está basada en un conjunto de rangos w_i asignado a las observaciones. Los rangos son funciones de los logaritmos de las funciones de supervivencia y se estima la función de supervivencia en tiempo $t_{(i)}$ usando

$$-e(t_{(i)}) = \sum_{j \leq t_{(i)}} \frac{m_{(j)}}{r_{(j)}}. \quad (2.18)$$

Los rangos sugeridos por Peto y Peto (1970) son $w_i = 1 - e(t_{(i)})$ para una observación no censurada $t_{(i)}$ y $-e(T)$ para una observación censurada en T . En la práctica, para una observación censurada t_i^+ , $w_i = e(t_{(j)})$, donde $t_{(j)}$ es la observación no censurada mas grande que $t_{(j)} \leq t_i^+$. Las observaciones censuradas reciben rangos negativos. Los rangos w suman cero para los dos grupos juntos. La *prueba logrank* esta basada en la suma S de los rangos w de los grupos. La varianza permutacional de S está dada por

$$Var(S) = \frac{n_1 n_2 \sum_{i=1}^{n_1+n_2} w_i^2}{(n_1 + n_2)(n_1 + n_2 - 1)}, \quad (2.19)$$

la cual puede ser re-escrita como

$$V = \left[\sum_{j=1}^k \frac{m_{(j)}(r_{(j)} - m_{(j)})}{r_{(j)}} \right] \frac{n_1 n_2}{(n_1 + n_2)(n_1 + n_2 - 1)}. \quad (2.20)$$

La prueba estadística $L = S / \sqrt{Var(S)}$ tiene una distribución normal estándar asintótica bajo la hipótesis nula. Si S es obtenida del grupo 1, la región crítica es $L < -Z_\alpha$ y si S es obtenida del grupo 2, la región crítica es $L > Z_\alpha$, donde α es el nivel de significancia para probar la hipótesis $H_0 : S_1 = S_2$ contra $H_1 : S_1 > S_2$.

3. OBJETIVOS

3.1 Objetivo General

- Proponer una prueba paramétrica para comparar concentraciones medias de contaminantes de poblaciones lognormales, conteniendo datos no detectados y covariables, basada en la función de verosimilitud y la reparametrización de los parámetros del modelo de regresión lognormal con variables indicadoras.

3.2 Objetivos Particulares

- Implementar el Algoritmo EM para reducir la complejidad de la optimización de la función de verosimilitud debida a los datos no detectados.
- Desarrollar un criterio de comparación basado en regiones e intervalos de confianza aproximados obtenidos por el método de Wald, para los parámetros y coeficientes del modelo de regresión lognormal con variables indicadoras.
- Describir poblaciones con datos no detectados, comparar medias bajo modelos de homogeneidad y heterogeneidad del parámetro de escala σ ; proponer una forma de selección del mejor modelo, y abordar el problema de comparación de medias con datos agrupados, a través de los modelos de efectos mixtos.
- Comparar la potencia de la prueba propuesta con la prueba no paramétrica logrank mediante simulación, con distintos tamaños de muestra y porcentajes de información censurada por la izquierda. Así también, observar la potencia de la prueba, al comparar poblaciones exponenciales y Gumbel.
- Desarrollar programas en R necesarios para el estudio de simulación y desarrollo de los ejemplos de aplicación.

4. METODOLOGÍA DESARROLLADA

A continuación se describen las herramientas y procedimientos desarrollados para dar cumplimiento a los objetivos planteados en el trabajo.

4.1. Aplicación del Algoritmo EM en la Familia Exponencial

El algoritmo EM se simplifica considerablemente al trabajar con una muestra aleatoria de la familia exponencial (Figura 4.1). Esto es, cuando la verosimilitud de los datos completos es:

$$f(x|\theta) = a(\theta)b(x)\exp\{\theta't(x)\}, \quad (4.1)$$

donde $t(x)$ es un vector de estadísticos mínimos suficientes y $\theta = (\theta_1, \dots, \theta_k)'$ es el vector paramétrico en la representación natural del modelo. En estas circunstancias,

$$Q(\theta|\theta^{(i-1)}) = \ln a(\theta) + E(\ln b(x)|y, \theta^{(i-1)}) + \theta'E(t(x)|y, \theta^{(i-1)}), \quad (4.2)$$

y derivando respecto a θ la función $Q(\theta|\theta^{(i-1)})$, el valor de $\theta^{(i)}$ será la solución a,

$$\partial(-\ln a(\theta))/\partial\theta = E[t(x)|y, \theta^{(i-1)}]. \quad (4.3)$$

$\theta \leftarrow \theta^{(0)}$ Repetir hasta convergencia { $t^{(i)} \leftarrow E(t(x) y, \theta)$ $\theta \leftarrow$ solución a $E(t(x) \varphi) = t^{(i)}$ }
--

Figura 5. Algoritmo EM en la familia exponencial.

La parte izquierda de (4.3) coincide con la esperanza de los estadísticos suficientes. Además, la segunda derivada con respecto a θ de $Q(\theta|\theta^{(i-1)})$ es $\partial(-E(t(x)|\theta))/\partial\theta$ que, intercambiando derivación e integración, resulta ser igual a $-Var(t(x))$. Por tanto es negativa definida, con lo que $\theta^{(i)}$ es el máximo global de $Q(\theta|\theta^{(i-1)})$.

Así pues, la Etapa M consiste en resolver, en θ , la ecuación $E(t(x)|\theta) = E(t(x)|y, \theta^{(i-1)})$

4.2. Obtención de la Matriz de Información Mediante el Algoritmo EM

El algoritmo EM no genera estimadores para la matriz de varianzas y covarianzas de los estimadores, razón por la cual se han hecho modificaciones al mismo con la finalidad de salvar este problema. Una de tales modificaciones es la hecha por Oakes (1999), y según Robert y Casella (2004) dicha modificación es simple y muy útil. Oakes (1999) logró mostrar que si $L(y;\theta)$ es la log-verosimilitud de la muestra entonces:

$$\frac{\partial^2 L(y;\theta)}{\partial\theta^2} = \left(\frac{\partial^2 Q(\theta';\theta)}{\partial\theta'^2} + \frac{\partial^2 Q(\theta';\theta)}{\partial\theta'\partial\theta} \right)_{\theta'=\theta} . \quad (4.4)$$

La varianza aproximada de $\hat{\theta}$ se calcula con:

$$Var\hat{\theta} \approx \left[\frac{\partial^2 L(y;\theta)}{\partial\theta^2} \right]^{-1} . \quad (4.5)$$

Con la aproximación en (4.5) se tiene una expresión para calcular, las varianzas-covarianzas de los estimadores de Máxima Verosimilitud (EMV) y consecuentemente los intervalos de confianza aproximados. La obtención de estas derivadas puede requerir un poco más de trabajo, pero en casi todos los casos no es muy difícil.

4.3. Construcción de Intervalos y Regiones de Confianza Aproximados

Los intervalos y regiones de confianza por aproximación normal están basadas en una aproximación cuadrática de la log-verosimilitud y son apropiados cuando la log-verosimilitud es aproximadamente cuadrática sobre la región de confianza. Con muestras grandes, bajo condiciones de regularidad (ver Anexo A.3), la log-verosimilitud es aproximadamente cuadrática y entonces la aproximación normal y los intervalos serán congruentes. El tamaño de muestra requerido para una buena aproximación no es fácilmente elegible debido a que depende del modelo, de la cantidad de información censurada y de la cantidad de interés en particular.

La aproximación normal de muestras grandes para la distribución de EMVs (Anexo A.4) puede ser usada para construir intervalos (regiones) de confianza aproximados para funciones escalares (vectoriales) de θ . En particular, una región aproximada del $100(1-\alpha)\%$ de confianza para θ es el conjunto de todos los valores de θ en el elipsoide

$$(\hat{\theta} - \theta)(\hat{\Sigma}_{\hat{\theta}})^{-1}(\hat{\theta} - \theta) \leq \chi_{(1-\alpha; r)}^2, \quad (4.6)$$

donde r es la longitud de θ . Esto se conoce como el Método de Wald, conocido también como el método de aproximación normal. Esta región de confianza esta basada en la distribución que resulta, asintóticamente, cuando es evaluado en el valor verdadero de θ , el estadístico de Wald

$$W(\theta) = (\hat{\theta} - \theta)(\hat{\Sigma}_{\hat{\theta}})^{-1}(\hat{\theta} - \theta) \sim \chi_r^2 \quad (4.7)$$

Más generalmente, sea $g(\theta)$ una función vectorial de θ . Entonces, una región aproximada del $100(1-\alpha)\%$ de confianza para un subconjunto r_1 -dimensional $g_1 = g_1(\theta)$, de la partición $g = [g_1(\theta), g_2(\theta)]$, es el conjunto de los valores de g_1 en el elipsoide

$$(\hat{g}_1 - g_1)' (\hat{\Sigma}_{\hat{g}_1})^{-1} (\hat{g}_1 - g_1) \leq \chi_{(1-\alpha; r)}^2, \quad (4.8)$$

donde $\hat{g}_1 = g_1(\hat{\theta})$ es el EMV de $g_1(\theta)$ y $\hat{\Sigma}_{\hat{g}_1}$ es el estimador local de la matriz de covarianza de \hat{g}_1 . El estimador $\hat{\Sigma}_{\hat{g}_1}$ puede ser obtenido de,

$$\hat{\Sigma}_{\hat{g}} = \left[\frac{\partial g(\theta)}{\partial \theta} \right] \hat{\Sigma}_{\hat{\theta}} \left[\frac{\partial g(\theta)}{\partial \theta} \right]'. \quad (4.9)$$

Bajo condiciones de regularidad, $\hat{\Sigma}_{\hat{\theta}} = (\hat{I}_{\theta})^{-1}$ es un estimador consistente de Σ_{θ} , donde

$$\hat{I}_{\theta} = - \left[\frac{\partial^2 \log L(\theta)}{\partial \theta \partial \theta'} \right]_{\theta = \hat{\theta}}. \quad (4.10)$$

es la matriz de información observada de Fisher (Anexo A.2). La utilización del algoritmo EM conduce a la aproximación de \hat{I}_{θ} por la ecuación (4.5). Esta región de confianza (o intervalo) esta basado en el resultado distribucional asintótico del subconjunto del estadístico de Wald, cuando evaluado en el valor verdadero de g_1 ,

$$W(g_1) = (\hat{g}_1 - g_1)' (\hat{\Sigma}_{\hat{g}_1})^{-1} (\hat{g}_1 - g_1) \sim \chi_{r_1}^2. \quad (4.11)$$

Como se muestra en Meeker y Escobar (1995), esta región de confianza aproximada (o intervalo) puede ser vista como una aproximación cuadrática para la log-verosimilitud perfil de $g_1(\theta)$ en \hat{g}_1 . Cuando $r_1 = 1$, $g_1 = g_1(\theta)$, es una función escalar de θ ; un intervalo de confianza aproximado del $100(1-\alpha)\%$ es obtenido de la formula familiar

$$\left[\underline{g}_1, \tilde{g}_1 \right] = \hat{g}_1 \pm z_{(1-\alpha/2)} \hat{eS}_{\hat{g}_1}, \quad (4.12)$$

donde, $\hat{es}(g_1) = \sqrt{\hat{Var}[g_1(\hat{\theta})]}$ es el estimador local para el error estándar de \hat{g}_1 y $z_{(1-\alpha/2)}$ es el $1-\alpha/2$ cuantil de la distribución normal estándar. Como un caso particular, un intervalo aproximado del $100(1-\alpha)\%$ de confianza para θ_i es,

$$[\underline{\theta}_i, \tilde{\theta}_i] = \hat{\theta}_i \pm z_{(1-\alpha/2)} \hat{es}_{\hat{\theta}_i}, \quad (4.13)$$

donde $\hat{es}_{\hat{\theta}_i}$ es la raíz cuadrada de la ii -ésima entrada de $\hat{\Sigma}_{\hat{\theta}}$, aproximada por (4.5) al implementar el algoritmo EM en el proceso de estimación máximo-verosímil.

4.3.1. Matriz de Varianzas-Covarianzas de los Parámetros

Los intervalos de confianza basados en la distribución normal aproximada por muestras grandes de los EMV requieren un estimador de la matriz de varianza-covarianza para los EMV de los parámetros. Para la distribución lognormal se calcula el estimador local $\hat{\Sigma}_{(\hat{\mu}, \hat{\sigma})}$ de $\Sigma_{(\mu, \sigma)}$ como la inversa de la matriz de información observada de Fisher, $\hat{I}_{(\hat{\mu}, \hat{\sigma})}$, osea,

$$\hat{\Sigma}_{\hat{\mu}, \hat{\sigma}} = \begin{bmatrix} \hat{Var}(\hat{\mu}) & \hat{Cov}(\hat{\mu}, \hat{\sigma}) \\ \hat{Cov}(\hat{\mu}, \hat{\sigma}) & \hat{Var}(\hat{\sigma}) \end{bmatrix} = \left[\begin{array}{cc} -\frac{\partial^2 \log L(\mu, \sigma)}{\partial \mu^2} & -\frac{\partial^2 \log L(\mu, \sigma)}{\partial \mu \partial \sigma} \\ -\frac{\partial^2 \log L(\mu, \sigma)}{\partial \sigma \partial \mu} & -\frac{\partial^2 \log L(\mu, \sigma)}{\partial \sigma^2} \end{array} \right]_{(\mu, \sigma) = (\hat{\mu}, \hat{\sigma})}^{-1} \quad (4.14)$$

La motivación para este estimador es una generalización de las ideas de curvatura de la verosimilitud. Las segundas derivadas parciales describen la curvatura de la $\log L(\mu, \sigma)$ en los EMV. Una mayor curvatura en la superficie de la log-verosimilitud implica mayor verosimilitud concentrada cerca de $\hat{\mu}, \hat{\sigma}$, y eso implica mejor precisión.

4.3.2. Intervalos de Confianza Aproximados para Funciones de μ y σ

Un intervalo de confianza aproximado para una función $g_1 = g_1(\mu, \sigma)$, puede estar basada en la distribución aproximada normal estándar $N(0,1)$ de $Z_{\hat{g}_1} = (\hat{g}_1 - g_1) / \hat{es}_{\hat{g}_1}$. Entonces un intervalo aproximado del $100(1-\alpha)\%$ de confianza para g_1 es $[\underline{g}_1, \tilde{g}_1] = \hat{g}_1 \pm z_{(1-\alpha/2)} \hat{es}_{\hat{g}_1}$. Donde, usando un caso especial de (4.9),

$$\hat{es}_{\hat{g}_1} = \left[\left(\frac{\partial g_1}{\partial \mu} \right)^2 \hat{Var}(\hat{\mu}) + 2 \left(\frac{\partial g_1}{\partial \mu} \right) \left(\frac{\partial g_1}{\partial \sigma} \right) \hat{Cov}(\hat{\mu}, \hat{\sigma}) + \left(\frac{\partial g_1}{\partial \sigma} \right)^2 \hat{Var}(\hat{\sigma}) \right]_{(\mu, \sigma) = (\hat{\mu}, \hat{\sigma})}^{1/2} \quad (4.15)$$

Debido a que σ es un parámetro positivo, es práctica común usar la log-transformación para obtener un intervalo de confianza. Aproximando la distribución de muestreo de $Z_{\log(\hat{\sigma})} = [\log(\hat{\sigma}) - \log(\sigma)] / \hat{es}_{\log(\hat{\sigma})}$ por una distribución $N(0,1)$, un intervalo aproximado del $100(1 - \alpha)\%$ de confianza para σ se obtiene como sigue,

$$\begin{aligned} [\log(\underline{\sigma}), \log(\tilde{\sigma})] &= z_{(1-\alpha/2)} \hat{es}_{\log(\hat{\sigma})} \pm \log(\hat{\sigma}) \\ &\Rightarrow \\ [\underline{\sigma}, \tilde{\sigma}] &= [\hat{\sigma}/w, \hat{\sigma} \times w] \end{aligned} \quad (4.16)$$

donde $w = \exp \left[z_{(1-\alpha/2)} \hat{es}_{\hat{\sigma}} / \hat{\sigma} \right]$ y $\hat{es}_{\hat{\sigma}} = \sqrt{\hat{Var}(\hat{\sigma})}$.

4.4. Aplicación de los Modelos de Regresión con Variables Explicatorias

Un modelo de regresión con variables explicatorias puede explicar porqué algunos especímenes mueren o fallan rápidamente y otros sobreviven a un largo tiempo. Si una variable explicatoria es ignorada, es posible que los estimadores estén sesgados. En

estudios de confiabilidad y supervivencia, las variables explicatorias pueden ser continuas, como el estrés, la temperatura o la dosis; discretas, como el número de tratamientos; categóricas, como el fabricante, localidad, género, etc. La idea de un modelo de regresión es expresar la distribución de tiempo a la falla o de vida como una función de k variables explicatorias $x = (x_1, \dots, x_k)$. Por ejemplo,

$$\Pr(Y \leq y; x) = F(y; x) = F(y). \quad (4.17)$$

Los modelos de regresión pueden provenir de la teoría físico/química, biológica/farmacológica, del ajuste de curvas por observaciones empíricas o alguna combinación de teoría y empirismo. Una clase importante de modelos de regresión permite que un vector de parámetros θ del modelo sea una función de las variables explicatorias. Generalmente se emplea una función con parámetros desconocidos que necesitan ser estimados de los datos. Por ejemplo, si x_i es una variable aleatoria escalar para la observación i , entonces la media de la distribución normal es

$$\mu_i = \beta_0 + \beta_1 x_i. \quad (4.18)$$

Se considera a x_i como una parte fija de los datos i . Cuando los valores de x_i son aleatorios los modelos y métodos estándares proveen inferencias condicionales sobre los valores fijos observados de las variables explicatorias.

4.5. Comparación Mediante Modelos de Regresión con Variables Indicadoras

Las variables consideradas en ecuaciones de regresión pueden tomar valores sobre rangos continuos. Sin embargo, ocasionalmente debe introducirse un factor de dos o más niveles distintos. Por ejemplo, los datos pueden originarse de tres máquinas, dos fábricas o seis operadores. En tales casos no se puede usar una escala continua, pero deben asignarse niveles en orden para tomar en cuenta que las máquinas, fábricas u

operadores pueden tener efectos determinísticos separados sobre la respuesta. Estas variables son llamadas *variables dummy* o *variables indicadoras*, y usualmente no están correlacionadas a ningún factor (Graybill, 1976).

En algunas situaciones es razonable usar un modelo de regresión con variables indicadoras para comparar grupos con un mismo σ y diferencias sólo en los valores de μ . El análisis puede ser hecho con una relación de regresión usando $\mu = \beta_0 + \beta_1 x$, donde $x = 0$ para un grupo y $x = 1$ para el otro. Así, sustituyendo x en el modelo se tiene,

$$\mu(0) = \beta_0, \quad \mu(1) = \beta_0 + \beta_1. \quad (4.19)$$

Además $t_q(1) - t_q(0) = \mu(1) - \mu(0) = \beta_1$, con lo que haciendo comparaciones con un σ común es menos ambiguo debido a que δ no depende del cuantil comparado. También pueden compararse grupos donde las σ son diferentes para los grupos, a través de la relación $\log(\sigma) = \gamma_0 + \gamma_1 x$ y una reparametrización del modelo de regresión como sigue,

$$\begin{aligned} \mu(0) &= \beta_0, & \mu(1) &= \beta_0 + \beta_1; \\ \log[\sigma(0)] &= \gamma_0, & \log[\sigma(1)] &= \gamma_0 + \gamma_1. \end{aligned} \quad (4.20)$$

Sin embargo, debe considerarse que la diferencia $D = t_p(1) - t_p(0)$ no resulta independiente de los cuantiles comparados, pero es una alternativa útil para comparar poblaciones donde se tiene heterogeneidad del parámetro de escala σ en el modelo.

4.6. Aplicación de los Modelos de Efectos Mixtos

Los modelos de efectos mixtos se utilizan para describir la relación entre una variable respuesta y covariables en los datos que están agrupados de acuerdo a uno o más factores de clasificación. Ejemplos de datos agrupados incluyen datos longitudinales, medidas repetidas, datos multiniveles y diseños en bloques. Las observaciones sobre

diferentes especímenes pueden ser consideradas independientes, no así las observaciones sobre un mismo espécimen. Generalmente, los modelos de efectos mixtos son potencialmente útiles cuando un investigador se enfrenta a datos agrupados, en los cuales las observaciones están correlacionadas dentro de los grupos pero independientes entre los diferentes grupos.

Asociando efectos aleatorios comunes a las observaciones del mismo nivel, factor de clasificación o agrupamiento, los modelos mixtos representan convenientemente la estructura de covarianzas inducida por el agrupamiento de datos (Laird y Ware, 1982; Pinheiro y Bates, 2004).

4.6.1. Definición del Modelo Lineal de Efectos Mixtos

En los modelos lineales de efectos mixtos, los efectos fijos y aleatorios ocurren linealmente. Estos modelos extienden los modelos lineales incorporando efectos aleatorios, los cuales pueden ser considerados como términos adicionales del error, para inducir la correlación entre observaciones del mismo grupo. Para un sólo nivel de agrupamiento el modelo lineal de efectos mixtos, descrito por Laird y Ware (1982), expresa el vector y_i para el i -ésimo grupo como

$$\begin{aligned} y_i &= X_i\beta + Z_ib_i + \varepsilon_i, \\ b_i &\sim N(0, \sigma_b^2\Psi), \quad \varepsilon_i \sim N(0, \sigma_\varepsilon^2 I), \end{aligned} \tag{4.21}$$

donde β es un vector de *efectos fijos*, b_i es el vector de *efectos aleatorios*; X_i, Z_i son conocidas como matrices de diseño de efectos fijos y aleatorios, y ε_i es el vector de errores *intra-grupo* con una distribución esférica Gaussiana. El vector de efectos aleatorios b_i y el error intra-grupo ε_i se asumen independientes entre los diferentes grupos e independientes de las observación del mismo grupo.

Una de las razones por las cuales los modelos de efectos mixtos son aplicables es que, natural y elegantemente, inducen las correlaciones entre medidas repetidas hacia un

mismo sujeto. Para el modelo (4.21), puede demostrarse que su correlación está dada por $\sigma_b^2 / (\sigma_b^2 + \sigma_\varepsilon^2)$. Ésto es, sean y_{ij} y $y_{ij'}$ dos medidas repetidas, entonces

$$\begin{aligned}
 \text{Corr}(y_{ij}, y_{ij'}) &= \frac{\text{Cov}(y_{ij}, y_{ij'})}{\sqrt{\text{Var}(y_{ij})\text{Var}(y_{ij'})}} = \frac{E[y_{ij}y_{ij'}] - E[y_{ij}]E[y_{ij'}]}{\sqrt{[\text{Var}(b_i) + \text{Var}(\varepsilon_{ij})][\text{Var}(b_i) + \text{Var}(\varepsilon_{ij'})]}} \\
 &= \frac{E[(\beta_j + b_i + \varepsilon_{ij})(\beta_{j'} + b_i + \varepsilon_{ij'})] - \beta_j\beta_{j'}}{(\sigma_b^2 + \sigma_\varepsilon^2)} = \frac{\beta_j\beta_{j'} + E[b_i^2] - \beta_j\beta_{j'}}{\sigma_b^2 + \sigma_\varepsilon^2} \quad (4.22) \\
 &= \frac{\sigma_b^2}{\sigma_b^2 + \sigma_\varepsilon^2}.
 \end{aligned}$$

Esta correlación, dependiendo de los valores de σ_b^2 y σ_ε^2 , pueden estar entre 0 y 1. De hecho, cuando la correlación entre mediciones de diferentes especimenes es cero, se asume que éstas son independientes.

Para el ajuste del modelo es más conveniente expresar la matriz de covarianzas en forma de un *factor de precisión relativa* Δ , el cual es una matriz que satisface

$$\frac{(\sigma_b^2\Psi)^{-1}}{1/\sigma_\varepsilon^2} = \Delta^T \Delta. \quad (4.23)$$

4.6.2. Estimación por Máxima Verosimilitud para Modelos Lineales de Efectos Mixtos

Considere el modelo (4.21). Los parámetros son β , σ^2 y los que determina Δ . El vector θ es usado para representar un conjunto restringido de parámetros que determinan Δ . Entonces, la verosimilitud para el modelo (4.21) es la densidad de probabilidad para los datos dados los parámetros, considerada como una función paramétrica con datos como valores fijos, no como una función de datos con parámetros fijos. Esto es,

$$L(\beta, \theta, \sigma^2 | y) = p(y | \beta, \theta, \sigma^2), \quad (4.24)$$

donde L es la verosimilitud, p es una densidad de probabilidad y y es vector de respuesta N -dimensional completo, con $N = \sum_{i=1}^M n_i$. Debido a que los efectos aleatorios no observables b_i son parte del modelo, se debe integrar la densidad condicional de los datos dado los efectos aleatorios con respecto a la densidad marginal de los efectos aleatorios para obtener la densidad marginal para los datos. Puede usarse la independencia de b_i y ε_i para expresarla como

$$L(\beta, \theta, \sigma^2 | y) = \prod_{i=1}^M \int p(y_i | b_i, \beta, \sigma^2) p(b_i | \theta, \sigma^2) db_i, \quad (4.25)$$

donde las densidades condicionales de y_i son normales multivariadas.

$$p(y_i | b_i, \beta, \sigma^2) = \frac{\exp(-\|y_i - X_i \beta - Z_i b_i\|^2 / 2\sigma^2)}{(2\pi\sigma^2)^{n_i/2}} \quad (4.26)$$

Y la densidad marginal de b_i es también normal multivariada

$$p(b_i | \theta, \sigma^2) = \frac{\exp(-\|\Delta b_i\|^2 / 2\sigma^2)}{(2\pi\sigma^2)^{q/2} \text{abs}|\Delta|^{-1}}, \quad (4.27)$$

donde $|\Delta|$ es el determinante de la matriz Δ .

5. ESTUDIO DE SIMULACIÓN Y FUNCIONES DE POTENCIA

Se presenta un estudio de simulación para comparar las funciones de potencia de la prueba paramétrica propuesta con la prueba no paramétrica *logrank* muy utilizada en análisis de supervivencia para comparar poblaciones con muestras censuradas. La prueba *logrank* se describe con detalles en la Sección 2.10 y para este estudio es implementada en *R*, bajo la función `survdiff`.

Se observó únicamente el caso de σ común en la comparación de medias, siguiendo la estructura de regresión lognormal con variables indicadoras para los log-cuantiles poblacionales, de donde

$$\mu_1 = \mu(0) = \beta_0, \quad \mu_2 = \mu(1) = \beta_0 + \beta_1, \quad \sigma_1 = \sigma_2 = \sigma.$$

Por lo que las diferencias de cuantiles (y medias) poblacionales, recaerá únicamente en $\log y_p(1) - \log y_p(0) = \mu(1) - \mu(0) = \beta_1$, que no depende del cuantil comparado. Entonces, para la observación y comparación de las funciones de potencias en ambas pruebas, paramétricas y no paramétricas, se propone el siguiente juego de hipótesis:

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0.$$

En este estudio se observaron y compararon simultáneamente las potencias de las pruebas, propuesta y *logrank*, a través de simulación computacional con 1,000 muestras de tamaños 15, 30 y 100; y distintos porcentajes de información censurada (25, 50 y 70%); esto para conocer la sensibilidad de la prueba al tamaño de muestra y de información censurada. También se observó si la prueba muestra algún grado de robusticidad al relajar el supuesto de distribución de origen. En el Anexo B.1, se presentan los programas en *R* que fueron desarrollados y utilizados para la obtención y discusión de resultados.

Se muestran tres gráficos, Figuras 6, 7 y 8, que ilustran las ventajas de la prueba propuesta, sobre la prueba logrank, en la comparación de medias poblacionales lognormal. Las Figuras 10 y 12 muestran la potencia de la prueba propuesta, relajando los supuestos de distribución; comparando poblaciones exponenciales y Gumbel. Para ambas distribuciones no lognormales, la prueba funciona aceptablemente.

5.1 Pruebas Propuesta y Logrank Comparando Muestras Lognormales

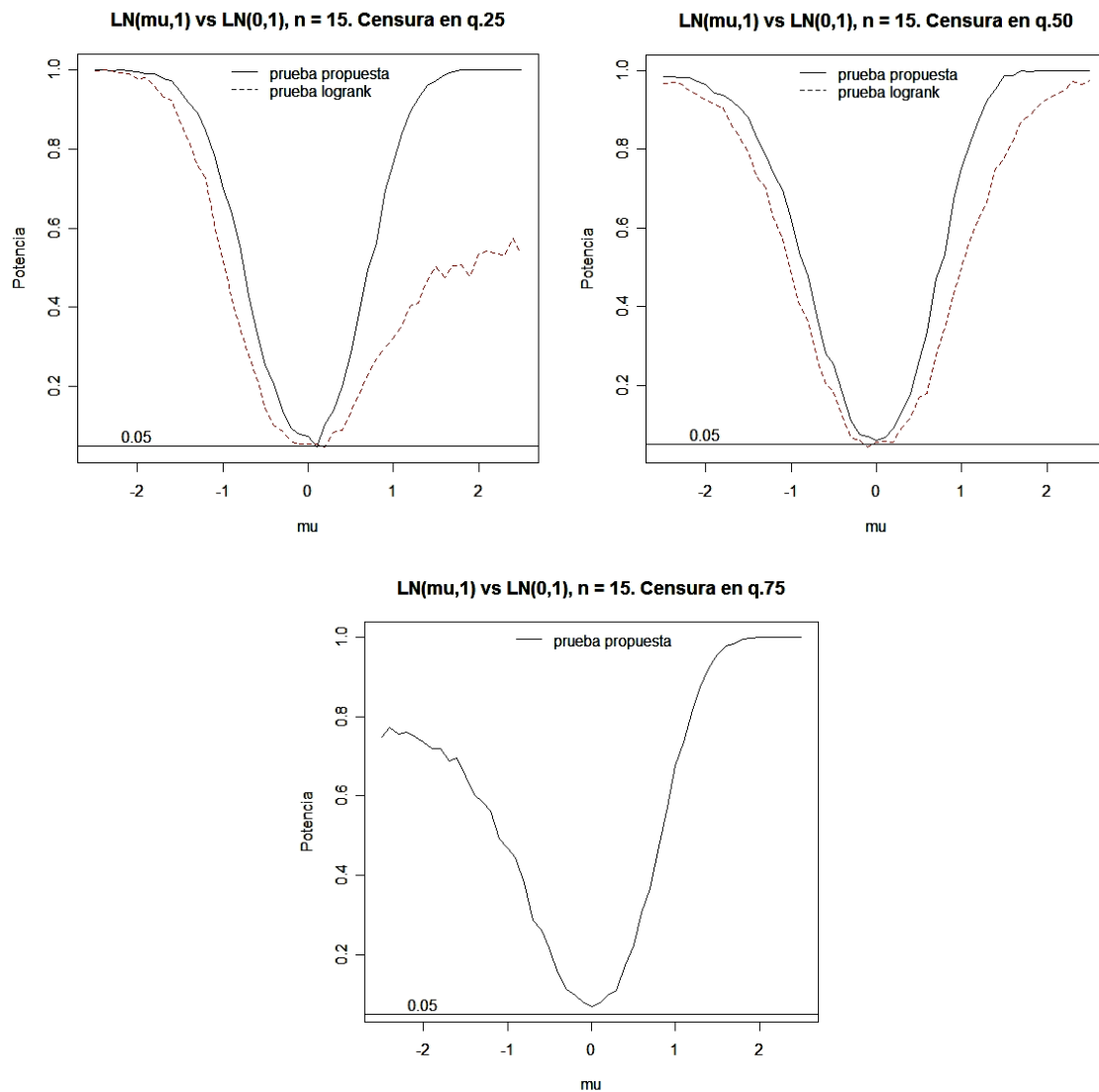


Figura 6. Pruebas propuesta y logrank comparando muestras lognormales de tamaño 15.

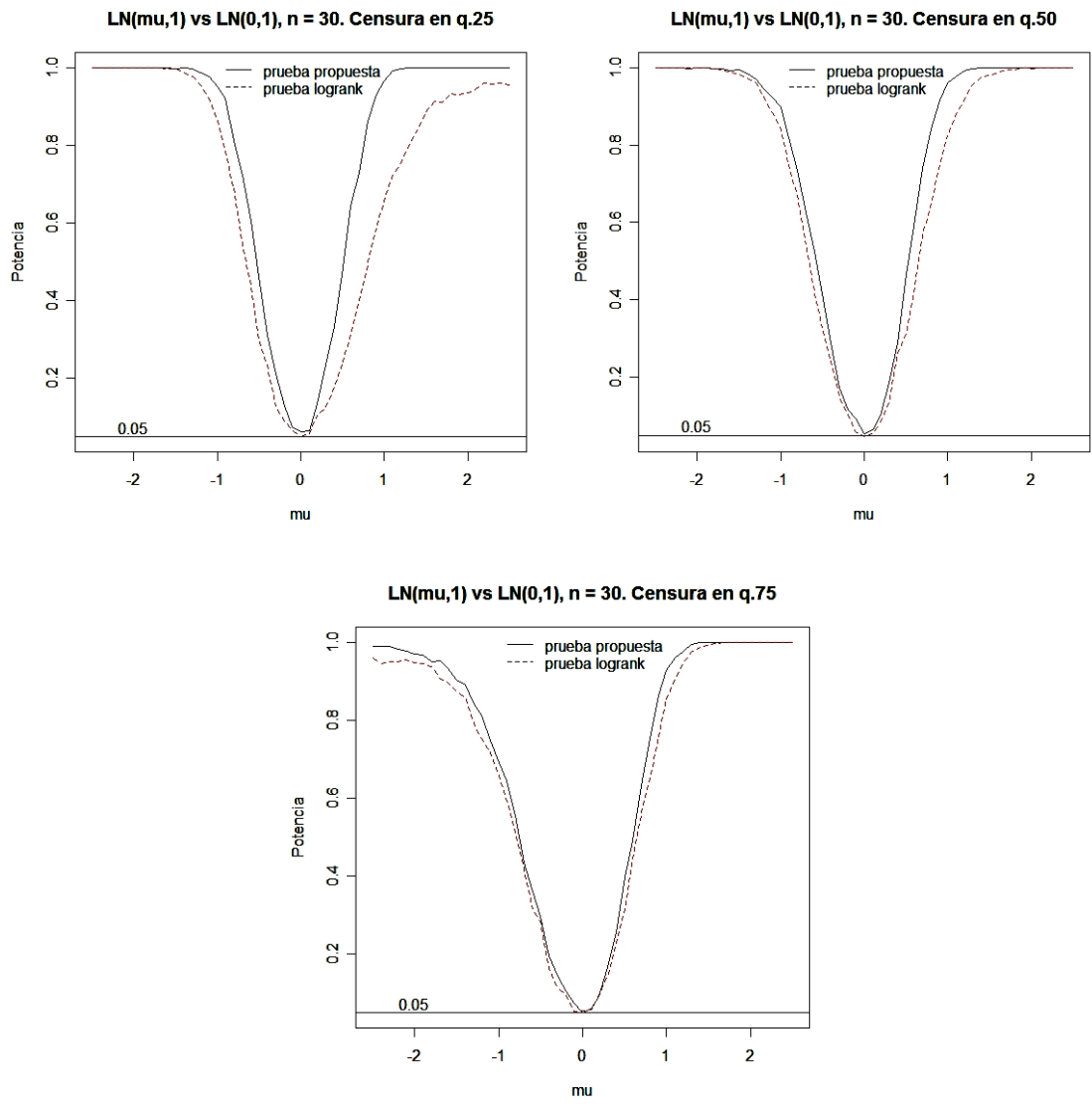


Figura 7. Pruebas propuesta y logrank comparando muestras lognormales de tamaño 30.

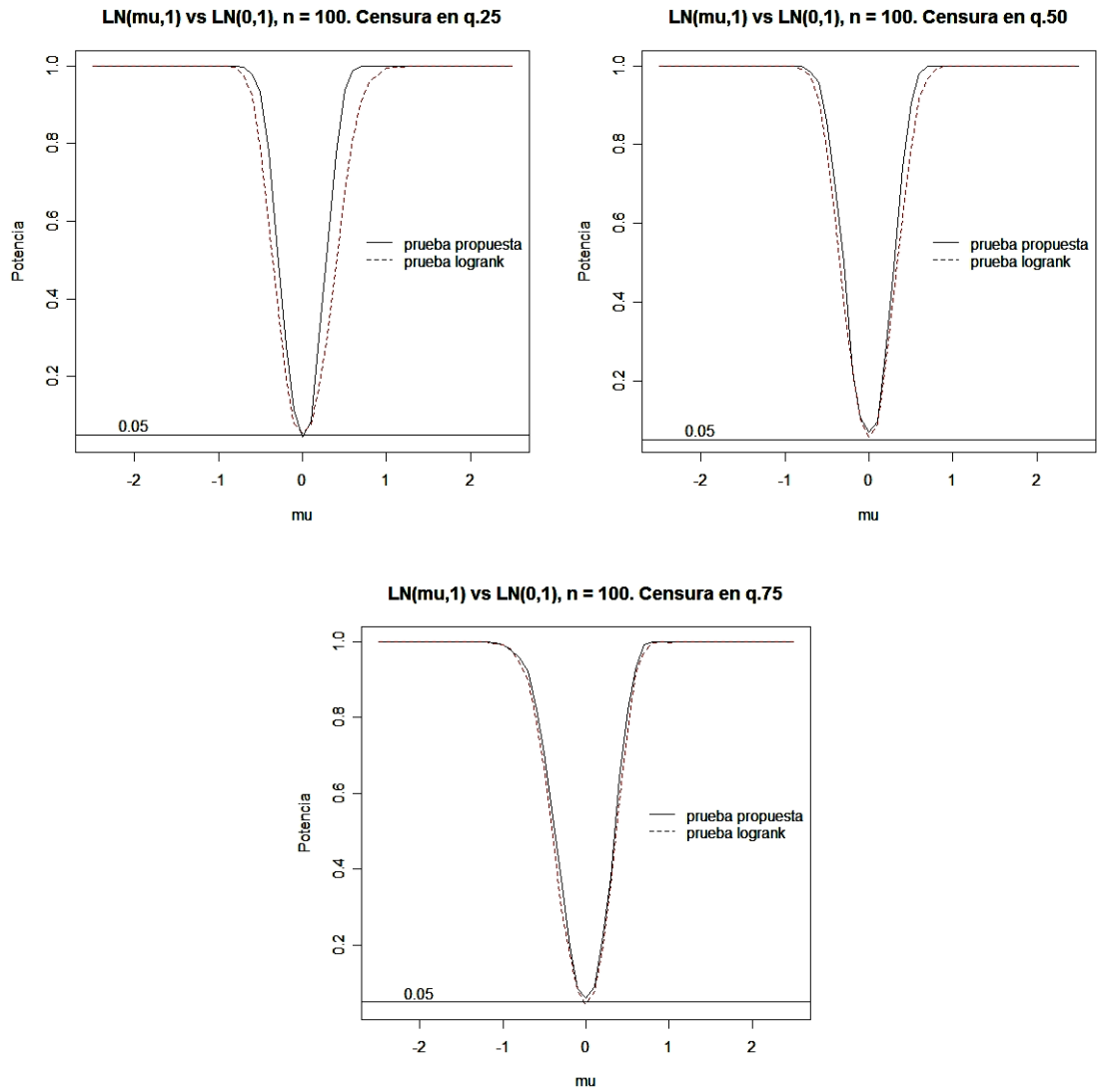


Figura 8. Pruebas propuesta y logrank comparando muestras lognormales de tamaño 100

5.2 Prueba Propuesta Comparando Muestras Exponenciales

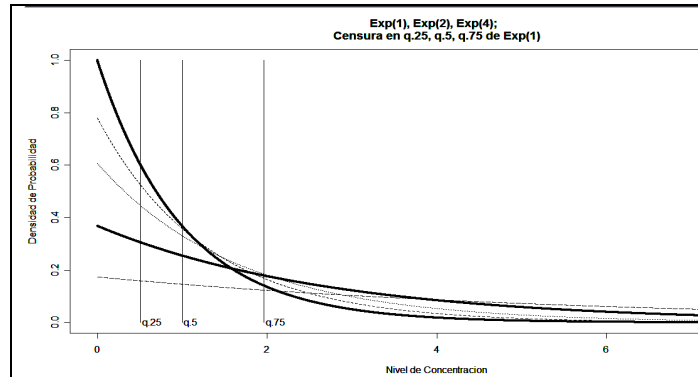


Figura 9. Gráficos de funciones de densidad para variables exponenciales

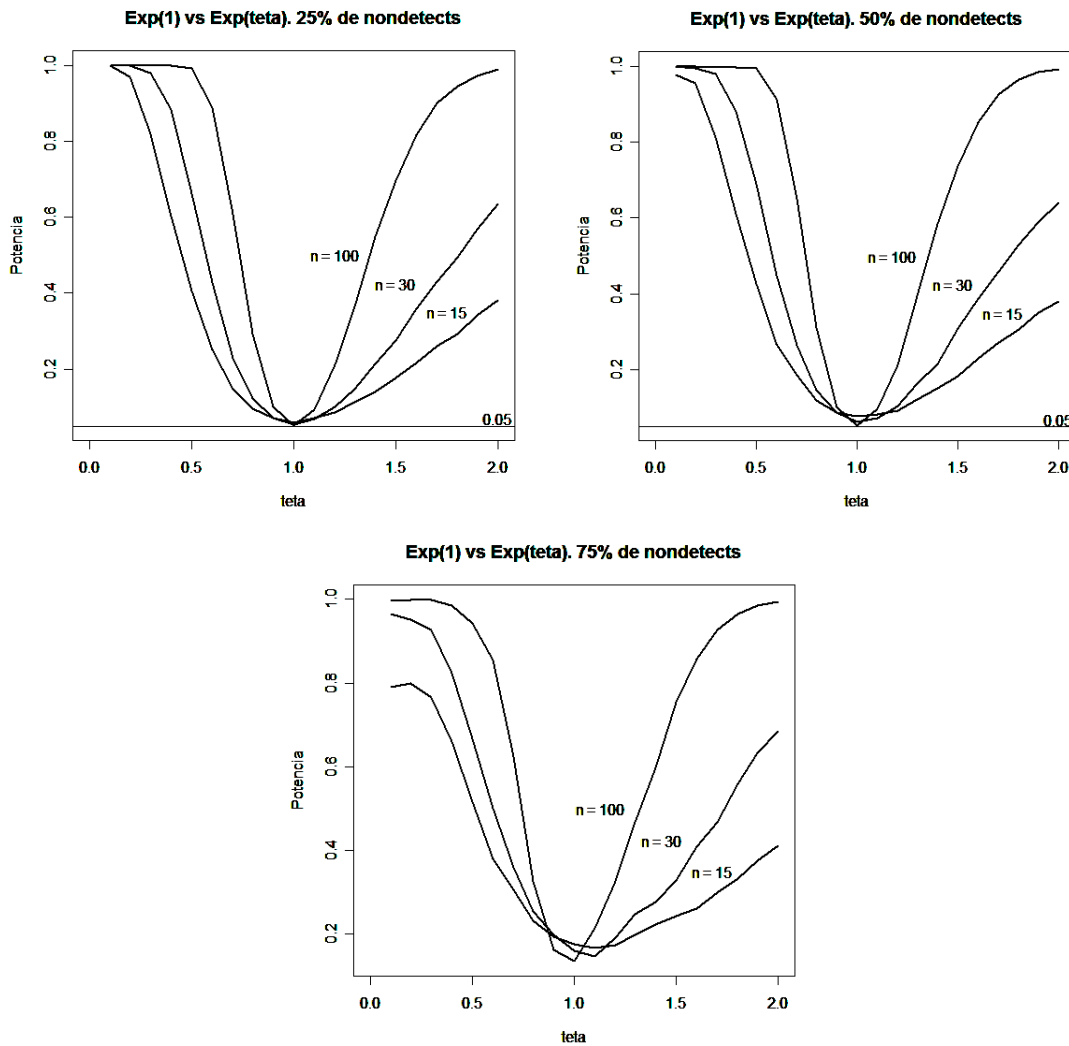


Figura 10. Potencia de la prueba propuesta comparando poblaciones exponenciales.

5.3 Prueba Propuesta Comparando Muestras Gumbel

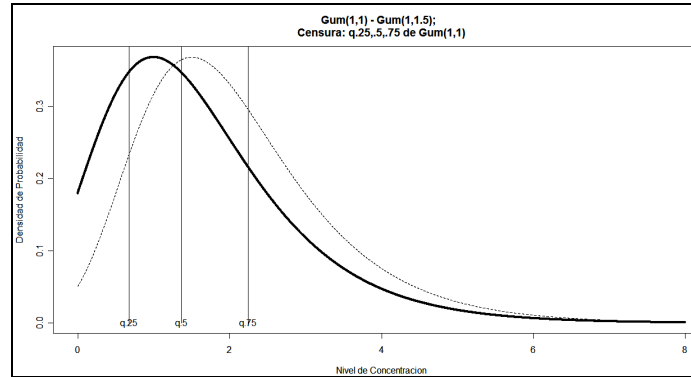


Figura 11. Gráficos de funciones de densidad para variables Gumbel

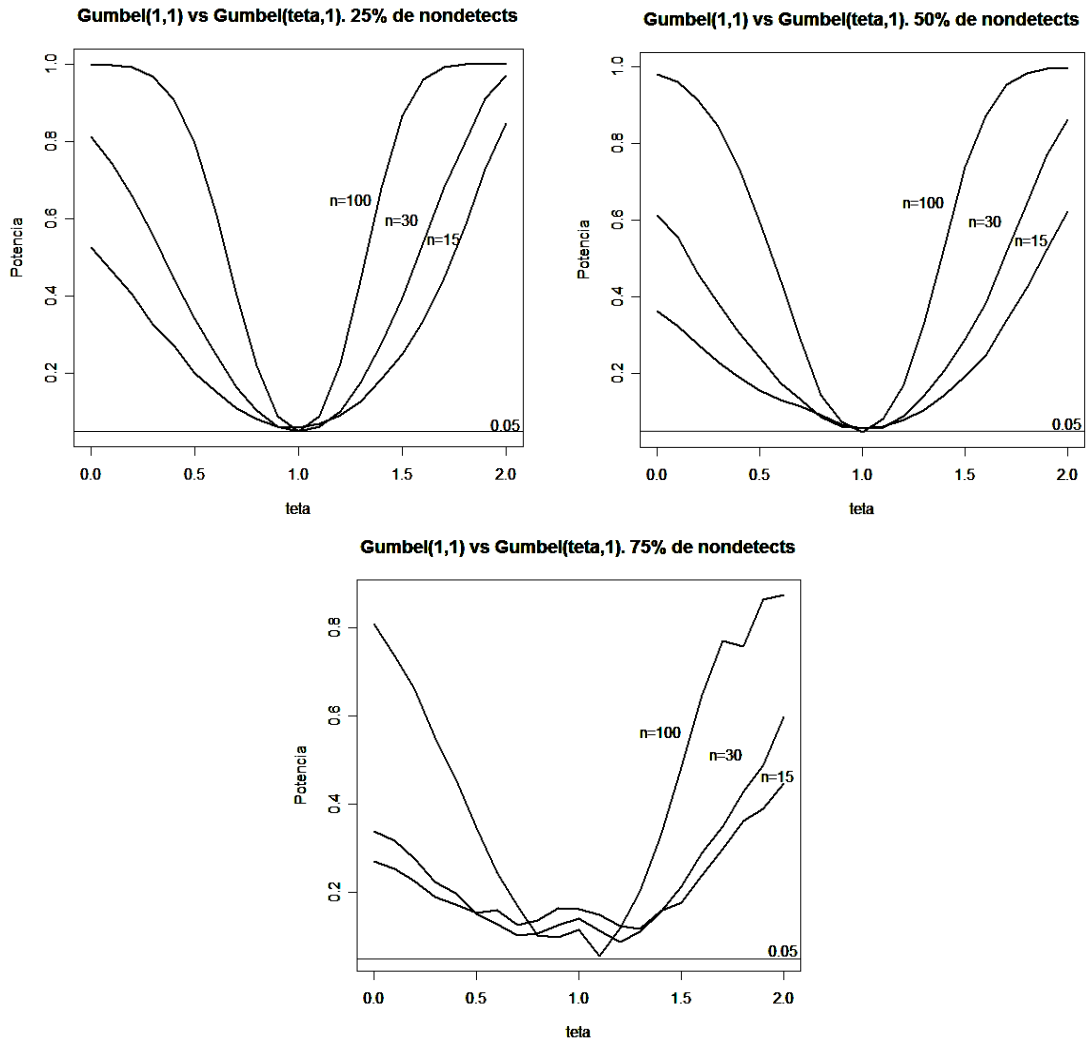


Figura 12. Potencia de la prueba propuesta comparando poblaciones Gumbel.

5.4 Resultados y Discusión del Estudio de Potencia

De acuerdo con la Figura 6, se observó que la función de potencia para la prueba paramétrica que se propone en este trabajo, tuvo mejor desempeño que las potencias para la prueba no paramétrica logrank en tamaños de muestra de 15 observaciones. En particular, censurando por la izquierda el 75% de los datos, la prueba logrank no pudo detectar diferencias, por el contrario la prueba propuesta mantiene un comportamiento asintótico aceptable. Para muestras de tamaño 30, se observó que la potencia de la prueba propuesta mantiene un mejor comportamiento que la prueba logrank, Figura 7; sin embargo a medida que los tamaños de muestra aumentan y la proporción de censura es menor, ambas pruebas tienden a coincidir, concordando con los resultados obtenidos para tamaños de muestras de 100 observaciones (Figura 8), donde ambas potencias son casi idénticas. Esto último debido a que se analizan muestras grandes, donde la asintoticidad es esperada en ambas pruebas.

Por otro lado, relajando los supuestos para la prueba paramétrica propuesta, se observó que la potencia de la prueba mantiene un desempeño aceptable para detectar diferencias en muestras exponenciales; la Figura 10 muestra los resultados obtenidos para esta aseveración, donde se aprecia que la prueba propuesta funciona mejor para proporciones de censura y tamaños de muestras moderados. Finalmente, para conocer el comportamiento de la función de potencia de esta prueba cuando se comparan poblaciones de distribuciones con cola pesada como las de Gumbel, se observó que la prueba propuesta mantiene una aceptable asintoticidad (Figura 12), sin embargo, el comportamiento no es mejor que la observada para las muestras aleatorias de distribuciones de cola liviana como las exponenciales (Figura 10).

En general, la función potencia de la prueba paramétrica propuesta tuvo mejor desempeño que la prueba no paramétrica logrank. Así también, la prueba propuesta tuvo un desempeño aceptable en la comparación de muestras no lognormales, exponenciales y de Gumbel; sin embargo, es importante señalar una notable asimetría en todas las funciones de potencia estimadas, lo cual será estudiado posteriormente.

6. APLICACIONES DE LA METODOLOGÍA EN DATOS AMBIENTALES

Las aplicaciones que se presentan a continuación, fueron orientadas a la solución de problemas reportados en estudios de ciencias ambientales y por organismos reguladores como la Environmental Protection Agency (EPA). La metodología se desarrolló bajo la función de verosimilitud, para extraer la información contenida en los datos exactos y no detectados. El análisis se centró en la obtención de estimaciones de MV para los parámetros del modelo lognormal, los coeficientes de los modelos de log-regresión con variables indicadoras y su matriz de covarianzas, obteniéndose intervalos de confianza aproximados para establecer un criterio de comparación de medias. En cada ejemplo se enfatizó las ventajas de implementar el algoritmo EM.

6.1. Inferencia Sobre Una Muestra Aleatoria Lognormal

Uno de los principales objetivos del análisis estadístico de una sola muestra aleatoria con datos no detectados, es caracterizar e identificar el modelo que describe la población de origen. Con la finalidad de tener un registro histórico, una referencia comparativa con niveles permisibles o para cuantificar un posible riesgo en la zona.

Ejemplo 1. Helsel (2005) presenta las concentraciones de arsénico reportadas por Tomlinson (2003), en un arroyo urbano Manoa Stream en Kanawai Field, Hawai, observándose tres límites de detección 0.9, 1 y 2 $\mu\text{g/L}$, con alrededor del 50% de datos no detectados denotados por el signo <, Cuadro 1. Suponiendo una distribución lognormal, se obtuvieron las estimaciones de los parámetros y la concentración media de arsénico, para caracterizar la población de origen y hacer inferencia.

Cuadro 2. Concentraciones de arsénico en Manoa Stream, Kanawai Field.

Arsénico ($\mu\text{g/L}$)	<1, <1, 1.7, <1, <1, <2, 3.2, <2, <2, 2.8, <2, <2, <2, <2, <2, 0.7, 0.9, 0.5, 0.5, <0.9, 0.5, 0.7, 0.6, 1.5
---------------------------------	--

6.1.1 Gráfico de Probabilidad Lognormal

La Figura 13 muestra un gráfico de probabilidad lognormal, obtenido utilizando SPLIDA (2004) a través de la relación $\{\log(y_p) = \mu + \Phi^{-1}(p)\sigma \text{ vs } \Phi^{-1}(p)\}$. El gráfico indica que es razonable emplear la distribución lognormal y se obtienen estimaciones gráficas.

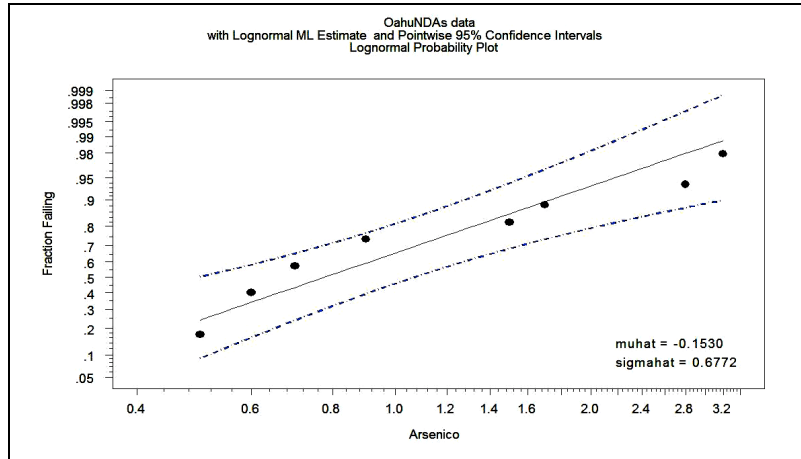


Figura 13. Gráfico de probabilidad lognormal para los datos del Cuadro 2.

Ahora, el análisis de la información estriba en la inferencia obtenida a través de la verosimilitud para una muestra aleatoria lognormal conteniendo datos no detectados,

$$L(\mu, \sigma) = \prod_{i=1}^n \left[\frac{1}{\sigma x_i} \phi\left(\frac{\log x_i - \mu}{\sigma}\right) \right]^{\delta_{ij}} \left[\Phi\left(\frac{\log x_i - \mu}{\sigma}\right) \right]^{1-\delta_{ij}} \quad (6.1)$$

6.1.2 Optimización Directa de la Función de Verosimilitud

El primer enfoque, para el análisis de datos, se realizó optimizando directamente la función (6.1), a través de programa en R descrito en B.2.1. De donde se obtuvieron las siguientes estimaciones puntuales para los parámetros del modelo lognormal y de la concentración media E (valor esperado),

$$\hat{\mu} = -0.153, \hat{\sigma} = 0.677, \hat{E} = 1.079.$$

De igual modo, a través de la matriz de información observada de Fisher, se obtuvo una estimación local para la matriz de covarianzas,

$$\hat{\Sigma}_{\hat{\mu}, \hat{\sigma}} = \begin{bmatrix} \hat{Var}(\hat{\mu}) & \hat{Cov}(\hat{\mu}, \hat{\sigma}) \\ \hat{Cov}(\hat{\mu}, \hat{\sigma}) & \hat{Var}(\hat{\sigma}) \end{bmatrix} = \begin{bmatrix} 0.029 & -0.008 \\ -0.008 & 0.016 \end{bmatrix}. \quad (6.2)$$

Con lo anterior y el método de Wald, se construyeron intervalos aproximados del 95% de confianza para los parámetros y la concertación media (función paramétrica).

Primero para μ , de (6.2) se tiene que $es_{\hat{\mu}} = \sqrt{\hat{Var}(\hat{\mu})} = \sqrt{0.029 \times 10^{-2}} = 0.169$, con lo que

$$\begin{aligned} [\underline{\mu}, \tilde{\mu}] &= \hat{\mu} \pm z_{(1-\alpha/2)} es_{\hat{\mu}} = -0.153 \pm 1.96 \times 0.169 \\ &= [-0.484, 0.127] \end{aligned} \quad (6.3)$$

Para σ , $es_{\hat{\sigma}} = 0.127$, de donde $w = \exp\left(z_{(1-\alpha/2)} es_{\hat{\sigma}} / \hat{\sigma}\right) = \exp(1.96 \times 0.127 / 0.677) = 1.283$, así

$$\begin{aligned} [\underline{\sigma}, \tilde{\sigma}] &= [\hat{\sigma}/w, \hat{\sigma} \times w] = [0.677/1.283, 0.677 \times 1.283] \\ &= [0.469, 0.978] \end{aligned} \quad (6.4)$$

Ahora, para construir un intervalo de confianza aproximado para la función paramétrica

$E = \exp\left(\mu + \frac{1}{2}\sigma^2\right)$, primero se calculó su error estándar, es decir, $es_{\hat{E}}$,

$$\begin{aligned} es_{\hat{E}} &= \left[\hat{Var}(\hat{E}) \right]^{1/2} = \left[\left(\frac{\partial E}{\partial \mu} \right)^2 \hat{Var}(\mu) + 2 \left(\frac{\partial E}{\partial \mu} \right) \left(\frac{\partial E}{\partial \sigma} \right) \hat{Cov}(\mu, \sigma) + \left(\frac{\partial E}{\partial \sigma} \right)^2 \hat{Var}(\sigma) \right]_{(\mu, \sigma) = (\hat{\mu}, \hat{\sigma})}^{1/2} \\ &= \left[\exp(2\hat{\mu} + \hat{\sigma}^2) \hat{Var}(\hat{\mu}) + 2\hat{\sigma} \exp(2\hat{\mu} + \hat{\sigma}^2) \hat{Cov}(\hat{\mu}, \hat{\sigma}) + 4\hat{\sigma}^2 \exp(2\hat{\mu} + \hat{\sigma}^2) \hat{Var}(\hat{\sigma}) \right]^{1/2} \\ &= 0.236, \end{aligned}$$

de donde,

$$[\underline{E}, \tilde{E}] = \hat{E} \pm z_{(1-\alpha/2)} \hat{es}_{\hat{E}} = 1.079 \pm 1.96 \times 0.236 = [0.617, 1.542].$$

6.1.3 Optimización de la Función de Verosimilitud Vía Algoritmo EM

El empleo del algoritmo EM reduce la complejidad del proceso de optimización de la función de verosimilitud. En B.2.1 se muestra la implementación de este método en R para la solución del problema en el Ejemplo 1.

Siguiendo un procedimiento análogo al implementado en 6.1.2, se obtuvieron las estimaciones puntuales y por intervalo de confianza para los parámetros de modelo, la concentración media y la matriz de varianza-covarianza, es decir,

$$\hat{\mu} = -0.253, \quad \hat{\sigma} = 0.515, \quad \hat{E} = 0.887,$$

$$\hat{\Sigma}_{\hat{\mu}, \hat{\sigma}} = \begin{bmatrix} 1.060 \times 10^{-2} & 1.132 \times 10^{-6} \\ 1.132 \times 10^{-6} & 5.300 \times 10^{-3} \end{bmatrix}$$

Ahora la inferencia por intervalos de confianza se muestra a continuación,

$$[\underline{\mu}, \tilde{\mu}] = [-0.459, -0.046],$$

$$[\underline{\sigma}, \tilde{\sigma}] = [0.391, 0.680].$$

Por último, para la concentración media de arsénico E , $\hat{es}_{\hat{E}} = \sqrt{\hat{Var}(\hat{E})} = 0.113$, de donde

$$[\underline{E}, \tilde{E}] = [0.665, 1.108].$$

Resultados y Discusión. El Cuadro 3 muestra los resultados del análisis de datos del Cuadro 2, optimizando la función de verosimilitud con el enfoque de análisis directo y vía el algoritmo EM.

Cuadro 3. Estimaciones por MV para los parámetros de los datos de arsénico

Parámetro	EMV		Error Estándar		I.C. Aprox. del 95% de confianza	
	Directo	Vía EM	Directo	Vía EM	Directo	Vía EM
μ	-0.153	-0.253	0.169	0.105	-0.484, 0.178	-0.459, -0.046
σ	0.677	0.515	0.127	0.073	0.469, 0.978	0.391, 0.680
Media	1.079	0.887	0.236	0.113	0.617, 1.542	0.665, 1.108
$\log L(\beta_0, \alpha, \sigma)$	-18.374	-17.632				

Del Cuadro 3 y comparando ambas formas de optimización, se observó que el algoritmo EM produjo una reducción de los errores estándar y los intervalos, obteniéndose a si un mejor ajuste para modelo ($\log L = -17.632$). Entonces, para inferencias o decisiones respecto al problema, se toman las estimaciones por MV obtenidas vía EM, es decir, $\hat{\mu} = -0.253$, $\hat{\sigma} = 0.515$, $E = 0.887$. Con lo anterior una estimación para la concentración media de arsénico en el Manoa Stream es de 0.887 $\mu\text{g/L}$, con un intervalo aproximado del 95% de confianza de [0.665, 1.108].

6.2 Inferencia Sobre Una Muestra Lognormal Con Una Covariable Continua

En monitoreo de calidad del agua, aire y exposición ambiental en centros de trabajo, las mediciones de concentraciones de contaminantes se realizan a través de periodos de tiempo. Las predicciones y estimaciones sobre valores medios en tiempos o periodos futuros son problemas a resolver en el análisis de este tipo de información.

Ejemplo 2. Los datos en la Cuadro 4 son dosis de radiación gamma trimestrales de un registro de 1956 a 1965 para un trabajador de la planta Oak Ridge Y-12, Frome and Watkins (2004). Se asume que las dosis siguen una distribución lognormal. Los valores denotados por < 30 indican dosis no detectadas a un límite de detección de 30. Se desea estimar la dosis durante un trimestre para un trabajador que no se observado, tomando como referencia el primer trimestre de 1961.

Cuadro 4. Dosis de radiación gamma $\mu\text{Sv}\times 100$

Año	1956	1957	1958	1959	1960	1961	1962	1963	1964	1965
Trimestre 1	< 30	110	16	103	15	2	15	< 30	< 30	3
Trimestre 2	< 30	16	46	64	60	53	56	< 30	< 30	4
Trimestre 3	< 30	< 30	99	36	29	53	44	4	< 30	5
Trimestre 4	52	< 30	93	35	75	89	23	4	< 30	23

Fuente: Frome and Watkins (2004). Statistical Analysis of Data with Non-Detectable Values

En la Figura 14 se muestran los gráficos de dispersión de la dosis y de la log-dosis de los datos del Cuadro 4. Se observó una tendencia negativa a través del tiempo; sin embargo, es una apreciación visual que complementara el análisis estadístico. El símbolo ∇ denota un dato no detectado.

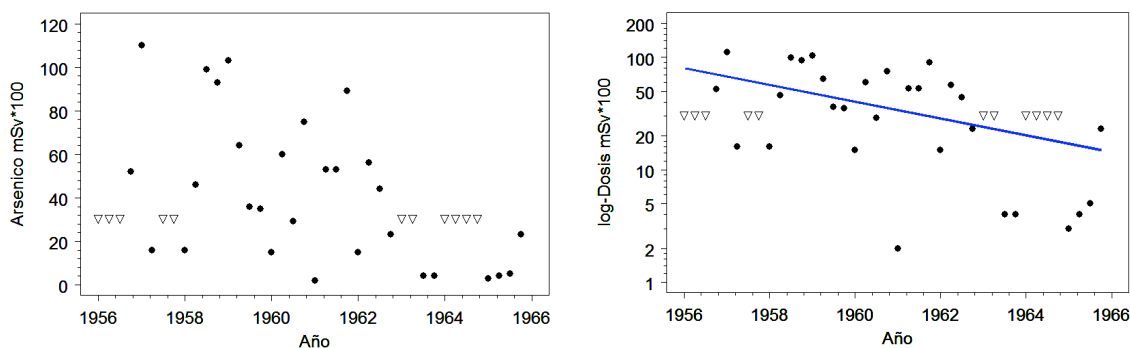


Figura 14. Gráficos de dispersión para las dosis y log-dosis de los datos del Cuadro 4.

El modelo de regresión lognormal para la inferencia es,

$$E[\log x_i] = \mu_i = \beta_0 + \alpha (\text{año} - 1961).$$

Donde β_0 es el intercepto de la log-dosis al primer trimestre de 1961 y α el cambio por año en x . Entonces, asumiendo lognormalidad, la inferencia sobre la muestra con datos no detectados se fundamenta en la función de verosimilitud siguiente,

$$L(\beta_0, \alpha, \sigma) = \prod_{i=1}^{40} \left[\frac{1}{\sigma x_i} \phi\left(\frac{\log x_i - \mu_i}{\sigma}\right) \right]^{\delta_i} \left[\Phi\left(\frac{\log x_i - \mu_i}{\sigma}\right) \right]^{1-\delta_i}, \quad (6.5)$$

donde y_i denota las dosis, $\delta_i = 0$ para datos no detectados y $\delta_i = 1$ para dosis exactas.

6.2.1. Optimización Directa de la Función de Verosimilitud

El primer enfoque de análisis de datos se realizó mediante la optimización directa de la función de verosimilitud (6.5), utilizando el programa *R* en B.2.1. De donde se obtuvieron las estimaciones para los parámetros del modelo de regresión lognormal,

$$\hat{\beta}_0 = 2.681, \quad \hat{\sigma} = 1.046, \quad \hat{\alpha} = -0.115.$$

La estimación obtenida para la matriz de varianzas-covarianzas, es la siguiente

$$\hat{\Sigma}_{\hat{\beta}_0, \hat{\sigma}, \hat{\alpha}} = \begin{bmatrix} \hat{Var}(\hat{\beta}_0) & \hat{Cov}(\hat{\beta}_0, \hat{\sigma}) & \hat{Cov}(\hat{\beta}_0, \hat{\alpha}) \\ \hat{Cov}(\hat{\sigma}, \hat{\beta}_0) & \hat{Var}(\hat{\sigma}) & \hat{Cov}(\hat{\sigma}, \hat{\alpha}) \\ \hat{Cov}(\hat{\alpha}, \hat{\beta}_0) & \hat{Cov}(\hat{\alpha}, \hat{\sigma}) & \hat{Var}(\hat{\alpha}) \end{bmatrix} = \begin{bmatrix} 0.032 & -0.009 & 0.002 \\ -0.009 & 0.019 & 0.001 \\ 0.002 & 0.001 & 0.003 \end{bmatrix} \quad (6.6)$$

Con lo anterior y el método de Wald, se construyeron los intervalos aproximados del 95% de confianza para los parámetros.

$$\begin{aligned} [\underline{\beta}_0, \tilde{\beta}_0] &= \hat{\beta}_0 \pm z_{(1-\alpha/2)} \hat{es}_{\hat{\beta}_0} = [2.331, 3.031], \\ [\underline{\sigma}, \tilde{\sigma}] &= \left[\hat{\sigma} / \exp\left(z_{(1-\alpha/2)} \hat{es}_{\hat{\sigma}} / \hat{\sigma}\right), \hat{\sigma} \times \exp\left(z_{(1-\alpha/2)} \hat{es}_{\hat{\sigma}} / \hat{\sigma}\right) \right] = [0.809, 1.351], \\ [\underline{\alpha}, \tilde{\alpha}] &= \hat{\alpha} \pm z_{(1-\alpha/2)} \hat{es}_{\hat{\alpha}} = [-0.220, -0.009]. \end{aligned} \quad (6.7)$$

6.2.2 Optimización de la Función de Verosimilitud Vía EM

Usando EM se obtuvieron $\hat{\beta}_0 = 2.922$, $\hat{\sigma} = 0.952$, $\hat{\alpha} = -0.194$ y la diagonal principal de la matriz de covarianzas, necesaria para obtener los intervalos de confianza aproximados,

$$Diag\left(\hat{\Sigma}_{\hat{\beta}_0, \hat{\sigma}, \hat{\alpha}}\right) = \left[\hat{Var}(\hat{\beta}_0) \quad \hat{Var}(\hat{\sigma}) \quad \hat{Var}(\hat{\alpha}) \right] = [0.153 \quad 0.110 \quad 0.052].$$

Ahora la inferencia por intervalos de confianza se muestra a continuación,

$$[\hat{\beta}_0, \tilde{\beta}_0] = [2.622, 3.221], \quad [\hat{\sigma}, \tilde{\sigma}] = [0.759, 1.194], \quad [\hat{\alpha}, \tilde{\alpha}] = [-0.297, -0.091].$$

Resultados y Discusión. Los resultados del análisis de los datos se muestran en el Cuadro 5, mediante la optimización directa de la verosimilitud (6.5) y vía EM.

Cuadro 5. EMV para los parámetros del modelo lognormal para las dosis

Parámetro	EMV		Error Estándar		I.C. Aprox. del 95% de confianza	
	Directo	Vía EM	Directo	Vía EM	Directo	Vía EM
β_0	2.681	2.922	0.179	0.153	2.331, 3.031	2.622, 3.221
α	-0.115	-0.194	0.054	0.052	-0.220, -0.009	-0.297, -0.091
σ	1.046	0.952	0.137	0.110	0.809, 1.351	0.759, 1.194
$\log L(\beta_0, \alpha, \sigma)$	-56.314	-53.775				

De los resultados anteriores y comparando los enfoques de optimización, se observó que el algoritmo EM redujo los errores estándar e intervalos. Esto representa un mejor ajuste para modelo, lo cual puede observarse en el aumento de la $\log L$ vía EM. Entonces para hacer inferencia se emplearan los EMV vía EM, es decir,

$$\hat{\mu}_i = \hat{\beta}_0 + \hat{\alpha} \times (\text{año} - 1961) = 2.922 - 0.194 \times (\text{año} - 1961); \quad \hat{\sigma} = 0.952.$$

De donde, para este contexto, la estimación de la dosis media para un trabajador en un trimestre de un año cualquiera, se obtendría de

$$E = \exp(383.809 - 0.194 \times \text{año}).$$

6.3 Comparación de Dos Poblaciones Lognormales Independientes

La comparación de dos grupos es diseño básico en muchos estudios ambientales. En algunos casos, un grupo tratamiento es comparado con uno control. El grupo control representa el conocimiento previo y el grupo tratamiento representa condiciones donde, por ejemplo, se sospecha que las concentraciones de contaminantes son más altas. En otros casos, los dos grupos son inspeccionados para saber si uno es mejor o peor que el otro. El interés es propiamente saber si los niveles de contaminantes en los grupos son iguales o diferentes (Helsel, 2005).

Ejemplo 3. Millard y Deverel (1988) reportan niveles de cobre en mantos freáticos, muestreados en la Zona Alluvial Fan y la Zona Basin-Trough, dos áreas del valle San Joaquín en California. Estos datos son presentados en la Cuadro 6. Aproximadamente 20% de los datos son reportados como no detectados, denotados por el signo <. Se desea comparar las concentraciones medias de cobre, a fin de conocer si son iguales o no en ambas poblaciones.

Cuadro 6. Datos de Concentraciones de Cobre para las dos zonas del Valle San Joaquín

Alluvial Fan Zone	<1, <1, <1, <1, <5, <5, <5, <5, <5, <5, <5, <5, <10, <10, <10, <20, <20, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 4, 5, 5, 5, 7, 7, 7, 8, 9, 10, 11, 12, 16, 20
Basin-Trough Zone	<1, <1, <2, <2, <5, <5, <5, <5, <5, <10, <10, <10, <10, <15, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 5, 6, 6, 8, 9, 9, 12, 14, 15, 17, 23

Fuente: Millard and Deverel (1988), Nonparametric statistical methods for comparing two sites based on data with multiple nondetect limits.

La Figura 15 muestra dos gráficos de probabilidad lognormal, verificándose el supuesto de lognormalidad para las concentraciones de cobre en cada muestra, necesario para el análisis de los datos. Esto coincide con la aseveración de Millard y Deverel (1988).

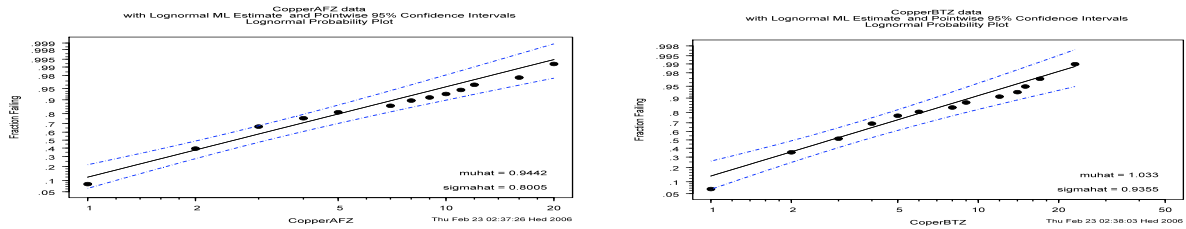


Figura 15. Gráficos de probabilidad lognormal para los datos del Cuadro 6.

6.3.1 Modelo 1. Homogeneidad del Parámetro de Escala σ .

La función log-cuantil del modelo de regresión lognormal, Meeker y Escobar (1998), incluyendo la variable indicadora x , para comparar dos muestras con σ común, es

$$\log [y_p(x)] = \mu(x) + \Phi^{-1}(p)\sigma = \beta_0 + \beta_1 x + \Phi^{-1}(p)\sigma. \quad (6.8)$$

Φ es la función de distribución de $N(0,1)$, $x = 0$ para una muestra y $x = 1$ para la otra.

Ahora, la verosimilitud para las dos muestras lognormales independientes, de tamaños n_1 y n_2 , respectivamente; con observaciones exactas y no detectados tiene la forma

$$L(\beta_0, \beta_1, \sigma) = \prod_{i=1}^2 \left\{ \prod_{j=1}^{n_i} \left[\frac{1}{\sigma y_{ij}} \phi \left(\frac{\log y_{ij} - \mu_i}{\sigma} \right) \right]^{\delta_{ij}} \left[\Phi \left(\frac{\log y_{ij} - \mu_i}{\sigma} \right) \right]^{1-\delta_{ij}} \right\}. \quad (6.9)$$

La comparación se realizó bajo el modelo $\mu(x) = \beta_0 + \beta_1 x$. Sustituyendo la variable indicadora x , $\mu_1 = \mu(0) = \beta_0$ para la primera muestra, y $\mu_2 = \mu(1) = \beta_0 + \beta_1$ para la otra, de donde $D = t_p(1) - t_p(0) = \mu(1) - \mu(0) = \beta_1$, que no depende de ningún cuantil. Así, si un intervalo de confianza aproximado para β_1 contiene al cero, las medias no difieren.

Resultados y Discusión. La verosimilitud (6.9) es optimizada, bajo la optimización directa y vía EM, obteniéndose los EMVs para los parámetros del modelo y sus intervalos de confianza, de donde se observara el respectivo a β_1 para determinar si existe diferencia significativa entre las medias. El Cuadro 7 muestra las estimaciones e intervalos, ilustrando las bondades de emplear el algoritmo EM.

Cuadro 7. EMV para parámetros en los datos de Cobre (σ común)

Parámetro	EMV		Error Estándar		I.C. Aprox. del 95% de confianza	
	Directo	Vía EM	Directo	Vía EM	Directo	Vía EM
β_0	0.606	1.050	0.117	0.114	0.376, 0.835	0.826, 1.274
β_1	0.056	-0.116	0.160	0.151	-0.259, 0.370	-0.413, 0.181
σ	0.929	0.800	0.065	0.065	0.810, 1.067	0.682, 0.939

	μ_1	μ_2	σ	E_1	E_2	$\log L_1(\beta, \sigma)$
Directo	0.661	0.606	0.929	2.984	2.822	-168.515
Vía EM	0.933	1.050	0.800	3.503	3.934	-135.336

Del Cuadro 7, es fácil ver que la implementación del algoritmo EM reduce el Error Estándar para los parámetros, provee intervalos de confianza más estrechos y consecuentemente un mejor ajuste para el Modelo 1, dado que el valor de $\log L$ es mayor al emplear EM. Note que el EMV vía EM para β_1 es un valor cercano a 0, -0.116, y su intervalo aproximado del 95% de confianza contiene el 0, (-0.413, 0.181),

de modo que la diferencia es estadísticamente nula. Así, a la luz de los datos, ambas concentraciones medias de cobre, E_1 y E_2 , son idénticas.

6.3.2 Modelo 2. Heterogeneidad del Parámetro de Escala σ .

La comparación se realizó reparametrizando ambos parámetros del modelo lognormal mediante $\mu_i = \beta_0 + \beta_1 x$ y $\log(\sigma_i) = \log[\sigma(x)] = \gamma_0 + \gamma_1 x$, de donde, $\mu_1 = \beta_0$, $\log(\sigma_1) = \gamma_0$ y $\mu_2 = \beta_0 + \beta_1$, $\log(\sigma_2) = \gamma_0 + \gamma_1$. Entonces, redefiniendo (6.8) y optimizando (6.9) redefinida se obtienen las estimaciones necesarias para construir regiones de confianza simultáneamente y con ellos averiguar si β_1 y γ_1 son ceros, lo que aseveraría una diferencia significativa entre las medias poblacionales. Esto es, las poblaciones tienen concentraciones medias idénticas si $W(0) = (\hat{\theta} - 0) (\hat{\Sigma}_{\hat{\theta}})^{-1} (\hat{\theta} - 0) \leq \chi^2_{(1-\alpha;2)}$.

Resultados y Discusión. De 6.3.3, la comparación de medias se realiza a través de una región de confianza aproximada para β_1 y γ_1 , observando si son nulos simultáneamente. El Cuadro 8 muestra las estimaciones.

Cuadro 8. EMV para los parámetros en los datos de cobre con heterogeneidad de σ

Parámetro	EMV		Error Estándar		
	Directo	Vía EM	Directo	Vía EM	
β_0	0.592	0.906	0.124	0.101	
β_1	0.084	0.038	0.170	0.137	
γ_0	- 0.039	- 0.202	0.099	0.088	
γ_1	- 0.072	- 0.102	0.141	0.124	
	μ_1	μ_2	σ_1	σ_2	$\log L_2(\beta, \gamma)$
Directo	0.676	0.592	0.962	0.895	- 168.386
Vía EM	0.944	0.906	0.817	0.738	- 151.575

De acuerdo con el Cuadro 8 y de B.2.3 se muestra que el uso del algoritmo EM mejora las estimaciones y el ajuste del modelo. De donde, vía EM, $\hat{\beta}_1 = 0.038$, $\hat{es}_{\hat{\beta}_1} = 0.137$, $\hat{\gamma}_1 = 0.038$, $\hat{es}_{\hat{\gamma}_1} = 0.137$ y $Cov(\hat{\beta}_1, \hat{\gamma}_1) = -1.425 \times 10^{-6}$; con lo que, al 95% de confianza,

$$W(0) = 1.816 \times 10^{-4} \leq 5.991 = \chi^2_{(95;2)}.$$

Por lo tanto, las concentraciones medias de cobre no difieren significativamente.

6.3.3. Selección del Mejor Modelo

En ambos modelos, 6.3.1 y 6.3.3, el algoritmo EM mejora la estimación de parámetros, varianzas e intervalos de confianza; además de acuerdo a los resultados, ambos modelos revelan que las medias no difieren. Sin embargo una forma de seleccionar el mejor modelo, es a través de una prueba de razón de verosimilitudes,

$$-2(\log L_{EM_1} - \log L_{EM_2}) = 32.478 > 3.84 = \chi^2_{(0.95;1)}.$$

De donde se aprecia que el Modelo 2 es el más apropiado para describir estos datos.

6.4 Comparación y Análisis de Tendencias en un Proceso de Remediación

La comparación de niveles medios de contaminantes de distintas zonas, el monitoreo de la calidad de agua en procesos de remediación es de gran importancia, considerando además el efecto de una covariable fija como el tiempo de monitoreo. El propósito puede ser observar el incremento de la contaminación como el resultado de cambios de uso de suelo o acumulación de desechos tóxicos.

Ejemplo 4. En McBean y Rovers (1998) se presentan datos de un proceso de remediación, donde se aplicaron técnicas de remediación de Octubre a Noviembre de 1980 y se monitoreó la calidad de agua de 1981 a 1982, ver Cuadro 9. Esta información se utiliza con la finalidad de ilustrar la metodología propuesta en procesos de remediación. Se toman los datos, censurando artificialmente a 1 ppm las mediciones no disponibles y las observaciones menores a 1. Se desea saber si el proceso de remediación mejora a calidad del agua y conocer el efecto del tiempo.

Cuadro 9. Niveles de contaminantes (ppm) antes y después del proceso de remediación

Fecha de Muestreo	Registros Preremediales	Fecha de Muestreo	Registros Postremediales
10/79	4.3	10/81	<1
11/79	16	11/81	1.5
12/79	6.1	12/81	1.3
1/80	<1	1/82	<1
2/80	<1	2/82	<1
3/80	2.66	3/82	2.1
4/80	3.00	4/82	1.1
5/80	4.42	5/82	<1
6/80	5.74	6/82	1.8
7/80	1.40	7/82	1.2
8/80	1.49	8/82	<1
9/80	2.3	9/82	1.1

Fuente: McBean and Rovers (1998). Statistical procedures for Analysis of Environmental Monitoring Data and Risk Assessment.

En la Figura 16 se muestran los gráficos de dispersión de las niveles de contaminantes antes y después del proceso de remediación, respectivamente; puede apreciarse una disminución de los niveles, sin embargo, el análisis de datos bajo la metodología propuesta será determinante.

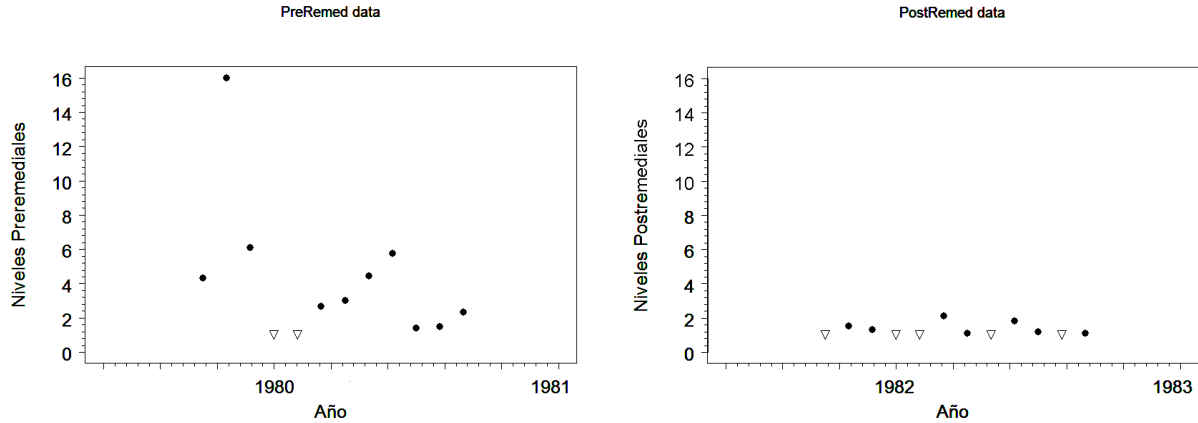


Figura 16. Gráficos de dispersión de contaminantes antes y después de la remediación.

Para una muestra lognormal y en presencia de una covariable continua t , se tiene que $E[\log(y)] = \mu = \beta + \alpha_0 t$, donde α_0 representa el intercepto del modelo log-lineal para el contaminante en la primera unidad de t y α_1 representa el cambio por unidad de t . Ahora, la función log-cuantil para el modelo de regresión correspondiente es,

$$\log[y_p(x)] = \mu(x) + \Phi^{-1}(p)\sigma = \beta_0 + \beta_1 x + \alpha_1 t + \Phi^{-1}(p)\sigma \quad (6.10)$$

Sean y_1, y_2 dos muestras lognormales con σ común, en presencia de una covariable continua t , entonces, su función de verosimilitud con datos exactos y no detectados es

$$L(\beta_0, \beta_1, \alpha_0, \sigma) = \prod_{i=1}^2 \left\{ \prod_{j=1}^{n_i} \left[\frac{1}{\sigma y_{ij}} \phi\left(\frac{\log y_{ij} - \mu_i}{\sigma}\right) \right]^{\delta_{ij}} \left[\Phi\left(\frac{\log y_{ij} - \mu_i}{\sigma}\right) \right]^{1-\delta_{ij}} \right\} \quad (6.11)$$

Donde $\mu_i = \mu(x) = \beta_0 + \beta_1 x + \alpha_0 t$. Sustituyendo x , $\mu_1 = \beta_0 + \alpha_0 t$ y $\mu_2 = \beta_0 + \beta_1 + \alpha_0 t$. La inferencia sobre la diferencia de medias dependerá solo de β_1 , ya que $\mu_2 - \mu_1 = \beta_1$ no depende de ningún cuantil, ni de t . Así, un intervalo de confianza aproximado para β_1 que contenga al 0, aseveraría que las medias no difieren significativamente. El efecto de tendencia puede cuantificarse a través $\hat{\alpha}_0$ y de su intervalo de confianza.

Resultados y Discusión. La comparación se realizó ajustando el modelo de regresión lognormal correspondiente (6.10). Entonces, optimizando (6.11), se obtienen los EMV para los parámetros del modelo y sus intervalos de confianza aproximado. Ahora, bajo el supuesto de sigma común y una covariable fija, la diferencia significativa entre las medias de ambos periodos fue observada únicamente a través del intervalo obtenido para β_1 . Por otro lado, el efecto de la covariable fija sobre los datos será medido a través del estimador y el intervalo para α_0 . El Cuadro 10 muestra los EMV, los IC Aprox. del 95% de confianza para los parámetros y los valores de la $\log L$.

Cuadros 10. EMV para los parámetros de los datos en el proceso remediación (σ común)

Parámetro	EMV		Error Estándar		IC Aprox. del 95% de confianza	
	Directo	Vía EM	Directo	Vía EM	Directo	Vía EM
β_0	- 0.378	1.189	0.439	0.358	- 1.239, 0.483	- 0.514, 0.891
β_1	0.559	1.073	0.358	0.304	- 0.143, 1.261	0.477, 1.670
α_0	0.018	- 0.037	0.056	0.044	- 0.092, 0.128	- 0.123, 0.050
σ	1.022	0.745	0.181	0.101	0.723, 1.445	0.572, 0.971

	μ_1	μ_2	σ	α_0	$\log L_1(\beta, \delta, \sigma)$
Directo	0.181	- 0.378	1.022	0.018	- 42.625
Vía EM	1.262	0.189	0.745	- 0.037	- 25.397

Observe en el Cuadro 10 que la implementación del algoritmo EM redujo la complejidad de la optimización de (6.11), obteniéndose un mejor ajuste para el modelo. Observe que el EMV vía EM para β_1 es un número positivo, 1.073, con un intervalo de confianza aproximado que no contiene el cero, (0.477, 1.670), de modo que la diferencia de medias es estadísticamente significativa.

Así que, a la luz de los datos, $\mu_{pre} > \mu_{pos}$, con lo que se asevera que el proceso de remediación mejoró la calidad del agua. Por otro lado, observe que el parámetro α_0 , correspondiente al tiempo, tiene un EMV vía EM de -0.037 , interpretándose como una ligera tendencia negativa, sin embargo su intervalo de confianza aproximado indicó que estadísticamente no existe tendencia significativas.

6.5 Comparación de Dos Muestras con Datos Agrupados

Al comparar los niveles medios de contaminantes en dos o más zonas, con frecuencia se tienen varias localidades dentro de éstas, lo que conduce a la obtención de datos agrupados o medidas repetidas. Esto produce efectos aleatorios, asociados al diseño de muestreo de las unidades individuales de la población y al diseño experimental para la obtención de mediciones (Pinheiro y Bates, 2000).

Para la comparación de dos muestras con datos agrupados y suponiendo homogeneidad de σ , se emplea (6.12) y reparametriza el efecto fijo del parámetro μ como un modelo de regresión simple con variable indicadora x . Esto es posible a través de un modelo log-lineal para las observaciones y_{ijk} , que corresponde a la i -ésima muestra del j -ésimo pozo en la k -ésima unidad de tiempo.,

$$\log(y_{ijk}) = \mu(x) + \alpha_0 t + b_i + \varepsilon_{ijk} = \beta_0 + \beta_1(x) + \alpha_0 t + b_i + \varepsilon_{ijk} \quad (6.12)$$

donde β_0 , β_1 , α_0 , son efectos fijos, b_i corresponde al efecto aleatorio debido al agrupamiento de los datos y ε_{ijk} es error para cada observación. Como en modelos anteriores, debido a la variable indicadora x , $\mu_1 = \beta_0 + \alpha_0 t$ y $\mu_2 = \beta_0 + \beta_1 + \alpha_0 t$. Entonces, la inferencia sobre la diferencia de medias dependerá solo del intervalo de confianza para β_1 . Este criterio será mas eficiente en la medida que la variabilidad de b_i sea mínima. La estimación se realiza empleando el algoritmo EM y el método general por máxima verosimilitud presentado por Pinheiro y Bates (2000), implementado en R bajo la función `lme`.

Ejemplo 5. Durante cinco meses se realizó un monitoreo sobre un sitio en riesgo de contaminación. Con la finalidad de observar si existe evidencia de contaminación, EPA (1992) reporta las concentraciones de tolueno en ppm de dos pozos de este sitio, para compararlas con las concentraciones obtenidas del mismo metal en tres pozos de un sitio control, Cuadro 11. Se observan un número considerable de datos no detectados.

Cuadro 11. Concentraciones de tolueno para dos muestras en un periodo de 5 meses

Mes	Sitio en Riesgo		Sitio Control		
	Pozo 1	Pozo 2	Pozo 1	Pozo 2	Pozo 3
1	< 5	< 5	< 5	< 5	< 5
2	7.5	< 5	12.5	13.7	20.1
3	< 5	< 5	8.0	15.3	35.0
4	< 5	< 5	< 5	20.2	28.2
5	6.4	< 5	11.2	25.1	19.0

Fuente: EPA 1992. Statistical Training Course for Ground-Water Monitoring Data Analysis.

Resultados y Discusión. Siguiendo el criterio de comparación definido en 6.4, se ajustó el modelo lineal mixto (6.12), empleando el algoritmo EM para la optimización de la función de verosimilitud y para obtener las varianzas de los estimadores de los parámetros. Utilizando la función `lme` (linear mixed effects) de *R* y anidando el algoritmo EM en una rutina de cómputo en *R* (Anexo B.2.5), se obtuvieron las estimaciones por MV para los parámetros de efectos fijos y aleatorios, ver Cuadro 12.

Cuadro 12. Estimaciones por MV para los parámetros de concentraciones de tolueno

Parámetro	EMV	Error Estándar	I.C. Aprox. del 95% de Confianza
β_0	1.792	0.348	1.110, 2.474
β_1	- 1.353	0.284	- 1.911, - 0.796
α_0	0.210	0.098	0.019, 0.401

	μ_1	μ_2	σ_b	σ_ε	$\log L(\beta, \gamma, \delta)$
EMV	0.439	1.792	0.104	0.690	- 27.664

De acuerdo con los resultados mostrados en el Cuadro 12, se observó que la estimación de β_1 es negativa (-1.353), con un intervalo de confianza aproximado que no contiene el cero, de modo que existe evidencia estadísticamente significativa de que ambas concentraciones medias son diferentes. En el contexto del problema, el sitio control tiene una mayor concentración media de tolueno. Así que, con estas muestras de datos, no existe evidencia de contaminación en el sitio en riesgo. Por otro lado, la estimación del efecto fijo de la covariable Mes, es 0.210, con un intervalo de confianza que no contiene el cero; es decir, hay evidencia de un efecto significativo y ascendente del tiempo en la concentración media del contaminante. Finalmente, es importante mencionar que el efecto aleatorio debido al agrupamiento, presentó una variabilidad estimada de 0.690, por lo que la correlación intra-grupo estimada, $\hat{\sigma}_b^2 / (\hat{\sigma}_b^2 + \hat{\sigma}_\varepsilon^2)$, fue de 0.022, cantidad cercana a cero.

7. CONCLUSIONES

La metodología desarrollada en este trabajo, basada en máxima verosimilitud y una modificación del algoritmo EM, resultó versátil, general y simple para describir y comparar poblaciones lognormales mediante modelos de regresión lognormal. El criterio de comparación mediante intervalos de confianza aproximados resultó simple y de fácil comprensión. La comparación de muestras con homogeneidad de σ es simple a través de la funciones log-cuantil y, para la heterogeneidad, las regiones de confianza por el método de Wald, es una alternativa eficiente. La metodología desarrollada se fundamentó en la teoría de probabilidad y estadística matemática dadas en el Anexo A.

Respecto al estudio de simulación, se observó que en tamaños de muestras moderados, la potencia de la prueba propuesta es mejor que la prueba logrank, al comparar poblaciones lognormales. En tamaños de muestras grandes, ambas pruebas tienen casi la misma potencia. Por otro lado, la prueba paramétrica propuesta funciona aceptablemente al comparar muestras no lognormales, exponenciales y Gumbel, indicando cierta robusticidad ante la relajación del supuesto de distribución de origen. La asimetría observada en las potencia estimadas ha motivado a un estudio posterior.

En los Ejemplos de aplicación, 1 y 2, se describió una sola población con datos no detectados y una covariable continua; la inferencia no fue difícil en cada caso y resultó simple su interpretación. En el Ejemplo 3 se ilustró la comparación de dos poblaciones bajo homogeneidad y heterogeneidad del parámetro σ , reparametrizando ambos parámetros, μ y σ . En este caso, el criterio de comparación resultó eficiente al observar un intervalo de confianza aproximado y el estadístico de Wald, respectivamente; también se presentó una forma de seleccionar el mejor modelo a través una prueba de razón verosimilitud, la cual no fue difícil de implementar e interpretar. Por otro lado, en el Ejemplo 4, las fechas de monitoreo como covariable continua, fueron agregadas sin dificultad en el modelo de regresión lognormal, obteniéndose estimadores e intervalos de confianza de fácil interpretación. En cada caso la metodología desarrollada se aplicó exitosamente y se cumplieron los objetivos respectivos.

El enfoque de modelos lineales mixtos, presentado en el Ejemplo 5, permitió extender el método de comparación a poblaciones con efectos aleatorios debido a datos agrupados o esquemas de muestreo, manteniéndose el criterio de comparación basado en los intervalos de confianza aproximados.

La metodología desarrollada, bajo la prueba paramétrica propuesta, el estudio de simulación y el desarrollo de los ejemplos de aplicación, se implementaron computacionalmente en R. Los programas y algoritmos respectivos se muestran en el Anexo B, con la finalidad de discutir, con futuros estudios, el cómputo y la algoritmia.

En los ejemplos de aplicación, el algoritmo EM mejoró sustancialmente la optimización de las funciones de verosimilitud respectivas que contienen datos no detectados, con lo que se observó un mejor ajuste, errores estándar menores e intervalos de confianza más estrechos, comparado con los obtenidos al optimizar directamente la función de verosimilitud. En los ejemplos mostrados se compararon únicamente dos poblaciones, sin embargo, el procedimiento puede extenderse a tres o más poblaciones lognormales, debido a que la implementación e interpretación no es difícil.

Los resultados del presente proyecto de investigación, manifestados en las conclusiones anteriores, generaron las publicaciones siguientes:

- Censored Data Analysis on Engineering and Biological Sciences, *Revista de Matemática: Teoría y Aplicaciones* of MathScinet and Zentralblath, Centro de Investigación en Matemática Pura y Aplicada, CIMPA – Universidad de Costa Rica.
- Statistical procedures to compare mean pollutant concentrations of lognormal populations containing nondetects data, *Environmental Modeling and Assessment* of Springer Journals Research.

Los contenidos de estas publicaciones se incluyen muestran en el Anexo C.

8. REFERENCIAS

- Ahn, H., 1998. Estimating the mean and variance of censored phosphorus concentrations in Florida rainfall: *Journal of the American Water Resources Association* 34, 583 – 593.
- Aitchinson, J., and J. A. C. Brown, 1969. *The lognormal Distribution*. Cambridge University Press, Cambridge..
- Akritas, M. G., 1994. Statistical analysis of censored environmental data: Chapter 7 of the *Handbook of Statistics, Volume 12*, edited by G.P. Patil and C.R. Rao. North-Holland, Amsterdam, 927 pp.
- Army, 1998. Evaluation of dredged material proposed for discharge in waters of the US – Testinfg manual: Army Corps of Engineers published as EPA-823-B-98-004.
- Bermudez, J.D., 2002. *Técnicas Numéricas en Estadística*. Notas de Curso de Doctorado, Departament D'Estadística I Investigació Operativa, Universitat De Valencia. Valencia, España.
- Billingsley, P., 1986. *Probability and Measure*, 2nd Ed. New York: John Wiley & Sons
- Buckley, T.J., J. Liddle, D.L. Ashley, D.C. Paschal, V.W. Burse, L.L. Needham, and G. Akland, 1997. Environmental and biomarker measurements in nine homes in the lower Rio Grande valley: Multiyear results for pesticides, metals, PAHs, and VOCs; *Environmental Pollution* 23, 705 – 722.
- Clarke, J.U., 1998. Estimation of censored data methods to allow statistical comparisons among very small samples with below detection limit observations: *Environmental Science and Technology* 32, 177 – 183.
- Cox, D.R., and Snell, E. J. 1981. *Applied Statistics*, New York: Chapman and Hall Inc.
- Dempster, A.P., Laird, N., & Rubin, D.B. 1977. Maximum likelihood for incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B39*, 1–38.
- Díaz-Francés, E., and Sprott, D.A. 2000. The use of the likelihood function in the analysis of environmental data. *Environmetrics* 11, 75–79.
- El-Shaarawi, A. H., and Viveros, R. 1997. Inferences about the mean in log-regression with environmental applications. *Environmetrics* 8, 569–582.

- EPA 1992. Statistical Training Course for Ground-Water Monitoring Data Analysis. EPA530-R-93-003. Office of Solid Waste. U.S. Environmental Protection Agency, Washington, DC.
- EPA 1998. Guidance for data quality assessment. Practical methods for data analysis; US Environmental Protection Agency EPA/600/R-96/084.
- Flury, B., & Zoppè, A. 2001. Exercise in EM. *The Am. Statist.* 54, pp. 207 – 209.
- Frome, E. L., and Watkins, J. W., 2004. Statistical Analysis of Data with Non-Detectable Values, Reports DE-AC05-00OR22725 Oak Ridge National Laboratory, U.S. Department Of Energy, Oak Ridge, Tennessee, <http://www.osti.gov/bridge>
- Galton, F. 1879. The geometric mean in vital and social statistics, *Proceedings of the Royal Society of London*, 29, 365-367.
- Gilbert, R.O. 1987. *Statistical Methods for Environmental Pollution Monitoring*. Wiley, New York. 320p.
- Gleit, A., 1985. Estimation for small normal data sets with detection limits. *Environmental Science and Technology* 19, 1201 – 1206.
- Graybill, F.A. 1976. *Theory and Application of the Linear Model*. Duxbury
- Harris, M. L., J.E. Elliot, R. W. Butler, and L. K. Wilson, 2003. Reproductive success and chlorinated hydrocarbon contamination of resident great blue herons from coastal British Columbia, 1977 - 2000: *Environmental Pollution* 121, 207 – 227.
- Heyde, C. C. 1963. On a property of the lognormal distribution, *Journal of the Royal Statistical Society, Series B*, 25, 392 – 393.
- Helsel, D. R. 1990. Less than obvious: Statistical treatment of data below the detection limit: *Environmental Science and Technology* 24, pp. 1766 – 1774.
- Helsel, D. R., and Gilliom, R. J. 1986. Estimation of distributional parameters for censored trace level water quality data, 2. Verification and applications: *Water Resources Research* 22, 147 – 155.
- Helsel, D. R. and T. A. Cohn, 1988. Estimation of descriptive statistics for multiply censored water quality data: *Water Resources Research* 24, 1997 – 2004.
- Helsel, D.R., 2005. *Nondetects And Data Analysis, Statistics for Censored Environmental Data*, Wiley-Interscience, New York, 251 pp.

- Hobbs, K. E., D. C. G. Muir, E. W. Bornb, R. Dietzc, T. Haugd, T. Metcalfee, C. Metcalfee and N. Oien, 2003. Levels and patterns of persistent organochlorines in mink whale (*Balaenoptera acutorostrada*) stocks from the North Atlantic and European Arctic; *Environmental Pollution* 121, 239 – 252.
- Huybrechts, T., O. Thas, J. Dewulf, and H. Van Langenhove, 2002. How to estimate moments and quantiles of environmental data sets with non-detected observations? A case study on volatile organic compounds in marine water samples; *Journal of chromatography* 975, 123 – 133.
- Isobe, T., E. D. Feigelson, and P. I. Nelson, 1986. Statistical methods for astronomical data with upper limits. *Astrophysical Journal* 306, 490 – 507.
- Johnson, N.L., and S. Kotz, 1970. *Continuous Univariate Distributions I*. Houghton Mifflin, Boston.
- Kolpin, D. W., E. T. Furlong, M. T. Meyer, E. M. Thurman, S. D. Zaugg, L. Barber, and H. T. Buxton, 2002. Pharmaceuticals, hormones, and other organic waste-water contaminants in U. S. streams, 1999 – 2000: A national reconnaissance; *Environmental Science and Technology* 36, 1202 – 1211.
- Kroll, C.N. and J.R. Stedinger, 1996. Estimation of moments and quantiles using censored data; *Water Resources Research* 32, 1005 – 1012.
- Laird, N.M., and Ware, J.H. 1982. Random-effects models for longitudinal data. *Biometrics* 38, 963 – 974.
- Lambert, D., Peterson, B., & Terpenning, I. 1991. Nondetects, Detection Limits, and the Probability of Detection. *Journal of the American Statistical Association*, Vol. 86, No. 414, 266 – 277.
- Lee, E. T., and J. W. Wang, 2003. *Statistical Methods for Survival Data Analysis*, Third edition. Wiley, New York, 534 pp.
- Lyn, H. S., 2001. Maximum likelihood inference for left-censored HIV RNA data, *Statistical in Medicine* 20, 33 – 45
- Lundgren, R. F., and T. J. Lopes, 1999. Occurrence, distribution, and trends of volatile organic compounds in the Ohio River and its major tributaries, 1987 – 96; U.S. Geological Survey Water-Resources Investigations Report 99 – 4257, 89 pp.

- McBean, E.A., and Rovers, F.A. 1998. Statistical procedures for Analysis of Environmental Monitoring Data and Risk Assessment. Prentice Hall. NJ. 336p.
- Meeker, W.Q., and Escobar, L.A. 1995. Teaching about approximate confidence regions based on maximum likelihood estimation. *The Am. Statist.*, 49, 48-53.
- Meeker, W.Q., and Escobar, L.A. 1998. *Statistical Methods for Reliability Data*, 450 – 454. New York: Wiley, 680p.
- Miesch, A., 1967. Methods of computation for estimating geochemical abundance: U.S. Geological Survey professional Paper 574 – B. 15 pp.
- Millard, S. P. and S. J. Deverel, 1988. Nonparametric statistical methods for comparing two sites based on data with multiple nondetect limits: *Water Resources Research* 24, 2087 – 2098.
- Millard, S. P., and Neerchal, N. K. 2001. *Environmental Statistics with SPLUS*. FL: CRC
- Navy 1999. Handbook for statistical analysis of environmental background data, Department of the Navy, Southwest Division, Naval Facilities Engineering Command, San Diego, CA.
- Nehls, G.J. and G.G. Akland, 1973, Procedures for handling aerometric data; *Journal of the Air Pollution Control Association* 23, 180 – 184.
- Oakes, D. 1999. Direct calculation of the information matrix via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, 61, 479 – 482.
- Ott, W. R. 1995. *Environmental Statistics and Data Analysis*, FL: CRC Press.
- Owen, W., and T. DeRouen, 1980. Estimation of the mean of lognormal data containing zeros and left-censored values, with applications to the measurement of worker exposure to air contaminants: *Biometrics* 36, 707 – 719.
- Perkins, J. L., G. N. Cutter, and M.S. Cleveland, 1990, Estimating the mean, variance, and confidence limits from censored (<limit of detection), lognormally-distributed exposure data: *American Industrial Hygiene Association Journal* 51, 416 – 419.
- Peto, R., and Peto, J., 1972. Asymptotically Efficient Rank Invariant Procedures. *Journal of the Royal Statistical Society, Series A*, 135, 185 - 207.
- Pinheiro, J. C., and Bates, D. M. 2000. *Mixed-Effects Models in S and S-PLUS*, New York: Springer

- R Development Core Team. 2006. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- Rao, S. T., J. Y. Ku, and K. S. Rao. 1991. Analysis of toxic air contaminant data containing concentrations below the limit of detection: *Journal of the Air and Waste Management Association* 41, 442 – 448.
- Robert, C. P., and G. Casella. 2004. *Monte Carlo Statistical Methods*. 2nd Ed. Springer, N.Y. 645p.
- Savage, I. R., 1956. Contributions to the Theory of Rank order Statistics: The Two Sample Case. *Annals of Mathematical Statistics*, 27, 590 - 615.
- She, N., 1997. Analyzing censored water quality data using a non-parametric approach: *Journal of the American Water Resources Association* 33, 615 – 624.
- Singh, A. and J. Nocerino, 2002. Robust estimation of mean and variance using environmental data sets with below detection limit observations; *Chemometrics and Intelligent Laboratory Systems* 60, 69 – 86.
- Slyman, D. J., A. de Peyster, and R. R. Donohoe, 1994. Hypothesis testing with values below detection limit in environmental studies: *Environmental Science and Technology* 28, 898–902.
- SPLIDA 2004. S-PLUS Life Data Analysis. W. Q. Meeker Department of Statistics Iowa State University Ames, Iowa, USA. URL <http://www.public.iastate.edu/~splida>.
- Till, A. E., 2003. Comment on pharmaceuticals hormones, and organic wastewater contaminants in US streams, 1999 – 2000; *Environmental Science and Technology* 37, 1052 – 1053.
- Tanner, M. R. 1994. *Tools for Statistical Inference. Methods for the Exploration of Posterior Distributions and Likelihood Functions*. 2a Ed. Springer-Verlag. N.Y.
- Tomlinson, M.S. 2003. *Effects of Ground-Water/Surface-Water Interactions and Land Use on Water Quality*. Written communication in advance of a USGS report.
- Wartmann, R. 1956. Anwendung der logarithmischen Normalverteilung, *Mitteilungsblatt für Mathematische Statistik*, 8, 83 – 91.
- Wen, X. H., 1994. Estimation of statistical parameters for censored lognormal hydraulic conductivity measurements: *Mathematical Geology* 26, 717 – 731.

ANEXO A. ALGUNOS RESULTADOS DE TEORÍA ESTADÍSTICA

A.1 Difusión del Error Estadístico - El Método Delta

En esta sección se ilustra como calcular el valor esperado, varianza y covarianzas de funciones de estimadores para los parámetros. Sea $g(\theta)$ una función escalar de los parámetros $\theta=(\theta_1, \dots, \theta_r)'$ y sean $\hat{\theta}$ y $g(\hat{\theta})$ los estimadores de θ y $g(\theta)$, respectivamente. El objetivo es obtener expresiones aproximadas para $E[g(\hat{\theta})]$ y $Var[g(\hat{\theta})]$ como una función de $E(\hat{\theta}_i)$, $Var(\hat{\theta}_i)$ y $Cov(\hat{\theta}_i, \hat{\theta}_j)$.

El caso simple es cuando $g(\hat{\theta})$ es una función lineal de $\hat{\theta}_i$, es decir, $g(\hat{\theta}) = a_0 + \sum_{i=1}^r a_i \hat{\theta}_i$, donde las a_i , $i = 1, \dots, r$, son constantes. Convenientemente, $g(\hat{\theta})$ se expresa como

$$g(\hat{\theta}) = a_0 + \sum_{i=1}^r a_i \hat{\theta}_i = b_0 + \sum_{i=1}^r b_i [\hat{\theta}_i - E(\hat{\theta}_i)], \quad (\text{A.1})$$

donde $b_0 = a_0 + \sum_{i=1}^r a_i E(\hat{\theta}_i)$ y $b_i = a_i$. Cálculos simples con esperanzas y varianzas dan

$$E[g(\hat{\theta})] = b_0,$$
$$Var[g(\hat{\theta})] = \sum_{i=1}^r b_i^2 Var(\hat{\theta}_i) + r \sum_{i=1}^r \sum_{j=1, j \neq i}^r b_i b_j Cov(\hat{\theta}_i, \hat{\theta}_j).$$

Cuando $g(\hat{\theta})$ es no lineal de $\hat{\theta}_i$ y $g(\hat{\theta})$ pueden ser aproximada por una función lineal de los $\hat{\theta}_i$ en la región con verosimilitud significativa. El procedimiento se llama “método delta” o “difusión de error estadístico” y aquí se describe una versión simplificada.

Cuando $g(\theta)$ tiene segundas derivadas parciales continuas, una expansión en series de Taylor de primer orden de $g(\theta)$ al rededor de $\mu = [E(\hat{\theta}_1), \dots, E(\hat{\theta}_r)]$ esta dada por

$$g(\hat{\theta}) \approx g(\mu) + \sum_{i=1}^r \frac{\partial g(\theta)}{\partial \theta_i} [\hat{\theta}_i - E(\hat{\theta}_i)], \quad (\text{A.2})$$

donde las derivadas parciales de $g(\theta)$ con respecto a los θ_i son evaluados en μ .

Observe que la ecuación (A.2) se parece a la ecuación (A.1) con

$$b_0 = g(\mu) \quad \text{y} \quad b_i = \frac{\partial g(\theta)}{\partial \theta_i}, \quad i = 1, \dots, r.$$

Consecuentemente,

$$\begin{aligned} E[g(\hat{\theta})] &= g(\mu), \\ \text{Var}[g(\hat{\theta})] &\approx \sum_{i=1}^r \left[\frac{\partial g(\theta)}{\partial \theta_i} \right]^2 \text{Var}(\hat{\theta}_i) + \sum_{i=1}^r \sum_{\substack{j=1 \\ j \neq i}}^r \left[\frac{\partial g(\theta)}{\partial \theta_i} \right] \left[\frac{\partial g(\theta)}{\partial \theta_j} \right] \text{Cov}(\hat{\theta}_i, \hat{\theta}_j). \end{aligned} \quad (\text{A.3})$$

Cuando los $\hat{\theta}_i$ no están correlacionados o las $\text{Cov}(\hat{\theta}_i, \hat{\theta}_j)$, $i \neq j$, son pequeñas comparadas con las $\text{Var}(\hat{\theta}_i)$, el último término del lado derecho de (A.3) es omitido.

Las mismas ideas aplican para una función vectorial. Por ejemplo, si $g_1(\theta)$ y $g_2(\theta)$ son dos funciones escalares entonces

$$\begin{aligned} \text{Cov}[g_1(\hat{\theta}), g_2(\hat{\theta})] &\approx \sum_{i=1}^r \left[\frac{\partial g_1(\theta)}{\partial \theta_i} \right] \left[\frac{\partial g_2(\theta)}{\partial \theta_i} \right] \text{Var}(\hat{\theta}_i) \\ &\quad + \sum_{i=1}^r \sum_{\substack{j=1 \\ j \neq i}}^r \left[\frac{\partial g_1(\theta)}{\partial \theta_i} \right] \left[\frac{\partial g_2(\theta)}{\partial \theta_j} \right] \text{Cov}(\hat{\theta}_i, \hat{\theta}_j). \end{aligned} \quad (\text{A.4})$$

En general, para una función vectorial $g(\theta)$ de los parámetros tales como las segundas derivadas parciales con respecto a los elementos de θ son continuas

$$Var[g(\hat{\theta})] \approx \left[\frac{\partial g(\theta)}{\partial \theta_i} \right] Var(\hat{\theta}) \left[\frac{\partial g(\theta)}{\partial \theta} \right],$$

donde $\partial g(\theta)/\partial \theta = [\partial g_1(\theta)/\partial \theta, \partial g_2(\theta)/\partial \theta, \dots]$ es la matriz de vectores gradientes de las primeras derivadas parciales de $g(\theta)$ con respecto a θ y

$$Var(\hat{\theta}) = \begin{bmatrix} Var(\hat{\theta}_1) & Cov(\hat{\theta}_1, \hat{\theta}_2) & \dots & Cov(\hat{\theta}_1, \hat{\theta}_r) \\ & Var(\hat{\theta}_2) & \dots & Cov(\hat{\theta}_2, \hat{\theta}_r) \\ & & \ddots & \vdots \\ simétrica & & & Var(\hat{\theta}_r) \end{bmatrix}$$

ambas varianzas evaluadas en θ .

El método delta provee buenas aproximaciones para $E[g(\hat{\theta})]$ y $Var[g(\hat{\theta})]$. Sin embargo, debe tenerse cuidado en aplicar este método debido a que la aproximación depende de la validez de la aproximación de Taylor y la magnitud del residuo de la aproximación

A.2 Matriz de Información de Fisher

Sea $\log L(\theta) = \sum_{i=1}^n \log L_i(\theta)$ la log-verosimilitud total para un modelo y datos especificados que consisten de n observaciones, no necesariamente idénticamente distribuidas. Se entiende que $\log L_i(\theta)$ es la contribución de la i -ésima observación a la $\log L(\theta)$. Cuando existe, $\hat{\theta}$ es el valor de θ que maximiza $\log L(\theta)$, conocido como el estimador de Máxima Verosimilitud (EMV). Sea $I(\theta)$ la cantidad promedio (o límite) por muestras grandes, de la información debida a la observación. Entonces en general,

$$I(\theta) = \lim_{n \rightarrow \infty} \left\{ \frac{1}{n} E \left[- \frac{\partial^2 \log L(\theta)}{\partial \theta \partial \theta'} \right] \right\} = \lim_{n \rightarrow \infty} \left\{ \frac{1}{n} \sum_{i=1}^n E \left[- \frac{\partial^2 \log L_i(\theta)}{\partial \theta \partial \theta'} \right] \right\}, \quad (A.5)$$

donde el valor esperado es con respecto a los datos aún no observados. Para muestras grandes, la matriz $I_\theta = nI(\theta)$ cuantifica aproximadamente la cantidad de información que “se espera” conseguir con los datos. Intuitivamente, esto puede ser visto con las segundas derivadas más grandes de $\log L(\theta)$ que indiquen mas curvatura en la log-verosimilitud, implicando que la verosimilitud está mas concentrada alrededor de su máximo.

Para una gran clase de situaciones y de modelos, incluyendo aquellos con observaciones independientes e idénticamente distribuidas, I_θ se simplifica en la bien conocida matriz de información de Fisher para θ ,

$$I_\theta = E \left[-\frac{\partial^2 \log L(\theta)}{\partial \theta \partial \theta'} \right] = \sum_{i=1}^n E \left[-\frac{\partial^2 \log L_i(\theta)}{\partial \theta \partial \theta'} \right]. \quad (\text{A.6})$$

I_θ es frecuentemente conocida como la matriz de información observada de Fisher. Cuando se tienen datos disponibles, se puede calcular la matriz local de observación observada para θ como

$$\hat{I}_\theta = -\frac{\partial^2 \log L(\theta)}{\partial \theta \partial \theta'} = \sum_{i=1}^n \left[-\frac{\partial^2 \log L_i(\theta)}{\partial \theta \partial \theta'} \right]. \quad (\text{A.7})$$

donde las derivadas son evaluadas en $\theta = \hat{\theta}$.

En la Sección A.5, se explica que, bajo condiciones de regularidad, $n\Sigma_{\hat{\theta}} = nI(\hat{\theta})^{-1}$ es la matriz de covarianza para la distribución asintótica de $\sqrt{n}(\hat{\theta} - \theta)$ y una estimación de I_θ puede ser usada para estimar la variabilidad de muestreo en $\hat{\theta}$.

A.3 Condiciones de Regularidad para Distribuciones de Localización-Escala

Cada resultado asintótico, tales como las distribuciones asintóticas de un estimador o sus propiedades asintóticas específicas, requieren su propio conjunto de condiciones sobre el modelo. Por ejemplo, bajo ciertas condiciones, es posible mostrar que los EMV son asintóticamente normal y, con condiciones adicionales, también son asintóticamente eficientes. El modelo en cuestión incluye el modelo de probabilidad y el modelo del proceso de inspección y las características del proceso de censura.

Para un conjunto grande de casos conocidos como “regulares”, existen resultados asintóticos útiles que se aplican cuando la función de distribución de Y , $f(y; \theta)$ (o transformación monótona de Y) satisface ciertas condiciones discutidas abajo.

Cuando Y (o una transformación de Y como $Y = \log Y$) es de localización-escala con densidad, $f_y(y; \theta) = (1/\sigma)\phi[(y - \mu)/\sigma]$, $\theta = (\mu, \sigma)$, $-\infty < y < \infty$, $-\infty < \mu < \infty$, $\sigma > 0$, la condiciones “de regularidad” pueden ser expresadas como sigue:

- $\phi(z) > 0$ para todo $-\infty < z < \infty$.
- El siguiente limite se cumple

$$\lim_{z \rightarrow \pm\infty} \left[z^2 \times \frac{\partial \phi(z)}{\partial z} \right] = 0.$$

- La segunda derivada $\partial^2 \phi(z) / \partial z^2$ es continua.
- La matriz

$$E \left\{ \frac{\partial^2 \log [\phi(z)]}{\partial \theta \partial \theta'} \right\},$$

es positiva definida y todos sus elementos son finitos.

Estas condiciones se satisfacen para la distribución normal (lognormal), de valor extremo (Weibull) y logística (loglogística). Pero no se cumplen para las mismas distribuciones con parámetro de corrimiento, debido a que $f_Y(y; \theta) > 0$ depende de θ .

A.4 Convergencia en Distribución

La convergencia en distribución es un importante concepto para describir el comportamiento de estimadores en muestras grandes. Por ejemplo, frecuentemente se está interesado en las propiedades estadísticas de los estimadores de MV $\hat{\theta}_n$ del escalar θ cuando el tamaño de muestra se incrementa. En este caso una aproximación común es considerar las razones estudentizadas

$$Z_n = Z_n(\theta) = \frac{\hat{\theta}_n - \theta}{se_{\hat{\theta}_n}}, \quad n = 2, \dots, \quad (\text{A.8})$$

donde \hat{se}_{θ_n} es un estimador consistente de $se_{\hat{\theta}_n}$. La distribución de Z_n es complicada, dependiendo del modelo, parámetros y tamaño de muestra. Pero bajo condiciones de regularidad (Sección A.3), si $Z_n(\theta)$ es evaluada en el valor verdadero θ , entonces,

$$\lim_{n \rightarrow \infty} F_{Z_n}(z) = \Phi(z), \quad \forall z.$$

Entonces para n finita, puede usarse la aproximación

$$\begin{aligned} \Pr \left[z_{(\alpha/2)} < Z_n < z_{(1-\alpha/2)} \right] &= F_{Z_n} \left[z_{(1-\alpha/2)} \right] - F_{Z_n} \left[z_{(\alpha/2)} \right] \\ &\approx \Phi \left[z_{(1-\alpha/2)} \right] - \Phi \left[z_{(\alpha/2)} \right] = 1 - \alpha. \end{aligned}$$

La adecuación de esta aproximación ha sido estudiada para cada problema; pero, en general, funciona bien para una gran clase de problemas y tamaño de muestras

moderadas a grandes. Mas generalmente, se dice que Z_n converge en distribución a la variable aleatoria continua V si

$$\lim_{n \rightarrow \infty} F_{Z_n}(z) = F_V(z), \quad \forall z,$$

donde $F_V(z)$ es la distribución de V . Entonces se puede usar la distribución límite de F_V para aproximar las probabilidades para n finitas como sigue:

$$\Pr(a < Z_n \leq b) = F_{Z_n}(b) - F_{Z_n}(a) \approx F_V(b) - F_V(a),$$

donde a y b son constantes específicas. Esta aproximación puede acers tan cercana como se desee tomando valores grandes de n . Esas ideas de convergencia en distribución se generalizan a variables aleatorias vectoriales (Billingsley, 1986).

Para otros ejemplos, sean $\hat{\theta}_n = (\hat{\theta}_{1n}, \hat{\theta}_{2n})$ los EMV de un vector $\theta = (\theta_1, \theta_2)$ con una muestra de tamaño n y el supuesto de que las condiciones de regularidad se cumplen.

La verosimilitud perfil de θ_1 es

$$R_n(\theta_1) = \max_{\theta_2} \left[\frac{L(\theta_1, \theta_2)}{L(\hat{\theta}_n)} \right]$$

El estadístico de razón de log-verosimilitud para el subconjunto de parámetros es $-2 \log [R_n(\theta_1)]$. El cual, evaluado en el valor verdadero de θ , converge en distribución a la distribución $\chi_{r_1}^2$, donde r_1 es la dimensión de θ_1 .

El subconjunto de parámetros para el “estadístico de Wald” es

$$W_n(\theta_1) = (\hat{\theta}_{1n} - \theta_1) \left(\hat{\Sigma}_{\hat{\theta}_{1n}} \right)^{-1} (\hat{\theta}_{1n} - \theta_1).$$

$W_n(\theta_1)$, evaluada en el valor verdadero de θ_1 , converge en distribución $\chi_{r_1}^2$.

A.5 Distribución Asintótica de los Estimadores de Máxima Verosimilitud

Se asume que los estimadores de máxima verosimilitud (EMV) de θ están basados en n observaciones y que las condiciones de regularidad, dadas en la Sección A.3, se cumplen. Entonces puede ser mostrado que $\sqrt{n}(\hat{\theta} - \theta)$ converge en distribución a una distribución normal multivariada con media cero y matriz de covarianza $I^{-1}(\theta)$ donde $I(\theta)$ está definido en la Sección A.4. En una convencional redacción casual, se dice que $\hat{\theta}$ es aproximadamente normal con media θ y matriz de covarianza $\Sigma_{\hat{\theta}} = I_{\theta}^{-1}$, donde $I_{\theta} = nI(\theta)$. La teoría estadística asintótica para muestras grandes muestra que bajo condiciones de regularidad estándar, los elementos de $\Sigma_{\hat{\theta}}$ son del orden de n^{-1} . Esto puede ser visto notando que $n\Sigma_{\hat{\theta}}$ no depende de n , siguiendo de la definición de $I(\theta)$ en la Sección A.4.

En general, el interés está en la inferencia sobre funciones de θ . Por ejemplo, considere un vector de funciones $g(\theta)$ de los parámetros tales que todas las segundas derivadas con respecto a los elementos de θ son continuas. El estimador de MV de $g(\theta)$ es $\hat{g} = g(\hat{\theta})$. En muestras grandes, $g(\hat{\theta})$ es aproximadamente normal con media $g(\theta)$ y matriz de varianza-covarianza

$$\Sigma_{\hat{g}} = \left[\frac{\partial g(\theta)}{\partial \theta} \right] \Sigma_{\hat{\theta}} \left[\frac{\partial g(\theta)}{\partial \theta} \right]^T. \quad (\text{A.9})$$

La aproximación se basa en el supuesto de que $g(\hat{\theta})$ es aproximadamente lineal en $\hat{\theta}$, alrededor de θ . La aproximación es mejor en muestras grandes debido a que la variación en $\hat{\theta}$ es más pequeña y entonces la región sobre la cual $\hat{\theta}$ varía, es correspondientemente más pequeña. Si esta región es suficientemente pequeña, la aproximación lineal será adecuada, como lo muestra el método delta (ver Sección A.1).

Para una función escalar g y θ , la fórmula se simplifica a

$$\text{Avar} \left[g \left(\hat{\theta} \right) \right] = \left[\frac{\partial g \left(\theta \right)}{\partial \theta} \right]^2 \text{Avar} \left[\hat{\theta} \right],$$

donde “Avar” es la función de la varianza asintótica. Por ejemplo, si θ es positiva y es la función logarítmica, la varianza asintótica de $\log \left(\hat{\theta} \right)$ es $\text{Avar} \left[\log \left(\hat{\theta} \right) \right] = \text{Avar} \left[\hat{\theta} \right] / \theta^2$.

Ahora, bajo condiciones de regularidad, $\hat{\Sigma}_{\hat{\theta}} = \left(\hat{I}_{\theta} \right)^{-1}$ es un estimador consistente de Σ_{θ} , donde \hat{I}_{θ} está definido en la ecuación (A.7). Este estimador local de Σ_{θ} es obtenido estimando la curvatura esperada en (A.6) por la curvatura observada en (A.7).

El estimador local de la matriz de varianzas-covarianzas de $\hat{g} = g \left(\hat{\theta} \right)$ puede ser obtenido sustituyendo $\hat{\Sigma}_{\hat{\theta}}$ por $\Sigma_{\hat{\theta}}$ en la ecuación (A.9) dando

$$\hat{\Sigma}_{\hat{g}} = \left[\frac{\partial g \left(\theta \right)}{\partial \theta} \right] \hat{\Sigma}_{\hat{\theta}} \left[\frac{\partial g \left(\theta \right)}{\partial \theta} \right]. \quad (\text{A.10})$$

Las derivadas se evalúan en $\theta = \hat{\theta}$. Para g , función escalar, y θ la fórmula se simplifica

$$\hat{V}ar \left[g \left(\hat{\theta} \right) \right] = \left[\frac{\partial g \left(\theta \right)}{\partial \theta} \right]^2 \hat{\Sigma}_{\hat{\theta}} = \left[\frac{\partial g \left(\theta \right)}{\partial \theta} \right]^2 \hat{V}ar \left(\hat{\theta} \right).$$

Por ejemplo, si θ es positiva y $g(\theta)$ es la función logarítmica, el estimador local de la varianza de $\log \left(\hat{\theta} \right)$ es $\hat{V}ar \left[\log \left(\hat{\theta} \right) \right] = \hat{V}ar \left(\hat{\theta} \right) / \hat{\theta}^2$ y $\hat{se} \left[\log \left(\hat{\theta} \right) \right] = \hat{se} \left(\hat{\theta} \right) / \hat{\theta}$.

Por otro lado, suponga que se quiere estimar θ_1 , de la partición $\theta = (\theta_1, \theta_2)$. Sea r_1 la longitud de θ_1 . La verosimilitud perfil para θ_1 es

$$R(\theta_1) = \max_{\theta_2} \left[\frac{L(\theta_1, \theta_2)}{L(\hat{\theta})} \right]. \quad (\text{A.11})$$

Cuando la longitud de θ_2 es 0 (como el caso de la distribución exponencial) la ecuación (A.11) es una verosimilitud relativa para $\theta = \theta_1$. De otro modo se tiene una verosimilitud relativa maximizada para θ_1 . En cualquier caso, $R(\theta_1)$ es comúnmente conocida como una verosimilitud perfil debido a que esta provee una vista del perfil de $L(\theta)$ a lo largo de un alineamiento perpendicular al eje de θ_1 .

Cuando θ_1 es de longitud 1, $R(\theta_1)$ es una curva proyectada sobre un plano; ahora, si θ_1 es de longitud 2 o más, $R(\theta_1)$ es una superficie proyectada en un hiperplano. En todo caso la proyección está en una dirección perpendicular a los ejes coordenados de θ_1 . Cuando θ_1 es de longitud 1 o 2, es útil desplegar $R(\theta_1)$ gráficamente.

Asintóticamente, $\log R_n(\theta_1) = -2 \log [R(\theta_1)]$ evaluado en el valor verdadero de θ_1 , tiene una distribución $\chi_{r_1}^2$. Para hacer una prueba de significancia de razón de verosimilitud, podríamos rechazar la hipótesis nula $H_0: \theta = \theta_0$, al nivel α de significancia, si

$$-2 \log [R(\theta_0)] > \chi_{(1-\alpha; r_1)}^2.$$

APÉNDICE B. PROGRAMAS EN R

B.1. Estudio de Simulación y Funciones de Potencia

B.1.1. Prueba Propuesta y Logrank Comparando Poblaciones Lognormales.

Prueba Propuesta Utilizando el Algoritmo EM

ALGORITMO EM

```
fillconEM<-function(wlns, censuraw)
{
  wnd<-rep(censuraw, length(wlns[which(wlns<=censuraw)]))
  wobs<-wlns[which(wlns>censuraw)]
  w<--log(c(wobs, wnd))
  censura<-c(rep(0, length(wobs)), rep(1, length(wnd)))
  regre<-lm(w ~ 1)
  lw<-length(w)
  b<-coef(regre)
  sigma<-sqrt(deviance(regre)/(lw-1))
  sigma1<-0
  b1<-0
  eps<-1
  iter<-0
  while(eps>0.005)
  {
    media<-b
    z<-(w-media)/sigma
    sesgo<-sigma*dnorm(z)/(1-pnorm(z))
    wnuevo<-(media+sesgo)*censura+w*(1-censura)
    regre<-lm(wnuevo ~ 1)
    b<-coef(regre)
    suma<-sum((z*sesgo*sigma+sigma^2-sesgo^2)*censura)
    sigma<-sqrt((deviance(regre)+suma)/lw)
    eps<-max(abs(b-b1), abs(sigma-sigma1))
    b1<-b
    sigma1<-sigma
    iter<-iter+1
  }
  return(exp(-wnuevo))
}
```

FUNCION DE VEROSIMILITUD COMPLETA PARA AMBAS MUESTRAS

```
vero<-function(beta)
{
  zx<-(log(x)-(beta[1]+beta[3]))/beta[2]
  zy<-(log(y)-beta[1])/beta[2]

  logL<-sum(log((1/beta)*dnorm(zx)))+sum(log((1/beta[2])*dnorm(zy)))
  return(-1*logL)
}
```

FUNCION PARA CALCULAR P-VALORES

```

alfas<-function(tmuestra)
{
niveles<-c(1:length(miu))
k<-0
for (i in miu)
  {
  k<-k+1
  estimapv<-function(j)
  {
  xlms<-rlnorm(tmuestra,i,1)
  ylms<-rlnorm(tmuestra)

  x<<-fillconEM(xlms,qlnorm(quantil,i,1))
  y<<-fillconEM(ylms,qlnorm(quantil))

  bg<-optim(c(0,0,i),vero, hessian=TRUE)
  es<-sqrt(diag(ginv(bg$hessian)))

  return(iffelse((LI[2]<0&0<LS[2]),1,0))
  }
  niveles[k]<-1-(sum(sapply(1:similar,estimapv))/similar)
  }
return(niveles)
}

similar<-1000
miu<-seq(-2.5,2.5,by=.1)
tmuestra<-c(15,30,100)

quantil<- .25
write.table(sapply(tmuestra,alfas),file = "c:/FUM/2LNSEMq25.csv", sep = ",",
col.names =NA,qmethod = "double")

quantil<- .5
write.table(sapply(tmuestra,alfas),file = "c:/FUM/2LNSEMq5.csv", sep = ",",
col.names =NA,qmethod = "double")

quantil<- .75
write.table(sapply(tmuestra,alfas),file = "c:/FUM/2LNSEMq75.csv", sep = ",",
col.names =NA,qmethod = "double")

```

Prueba Logrank

ALGORITMO PARA LA PRUEBA LOGRANK

```
alfaslr<-function(tmuestra)
{
niveles<-c(1:length(miu))
k<-0
for (i in miu)
  {
  k<-k+1
  estimapv <- function(j)
  {
  xrand<-rlnorm(tmuestra,i,1)           # Muestra 1, LN(mu,1)
  censurax<-qlnorm(censura,i,1)
  xn<-xrand[which(xrand<=censurax)]
  xd<-xrand[which(xrand>censurax)]
  x<-c(rep(censurax,length(xn)),xd)
  censor1<-c(rep(2,length(xn)),rep(1,length(xd)))

  yrand<-rlnorm(tmuestra)              # Muestra 2, LN(0,1)
  censuray<-qlnorm(censura,i,1)
  yn<-yrand[which(yrand<=censuray)]
  yd<-yrand[which(yrand>censuray)]
  y<-c(rep(censuray,length(yn)),yd)
  censor2<-c(rep(2,length(yn)),rep(1,length(yd)))

  concentras<-c(x,y)                   # Prueba LOG-RANK
  censor<-c(censor1,censor2)
  tratam<-c(rep(1,length(x)),rep(2,length(y)))
  compara<-survdif(Surv(concentras,censor) ~ tratam)

  chicuad<-compara$chisq
  return(chicuad)
  }
  out <- sapply(1:simula,estimapv)
  niveles[k]<-1-(length(which(out<qchisq(0.95,1)))/simula)
  }
return(niveles)
}

simula<-1000
miu<-seq(-2.5,2.5,by=.1)
tmuestra<-c(15,30,100)

censura<- 0.25
write.table(sapply(tmuestra,alfaslr), file = "c:/FUM/LgRnk2MstrsLnsq25.csv",
sep = ",", col.names =NA,qmethod = "double")

censura<- 0.5
write.table(sapply(tmuestra,alfaslr), file = "c:/FUM/LgRnk2MstrsLnsq50.csv",
sep = ",", col.names =NA,qmethod = "double")

censura<- 0.75
write.table(sapply(tmuestra,alfaslr), file = "c:/FUM/LgRnk2MstrsLnsq75.csv",
sep = ",", col.names =NA,qmethod = "double")
```

B.1.2. Prueba Propuesta Comparando Poblaciones Exponenciales

ALGORITMO EM

```
fillconEM<-function(wlns, censuraw)
{
  wnd<-rep(censuraw, length(wlns[which(wlns<=censuraw)]))
  wobs<-wlns[which(wlns>censuraw)]
  w<--log(c(wobs, wnd))
  censura<-c(rep(0, length(wobs)), rep(1, length(wnd)))
  regre<-lm(w ~ 1)
  lw<-length(w)
  b<-coef(regre)
  sigma<-sqrt(deviance(regre)/(lw-1))
  sigma1<-0
  b1<-0
  eps<-1
  iter<-0
  while(eps>0.005)
  {
    media<-b
    z<-(w-media)/sigma
    sesgo<-sigma*dnorm(z)/(1-pnorm(z))
    wnuevo<-(media+sesgo)*censura+w*(1-censura)
    regre<-lm(wnuevo ~ 1)
    b<-coef(regre)
    suma<-sum((z*sesgo*sigma+sigma^2-sesgo^2)*censura)
    sigma<-sqrt((deviance(regre)+suma)/lw)
    eps<-max(abs(b-b1), abs(sigma-sigma1))
    b1<-b
    sigma1<-sigma
    iter<-iter+1
  }
  return(exp(-wnuevo))
}
```

FUNCION DE VEROSIMILITUD COMPLETA PARA AMBAS MUESTRAS EXPONENCIALES

```
vero<-function(beta)
{
  zx<-(log(x)-(beta[1]+beta[3]))/beta[2]
  zy<-(log(y)-beta[1])/beta[2]

  logL<-sum(log((1/beta)*dnorm(zx)))+sum(log((1/beta[2])*dnorm(zy)))
  return(-1*logL)
}
```

FUNCION PARA CALCULAR P-VALORES

```

alfas<-function(tmuestra)
{
niveles<-c(1:length(miu))
k<-0
for (i in miu)
  {
  k<-k+1
  estimapv<-function(j)
  {
  xlms<-rexp(tmuestra,i)
  ylms<-rexp(tmuestra)

  x<<-fillconEM(xlms,qexp(quantil,i))
  y<<-fillconEM(ylms,qexp(quantil))

  bg<-optim(c(0,0,i),vero, hessian=TRUE)
  es<-sqrt(diag(ginv(bg$hessian)))

  return(iffelse((LI[2]<0&0<LS[2]),1,0))
  }
  niveles[k]<-1-(sum(sapply(1:similar,estimapv))/similar)
  }
return(niveles)
}

similar<-1000
miu<-seq(-2.5,2.5,by=.1)
tmuestra<-c(15,30,100)

quantil<- .25
write.table(sapply(tmuestra,alfas),file = "c:/FUM/2ExpsEMq25.csv", sep = ",",
col.names =NA,qmethod = "double")

quantil<- .5
write.table(sapply(tmuestra,alfas),file = "c:/FUM/2ExpsEMq5.csv", sep = ",",
col.names =NA,qmethod = "double")

quantil<- .75
write.table(sapply(tmuestra,alfas),file = "c:/FUM/2ExpsEMq75.csv", sep = ",",
col.names =NA,qmethod = "double")

```

B.1.3. Prueba Propuesta Comparando Poblaciones Gumbel

ALGORITMO EM

```
fillconEM<-function(wlns, censuraw)
{
  wnd<-rep(censuraw, length(wlns[which(wlns<=censuraw)]))
  wobs<-wlns[which(wlns>censuraw)]
  w<--log(c(wobs, wnd))
  censura<-c(rep(0, length(wobs)), rep(1, length(wnd)))
  regre<-lm(w ~ 1)
  lw<-length(w)
  b<-coef(regre)
  sigma<-sqrt(deviance(regre)/(lw-1))
  sigma1<-0
  b1<-0
  eps<-1
  iter<-0
  while(eps>0.005)
  {
    media<-b
    z<-(w-media)/sigma
    sesgo<-sigma*dnorm(z)/(1-pnorm(z))
    wnuevo<-(media+sesgo)*censura+w*(1-censura)
    regre<-lm(wnuevo ~ 1)
    b<-coef(regre)
    suma<-sum((z*sesgo*sigma+sigma^2-sesgo^2)*censura)
    sigma<-sqrt((deviance(regre)+suma)/lw)
    eps<-max(abs(b-b1), abs(sigma-sigma1))
    b1<-b
    sigma1<-sigma
    iter<-iter+1
  }
  return(exp(-wnuevo))
}
```

FUNCION DE VEROSIMILITUD COMPLETA PARA AMBAS MUESTRAS GUMBEL

```
vero<-function(beta)
{
  zx<-(log(x)-(beta[1]+beta[3]))/beta[2]
  zy<-(log(y)-beta[1])/beta[2]

  logL<-sum(log((1/beta)*dnorm(zx)))+sum(log((1/beta[2])*dnorm(zy)))
  return(-1*logL)
}
```

FUNCION PARA CALCULAR P-VALORES

```
alfas<-function(tmuestra)
```

```

{
niveles<-c(1:length(miu))
k<-0
for (i in miu)
  {
  k<-k+1
  estimapv<-function(j)
  {
  xrand<-rgumbel(tmuestra,i,1)
  yrand<-rgumbel(tmuestra,1,1)

  x<<-fillconEM(xlns, qgumbel(quantil,i,1))
  y<<-fillconEM(ylns, qgumbel(quantil,1,1))

  bg<-optim(c(0,0,i),vero, hessian=TRUE)
  es<-sqrt(diag(ginv(bg$hessian)))

  return(iffelse((LI[2]<0&0<LS[2]),1,0))
  }
  niveles[k]<-1-(sum(sapply(1:similar,estimapv))/similar)
  }
return(niveles)
}

similar<-1000
tmuestra<-c(15,30,100)
location<-seq(.5,1.5,by=.1)

quantil<-.25
write.table(sapply(tmuestra,alfas), file = "c:/FUM/2GumbsEMq25.csv", sep =
",", col.names =NA,qmethod = "double")

quantil<-.50
write.table(sapply(tmuestra,alfas), file = "c:/FUM/2GumbsEMq50.csv", sep =
",", col.names =NA,qmethod = "double")

quantil<-.75
write.table(sapply(tmuestra,alfas), file = "c:/FUM/2GumbsEMq75.csv", sep =
",", col.names =NA,qmethod = "double")

```

B.2 Ejemplos en Aplicación de los Métodos

B.2.1 Una Muestra Aleatoria Lognormal

Optimización Directa de la Verosimilitud

```
# Observaciones exactas de Arsénico
(x<-c(1.7,3.2,2.8,0.7,0.9,0.5,0.5,0.5,0.7,0.6,1.5))

# Observaciones no detectadas
(xnd<-c(1,1,1,1,2,2,2,2,2,2,2,2,.9))

# Función de verosimilitud con datos censurados por la izquierda
verocen<-function(beta){
  -sum(dnorm(log(x),beta[1],beta[2],log=TRUE)
  + pnorm(log(xnd),beta[1],beta[2],log.p=TRUE))
}

# Estimación mediante optimización (optim) de la verosimilitud
(bg<-optim(c(-.25,.5),verocen,hessian=TRUE)$par)
# -0.1530665  0.6772659

# Matriz de información observada de Fisher, estimador de  $\Sigma$ 
(ginv(bg$hessian))
#           [,1]           [,2]
# [1,]  0.028524105 -0.007510195
# [2,] -0.007510195  0.016077341

# Errores estándar estimados para  $\mu$  y  $\sigma$ ,
(es<-sqrt(diag(ginv(bg$hessian))))
# 0.1689479 0.1268487

# Intervalo aproximado del 95% de confianza para  $\mu$  y  $\sigma$ 
(LIySmu<- c(bg$par[1]-1.96*es[1], bg$par[1]+1.96*es[1]))
# -0.4840925  0.1779595

(LIySsigma<- c(bg$par[2]/exp(1.96*es[2]/bg$par[2]),
               bg$par[2]*exp(1.96*es[2]/bg$par[2])))
# 0.4691700 0.9776607

# Logverosimilitud estimada
(verocen(bg$par))
# - 18.37448

# Estimador de la concentración media
Media<-exp(bg$par[1]+.5*( bg$par[2])^2)
# 1.079263
```

Optimización de la Verosimilitud Vía EM

```
# Algoritmo EM

em.fill<-function(w,censura,mes)
{
  regre<-lm(w ~ 1)
  lw<-length(w)
  b<-coef(regre)
  sigma<-sqrt(deviance(regre)/(lw-1))
  sigma1<-0
  b1<-0
  eps<-1
  while(eps>0.0001)
  {
    media<-predict(regre,type="response")
    z<-(w-media)/sigma
    sesgo<-sigma*dnorm(z)/(1-pnorm(z))
    wnuevo<-(media+sesgo)*censura+w*(1-censura)
    regre<-lm(wnuevo ~ 1)
    b<-coef(regre)
    suma<-sum((z*sesgo*sigma+sigma^2-sesgo^2)*censura)
    sigma<-sqrt((deviance(regre)+suma)/(lw))
    eps<-max(abs(mean(b)-mean(b1)),abs(sigma-sigma1))
    b1<-b
    sigma1<-sigma
  }
  return(wnuevo)
}

# Concentraciones de arsénico

x<-c(1,1,1.7,1,1,2,3.2,2,2,2.8,2,2,2,2,2,0.7,0.9,0.5,0.5,0.9,0.5,
      0.7,0.6,1.5)

# Indicador de Censura

censorx<-c(1,1,0,1,1,1,0,1,1,0,1,1,1,1,1,0,0,0,0,1,0,0,0,0)

# Datos aumentados o artificiales

( datart<-exp( -( em.fill( -log(x),censorx) ) ) )

# Estimacion preliminar para  $\mu$  via glm

summary(rg<-glm(log(datart) ~ 1))

#           Estimate Std. Error t value Pr(>|t|)
# (Intercept)  -0.2527      0.1052  -2.402  0.0248 *

# Estimacion preliminar para  $\sigma$ 

(sigma<-sqrt(deviance(rg)/(length(datart)-1)))
# 0.5153089
```

```

# Logverosimilitud estimada

logLik(rg)
# -17.63208

# Estimación final para  $\mu$  y  $\sigma$ , vía optim*

veroEM<-function(beta) {
  -sum( dnorm(log(datart),beta[1],beta[2],log=TRUE) )}

(bg<-optim(c(-.2,.5),veroEM,hessian=TRUE)$par)
# [1] -0.2526516  0.5153089

# Matriz de información observada de Fisher, estimador de  $\Sigma$ 

(ginv(bg$hessian))
#           [,1]           [,2]
# [1,] 1.106704e-02 1.132199e-06
# [2,] 1.132199e-06 5.299909e-03

# Errores estándar estimados para  $\mu$  y  $\sigma$ 

(es<-sqrt(diag(ginv(bg$hessian))))
# 0.10526626 0.07280048

# Intervalo aproximado del 95% de confianza para  $\mu$  y  $\sigma$ 

(LIyS<- c(rg$coef[1]-1.96*0.1052,rg$coef[1]+1.96*0.1052))
# -0.45887748 -0.04649348

(LIySsigma<- c(sigma/exp(1.96*es[2]/sigma),sigma*exp(1.96*es[2]/sigma)))
# 0.3906714 0.6797101

# Estimador de la concentración media

Media<-exp(bg$par[1]+.5*( bg$par[2])^2)
# 0.8870269

```

B.2.2. Ejemplo 2. Una Muestra Con Una Covariable Continua

Optimización Directa de la Verosimilitud

```

# Dosis exactas de radiación gamma y año de registro
x<-c(52,110,16,16,46,99,93,103,64,36,35,15,60,29,75,2,53,53,89,15,
     56,44,23,4,4,3,4,5,23)

aniox<-c(1956,1957,1957,1958,1958,1958,1958,1959,1959,1959,1959,
        1960,1960,1960,1960,1961,1961,1961,1961,1962,1962,1962,1962,
        1963,1963,1965,1965,1965,1965)

# Dosis no detectadas y año de registro
xnd<-c(30,30,30,30,30,30,30,30,30,30,30,30)

anioxnd<-c(1956,1956,1956,1957,1957,1963,1963,1964,1964,1964,1964)

# Función de verosimilitud para datos censurados por la izquierda
verocen<-function(beta){
  -sum(dnorm(log(x),beta[1] + beta[3]*(aniox-1961),beta[2],log=TRUE)
      + pnorm(log(xnd),beta[1]+beta[3]*(anioxnd-1961),beta[2],log.p=TRUE))
}

# Estimación mediante optimización (optim) de la verosimilitud
(bg<-optim(c(3,1,-.17),verocen,hessian=TRUE))
# 2.6807039 1.0458352 -0.1147044

# Matriz de información observada de Fisher, estimador de  $\Sigma$ 
(es<-ginv(bg$hessian))

# 0.031874720 -0.0086718245 0.0024617474
# -0.008671824 0.0186985449 0.0005428727
# 0.002461747 0.0005428727 0.0029063766

# Errores estándar estimados para  $\beta_0$ ,  $\alpha$  y  $\sigma$ .
(es<-sqrt(diag(ginv(bg$hessian))))
# 0.17853493 0.13674262 0.05391082

# Intervalo aproximado del 95% de confianza para  $\beta_0$ ,  $\alpha$  y  $\sigma$ 
(LIbetayalfa<- c(bg$par[1]-1.96*es[1],bg$par[3]-1.96*es[3]))
# 2.3309193 -0.2203432

(LSbetayalfa<- c(bg$par[1]+1.96*es[1],bg$par[3]+1.96*es[3]))
# 3.030776220 -0.009012763

(LIySsigma<- c(bg$par[2]/exp(1.96*es[2]/ bg$par[2]),
              bg$par[2]*exp(1.96*es[2]/ bg$par[2])))
# 0.8094069 1.3513244

# Logverosimilitud estimada

```



```
(verocen(bg$par))  
# - 56.31422
```

Optimización de la Verosimilitud Vía EM

```
# Algoritmo EM  
  
em.fill<-function(w,censura,mes)  
{  
  regre<-lm(w ~ mes)  
  lw<-length(w)  
  b<-coef(regre)  
  sigma<-sqrt(deviance(regre)/(lw-2))  
  signal<-0  
  b1<-0  
  eps<-1  
  while(eps>0.0001)  
  {  
    media<-predict(regre,type="response")  
    z<-(w-media)/sigma  
    sesgo<-sigma*dnorm(z)/(1-pnorm(z))  
    wnuevo<-(media+sesgo)*censura+w*(1-censura)  
    regre<-lm(wnuevo ~ mes)  
    b<-coef(regre)  
    suma<-sum((z*sesgo*sigma+sigma^2-sesgo^2)*censura)  
    sigma<-sqrt((deviance(regre)+suma)/(lw))  
    eps<-max(abs(mean(b)-mean(b1)),abs(sigma-signal))  
    b1<-b  
    signal<-sigma  
  }  
  return(wnuevo)}  
  
# Dosis de radiación gamma  
  
x<-c(30,30,30,52,110,16,30,30,16,46,99,93,103,64,36,35,15,60,29,  
      75,2,53,53,89,15,56,44,23,30,30,4,4,30,30,30,30,3,4,5,23)  
  
# Años para las dosis  
  
aniox<-c(1956,1956,1956,1956,1957,1957,1957,1957,1958,1958,1958,  
          1958,1959,1959,1959,1959,1960,1960,1960,1960,1961,1961,1961,  
          1961,1962,1962,1962,1962,1963,1963,1963,1963,1964,1964,1964,  
          1964,1965,1965,1965,1965)  
  
# Indicador de Censura  
  
censorx<-c(1,1,1,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,  
            0,1,1,0,0,1,1,1,1,0,0,0,0)  
  
# Datos aumentados o artificiales  
  
( datart<-exp( -( em.fill( -log(x),censorx, aniox-1961 ) ) ) )  
  
# Estimacion de  $\beta_0$  y  $\alpha$  via glm
```

```

summary(rg<-glm(log(datart) ~ aniox-1961))

#           Estimate Std. Error t value Pr(>|t|)
# (Intercept)  2.92180   0.15283  19.118 < 2e-16 ***
# aniox-1961  -0.19407   0.05242  -3.702 0.000676 ***

# Estimacion de  $\sigma$ 

(sigma<-sqrt(deviance(rg)/(length(datart)-2)))
# 0.9522587

# Logverosimilitud estimada

logLik(rg)
# -53.77493

# Verosimilitud para la estimación de covarianzas de  $\beta_0$ ,  $\alpha$ ,  $\sigma$  y  $\Sigma$ 

veroEM<-function(beta) {
  -sum(dnorm(log(datart),beta[1]+beta[3]*mesx,beta[2],log=TRUE))}

(bg<-optim(c(2,1,-.2),veroEM,hessian=TRUE))
# 2.921799,0.9522587,-0.19407

(ginv(bg$hessian))

# 2.335994e-02 8.471865e-06 1.853514e-10
# 8.471865e-06 1.209309e-02 2.594377e-07
# 1.853514e-10 2.594377e-07 2.748447e-03

# Errores estándar estimados

(es<-sqrt(diag(ginv(bg$hessian))))
# 1.528396e-01 1.099686e-01 5.242563e-02

# Intervalo aproximado del 95% de confianza para  $\beta_0$  y  $\alpha$  y  $\sigma$ 

(LI<- c(rg$coef[1]-1.96*0.15283,rg$coef[2]-1.96*0.05242))
# 2.6222522 -0.2968161

(LS<- c(rg$coef[1]+1.96*0.15283,rg$coef[2]+1.96*0.05242))
# 3.2213458 -0.0913297

(LIySsigma<- c(sigma/exp(1.96*es[2]/sigma),sigma*exp(1.96*es[2]/sigma))
# 0.7593724 1.1941396

```

B.2.3 Ejemplo 3. Dos Poblaciones Lognormales Independientes

Modelo 1. Homogeneidad de σ


```

# - 168.516

# Estimadores para mu1 , mu2 y las concentraciones medias de ambas zonas.

(mu1ymu2<- c(bg$par[1]+bg$par[3], bg$par[1]))
# 0.6614678 0.60558

(E1<-exp(bg$par[1]+bg$par[3]+0.5* bg$par[2]^2))
# 2.984142

(E2<-exp(bg$par[1]+0.5* bg$par[2]^2))
# 2.821939

```

Inferencia Mediante Optimización Vía EM de la Verosimilitud

```

# Algoritmo EM

em.fill<-function(w, censura, muestra)
{
  regre<-lm(w ~ muestra)
  lw<-length(w)
  b<-coef(regre)
  sigma<-sqrt(deviance(regre)/(lw-2))
  sigma1<-0
  b1<-0
  eps<-1
  while(eps>0.0001)
  {
    media<-predict(regre, type="response")
    z<-(w-media)/sigma
    sesgo<-sigma*dnorm(z)/(1-pnorm(z))
    wnuevo<-(media+sesgo)*censura+w*(1-censura)
    regre<-lm(wnuevo ~ muestra)
    b<-coef(regre)
    suma<-sum((z*sesgo*sigma+sigma^2-sesgo^2)*censura)
    sigma<-sqrt((deviance(regre)+suma)/(lw))
    eps<-max(abs(mean(b)-mean(b1)), abs(sigma-sigma1))
    b1<-b
    sigma1<-sigma
  }
  return(wnuevo)
}

# Vector de muestras de concentraciones de cobre para ambas zonas
muestra<-c(rep(1, length(c(x, xn))), rep(0, length(c(y, yn))))

# Vector indicador de censura para los datos de ambas muestras
censura<-c(rep(0, length(x)), rep(1, length(xn)),
           rep(0, length(y)), rep(1, length(yn)))

# Datos aumentados o artificiales

```

```

(datart<-exp(-(em.fill(-log(c(xobs,xnd,yobs,ynd)),censura,muestra))))

# Estimacion preliminar para  $\beta_0$   $\beta_1$  y  $\sigma$ , via glm

(rg<-glm(log(datart) ~ muestra))

#           Estimate Std. Error t value Pr(>|t|)
# (Intercept)   1.0496    0.1143   9.182 2.54e-15 ***
# muestra        -0.1162    0.1514  -0.768   0.444

(sigma<-sqrt(deviance(rg)/(length(datart)-2)))
# 0.8001688

# Logverosimilitud estimada

logLik(rg)
# -135.3358

# Verosimilitud para la estimación de covarianzas de  $\beta_0$ ,  $\beta_1$  y  $\sigma$ 

veroEM<-function(beta) {
  -sum(dnorm(log(datart[1:65]),beta[1]+beta[3],beta[2],log=TRUE)
  +dnorm(log(datart[66:114]),beta[1],beta[2],log=TRUE))
}

# Errores estándar estimados

(bg<-optim(c(1,1,-.1),veroEM,hessian=TRUE))
(es<-sqrt(diag(ginv(bg$hessian))))
# 0.11432248 0.06543531 0.15143338

# Intervalo aproximado del 95% de confianza para  $\beta_0$  y  $\beta_1$  y  $\sigma$ 

(LI<- c(rg$coef[1]-1.96*.1143,rg$coef[2]-1.96*.1514))
# 0.8256135 -0.4129497

(LS<- c(rg$coef[1]+1.96*.1143,rg$coef[2]+1.96*.1514))
# 1.2736695 0.1805383

# (LIySsigma<-c(sigma/exp(1.96*es[2]/sigma),sigma*exp(1.96*es[2]/sigma)))
# 0.6816661 0.9392723

# Estimadores para los parámetros  $\mu_1$  y  $\mu_2$ 

(mulymu2<- c(rg$coef[1]+rg$coef[2], rg$coef[1]))
# 0.9334358 1.049642

# Estimadores para las concentraciones medias de ambas zonas.

(E1<-exp(bg$par[1]+bg$par[3]+0.5* sigma^2))
# 3.502829

(E2<-exp(bg$par[1]+0.5* sigma^2))
# 3.934473

```



```

# 0.6756477 0.59187

# Estimadores para los parámetros  $\sigma_1$  y  $\sigma_2$ 

(sigma1ysigma2<- c(exp(bg$par[2]), exp(bg$par[2]+ bg$par[4])))
# 0.9619128 0.8951652

```

Inferencia Mediante Optimización Vía EM de la Verosimilitud

```

# Algoritmo EM

em.fill<-function(w, censura)
{
  regre<-lm(w ~ 1)
  lw<-length(w)
  b<-coef(regre)
  sigma<-sqrt(deviance(regre)/(lw-1))
  sigma1<-0
  b1<-0
  eps<-1
  while(eps>0.0001)
  {
    media<-predict(regre,type="response")
    z<-(w-media)/sigma
    sesgo<-sigma*dnorm(z)/(1-pnorm(z))
    wnuevo<-(media+sesgo)*censura+w*(1-censura)
    regre<-lm(wnuevo ~ 1)
    b<-coef(regre)
    suma<-sum((z*sesgo*sigma+sigma^2-sesgo^2)*censura)
    sigma<-sqrt((deviance(regre)+suma)/(lw))
    eps<-max(abs(mean(b)-mean(b1)),abs(sigma-sigma1))
    b1<-b
    sigma1<-sigma
  }
  return(wnuevo)
}

# Vectores de concentraciones de cobre para cada zonas

x0<-c(x,xn); y0<-c(y,yn)

# Vectores indicadores de censura para los datos de ambas muestras

censurax<-c(rep(0,length(x)),rep(1,length(xn)))
censuray<-c(rep(0,length(y)),rep(1,length(yn)))

# Datos aumentados o artificiales

(datartx<-exp(-(em.fill(-log(x0),censurax))))
(datarty<-exp(-(em.fill(-log(y0),censuray))))

# Verosimilitud para la inferencia sobre  $\beta_0$ ,  $\gamma_0$ ,  $\beta_1$ , y  $\gamma_1$ .

veroEM<-function(beta){

```

```

    -sum(dnorm(log(datartx),beta[1]+beta[3],
    exp(beta[2]+beta[4]),log=TRUE)
    + dnorm(log(datarty),beta[1],exp(beta[2]),log=TRUE))
  }

Estimadores y errores estándar para  $\beta_0$ ,  $\gamma_0$ ,  $\beta_1$ , y  $\gamma_1$ .

(bg<-optim(c(0.592,-0.0388,0.084,-0.072),veroEM, hessian=TRUE))
# 0.90581762 -0.20211396 0.03845537 -0.10161051

(es<-sqrt(diag(ginv(bg$hessian))))
# 0.10133661 0.08771683 0.13656392 0.12404217

# Matriz de información observada de Fisher, estimador de  $\Sigma$ 

(ginv(bg$hessian))
#           [,1]           [,2]           [,3]           [,4]
# [1,] 1.026911e-02 -2.236751e-06 -1.026911e-02 2.236750e-06
# [2,] -2.236751e-06 7.694241e-03 2.236751e-06 -7.694241e-03
# [3,] -1.026911e-02 2.236751e-06 1.864970e-02 -1.424695e-06
# [4,] 2.236750e-06 -7.694241e-03 -1.424695e-06 1.538646e-02

# Intervalo aproximado del 95% de confianza para  $\beta_0$ ,  $\gamma_0$ ,  $\beta_1$ , y  $\gamma_1$ .

(LI<- bg$par-1.96*es)
# 0.7071979 -0.3740389 -0.2292099 -0.3447332

(LS<- bg$par+1.96*es)
# 1.10443737 -0.03018898 0.30612066 0.14151216

# Logverosimilitud estimada

(veroEM(bg$par))
# - 151.5746

# Estimadores para los parámetros  $\mu_1$  y  $\mu_2$ 

(mulymu2<- c(bg$par[1]+bg$par[3], bg$par[1]))
# 0.944273 0.9058176

# Estimadores para los parámetros  $\sigma_1$  y  $\sigma_2$ 

(sigma1<- c(exp(bg$par[2]), exp(bg$par[2]+ bg$par[4])))
# 0.8170018 0.7380642

```

B.2.4 Ejemplo 4. Comparación y Análisis de Tendencia

Optimización Directa de la Verosimilitud

```

# Niveles preremediales exactos y fecha de registro.
x<-c(4.3,16,6.1,2.66,3.0,4.42,5.74,1.40,1.49,2.3)

```



```

mesx<-c(1,2,3,6,7,8,9,10,11,12)

# Niveles preremediales no detectados y fecha de registro.
xn<-c(1,1); mesxn<-c(4,5)

# Niveles postremediales exactos y fecha de registro.
y<-c(1.5,1.3,2.1,1.1,1.8,1.2,1.1); mesy<-c(2,3,6,7,9,10,12)

# Niveles postremediales no detectados y fecha de registro.
yn<-c(1,1,1,1,1); mesyn<-c(1,4,5,8,11)

# Función de verosimilitud para las dos muestras censuradas
verocen<-function(beta) {
  -sum(dnorm(log(x),beta[1]+beta[3]+beta[4]*mesx,beta[2],log=TRUE)
  + pnorm(log(xn),beta[1]+beta[3]+beta[4]*mesxn,beta[2],log.p=TRUE)
  + dnorm(log(y),beta[1]+beta[4]*mesy,beta[2],log=TRUE)
  + pnorm(log(yn),beta[1]+beta[4]*mesyn,beta[2],log.p=TRUE))
}

# Estimación directa de la verosimilitud para estimar  $\beta_0$ ,  $\sigma$ ,  $\beta_1$  y  $\alpha_0$ 
(bg<-optim(c(0.18,0.74,1.07,-0.03),verocen,hessian=TRUE))
# -0.37788268 1.02208927 0.55874869 0.01775412

# Errores estándar estimados para  $\beta_0$ ,  $\sigma$ ,  $\beta_1$  y  $\alpha_0$ .
(es<-sqrt(diag(ginv(bg$hessian))))
# 0.4394422 0.1807269 0.3581816 0.0560013

# Intervalo aproximado del 95% de confianza para  $\beta_0$ ,  $\beta_1$ ,  $\alpha_0$  y  $\sigma$ .
(LI<- bg$par[-2]-1.96*es[-2])
# -1.23918932 -0.14328730 -0.09200844

(LS<- bg$par[-2]+1.96*es[-2])
# 0.4834240 1.2607847 0.1275167

# (LIySsigma<- c(bg$par[2]/exp(1.96*es[2]/bg$par[2]),
  bg$par[2]*exp(1.96*es[2]/bg$par[2])))
# 0.7227294 1.4454463

# Logverosimilitud estimada
(verocen(bg$par))
# - 42.62499

# Estimadores para los parámetros  $\mu_1$ ,  $\mu_2$  y  $\sigma$ 
(mulymu2<- c(bg$par[1]+bg$par[3],bg$par[1]))
# 0.180866 -0.3778827

```

```
(sigma<- bg$par[2])
# 1.022089
```

Optimización de la Verosimilitud Vía EM

```
# Algoritmo EM
```

```
em.fill<-function(w,censura,muestra,mes)
{
  regre<-lm(w ~ muestra + mes)
  lw<-length(w)
  b<-coef(regre)
  sigma<-sqrt(deviance(regre)/(lw-3))
  sigma1<-0
  b1<-0
  eps<-1
  while(eps>0.0001)
  {
    media<-predict(regre,type="response")
    z<-(w-media)/sigma
    sesgo<-sigma*dnorm(z)/(1-pnorm(z))
    wnuevo<-(media+sesgo)*censura+w*(1-censura)
    regre<-lm(wnuevo ~ muestra + mes)
    b<-coef(regre)
    suma<-sum((z*sesgo*sigma+sigma^2-sesgo^2)*censura)
    sigma<-sqrt((deviance(regre)+suma)/(lw))
    eps<-max(abs(mean(b)-mean(b1)),abs(sigma-sigma1))
    b1<-b
    sigma1<-sigma
  }
  return(wnuevo)
}
```

```
# Vector de concentraciones de para ambas muestras censuradas
```

```
muestra<-c(rep(1,length(x)),rep(0,length(y)))
```

```
# Vector indicador de censura para los datos
```

```
censura<-c(censorex,censory)
```

```
# Vector de fechas de registros de las concentraciones
```

```
mes<-c(mesx,mesy)
```

```
# Datos aumentados o artificiales
```

```
(datart<-exp(-(em.fill(-log(c(x,y)),censura,muestra,mes))))
```

```
# Estimacion  $\beta_0$ ,  $\beta_1$ ,  $\alpha_0$  y  $\sigma$ , vía glm
```

```
summary(rg<-glm(log(datart) ~ muestra + mes))
```

```
#           Estimate Std. Error t value Pr(>|t|)
# (Intercept)  0.18851    0.35826    0.526    0.604
```

```

# muestra      1.07324    0.30427    3.527    0.002 **
# mes         -0.03672    0.04407   -0.833    0.414

(sigma<-sqrt(deviance(rg)/(length(datart)-3)))
# 0.7452989

logLik(rg)
# -25.39687

# Verosimilitud para la estimación de covarianzas de  $\beta_0$ ,  $\sigma$ ,  $\beta_0$  y  $\alpha_0$ 

# veroEM<-function(beta) {
  -sum(dnorm(log(datart[1:12]),beta[1]+beta[3]
  +beta[4]*mesx,beta[2],log=TRUE)
  +dnorm(log(datart[13:24]),beta[1]+beta[4]*mesy,beta[2],log=TRUE))
}

# Errores estándar estimados

(bg<-optim(c(0.18851,0.7452989,1.07324,-0.03672),veroEM, hessian=TRUE))

(es<-sqrt(diag(ginv(bg$hessian))))
# 0.3351183 0.1006263 0.2846162 0.0412242

# Intervalo aproximado del 95% de confianza para  $\beta_0$ ,  $\beta_1$ ,  $\alpha_0$  y  $\sigma$ 

(LI<- c(rg$coef[1]-1.96*0.35826, rg$coef[2]-1.96*0.30427,
  rg$coef[3]-1.96*0.04407))
# -0.5136818 0.4768719 -0.1230938

(LS<- c(rg$coef[1]+1.96*0.35826, rg$coef[2]+1.96*0.30427,
  rg$coef[3]+1.96*0.04407))
# 0.89069741 1.66961027 0.04966058

# (LIySsigma<-c(sigma/exp(1.96*es[2]/sigma),sigma*exp(1.96*es[2]/sigma)))
# 0.5720100 0.9710851

# Estimadores para los parámetros  $\mu_1$  y  $\mu_2$ 

(mu1ymu2<- c(rg$coef[1]+rg$coef[2], rg$coef[1]))
# 1.261749 0.1885078

```

B.2.5. Ejemplo 5. Dos Muestras con Datos Agrupados

```

# Algoritmo EM

em.fill<-function(w,censura,pozo,mes)
{
  regre<-lm(w ~ pozo + mes)
  lw<-length(w)
  b<-coef(regre)
  sigma<-sqrt(deviance(regre)/(lw-3))
}

```

```

sigma1<-0
b1<-0
eps<-1
while (eps>0.0001)
{
  media<-predict (regre, type="response")
  z<-(w-media)/sigma
  sesgo<-sigma*dnorm(z)/(1-pnorm(z))
  wnuevo<-(media+sesgo)*censura+w*(1-censura)
  regre<-lm(wnuevo ~ pozo + mes)
  b<-coef(regre)
  suma<-sum((z*sesgo*sigma+sigma^2-sesgo^2)*censura)
  sigma<-sqrt((deviance(regre)+suma)/lw)
  eps<-abs(sigma-sigma1)
  b1<-b
  sigma1<-sigma
}
return(wnuevo)
}

# Concentraciones de Tolueno en pozos del sitio en riesgo
BW<-c(5,7.5,5,5,6.4,5,5,5,5,5)
PozoBW<-c(1,1,1,1,1,2,2,2,2,2)
MesBW<-c(1,2,3,4,5,1,2,3,4,5)
CensBW<-c(1,0,1,1,0,1,1,1,1,1)

# Concentraciones de Tolueno en pozos del sitio control
CW<-c(5,12.5,8.0,5,11.2,5,13.7,15.3,20.2,25.1,5,20.1,35.0,28.2,19.0)
PozoCW<-c(1,1,1,1,1,2,2,2,2,2,3,3,3,3,3)
MesCW<-c(1,2,3,4,5,1,2,3,4,5,1,2,3,4,5)
CensCW<-c(1,0,0,1,0,1,0,0,0,0,1,0,0,0,0)

# Datos artificiales o aumentados
xBW<-exp(-(em.fill(-log(BW),CensBW,PozoBW,MesBW)))
yCW<-exp(-(em.fill(-log(CW),CensCW,PozoCW,MesCW)))

# Se crea un dataframe con la información anterior
Zona<-c(rep(1,length(xBW)),rep(0,length(yCW)))
Pozos<-c(PozoBW,PozoCW)
Tolueno<-c(log(xBW),log(yCW))
Mes<-c(MesBW,MesCW)

(datosBWyCW<-data.frame(Zona,Pozos,Tolueno,Mes))

```

```

# Se ajusta el modelo loglineal de efectos mixtos, agrupados por pozos

(res<-lme(Tolueno ~ Zona + Mes, random = ~ 1 | Pozos, data = datosBWyCW))

#      AIC      BIC    logLik
# 65.32737 70.78258 -27.66369

#Random effects:
# Formula: ~1 | Pozos
#      (Intercept) Residual
#StdDev:  0.1041758 0.6897608

#Fixed effects: Tolueno ~ Zona + Mes
#      Value Std.Error DF   t-value p-value
#(Intercept)  1.7918096 0.3478133 20   5.151642 0.0000
#Zona        -1.3532705 0.2844619 20  -4.757299 0.0001
#Mes         0.2103036 0.0975469 20   2.155923 0.0434

# Intervalo aproximado del 95% de confianza para  $\beta_0$ ,  $\beta_1$  y  $\alpha_0$ .

(LIySb0<- c(1.7918096 - 1.96*0.3478133, 1.7918096 + 1.96*0.3478133))
# 1.110096 2.473524

(LIySb1<- c(-1.3532705 - 1.96*0.2844619, -1.3532705 + 1.96*0.2844619))
# -1.910816 -0.7957252

(LIySd<- c(0.2103036 - 1.96*0.0975469, 0.2103036 + 1.96*0.0975469))
# 0.01911168 0.4014955

# Estimadores para los parámetros  $\mu_1$  y  $\mu_2$ 

mulymu2<- c(1.7918096 - 1.3532705, 1.7918096)
# 0.4385391 1.7918096

```

ANEXO C. ARTÍCULOS DE INVESTIGACIÓN

1. ANÁLISIS DE DATOS CENSURADOS PARA INGENIERÍA Y CIENCIAS BIOLÓGICAS

(Censored Data Analysis on Engineering and Biological Sciences)

REVISTA DE MATEMÁTICA: TEORÍA Y APLICACIONES

MathScinet & Zentralblath,

CENTRO DE INVESTIGACIÓN EN MATEMÁTICA PURA Y APLICADA,

CIMPA – Universidad de Costa Rica.

Año 2007, Vol. 14, Num. 2, Páginas: 237 – 248.

**2. Statistical Procedures to Compare Mean Pollutant Concentrations of
Lognormal Populations Containing Nondetects Data
ENVIRONMENTAL MODELING AND ASSESSMENT**

ISI-Thompson - Springer Research Journals,

Diciembre 2007, En Revisión.

3. Statistical Theory for Handling Nondetects Data of Lognormal Populations.

En Borrador

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 of the 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

$$L(\theta) = cP(DATA; \theta) = cf(x_1; \theta)f(x_2; \theta) \dots f(x_n; \theta). \quad (1)$$

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.

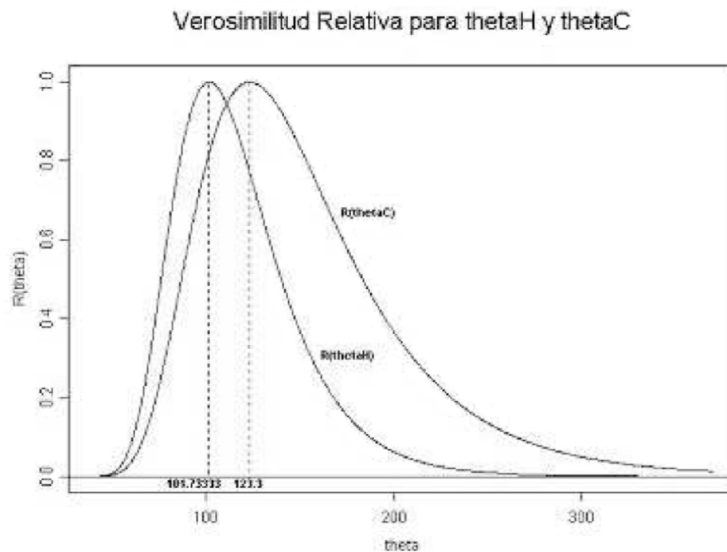


Figura 1: Verosimilitudes relativas y EMV para cada grupo.

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.

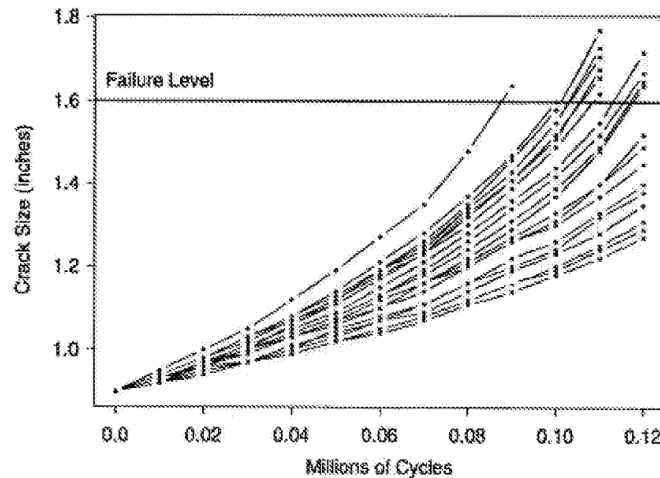


Figura 2: Trayectorias individuales de degradación.

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

$$L(\theta_\beta, \sigma_\epsilon | DATA) = \prod_{i=1}^{21} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \left[\prod_{j=1}^{n_i} \frac{1}{\sigma_\epsilon} \phi_{nor}(\zeta_{ij}) \right] f_\beta(\beta_{1i}, \beta_{2i}; \theta_\beta) d\beta_{1i}, \dots, d\beta_{2i}. \quad (7)$$

donde $\zeta_{ij} = \frac{[y_{ij} - a(t_{ij}, \beta_{1i}, \dots, \beta_{2i})]}{\sigma_\epsilon}$ 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].

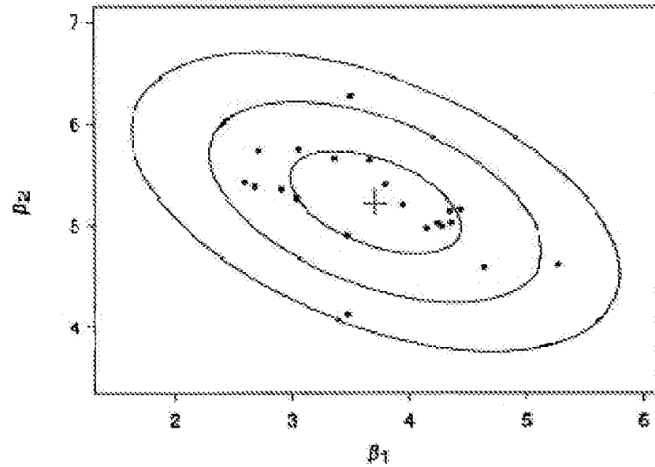


Figura 3: Contorno de la distribución normal multivariada para los parámetros β_1 y β_2 .

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

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

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,

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

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

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

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

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

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,

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

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:

$$\frac{\partial \log L_4(\mu_1, \mu_2, \sigma_1, \sigma_2)}{\partial \mu_i} = 0, \quad \frac{\partial \log L_4(\mu_1, \mu_2, \sigma_1, \sigma_2)}{\partial \sigma_i} = 0, \quad i = 1, 2. \quad (17)$$

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:

$$\begin{aligned} \chi_1^2 &= -2 [\ln L_1(\hat{\mu}, \hat{\sigma}) - \ln L_4(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2)] > \chi_{\alpha, 2}^2, \\ \chi_2^2 &= -2 [\ln L_2(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}) - \ln L_4(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2)] > \chi_{\alpha, 2}^2, \\ \chi_3^2 &= -2 [\ln L_1(\hat{\mu}, \hat{\sigma}) - \ln L_2(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma})] > \chi_{\alpha, 1}^2, \\ \chi_4^2 &= -2 [\ln L_3(\hat{\mu}, \hat{\sigma}_1, \hat{\sigma}_2) - \ln L_4(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2)] > \chi_{\alpha, 1}^2. \end{aligned} \quad (18)$$

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 mas 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
	<u>65</u>	40	1	<u>23</u>	<u>1</u>	40	3
		50	1		49	50	2
		<u>620</u>	<u>1</u>			60	2
			67			70	1
						<u>90</u>	<u>1</u>
							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.

Statistical procedures to compare mean pollutant concentrations of lognormal populations containing nondetects data

Fidel Ulín-Montejo^a, Humberto Vaquera–Huerta^a, Jorge D. Etchevers–Barra^b
and Sergio Pérez–Elizalde^a

^a Estadística. Campus Montecillo. Colegio de Postgraduados. 56230. Carretera México-Texcoco
km. 36.5. Montecillo, Estado de México, MEX.

E-mail: fidel@colpos.mx

^b Edafología. Campus Montecillo. Colegio de Postgraduados. 56230. Carretera México-Texcoco
km. 36.5. Montecillo, Estado de México, MEX.

jetchev@colpos.mx

Abstract Frequently, the comparison of populations containing pollutants is done by using nonparametric statistical methods. In practice, if a sample has nondetects data, these are omitted or substituted by arbitrary values. To this regard, the environmental regulation organisms need the risks to be characterized in terms of the mean concentration. From a parametric approach, this work takes on the problem of comparison of mean concentrations of lognormal populations using lineal regression models with indicator variables. The EM algorithm, likelihood function, and Wald's method are used to handle nondetects data, getting estimations, and building confidence regions. In addition, a lineal mixed effects models approach is shown to compare population means in presence of random effects. The proposed method was simple and efficient for two populations, but it can be extended. Three examples of environmental data are induced to show the use and performance of the proposed statistical procedures.

Key words: EM Algorithm, lognormal model, likelihood, Wald's method, mixed effects models, nondetects.

1. Introduction

Environmental statistics uses statistical theory and methods to solve relevant problems like the monitoring of the quality of water and groundwater, assessing zones on remediation process, and measuring the risk of contamination [15]. Large concentrations of environmental pollutants have adverse health effects, yet small concentrations may not be benign. To know the levels of environmental risk of these pollutants, the mentioned extraordinarily low levels of concentration can be identified and measured; sometimes the signal produced by the pollutant is too small for instrumentation to quantify or to discriminate, being thus reported as nondetects instead [8,10]. Such incomplete observations are named left-censored data. The censoring is a point of reference that indicates up to where it was possible to take a measurement of the variable of interest; in this case, the point of reference is the limit of detection, LD. On [7,14] some procedures to compare populations trough the median are shown, likewise on [5,16], where nondetects data are omitted or arbitrarily substituted; in both cases results and conclusions are limited. The EM algorithm [2,17], for maximum likelihood estimation from incomplete data, has been a very useful tool which has permeated modern statistics. It can be implemented computationally in statistical packages like R [22], and it can be modified according to problem context. On the other hand, in some cases, normality is assumed with a logarithmic transformation of data, and the methods based on normal distribution models are used to obtain inferences [7]. However, as it is discussed in [4], the results obtained must be reported as obtained measurements and not in the logarithmic transformed scale. For example, some environmental regulation and scientific organizations need the risks to be characterized in terms of the mean concentration of the pollutant. The pollutant concentration belongs to the class of extensive variables, since these are physically additive in a useful sense [1]. In spite of the form of the distribution, the mean concentration of the pollutant

has a physical interpretation, which cannot be said of a logarithmic transformation. In this sense, the environmental data can present characteristics that lead to very interesting inference problems. The goal is to obtain quantitative statements about unknown parameters, such as the approximate confidence intervals, which use all the parametric information contained in the sample. An estimator for maximum likelihood and its variance are exceptionally sufficient to specify or to reproduce the complete likelihood function [3]; so that, approximate confidence regions can be calculated with Wald's method [13].

The objective of this work is to propose a parametric alternative to compare pollutant mean concentrations of lognormal populations, making a parameterization on lognormal model parameters with an indicator regression model, in presence of nondetects data and covariates. A criterion of comparison is stated by means of approximate confidence regions and intervals. It first illustrates the procedure to compare independent populations. After that, populations containing a continuous covariate, and homogeneity and heterogeneity of the scale parameter σ models are considered; additionally, a way of selecting the best model using a likelihood-ratio test is presented here. Finally, with a mixed-effects model, the procedure is extended to population containing continuous and random covariates due to monitoring periods, sampling designs or grouped data. A modification of the EM algorithm is used to fit lognormal models, handle nondetects, obtaining variances of estimators and approximate regions of 95% confidence. The procedure is implemented in R and application examples are developed using data from environmental studies, showing the versatility and advantages of this parametric procedures.

2. Methodology

The main objective of likelihood inference is to fit models to data by entertaining model-parameter combinations for which the probability of the data is large. Likelihood methods provide general and versatile tools for fitting models to data. The methods can be applied with a wide variety of parametric models with censored data and explanatory variables (i.e., regression analysis). The large-sample likelihood theory guarantees that these methods yield the most accurate estimates. These properties are approximate in moderate and small sample sizes. Several studies have shown that likelihood methods generally perform as well as others available [13].

2.1 Likelihood function and likelihood-ratio test

The likelihood function is either equal to or approximately proportional to the probability of the data. Then, for a given set of data and specified probability model $F(y;\theta)$, the likelihood is viewed as a function of the unknown model parameters θ . The total likelihood for a left-censored sample, containing n independent observations, is

$$L(\theta) = C \prod_{i=1}^n L_i(\theta; data_i) = C \prod_{i=1}^n [f(y_i; \theta)]^{\delta_i} [F(y_i; \theta)]^{1-\delta_i} \quad (1)$$

where $L_i(\theta; data_i)$ is the probability of the observation i , $data_i$, is the data for observation i , $\delta_i = 1$ if y_i is an exact observation, $\delta_i = 0$ if y_i is a left-censored observation. C is a constant depending on the sampling inspection scheme but not on the parameters θ . The θ that maximizes $L(\theta)$ provides a maximum likelihood estimate (ML Estimator) of $F(y;\theta)$, it is denoted by $\hat{\theta}$.

Now, a likelihood-ratio significance test evaluates if a general model fit data better than a restricted model. Then, a restricted model fit data as well as a general model if, asymptotically,

$$-2\log\left[L_1(\theta_0)/L_2(\hat{\theta})\right] \sim \chi_v^2. \quad (2)$$

Where $L_2(\hat{\theta})$ is the maximum likelihood of the general model on $\hat{\theta}$, $L_1(\theta_0)$ is the maximum likelihood of restricted model on θ_0 , and χ_v^2 is a chi-squared random variable with v degrees of freedom. v is the difference between the dimension of $\hat{\theta}$ and the dimension of θ_0 .

2.2 EM algorithm

The EM algorithm [2,6], is a powerful tool for computing maximum likelihood estimates with incomplete data. ‘‘Incomplete’’ is a generic word that, according to the situation, can assume different meanings: missing values, unknown components, censored observations, latent variables, and so on. A brief description of the EM algorithm follows:

Let y denote the observed data and x the unknown data, θ the parameter of interest and $\ell_c(\theta; y, x)$ the complete-data log-likelihood, defined for all θ en a parameter space Ω . Starting with an initial parameter $\theta^{(0)} \in \Omega$, the EM algorithm repeats the following two steps until convergence.

- E-step: compute $\ell^{(j)}(\theta) = E_{x|y, \theta^{(j-1)}} [\ell_c(\theta; y, x)]$, where the expectation is taken with respect to the conditional distribution of the missing data x given the observed data y , and the current numerical value $\theta^{(j-1)}$; is used in evaluating the expected value.
- M-step: find $\theta^{(j)} \in \Omega$ that maximizes $\ell^{(j)}(\theta)$.

2.3 Information matrix via the EM algorithm

The algorithm EM does not generate estimators for the matrix of variances and covariances of the ML estimators; because of this it has been modified to solve this problem. A modification was done by [17], and according to [21] this modification is simple and very useful. [17] proved that, if $\ell(\theta; y)$ is the log-likelihood of the sample, then:

$$\frac{\partial^2 \ell(\theta; y)}{\partial \theta^2} = \left(\frac{\partial^2 \ell^{(j)}(\theta)}{\partial \theta^{(j)2}} + \frac{\partial^2 \ell^{(j)}(\theta)}{\partial \theta^{(j)} \partial \theta} \right)_{\theta^{(j)} = \theta}. \quad (3)$$

The approximate variance of $\hat{\theta}$ is calculated with:

$$Var(\hat{\theta}) \approx [\partial^2 \ell(\theta; y) / \partial \theta^2]^{-1}. \quad (4)$$

Thus making it possible to obtain approximate confidence regions and intervals.

2.4 Approximate confidence regions and intervals

The large-sample normal approximation for the distribution of ML estimators can be used to compute approximate confidence regions for θ . This is sometimes known as "Wald's method", also known as the normal-approximation method [13]. In particular, an approximate $100(1-\alpha)\%$ confidence region for $\theta = (\theta_1, \theta_2, \dots, \theta_r)$ is the set of all values of θ in the ellipsoid

$$W(\theta) = (\hat{\theta} - \theta) (\hat{\Sigma}_{\hat{\theta}})^{-1} (\hat{\theta} - \theta) \leq \chi_{(1-\alpha; r)}^2, \quad (5)$$

where r is the dimension of θ , $\hat{\Sigma}_{\hat{\theta}}$ is the local estimate of the covariance matrix of $\hat{\theta}$ obtained from equation (4). $\chi_{(1-\alpha; r)}^2$ is the $1-\alpha$ quantil of a χ_r^2 variable. Then, an approximate $100(1-\alpha)\%$ normal-approximation confidence interval for θ_i is obtained from the familiar formula

$$[\underline{\theta}_i, \tilde{\theta}_i] = \hat{\theta}_i \pm z_{(1-\alpha/2)} se_{\hat{\theta}_i}, \quad (6)$$

where $se_{\hat{\theta}_i}$ is the square root of ii^{th} entry in the estimate of (4), $z_{(1-\alpha/2)}$ is the $1-\alpha/2$ quantil of the normal standard variable. This normal-approximation confidence interval can be viewed as a quadratic approximation for the log profile likelihood of θ_i in $\hat{\theta}_i$ [12].

3. Proposed Statistical Models

3.1 Lognormal probability distribution

Researchers on environmental sciences have reported that pollutant concentrations in air and soil, and concentrations of metal residues in rivers, have lognormal distribution [7,18,19]. If y is a random two-parameter lognormal variable, it has a probability density function,

$$f(y; \mu, \sigma) = \frac{1}{y\sigma\sqrt{2\pi}} \exp \left\{ -\frac{[\log(y) - \mu]^2}{2\sigma^2} \right\}, \quad (7)$$

$$0 < y < \infty, -\infty < \mu < \infty, \sigma > 0.$$

The median of y , $M = e^\mu$ depends only on μ ; contrastingly its mean, $E = e^{\mu + \sigma^2/2}$ depends on μ and σ , because of this, a simultaneous analysis involving both parameters is necessary when comparing population means.

3.2 Comparison of means of two independent lognormal populations

The quantile function for the log-location-scale lognormal distribution regression model, with a indicator variable x to compare two population means, is given by,

$$\log [y_p(x)] = \mu(x) + \Phi^{-1}(p)\sigma = \beta_0 + \beta_1 x + \Phi^{-1}(p)\sigma. \quad (8)$$

Φ is the cumulative distribution function of a normal standard; regarding the indicator variable, $x = 0$ for one population and $x = 1$ for the other one [13].

3.2.1 Homogeneity of the scale parameter σ model

The likelihood for two independent samples, of sizes n_1 and n_2 , containing left-censored and exact observations, has the form

$$L(\beta_0, \beta_1, \sigma) = \prod_{i=1}^2 \left\{ \prod_{j=1}^{n_i} \left[\frac{1}{\sigma y_{ij}} \phi(z_{ij}) \right]^{\delta_{ij}} \left[\Phi(z_{ij}) \right]^{1-\delta_{ij}} \right\}, \quad (9)$$

where ϕ is the density function of normal standard, $z_{ij} = [\log y_{ij} - \mu_i] / \sigma$, $\mu_i = \mu(x) = \beta_0 + \beta_1 x$, $\delta_{ij} = 1$ for an exact observation and $\delta_{ij} = 0$ for nondetects in sample i . The analysis of comparison of both means can be done by substituting x in $\mu(x) = \beta_0 + \beta_1 x$, then $\mu_1 = \mu(0) = \beta_0$ and $\mu_2 = \mu(1) = \beta_0 + \beta_1$. Now, $\log[y_p(1)] - \log[y_p(0)] = \mu(1) - \mu(0) = \beta_1$, which does not depend on any quantil. This way, if an approximate confidence interval for β_1 contains a zero value, it is concluded that no significant difference exists between both population means.

3.2.2 Heterogeneity of the scale parameter σ model

Here, the comparison is done adding indicator variables in regression models for both parameters, rewriting (8) and optimizing (9) with the new parameterization: $\mu_i = \beta_0 + \beta_1 x$, $\log(\sigma_i) = \log[\sigma(x)] = \gamma_0 + \gamma_1 x$. Then, $\mu_1 = \beta_0$, $\log(\sigma_1) = \gamma_0$ for the first sample and $\mu_2 = \beta_0 + \beta_1$, $\log(\sigma_2) = \gamma_0 + \gamma_1$, for the second one. An approximate confidence region will allow, simultaneously, analyzing if β_1 and γ_1 are both zeroes, to asses the significant difference between population means. This is, the populations have identical pollutant concentration means if $W(0) = (\hat{\theta} - 0) (\hat{\Sigma}_{\hat{\theta}})^{-1} (\hat{\theta} - 0) \leq \chi^2_{(1-\alpha; 2)}$, where $\hat{\theta} = (\hat{\beta}, \hat{\gamma})$.

3.3 Comparison of means of two lognormal populations in presence of a continuous covariable

For a random lognormal sample y in presence of a continuous covariable t , it is given that $E[\log(y)] = \mu = \beta + \alpha_1 t$, where α_1 represents the intercept of the loglineal model for pollutants in the first unit of t , and α_1 represents the change per unit of t . The log-quantil function for the corresponding regression model has the following form,

$$\log[y_p(x)] = \mu(x) + \Phi^{-1}(p)\sigma = \beta_0 + \beta_1 x + \alpha_1 t + \Phi^{-1}(p)\sigma \quad (10)$$

Given y_1, y_2 two random lognormal samples with common σ , in presence of a continuous covariable t . Then, the likelihood for y_1, y_2 with exact and nondetects data is

$$L(\beta_0, \beta_1, \alpha_0, \sigma) = \prod_{i=1}^2 \left\{ \prod_{j=1}^{n_i} \left[\frac{1}{\sigma y_{ij}} \phi(z_{ij}) \right]^{\delta_{ij}} \left[\Phi(z_{ij}) \right]^{1-\delta_{ij}} \right\}, \quad (11)$$

where $z_{ij} = [\log y_{ij} - \mu_i] / \sigma$. $\mu_i = \mu(x) = \beta_0 + \beta_1 x + \alpha_1 t$. It substitutes x by 1 for the one sample and 0 for the other one, giving $\mu_1 = \beta_0 + \alpha_1 t$ and $\mu_2 = \beta_0 + \beta_1 + \alpha_1 t$. The inference on the difference of means will depend only on β_1 , since $\mu_2 - \mu_1 = \beta_1$ does not depend on any quantil, nor on t . Thus, an approximated confidence interval for β_1 that contains a zero value, would affirm that no significant difference exists between the population means. In the same way, the effect of trend of data can be quantified through $\hat{\alpha}_1$ and its approximated confidence interval.

3.4 Comparison of means of two lognormal populations using a mixed-effects model

Mixed-effects models are used to describe relationships between a response variable and some covariates in data that are grouped according to one or more classification factors, for example longitudinal information, repeated measures, and block designs. By associating random effects common to observations sharing the same level or classification factor, mixed-effects models flexibly represent the covariance structure induced by the grouping of the data [9,20].

3.4.1 Linear Mixed-Effects Model

Linear mixed-effects models are mixed-effects models in which both the fixed and the random effects occur linearly in the model function. They extend linear models by incorporating random effects, which can be regarded as additional error terms, to account for correlation among observations within the same group. The linear mixed-effects model, described by [9,20], is defined as,

$$Y = X\beta + Zb + \varepsilon; \quad b \sim N(0, \Psi), \quad \varepsilon \sim N(0, \sigma^2 I), \quad (12)$$

where β is the vector of fixed effects, b is the vector of random effects; X, Z are known regressor matrices, and ε is the within-group error vector with a spherical Gaussian distribution. The random effects b and the within-group error ε are assumed to be independent for different groups and to be independent of each other for the same group. The parameter estimation is obtained using the general method for maximum likelihood presented in [20] and implemented in R using the `lme` function [22].

3.4.2 Model and Comparison Criterion

For the comparison of two population means, and under the assumption of homogeneity of σ , the model (12) is used and it parameterizes μ with an indicator-variable regression model. This is possible through a loglinear model for the observations,

$$\log(y) = \mu(x) + \alpha_1 t + b_0 = \beta_0 + \beta_1(x) + \alpha_1 t + b_0, \quad (13)$$

From Table 2 it is observed that the ML estimator via EM for β_1 is close to zero, -0.116 , and its confidence interval contains a zero value, $(-0.413, 0.181)$, so that the difference is statistically null. Then, according to the context, the means concentrations of copper are equal for both zones.

Model 2. Heterogeneity of σ

From 3.2.2, the comparison of means is done through an approximate confidence region for β_1 and γ_1 , assessing if both are simultaneously null. Estimations are shown in Table 3.

Table 3. ML estimators for parameters in copper data with the σ heterogeneity model

Parameter	ML Estimators		Standard error	
	Direct	Via EM	Direct	Via EM
β_0	0.592	0.906	0.124	0.101
β_1	0.084	0.038	0.170	0.137
γ_0	-0.039	-0.202	0.099	0.088
γ_1	-0.072	-0.102	0.141	0.124

	μ_1	μ_2	σ_1	σ_2	$\log L_2(\beta, \gamma)$
Direct	0.676	0.592	0.962	0.895	-168.386
Via EM	0.944	0.906	0.817	0.738	-151.575

Table 3 shows that the use of EM improves the standard errors, and the fitting of the model. Via EM, $\hat{\beta}_1 = 0.038$ and its standard error is 0.137; for γ_1 , -0.102 and 0.124, respectively. Then, $W(0) = 1.816 \times 10^{-4} \leq 5.991 = \chi_{(0.95; 2)}^2$. Therefore, with a 95% of confidence, no significant differences exist between the means of copper concentrations.

Discussion In both models, the EM algorithm improves the estimation, the standard errors, and confidence intervals, while reducing the complexity of the optimization of (9). Then, according to the results, Model 1 and Model 2 reveal that the mean contaminant concentrations do not differ between the two zones; nevertheless, to illustrate how to select the best model, a likelihood ratio test is applied, where $-2(\log L_{EM1} - \log L_{EM2}) = 32.478 > 3.84 = \chi_{(0.95; 1)}^2$. Wherefrom, it is appreciated that Model 2 is the more appropriate to describe these particular data.

4.2 Comparison of two population means and trend analysis on a remediation process.

An important objective of many environmental monitoring programs is to detect changes in trend in pollution levels over time. The comparison of mean concentrations and the monitoring of the water quality in remediation processes are very important, considering also the effects caused by fixed covariates like the monitoring time.

Example It reports data of arsenic concentrations in a remediation process, from October to November 1980, water quality was monitored from 1981 to 1982 [11], Table 4. To illustrate the proposed method, observations bellow 1 and missing values was censored artificially by 1 ppm. The aim is to determine if the remediation processes improve water quality and the effect of time. The homogeneity of the scale parameter σ is assumed for both lognormal populations.

Table 4. Arsenic concentrations (mg/L) pre and post remediation process

Date of sampling	Preremedial records	Date of sampling	Postremedial records
10/79	4.3	10/81	<1
11/79	16	11/81	1.5
12/79	6.1	12/81	1.3
1/80	<1	1/82	<1
2/80	<1	2/82	<1
3/80	2.66	3/82	2.1
4/80	3.00	4/82	1.1
5/80	4.42	5/82	<1
6/80	5.74	6/82	1.8
7/80	1.40	7/82	1.2
8/80	1.49	8/82	<1
9/80	2.3	9/82	1.1

McBean & Rovers (1998) arsenic in remedial process

Homogeneity of σ and continuous covariate model

The comparison is done fitting model (10). Then, optimizing (11), the ML estimators are obtained for the parameters and their approximated confidence intervals. Then, under the assumption of common σ and a continuous covariate t , the difference between the means of both periods will be observed only through the interval obtained for β_1 . On the other hand, the effect of t on the information will be measured through the estimator and the interval for α_1 . Table 5 shows the ML estimators, the confidence intervals for the parameters and the values of $\log L$.

Table 5. ML estimators for concentration parameters in the remedial process

Parameter	ML Estimator		Standard Error		Approx. Interval 95% Confidence	
	Direct	Via EM	Direct	Via EM	Direct	Via EM
β_0	-0.378	1.189	0.439	0.358	-1.239, 0.483	-0.514, 0.891
β_1	0.559	1.073	0.358	0.304	-0.143, 1.261	0.477, 1.670
α_1	0.018	-0.037	0.056	0.044	-0.092, 0.128	-0.123, 0.050
σ	1.022	0.745	0.181	0.101	0.668, 1.376	0.548, 0.943

	μ_1	μ_2	σ	α_1	$\log L_1(\beta, \delta, \sigma)$
Direct	0.181	-0.378	1.022	0.018	-42.625
Via EM	1.262	0.189	0.745	-0.037	-25.397

Discussion It is observed in Table 5 that the EM algorithm improves the fitting of (10), obtaining better estimations for approximate confidence parameters and intervals. Note that, via EM, $\hat{\beta}_1=1.073$, with an interval that does not contain a zero value, (0.477, 1.670), so that the difference of means is significant. Therefore, according to the results, $\mu_{pre} > \mu_{pos}$, so it is confirmed that the remediation process improves water quality. On the other hand, it is observed that α_1 , corresponding to the time covariate t , has a ML estimator via EM of -0.037 , which is interpreted as a light trend to diminish, nevertheless its approximated confidence interval indicates that, statistically, no trend exists on the mean concentrations of the pollutant.

4.3 Comparison of two population means containing grouped data

Sometimes, when two or more sites are compared to know the differences among their mean contaminant concentrations, it happens that these sites have several sampling localities, which generate grouped or data in blocks. This generates random effects associated with the sampling design of the individual units of the sites and to the experimental design for obtaining data [20].

Example. The EPA [5] does a monitoring on a site in risk of contamination to compare its toluene concentrations to a control site. The aim is to observe if there is evidence of pollution. The information presents a great number of nondetects data, Table 6.

Table 6. Toluene concentrations data from two sites in a period of five months

Month	Groundwater site		Compliance sites		
	Well 1	Well 2	Well 1	Well 2	Well 3
1	< 5	< 5	< 5	< 5	< 5
2	7.5	< 5	12.5	13.7	20.1
3	< 5	< 5	8.0	15.3	35.0
4	< 5	< 5	< 5	20.2	28.2
5	6.4	< 5	11.2	25.1	19.0

US EPA (1992) toluene concentrations data ppm

Linear mixed-effects model for comparing two sites with grouped data

Following the criterion defined in 3.4.2, model (13) is fitted using the general method for maximum likelihood and the EM algorithm; to handle nondetects, obtain estimators for variance and standard errors for the estimators. Using `lme` in R and grouping according to the wells in each site, the inferences on the all parameters are obtained, as shown in Table 7

Table 7. ML estimators for parameters on toluene concentrations data

Parameter	ML Estimators via EM	Standard Error	Approx. Interval 95% confidence
β_0	1.792	0.348	1.110, 2.474
β_1	- 1.353	0.284	- 1.911, - 0.796
α_1	0.210	0.098	0.019, 0.401

MLE Via EM	μ_1	μ_2	σ_b	σ_ε	$\log L(\beta, \gamma, \delta)$
		0.439	1.792	0.104	0.690

Discussion It is observed that β_1 has a negative ML estimator, -1.353, with an interval that does not contain a zero value, (-1.911, -0.796), so that there is evidence that both mean concentrations are different. In the context, the control site has a higher toluene concentration. Therefore, no evidence of pollution exists in the site in risk of contamination. On the other hand, the estimator for the parameter of the month covariate is 0.210, with the interval (0.019, 0.401), so that, there is evidence of a significant and increasing effect of time on the pollutant concentrations. Finally, it is important to mention that the random effect, due to the grouping by wells σ_b , presents a relatively minor variability, 0.690, which makes it possible to consider the results and previous conclusions as acceptable.

5. Conclusions

The methodology used, which is based on maximum likelihood and a modification of the EM algorithm, was versatile and simple to compare lognormal populations through lognormal regression models. In example 1, a comparison of two populations with both assumptions, homogeneity and heterogeneity of σ , using a parameterization for two parameters lognormal models was shown. The criterion for comparison was efficient to observe an approximate confidence interval and Wald's region, respectively. Additionally, it illustrated a way to select the best model through a likelihood-ratio test. In example 2, monitoring dates were considered as a continuous covariate; it was simply added to a lognormal regression model, obtaining estimators and approximate confidence intervals with easy interpretation. An approach from linear mixed models, developed in example 3, permitted to extend the comparison methodology to populations containing random covariates caused by grouped data, maintaining the criterion based on approximate confidence intervals. In all application examples, the EM algorithm improved the optimization of respective likelihood functions containing nondetects. Thus a better fitting, lower standard errors, and narrowed approximate confidence intervals were observed compared to those obtained from direct optimization of likelihood. This work has presented methodology for comparing two populations. Nevertheless, the procedure can be extended to three or more lognormal populations, since the implementation and understanding are not difficult.

Acknowledgment The PhD studies, under which this work was conducted, was supported by the PROMEP-UJAT program. The financial support is highly appreciated. The authors thank Dr. Gustavo Ramírez-Valverde and Dr. Gabriel A. Rodríguez-Yam for their comments and suggestions.

References

1. Cox, D.R., & Snell, E. J. (1981). *Applied Statistics: Principles and Examples*, New York: Chapman and Hall, Inc.
2. Dempster, A.P., Laird, N., & Rubin, D.B. (1977). Maximum likelihood for incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B39*, 1–38.
3. Díaz-Francis, E., & Sprott, D.A. (2000). The use of the likelihood function in the analysis of environmental data. *Environmetrics 11*, 75–79.
4. El-Shaarawi, A. H., & Viveros, R. (1997). Inferences about the mean in log-regression with environmental applications. *Environmetrics 8*, 569–582.
5. EPA (1992). *Statistical Training Course for Ground-Water Monitoring Data Analysis*. EPA530-R-93-003. Office of Solid Waste. U.S. Environmental Protection Agency, Washington, DC.
6. Flury, B., & Zoppè, A. (2001). Exercise in EM. *The Am. Statist. 54*, pp. 207 - 209.
7. Gilbert, R.O. (1987). *Statistical Methods for Environmental Pollution Monitoring*. New York: Wiley.
8. Helsel, D. R. (2005). *Nondetects And Data Analysis*. New York: Wiley.
9. Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics 38*, 963 – 974.
10. Lambert, D., Peterson, B., & Terpenning, I. (1991). Nondetects, detection limits, and the probability of detection. *Journal of the American Statistical Association*, Vol. 86, No. 414, 266-277.
11. McBean, E.A., & Rovers, F.A. (1998). *Statistical Procedures for Analysis of Environmental Monitoring Data and Risk Assessment*. USA: Prentice Hall.
12. Meeker, W. Q., & Escobar, L. A. (1995). Teaching about approximate confidence regions based on maximum likelihood estimation. *The American Statistician*, 49, 48-53.
13. Meeker, W. Q., & Escobar, L. A. (1998). *Statistical Methods for Reliability Data*, 450–454. New York: Wiley
14. Millard, S. P., & Deverel, S. J. (1988). Nonparametric statistical methods for comparing two sites. *Water Resources Research 24*, 2087–2098.
15. Millard, S. P., & Neerchal, N. K. (2001). *Environmental Statistics with S-PLUS*. USA: CRC Press

16. Navy (1999). Handbook for statistical analysis of environmental background data, Department of the Navy, Southwest Division, Naval Facilities Engineering Command, San Diego, CA.
17. Oakes, D. (1999). Direct calculation of the information matrix via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, 61, 479 – 482.
18. Ott W.R. (1990) A physical explanation of the lognormality of pollutant concentrations. *Journal of the Air and Waste Management Association* 40, 1378-1383
19. Ott, W. R. (1995). *Environmental Statistics and Data Analysis*, USA: CRC Press.
20. Pinheiro, J.C., & Bates, D.M. (2004). *Mixed-Effects Models in S and S-PLUS*, New York: Springer
21. Robert, C. P., and G. Casella. 2005. *Monte Carlo Statistical Methods*. 2nd Ed. New York: Springer.
22. R Development Core Team. (2006). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.