ESTADÍSTICA EN VARIABLES CON CENSURA:
APLICACIÓN A DATOS MEDIOAMBIENTALES

Beatriz Quintanilla Casas
Máster en Bioinformática y Bioestadística
Área de Bioestadística

Consultor/a: **Silvia Teresa Vargas Castrillón**
Profesores responsables de la asignatura: **Alexandre Sánchez Pla, Antoni Pérez Navarro, Carles Ventura Royo, José Antonio Morán Moreno y María Jesús Marco Galindo**

Junio de 2017

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Estadística en variables con censura: aplicación a datos medioambientales*
Nombre del autor:	*Beatriz Quintanilla Casas*
Nombre del consultor/a:	*Silvia Teresa Vargas Castrillón*
Nombre del PRA:	*Alexandre Sánchez Pla, Antoni Pérez Navarro, Carles Ventura Royo, José Antonio Morán Moreno y María Jesús Marco Galindo*
Fecha de entrega (mm/aaaa):	06/2017
Titulación:	*Máster en Bioinformática y Bioestadística*
Área del Trabajo Final:	*Bioestadística*
Idioma del trabajo:	*Español*
Palabras clave	*Censura, mercurio, bioestadística*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados y conclusiones del trabajo.*

En el ámbito de la estadística, se denomina censura al fenómeno que afecta a una variable de estudio cuando su medición no puede ser cuantificada. El presente proyecto se focaliza en la censura que tiene lugar en los estudios de temática medioambiental, cuando el resultado es un valor inferior al límite de detección del método analítico que se emplea. Esto puede resultar un problema en su tratamiento estadístico, dado que usualmente los casos censurados son sustituidos y la aplicación de la metodología estadística estándar puede proporcionar resultados erróneos y llevar a conclusiones poco fiables. Por ello, la finalidad del proyecto es presentar la censura estadística desde una perspectiva global, así como los métodos estadísticos específicos para su tratamiento, mediante una revisión bibliográfica. Posteriormente, la información teórica proporcionada se traslada a la práctica realizando el tratamiento de datos censurados por la izquierda, los cuales se obtuvieron en un estudio real de evaluación de la exposición al mercurio.

Durante este proceso, se compararon los resultados obtenidos aplicando la sustitución de los valores censurados con los proporcionados por métodos estadísticos recomendados. Como resultado se ha comprobado que, aunque la sustitución de los valores censurados resulta más fácil de aplicar, en ocasiones puede proporcionar resultados poco fiables.

Sin embargo, también es necesario tener en cuenta que la aplicación de métodos más sofisticados puede comportar otro tipo de inconvenientes como la baja bondad de ajuste. Por tanto, es importante tener presente el objetivo del tratamiento de los datos censurados para decidir qué método estadístico aplicar.

Abstract (in English, 250 words or less):

In the area of statistical science, the phenomenon that appears when the measurement of the variable of study cannot be quantified is called censorship. The present project is focused in the censorship which takes places in environmental studies, when the result obtained is under detection limit. This could be a problem in the statistical treatment of the data, due to the censored values are usually substituted and the application of the standard statistical methods could provide wrong results and unreliable conclusions. For this reason, the aim of this project is to introduce the statistical censorship in a general way, as well as the statistical methods that are recommended to deal with it, doing an exhaustive review. Then, the theoretical information obtained is put into practice doing the statistical treatment of a left censored dataset that belongs to a real study for the assessment of people's exposition to environmental mercury.

During this process, the results obtained by two different strategies (substitution and specific methodology for censorship) were compared. As a result, it has been proven that, although the substitution of censored values is easier, sometimes it can provide unreliable results. However, it is also necessary to take into account that more sophisticated methods can entail other types of disadvantages such as low goodness of fit. Therefore, to keep in mind the purpose of the treatment of censored data is important to decide the correct statistical method for applying.

Índice

Lista de figuras

Lista de tablas

1
Introducción

1. CONTEXTO Y JUSTIFICACIÓN DEL TRABAJO

1.1 Descripción general

Desde el punto de vista de la estadística, se denomina censura al fenómeno que afecta a las variables de interés en un estudio cuando existe una limitación en la información que se dispone de ellas. Los datos censurados son aquellas observaciones que no pueden ser cuantificadas, únicamente se conoce que el valor se encuentra por debajo o por encima de un umbral determinado o bien incluido en un intervalo [1].

Tradicionalmente las variables con censura se han asociado a los estudios de supervivencia, dado que ésta circunstancia se presenta cuando el evento de interés no tiene lugar durante el tiempo de observación [2,3]. Sin embargo, también es común encontrar datos censurados en aquellos estudios de ámbito medioambiental ya que los resultados de las mediciones pueden superar el límite de detección/cuantificación de la técnica analítica o quedarse por debajo del mismo. [4,5]. Centrándonos en el segundo ámbito de estudio, resulta crucial poder tratar este tipo de datos censurados con el objetivo de extraer conclusiones fiables; sobre todo en aquellos estudios que tienen como finalidad evaluar una problemática medioambiental que afecta negativamente y de forma directa a la salud de la población.

Uno de los casos más conocidos y extendidos en todo el mundo es la problemática del mercurio, un metal empleado en diversas actividades industriales, así como en equipamiento eléctrico, termómetros, barómetros e incluso como material constituyente de prótesis dentales (amalgamas dentales). El mercurio, en sus distintas formas químicas, se encuentra en la atmosfera debido al uso de los combustibles fósiles a través de la cual llega a la tierra y al agua, siendo el medio acuático el más contaminado en este metal. De este modo, resulta comprensible que una de las mayores fuentes de exposición al mercurio sea, entre otras, la ingesta de pescado y productos del mar [6]. A causa de los efectos negativos que puede presentar una exposición aguda o crónica al mercurio, son varios los estudios que pretenden evaluar la concentración de dicho elemento en sangre [7,8,9] u orina [10,11] de una población específica con el objetivo de ajustar las políticas de salud pública referentes a este problema.

1.2 Justificación del TFM

La obtención de mediciones censuradas en estudios medioambientales es un problema habitual y, a su vez, controvertido. La aplicación de metodologías inadecuadas que se han empleado de forma tradicional durante años y que se siguen utilizando actualmente para el tratamiento estadístico de estos datos censurados, como es el caso de la sustitución de los valores censurados, puede conllevar la obtención de resultados

sesgados. Por este motivo, es necesario conocer los métodos estadísticos adecuados que pueden llevarse a cabo con el objetivo de obtener conclusiones válidas y fiables.

En el presente trabajo, se espera obtener una visión global y actualizada de la metodología estadística que se ha aplicado a los conjuntos de datos con casos censurados en estudios de temática medioambiental. Además, en base a la revisión bibliográfica efectuada, se pondrán en práctica los conceptos presentados sobre un conjunto de datos censurado con el objetivo de comparar tanto el desarrollo y aplicación de distintas técnicas estadísticas, así como los resultados que proporcionan cada una de ellas.

2. OBJETIVOS DEL TRABAJO

A continuación, se exponen los objetivos que se persiguen en el presente proyecto, los objetivos generales son más amplios y abarcan la totalidad del trabajo mientras que estos se desglosan en varios objetivos específicos para asegurar el desarrollo y cumplimiento de los primeros.

2.1 Objetivos generales

1. Analizar los distintos métodos estadísticos empleados en el tratamiento de datos con censura.
2. Aplicar las técnicas estadísticas adecuadas a un conjunto de datos real sobre la problemática medioambiental del mercurio.

2.2 Objetivos específicos

1. Identificar los distintos tipos de censura que se pueden encontrar en conjuntos de datos provenientes de estudios de temática diversa.
2. Exponer los posibles enfoques estadísticos que pueden aplicarse para el tratamiento de resultados censurados.
3. Discutir los resultados obtenidos en distintos estudios experimentales que han empleado metodologías estadísticas diferentes para el tratamiento de datos que presentan casos de censura.
4. Presentar un estudio real sobre la problemática medioambiental del mercurio.
5. Evaluar la presencia de casos censurados entre las variables de interés del estudio experimental.
6. Aplicar los métodos estadísticos adecuados y discutir su desarrollo y resultados obtenidos.

3. ENFOQUE Y MÉTODO SEGUIDO

El presente proyecto se divide en dos partes diferenciadas:

- La primera parte consiste en realizar una revisión bibliográfica exhaustiva sobre la temática de la censura estadística y, de un modo más específico, cómo afecta este fenómeno a los estudios realizados en el ámbito medioambiental. El propósito es conocer cómo detectar la censura en los conjuntos de datos y qué metodología estadística puede aplicarse para lograr un tratamiento estadístico correcto de los datos con casos censurados. Para ello, se emplean las bases de datos en línea como *Sciencedirect* o *Scopus* y los recursos en línea de la biblioteca de la UOC.
- La segunda parte del proyecto se centra en el tratamiento de unos datos reales extraídos de un estudio donde se mide la concentración de mercurio en orina en humanos. El software estadístico que se emplea es *Rstudio*, junto con los paquetes que se precisen en cada caso, y la herramienta *RMarkdown* que permite presentar el código elaborado en cada metodología estadística y los resultados que se han derivado.

4. PLANIFICACIÓN DEL TRABAJO

A continuación, se desglosan los objetivos específicos presentados anteriormente en tareas y se propone una planificación temporal para la ejecución de cada una de ellas; además se indican los hitos que marcan dicha temporización. En la Figura 1 se muestra el cronograma correspondiente al proyecto.

4.1 Tareas

Objetivo específico 1. Identificar los distintos tipos de censura que se pueden encontrar en conjuntos de datos provenientes de estudios de temática diversa.

Tarea 1.1 Búsqueda bibliográfica sobre el fenómeno de la censura en datos.

Tarea 1.2 Clasificación de los tipos de censura.

Tarea 1.3 Relacionar cada tipo de censura con los posibles ámbitos de estudio donde se pueden encontrar.

Objetivo específico 2. Exponer los posibles enfoques estadísticos que pueden aplicarse para el tratamiento de datos censurados.

Tarea 2.1 Revisión bibliográfica sobre la metodología estadística en variables con censura.

Tarea 2.2 Interpretación de los modelos explicados.

Objetivo específico 3. Discutir los resultados obtenidos en distintos casos que han empleado metodologías estadísticas diferentes para el tratamiento de datos con censura.

Tarea 3.1 Búsqueda de referencias bibliográficas de tratamiento de datos medioambientales con censura.

Tarea 3.2 Valoración de ventajas e inconvenientes que supone cada método, tanto en el desarrollo como en los resultados.

Objetivo específico 4. Presentar un estudio real sobre la problemática medioambiental del mercurio.

Tarea 4.1 Explicación de la evaluación de parámetros biológicos como indicadores de problemáticas medioambientales, especialmente el caso del mercurio

Tarea 4.2 Exposición del estudio: condiciones, población, muestra, objetivos, etc.

Tarea 4.2 Importación de los datos a R y adecuación del conjunto de datos.

Objetivo específico 5. Evaluar la presencia de datos censurados entre las variables de interés del estudio experimental.

Tarea 5.1 Selección de las variables de estudio y descripción de éstas.

Tarea 5.2 Identificación y descripción de los datos censurados.

Objetivo específico 6. Aplicar los métodos estadísticos adecuados y discutir su desarrollo y resultados obtenidos.

Tarea 6.1 Elección de la/s metodología/s estadística/s más adecuadas.

Tarea 6.2 Elaboración del código necesario para ejecutar cada uno de los métodos elegidos.

4.2 Cronograma

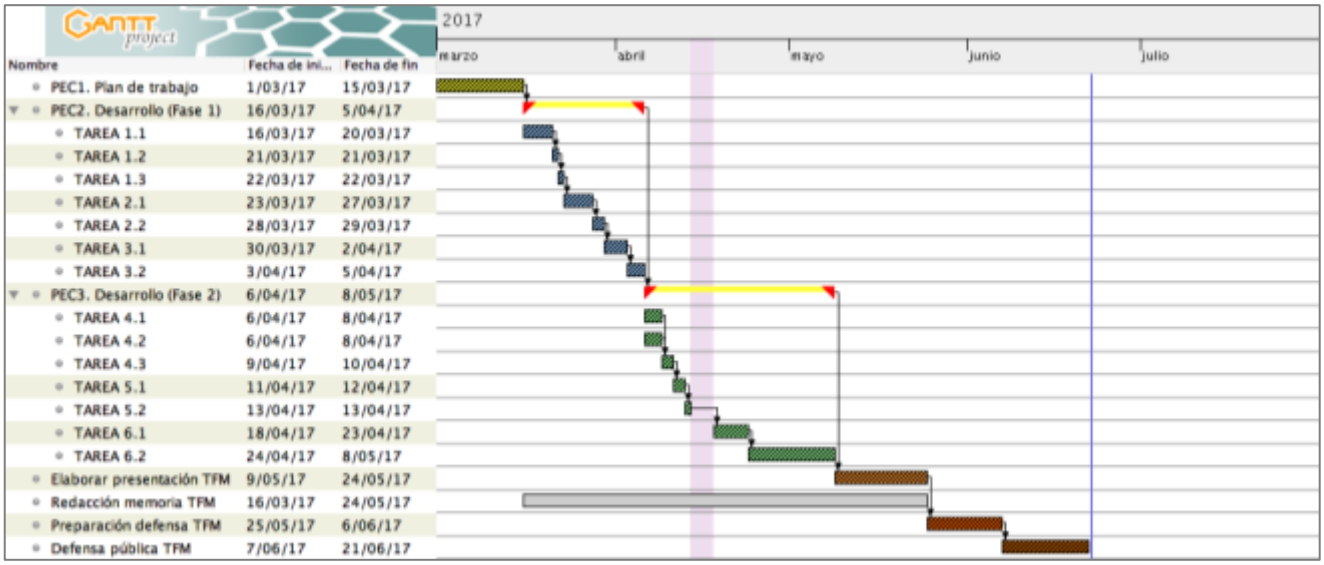


Figura 1. Diagrama de Gantt con la temporización de las tareas y los hitos del proyecto

5. BREVE SUMARIO DE PRODUCTOS OBTENIDOS

En el presente proyecto no se ha realizado el desarrollo de un producto, entendido como tal, sino que se han obtenido unos hitos de acuerdo con los objetivos marcados:

- Revisión bibliográfica sobre la censura estadística y los estudios en el ámbito medioambiental.
- Tratamiento estadístico de un conjunto de datos con censura por la izquierda obtenido en un estudio real de temática medioambiental en R: se presenta el código creado y aplicado, junto con su explicación a través de la herramienta *RMarkdown*.

6. BREVE DESCRIPCIÓN DE LOS OTROS CAPÍTULOS DE LA MEMORIA

Los capítulos que conforman la presente memoria del proyecto se dividen en los siguientes bloques temáticos:

- **Bloque 1. Introducción.**

Se trata del bloque actual, donde se puede encontrar una descripción y justificación del proyecto, así como los objetivos del mismo y la temporalización de las tareas que se han llevado a cabo.

- **Bloque 2. Revisión bibliográfica: la censura estadística y su tratamiento en estudios de ámbito medioambiental.**

Se presenta el tema de la censura estadística explicando su definición y los tipos que se pueden encontrar, además de la importancia de conocer los métodos estadísticos adecuados para su tratamiento según las características del conjunto de datos. Se realiza una revisión de artículos de temática medioambiental, concretamente aquellos que tratan sobre la problemática del mercurio y su relación con la censura estadística por la izquierda.

- **Bloque 3. Caso práctico: tratamiento estadístico datos medioambientales censurados.**

Se describe el conjunto de datos con casos censurados por la izquierda a tratar: en qué estudio se han obtenido los datos, descripción de la muestra y de las variables de interés. Posteriormente se realiza el tratamiento de los datos presentados con el software R.

- **Bloque 4. Conclusiones**

Se realiza una reflexión global del trabajo, tanto a nivel de planificación de las tareas y logro de los objetivos como a nivel de resultados obtenidos y aprendizaje efectuado.

- **Bloque 5. Glosario**

Se incluye una definición de los términos y acrónimos más relevantes utilizados en la memoria.

- **Bloque 6. Referencias**

Listado de referencias bibliográficas consultadas (artículos, libros, sitios web, etc.), ordenadas según el orden de aparición en la presente memoria.

2

Revisión bibliográfica.

La censura estadística y su tratamiento en estudios de ámbito medioambiental

1. LA ESTADÍSTICA: HERRAMIENTA INDISPENSABLE EN LA INVESTIGACIÓN CIENTÍFICA

Basándonos en los principios del método científico, la primera fase que se debe llevar a cabo para iniciar un proceso de investigación es necesariamente la observación de los hechos o fenómenos de interés. Después de esta etapa, se deben formular las hipótesis sobre el hecho observado y a continuación se procede a la fase de experimentación con el objetivo de recabar información de una forma controlada en función de la hipótesis formulada. La última etapa, aunque no por ello menos importante, sería la extracción de las conclusiones a partir del análisis de los datos obtenidos en la fase anterior; la aceptación o rechazo de la hipótesis planteada inicialmente permitirá avanzar en el conocimiento de la materia en cuestión.

Pese a que cada una de las etapas anteriores contribuyen de igual manera al éxito del proceso, históricamente se ha asociado la investigación científica de forma más directa con las primeras 3 fases. Tal y como describe el autor del libro *Statistics in scientific investigation: Its basis, Application and Interpretation* [12]: "La popular imagen de la investigación científica es de personas con batas blancas en un laboratorio, normalmente bajo la tutela de un genio excéntrico que va a realizar un descubrimiento trascendental". Sin embargo, el proceso que implica el análisis de los resultados obtenidos es tan importante como el diseño del estudio dado que las conclusiones finales que obtendrán a partir de éste. La herramienta que permite a los investigadores tratar los datos obtenidos y entender qué ha pasado durante el experimento es la estadística. El problema es que, como bien vuelve a explicar McPhellan [12]: "Hace veinte años, la típica introducción de la estadística a los científicos era una colección de métodos estadísticos recogidos en un recetario. Los científicos simplemente seleccionaban, entre los pocos métodos proporcionados, aquél que estaba descrito que podía servir en una situación concreta y aplicaban la fórmula para poder extraer conclusiones de acuerdo a la regla estipulada".

Este hecho se pone de manifiesto en la revisión de García-Berthou y Alcaraz [13], donde se comprueban los resultados proporcionados por estudios del área de medicina publicados en las revistas *Nature* y *British Medical Jorunal* durante el año 2001. Sorprendentemente, alrededor de un 11% de los artículos revisados en cada una de las revistas mostraban resultados incongruentes, los cuales suelen deberse a una incorrecta transcripción o redondeo. Además, los autores recomiendan proporcionar intervalos de confianza como una mejor alternativa a los contrastes de hipótesis, ya que se detecta un uso inadecuado del p valor obtenido en los contrastes para la posterior extracción de conclusiones.

A menudo, a causa del diseño experimental llevado a cabo, los datos obtenidos presentan características que pueden suponer una dificultad adicional en su posterior análisis estadístico. Uno de los problemas más frecuentes a los que se deben enfrentar los investigadores son los llamados datos con observaciones ausentes (*missing data*).
En ocasiones estas observaciones no son totalmente ausentes, sino que son incompletas; se dispone de información sobre la variable estudiada, pero ésta no es suficiente para aplicar la metodología estadística estándar de forma directa.

El presente proyecto se centra, precisamente, en uno de los tipos de observaciones incompletas más habituales en los estudios del área de ciencias biológicas conocidas como observaciones censuradas [14]. La bioestadística es la ciencia que permitirá tratar el conjunto de datos que presenten esta característica.

2. EL FENÓMENO DE LA CENSURA ESTADÍSTICA

En el ámbito de ciencias biológicas, uno de los diseños experimentales más empleados es el estudio de supervivencia o longitudinal. Su característica principal es que los sujetos de estudio son observados durante un tiempo estipulado (tiempo de seguimiento) y el objetivo es conocer el tiempo en el que tiene lugar el evento o suceso de interés (tiempo de ocurrencia) [2]. Dado que el diseño longitudinal puede aplicarse a distintos campos de investigación, el evento de estudio será el propio de cada disciplina: en medicina puede ser la muerte de un paciente que sufre una patología concreta o bien un episodio relacionado con la enfermedad, en ingeniería puede referirse al fallo de una máquina o de una de sus piezas y en economía podría ser el inicio de un nuevo empleo después de un período de desempleo [15].

Arrizabalaga [2], define que una observación en un estudio epidemiológico está censurada cuando el individuo no presenta el evento de interés durante el tiempo de seguimiento (duración del estudio). Aunque se trate de un estudio longitudinal, lógicamente éste debe tener un punto de inicio y final adaptado a los recursos disponibles. Dependiendo de si el evento ha tenido lugar antes del punto de inicio del estudio o bien no se ha ocasionado una vez dado por finalizado, el tiempo de censura será la fecha de inicio o final del estudio.

Es importante no confundir el fenómeno de la censura con otros tipos de datos incompletos; la confusión más común se produce entre las variables censuradas y las **truncadas**. El truncamiento tiene lugar cuando sólo aquellos sujetos que manifiestan el evento dentro de una ventana observacional son observados, del resto no se realiza ningún seguimiento y, por tanto, no se obtiene información sobre ellos [16].

El ejemplo más claro de truncamiento lo encontramos en el campo de la astronomía: en una parte del espacio, sólo los elementos suficientemente brillantes pueden ser observados desde la Tierra; aquellos cuya intensidad lumínica es inferior a un cierto nivel, no es posible saber de su existencia [15].

Dependiendo del ámbito de estudio, el diseño experimental y los resultados obtenidos para la variable, se pueden agrupar las observaciones censuradas en 3 grandes grupos: censuradas por la derecha, censuradas por la izquierda y censuradas en un intervalo. En los siguientes sub-apartados se explica cada uno de estos grupos de forma detallada.

2.1 Variables censuradas por la derecha

En epidemiología, una observación está censurada por la derecha cuando el evento de interés no tiene lugar durante el período de observación, de modo que no es posible determinar el tiempo de ocurrencia [17]. En función de la causa que ha originado la censura por la derecha, podemos encontrar distintos tipos [15,16]:

- Censura por la derecha de tipo I: la finalización del período de observación de los individuos tiene lugar a un tiempo predeterminado, el cual se fija durante el diseño del estudio; así el tiempo de censura es conocido y fijo. La censura de tipo I puede ser:
    - Fija: el tiempo de inicio y final del estudio (censura) es el mismo para todos los individuos.
    - Progresiva: los individuos se dividen en grupos y cada uno de los grupos tiene un tiempo de censura concreto.
    - Generalizada: cada uno de los individuos tiene un tiempo de entrada al estudio y un tiempo de censura específico.

- Censura por la derecha de tipo II: la finalización del período de observación de los sujetos no ocurre a un tiempo prefijado, sino que éste continúa hasta que ocurre el suceso de estudio en una proporción establecida de individuos respecto al total. Por tanto, el tiempo de censura no se conoce a priori, pero se trata de una variable fija dado que la proporción sí que se estipula durante el diseño del estudio.

- Censura por la derecha de tipo III: este grupo está formado por la censura aleatoria, también llamada no informativa. Se denomina así porqué el tiempo de censura lo determina un fenómeno aleatorio no esperado, que tiene lugar durante la consecución del estudio e impide seguir con la observación del individuo hasta el tiempo final. Estos sucesos no esperados pueden ser: la pérdida del sujeto sin más información, el abandono voluntario del estudio o la

experimentación de un evento de competencia con el evento de interés que obliga a eliminar al individuo del estudio. Por tanto, en todos estos casos, el tiempo de censura es aleatorio y tiene lugar antes del tiempo de finalización del estudio que se ha estipulado.

También es posible encontrar casos de censura por la derecha en estudios donde se realiza una medición analítica de la variable de interés y el valor obtenido supera un umbral determinado, por tanto, obtenemos una información incompleta sobre ésta no pudiendo ser cuantificada [1].

2.2 Variables censuradas por la izquierda

Si nos referimos a un estudio longitudinal del área de epidemiología, las observaciones censuradas por la izquierda serán aquellas en las que el evento de interés ha tenido lugar antes del punto de inicio del estudio. Así, el tiempo de censura en este caso será el tiempo de inicio del período de seguimiento ya que se conoce que el suceso ha ocurrido previamente, pero no puede saberse con exactitud cuándo (no es cuantificable) [16].

Tal y como se ha indicado en la censura por la derecha, la censura por la izquierda también la encontramos en mediciones analíticas de la variable. En este caso, el valor obtenido en la medición es inferior a un umbral determinado y, por este motivo no es cuantificable. Un ámbito de estudio donde este tipo de censura es recurrente es el orientado a la investigación medioambiental, dado que los instrumentos de medida tienen un límite de detección específico y a menudo se detectan observaciones que no lo alcanzan [1,4,5,15].

2.3 Variables censuradas en un intervalo

Las observaciones que presentan censura en un intervalo suelen ocasionarse en aquellos estudios donde las mediciones de la variable de interés se realizan de forma periódica, de modo que es posible que el suceso de estudio haya tenido lugar en un tiempo entre dos de las mediciones. En este caso, se sabe que el tiempo de ocurrencia se sitúa entre dos tiempos de censura (el máximo y el mínimo valor que conforman el intervalo), pero se desconoce el tiempo exacto y esto no permite su cuantificación [15,16].

Está documentado que uno de los tipos de estudios donde este fenómeno es más frecuente son los estudios de vida útil de alimentos, ya que el producto puede ser rechazado entre los tiempos de evaluación [17]. Aunque es cierto que en ellos también podríamos encontrar observaciones censuradas por la derecha (por ejemplo, si un individuo acepta el producto aun superando el tiempo máximo de almacenaje) o por la

izquierda (por ejemplo, si se testan productos de la competencia y se desconoce la fecha de producción, pero el individuo no lo acepta desde el primer momento) [18].

Pese a que este tipo de valores censurados deben proporcionarse en forma de intervalo, no deben confundirse con los **valores agrupados**. Zhang [3] indica que, aunque ambos valores se den entre un valor máximo y un mínimo se diferencian porqué los intervalos de los datos agrupados son constantes y no hay posibilidad de que se solapen los de distintos individuos; por el contrario, en el caso de la censura en intervalo el valor máximo y mínimo que forman el intervalo es propio para cada sujeto y pueden llegar a solaparse entre ellos.

A menudo es posible encontrar en un mismo estudio observaciones censuradas por la derecha, izquierda e incluso en un intervalo [16]. Cuando se presenta censura por la derecha y por la izquierda al mismo tiempo, normalmente se le denominan datos **doblemente censurados** [15].

3. METODOLOGÍA ESTADÍSTICA EN EL TRATAMIENTO DE DATOS CON VALORES CENSURADOS

Las técnicas estadísticas que se aplican para llevar a cabo el análisis de datos censurados son utilizadas de forma común en los ámbitos de investigación médica, así como en el sector industrial. Este conjunto de métodos estadísticos se denomina “análisis de supervivencia” cuando se refiere a datos médicos y “análisis de fiabilidad” si hablamos del ámbito de la ingeniería [1]. En ambos casos, el tiempo de ocurrencia es la variable de interés, de modo que el fundamento de estas técnicas es el mismo [17]. Es necesario destacar que omitir los casos censurados de un conjunto de datos, como podemos suponer, puede provocar un gran desvío en los resultados obtenidos y es señal de una incorrecta praxis en el análisis de los datos. Es preciso conocer los métodos estadísticos que pueden aplicarse con el objetivo de obtener unas conclusiones lo más verídicas posibles [18].

El artículo de Gijbels [15], expone los procedimientos estadísticos básicos que se emplean en el tratamiento de datos con casos censurados por la derecha. Si el propósito es realizar una estimación es posible emplear métodos paramétricos, como la Estimación de Máxima Verosimilitud (*Maximum Likelihood Estimation, MLE*), o métodos de estimación no paramétricos de una función de supervivencia o de una función del riesgo acumulada. Además, propone emplear dos posibles modelos de regresión para estos datos: el modelo de riesgos proporcionales (general o de Cox) y el modelo de tiempo de fallo acelerado. El primero de los modelos suele ser el más utilizado dado que la inferencia estadística bajo éste es más simple y está basada en la verosimilitud parcial.

Por otro lado, Zhang [3] describe que es posible realizar estimaciones de tipo no paramétrico a partir del método MLE no paramétrico o aplicando el algoritmo no paramétrico iterativo minoritario (*Iterative Convex Minorant, ICM*); así como una posible combinación de los anteriores (MLE-ICM). El problema es que se trata de algoritmos iterativos que resultan complicados de ejecutar y los programas estadísticos no incorporan estos métodos propiamente dichos; sin embargo, pueden aplicarse funciones desarrolladas para datos censurados por la derecha. De hecho, las técnicas relativas al análisis de supervivencia normalmente están enfocadas a la censura por la derecha, pero se han realizado generalizaciones aplicables a la censura en intervalo.

Tal y como se ha comentado en el apartado 2.2, los datos censurados por la izquierda son frecuentes en los estudios del ámbito de las ciencias medioambientales. Dado que este ámbito de estudio, el medioambiental, es precisamente el alcance del presente trabajo, la metodología estadística que se aplica a los datos con censura por la izquierda se explica de forma más detallada en el siguiente sub-apartado.

3.1 Tratamiento estadístico de datos censurados por la izquierda: datos medioambientales.

Como se ha comentado anteriormente, las observaciones censuradas por la izquierda son aquellas cuyos valores resultantes de la medición son inferiores al umbral o límite de detección (LD) del método analítico o el instrumento empleado; por este motivo, normalmente suelen presentarse como "<LD" o trazas. Hace algunos años era común interpretar estos valores como 0 a la hora de realizar el análisis estadístico de los datos, hecho que provocaba un gran desvío en los resultados [4]. Otra mala praxis muy empleada con los casos censurados era su eliminación del conjunto de datos, de modo que se perdía toda la información relativa a la observación como es la proporción de casos censurados respecto el total de observaciones. Es necesario recordar que las observaciones censuradas poseen información, aunque incompleta, y por ello es preciso emplear los métodos estadísticos que permitan extraer esta información e incorporarla al tratamiento de los datos [15].

Durante años, una práctica muy frecuente fue la sustitución (también llamada fabricación) de los valores censurados por una fracción del valor correspondiente al límite de detección. Esta práctica supone una adición invasiva de señal que no se encuentra en los datos originales, modificando así la información presente en los casos no censurados. Pese a que actualmente numerosas guías del ámbito medioambiental recomiendan el uso del análisis de supervivencia en datos censurados, la propia USEPA (*United States Environmental Protection Agency*) en el año 2009 permitía el uso de la sustitución para la estimación y contraste de hipótesis cuando la proporción de datos

censurados era inferior al 15% [1]. A partir de aquí, se plantea la cuestión de si existe una proporción de casos censurados en los datos adecuada con la que poder aplicar la sustitución. Helsel [1] opina que todo depende de la calidad que se precise en los resultados, por ello en aproximaciones semi-cuantitativas podría llegar a emplearse. Sin embargo, el hecho de que actualmente existan técnicas estadísticas específicas para este tipo de datos invita a su aplicación.

En el libro *Statistics for Censored Environmental Data Using Minitab and R* [1] se detalla la metodología estadística recomendada para datos medioambientales censurados por la izquierda y los criterios para reconocer qué técnicas son mejores en cada caso. A continuación, se muestra un resumen de estas metodologías que pueden aplicarse a datos con uno o más límites de detección:

- Métodos no paramétricos

Son la alternativa más sencilla a la sustitución, pero suponen una ventaja frente a ella ya que no añaden señal invasiva. Como bien indica su nombre, se emplea la posición de cada una de las observaciones respecto al conjunto de datos en lugar de utilizar parámetros. Además, no es necesario asignar una distribución a los datos y es posible aplicarlos directamente sobre ellos en caso de tener un único límite de detección; si se tiene más de uno es necesario re-censurar las observaciones al valor más alto.

Se presentan dos procedimientos posibles: el método binario y el ordinal.

  - Método binario: los datos son categorizados en dos clases (por encima y por debajo del LD). El inconveniente que presentan es que no emplean toda la información disponible en las observaciones no censuradas.

  - Método ordinal: encontramos el método de Mann-Whitney (*Rank-sum*) y el de Kruskal-Wallis. Respecto al anterior, estos métodos sí mantienen la información de rango de los datos no censurados y pueden aplicarse a los datos con un único LD de forma sencilla.

- Estimación de Máxima Verosimilitud (MLE)

Este método paramétrico emplea 3 tipos de información: los valores no censurados, la proporción de observaciones censuradas y la fórmula matemática de una distribución concreta. Es necesario evaluar si los datos se ajustan correctamente a la distribución asumida, por ello se recomienda aplicar este método sobre conjuntos de datos con un tamaño superior a 50 unidades.

Otros métodos paramétricos distintos a MLE serían los métodos basados en regresión de orden estadística (*Regression on Order Statistics, ROS*), cuyas variantes más utilizadas son la robusta (rROS) que asume una distribución log-normal y la GROS que asume una distribución gamma.

En el artículo publicado por Shoari *et al.* [5], se realizan unas simulaciones con el objetivo de evaluar la aplicación de distintos métodos estadísticos sobre datos medioambientales censurados por la izquierda con una distribución asimétrica y modelo inespecífico. Como resultado obtiene que el modelo MLE es más eficiente respecto a otros modelos (rROS y GROS) cuando los datos son poco asimétricos, se vuelve más robusto a medida que la proporción de observaciones censuradas se reduce y no hay diferencias respecto a otros métodos por lo que se refiere al tamaño de la muestra. Finalmente, se comprueba que el método MLE bajo la asunción de una distribución gamma, rROS y GROS son alternativas viables independientemente de la proporción de datos censurados.

- Métodos no paramétricos de análisis de supervivencia

Al tratarse de técnicas no paramétricas, como el primero de los casos explicados, se emplean las posiciones relativas de cada observación dentro del conjunto de datos en lugar de parámetros; además, no es necesario asignar una distribución a los datos. El hecho de que la mayoría de las mediciones en sistemas naturales posean distribuciones asimétricas, hace creer que realmente puede ser mejor aplicar procedimientos no paramétricos como los métodos de Kaplan-Meier (KM) o Turnbull.

Una de las desventajas de tratar con datos censurados por la izquierda, es que la mayoría de programas estadísticos están desarrollados para aplicar a datos censurados por la derecha ya que son más comunes en el ámbito de la epidemiología y la ingeniería. Así pues, en muchos casos será necesario transformar los datos censurados por la izquierda a datos censurados por la derecha para poder aplicarlos correctamente; este procedimiento se conoce como *flipping*, que se traduce como “volteamiento”.

Como se ha indicado anteriormente, existen criterios que se deben tener en cuenta para escoger el modelo estadístico indicado para el tratamiento de datos censurados por la izquierda. Dos de las características principales que deben ser tenidas en cuenta son el tamaño de la muestra y la proporción de datos censurados tal y como se muestra en la Tabla 1.

Tabla 1. Criterios para la elección de métodos estadísticos en función del tamaño de muestra y la proporción de datos censurados [Adaptación de la Ref. 1].

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="4">Tamaño de muestra</td></tr>
<tr><td colspan="2">< 50</td><td colspan="2">>50</td></tr>
<tr><td rowspan="5">Proporción datos censurados</td><td rowspan="4"><50%</td><td rowspan="2">Múltiples LD</td><td>KM</td><td rowspan="2">Múltiples LD</td><td>KM</td></tr>
<tr><td>Turnbull</td><td>Turnbull</td></tr>
<tr><td rowspan="2">Único LD</td><td>rRos</td><td rowspan="2">Único LD</td><td>rRos</td></tr>
<tr><td>rMLE</td><td>rMLE</td></tr>
<tr><td>50-80%</td><td colspan="2">rRos</td><td colspan="2" rowspan="2">MLE</td></tr>
<tr><td>>80%</td><td colspan="2">rMLE</td></tr>
<tr><td>>80%</td><td colspan="2">Proporción de observaciones no censuradas</td><td colspan="2">Estimaciones de los percentiles situados por encima del % casos censurados</td></tr>
</table>

El estudio realizado por Banta-Green *et al.* [4] sigue las recomendaciones presentes en la Tabla 1, las cuales están incluidas en la guía elaborada por Helsel [1]. El objetivo del estudio es evaluar la concentración de drogas presentes en el agua residual de algunas localidades y relacionar estos valores con su consumo; los datos obtenidos son censurados por la izquierda, aunque para cada tipo de sustancia la proporción de casos censurados varía. Así, para los datos con una censura moderada (<50%) emplean el método de KM, si la censura es media (entre 50 y 80%) aplican el método MLE con distribución log-normal, raíz cuadrada o normal; y si la censura es elevada (>80%) únicamente se indica la proporción de casos con valores no censurados (cuantificados).

La guía elaborada por Hustor y Juarez-Colunga [19] para el tratamiento de datos que incluyen casos censurados, incluye un diagrama de decisión para los métodos estadísticos a emplear según el tamaño de muestra y proporción de casos censurados (Figura 2). Tal y como se puede observar, la información contenida en ella coincide con la de la tabla 1 y resulta una forma fácil de reconocer el mejor método a emplear en función del conjunto de datos que se disponga y el objetivo que se persiga.

4. EVALUACIÓN DE METALES PESADOS EN MUESTRAS BIOLÓGICAS

Actualmente pueden encontrarse numerosas referencias sobre estudios de biomonitorización que se han llevado a cabo a lo largo de los años con el objetivo de evaluar la exposición de la población a diferentes contaminantes ambientales. Uno de los estudios más completos es el realizado por el organismo CDC (*Centers for Disease Control and Prevention, USA*) publicado en el año 2001 [20], donde se incluye la evaluación de 27 contaminantes ambientales en la población estadounidense:

13 metales pesados, metabolitos en orina de 7 ftalatos, 6 metabolitos en orina de pesticidas organofosforados y cotinina (un metabolito de la nicotina). Otro estudio de monitorización biológica llevado a cabo en población estadounidense fue el de Caldwell *et al.* [9], aunque en este caso se limita al análisis de mercurio en sangre.

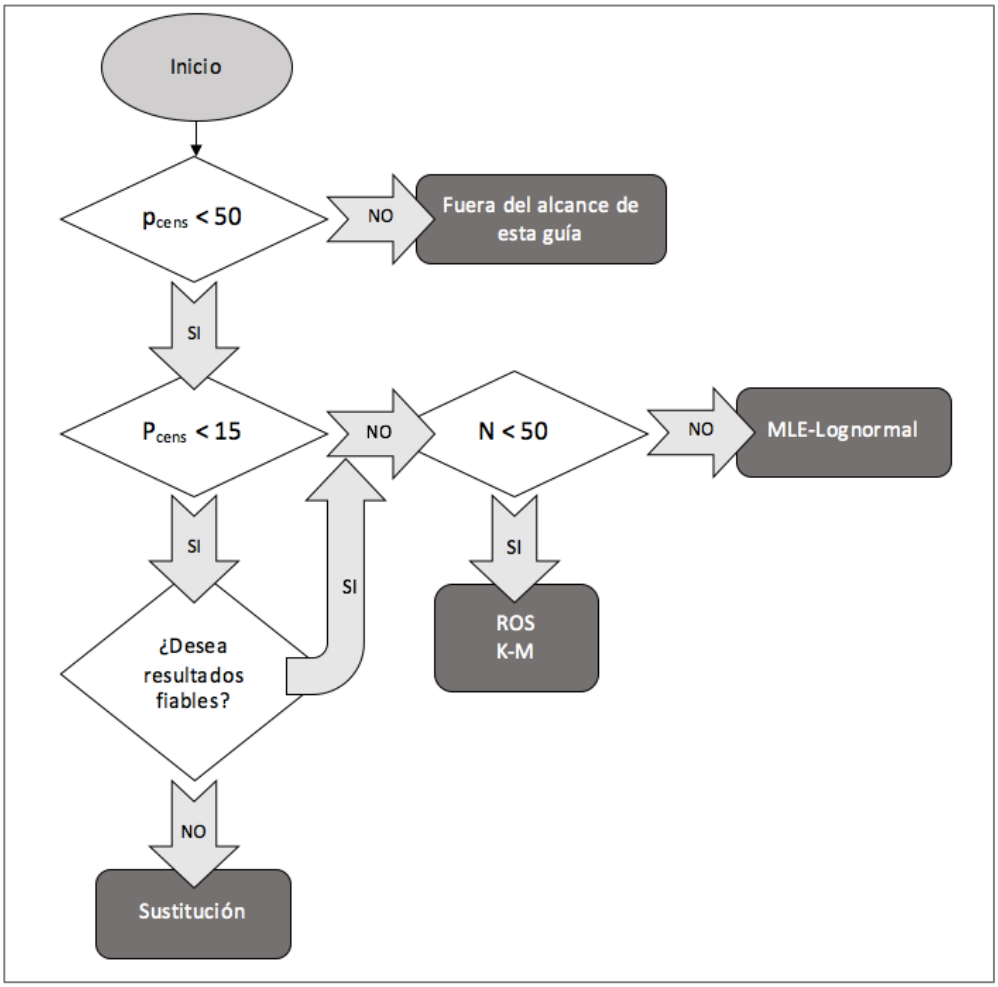


Figura 2. Diagrama de decisión sobre métodos estadísticos para datos que contienen casos no detectables (censurados) [Adaptación de la Ref.19].

A nivel europeo también se han elaborado estudios de biomonitorización sobre contaminantes ambientales. Por un lado, la *German Environmental Survey (GerES III)* analiza un amplio espectro de químicos ambientales en muestras de sangre [7] y orina [10] de la población alemana. Los estudios elaborados por Benes *et al.* [8] y por Batariova *et al.* [21] en la población de la República Checa se focalizan en la evaluación de metales pesados en sangre y orina. Otro ejemplo, lo encontramos en el estudio italiano publicado en el año 2002 sobre evaluación de mercurio en orina [11].

Los anteriores estudios de biomonitorización tiene como objetivo obtener información sobre los niveles de exposición a los contaminantes ambientales que se estudian, proporcionar unos valores de referencia en muestras biológicas para cada uno de estos químicos y evaluar el incremento o reducción de la concentración de contaminantes ambientales en la población, así como conocer las causas de estas tendencias.

En este tipo de estudios experimentales, las variables objeto de investigación se corresponden a medidas continuas cuyos valores se han obtenido a través de mediciones realizadas con aparatos ajustados y calibrados de forma correcta [22]. Tal y como se ha descrito en el apartado 3.1 del presente trabajo, es posible que una proporción de los valores obtenidos sean inferiores a los límites de detección (LD) de cada instrumento en función de la metodología analítica seguida y, por tanto, los datos pueden estar censurados por la izquierda. Si se analiza la metodología estadística que se ha seguido en el tratamiento de datos censurados por la izquierda de las anteriores referencias, observamos que la mayoría de ellos emplean la sustitución (o fabricación) de valores en base al LD en los casos censurados. Una de las opciones más frecuentes es la sustitución de los valores no detectables (<LD) por el valor equivalente a LD/2 [7,10,11,21]; además el artículo de Batariova *et al.* [21] indica que cuando la proporción de casos censurados supera el 50% del total de los datos no se realiza estadística descriptiva. Otra opción es la que emplea Caldwell *et al.* [9], dividir el LD entre la raíz cuadrada de 2 y emplear este valor para sustituir los casos <LD. La única referencia de las consultadas que emplea una metodología estadística específica para el tratamiento de datos censurado es Bleda-Hernández *et al.* [22].

4.1 La problemática medioambiental del Mercurio

Tal y como se ha visto en el anterior apartado, los metales pesados son unos de los contaminantes ambientales más evaluados en los estudios de biomonitorización a causa de los riesgos para la salud de estos pueden suponer; entre ellos, se encuentra el mercurio. La segunda parte del presente proyecto está centrada en el tratamiento de datos de un estudio de evaluación de mercurio en orina, por ello se dedica la última parte de la revisión bibliográfica a este metal. Cabe destacar que, aunque la finalidad de esta segunda parte sea la aplicación de metodología estadística adecuada en el tratamiento de datos medioambientales censurados por la izquierda y no la obtención de conclusiones desde un punto de vista biológico, por todos es sabido que el uso de la bioestadística implica un conocimiento de los fundamentos biológicos del estudio que se está llevando a cabo. Por este motivo, se presenta a continuación una breve explicación sobre la problemática medioambiental que entraña este metal, la cual ha sido extraída del programa de biomonitorización de la CDC [6].

El mercurio es un metal natural que puede encontrarse en 3 formas distintas: metálica (o elemental), orgánica e inorgánica. La forma metálica se emplea de forma muy habitual en usos industriales, formando parte de instrumentos (termómetros, barómetros, esfigmomanómetros) y también como elemento que conforma las amalgamas dentales. Además del riesgo de exposición ocupacional por inhalación de vapores en cada una de estas industrias, el mercurio metálico pasa a la atmosfera a través de la combustión de combustibles fósiles y la incineración de residuos sólidos. Posteriormente puede depositarse en la tierra y el agua, en ésta última los

microorganismos acuáticos transforman el mercurio metálico en la forma orgánica, más conocida como metil mercurio. El metil mercurio es la forma más tóxica y se bioacumula en los tejidos de los organismos terrestres y acuáticos que forman parte de la cadena alimentaria; de ahí que una de las fuentes de metil mercurio sea la dieta, sobre todo el pescado y marisco.

La cinética del mercurio depende de la forma química, siendo el metil mercurio la forma que se absorbe en mayor proporción en el tracto gastrointestinal (aproximadamente un 95%) seguido de la forma inorgánica (<15%) y por último el mercurio elemental, cuya absorción se produce mayoritariamente por vía inhalatoria. Respecto a la acumulación en tejidos, el metil mercurio es la única forma que puede atravesar la barrera hematoencefálica y alojarse en el cerebro, mientras que la forma elemental se deposita en los riñones; ambas formas pasan a la forma inorgánica una vez se encuentran en los tejidos. La excreción del mercurio puede producirse mayoritariamente por vía urinaria (mercurio metálico), fecal (metil mercurio) o ambas (mercurio inorgánico).

Los efectos que el mercurio puede causar en la salud son muy diversos y depende de la forma de éste, de características intrínsecas del individuo expuesto, de la dosis y de la duración de la exposición. Así, puede causar desde una neumonitis si hay una elevada exposición de forma aguda a vapores de mercurio metálico, hasta afectaciones en el sistema nervioso central en caso de intoxicación por metil mercurio.

La mayoría de estudios que evalúan la exposición de la población al mercurio realizan analíticas en sangre o en orina [9,11] dado que existen metodologías analíticas validadas y se trata de muestras biológicas que puede obtenerse de forma no invasiva y con relativa facilidad. En la actualidad es complicado encontrar unos valores de referencia de concentraciones normales de mercurio en sangre y orina, ya que la considerada "normalidad" depende de muchos factores; lo que sí se debe tener en cuenta es que el hecho de encontrar una cantidad medible de mercurio en sangre u orina no implica necesariamente que éste vaya a causar un efecto adverso en la salud [6]. En la tabla 2 se recogen algunos valores de referencia de normalidad.

Tabla 2. Valores normales de mercurio (Hg) en orina y sangre

REFERENCIA	ORINA	SANGRE
Ramírez, 2008 [23]	<20 μg Hg/L	<10 μg Hg/L
NTP 184 [24]		
Baselt, 1980	<10 μg Hg/L	20 μg Hg/L
Lauwerys, 1980	<5 μg Hg/g creatinina	<20 μg Hg/L
Göthe, 1985	-	<14 μg Hg/L
Merian, 1987	0.2-2 μg Hg/L	<0.2 – 1.9 μg Hg/L
Brodkin *et al.*, 2007 [25]	4nmol Hg/mmol creatinina (7.09 μg Hg/g creatinina)*	23nmol Hg/L (4.6 μg Hg/L)

*Cálculo elaborado con los siguientes datos: pm creatinina= 113.12g/mol; pa Hg=200.59g/mol

3

Caso práctico.
Tratamiento estadístico de datos medioambientales censurados

1. DESCRIPCIÓN DEL CONJUNTO DE DATOS

1.1 Descripción del estudio

El conjunto de datos empleado para realizar la aplicación práctica de los conocimientos teóricos adquiridos se ha obtenido de la encuesta de la salud NHANES (*National Health and Nutrition Examination Survey*), un programa de estudios de biomonitorización para evaluar el estado de salud y nutricional de la población adulta e infantil de los Estados Unidos [26]. El organismo coordinador de este programa es el NCHS (*National Center of Health Statistics*), el cual forma parte del CDC (*Centers for Disease Control and Prevention*). El objetivo de la encuesta es la obtención de la prevalencia de las enfermedades principales y los factores de riesgo de las mismas. La información obtenida se emplea tanto para el establecimiento de estándares nutricionales y de salud, como en futuros estudios epidemiológicos e investigación en ciencias de la salud que ayudan en el desarrollo de políticas de salud y diseño de programas de salud.

Esta encuesta de salud se puso en marcha a principios de la década de los 60 y su diseño combina entrevistas a los sujetos con exámenes físicos y analíticas. El contenido de ésta es muy amplio, ya que recoge enfermedades, condiciones médicas e indicadores de salud. Uno de los apartados de la encuesta trata la exposición a agentes medioambientales y, entre ellos, se encuentran los metales pesados. El conjunto de datos que se emplea en este proyecto es realmente un sub-conjunto de datos de la encuesta NHANES en el que se evalúa la concentración de mercurio en orina.

1.2 Descripción de la muestra

Se trata de un conjunto de 2755 individuos seleccionados entre la población adulta e infantil (a partir de los 6 años) residentes en los Estados Unidos.
Las muestras de orinas fueron correctamente recogidas, transportadas y almacenadas siguiendo los criterios de buenas prácticas descritos en la guía de la encuesta. Posteriormente fueron analizadas por ICP-MS (Inductively Coupled Plasma – Mass spectrometry). El límite de detección es constante para el analito en el conjunto de datos, siendo este valor 0.13μg/L.

1.3 Descripción de las variables de estudio

El conjunto de datos original contiene 7 variables que hacen referencia tanto a información general del individuo de estudio como a los resultados obtenidos en las pruebas analíticas de orina; se encuentran descritas en la tabla 3.

Tabla 3. Información sobre las variables contenidas en el conjunto de datos a tratar

VARIABLE	DESCRIPCIÓN	TIPO DE VARIABLE	CODIFICACIÓN [a]	UNIDADES
SEQN	Número de identificación	Cuantitativa discreta	-	-
RIAGENDR	Género	Factorial con 2 niveles	1: Masculino (M) 2: Femenino (F)	-
RIDAGEYR	Edad	Cuantitativa discreta	De 6 a 79 80: 80 años o más	Años
DMDMARTL	Estado civil [b]	Factorial con 8 niveles	1: Casado/a 2: Viudo/a 3: Divorciado/a 4: Separado/a 5: Soltero/a 6: Viviendo en pareja 77: Pregunta rechazada 99: No sabe	-
URXUHG	Nivel de mercurio en orina	Cuantitativa continua	0.09: <LD [c] Rango de 0.13 a 83.03	μg/L
URDUHGLC	Nivel de mercurio superior o inferior al límite de detección	Factorial con 2 niveles	0: Igual o superior al LD 1: Inferior al LD	-
URXUCR	Nivel de creatinina en orina	Cuantitativa continua	Rango de 5 a 546	mg/dL

[a] Cuando la variable aparece codificada con un punto (.) significa que se trata de un valor ausente (*missing value*)

[b] El estado civil no se proporciona para los individuos de 19 años o menores, aparece como valor ausente

[c] Valores de mercurio en orina que están por debajo del LD (no detectables); este valor corresponde a $\mathrm{LD}/\sqrt{2}$

2. TRATAMIENTO ESTADÍSTICO DEL CONJUNTO DE DATOS

El tratamiento del conjunto de datos descrito en el apartado anterior se lleva a cabo mediante el software estadístico RStudio [27]. Con el objetivo de facilitar la descripción del proceso que tiene lugar, se utiliza la herramienta RMarkdown de RStudio para crear un informe interactivo que permite mostrar el código utilizado, el resultado obtenido e incorporar las explicaciones necesarias. El diagrama de flujo que se presenta a continuación describe las etapas principales del tratamiento de datos (Figura 3).

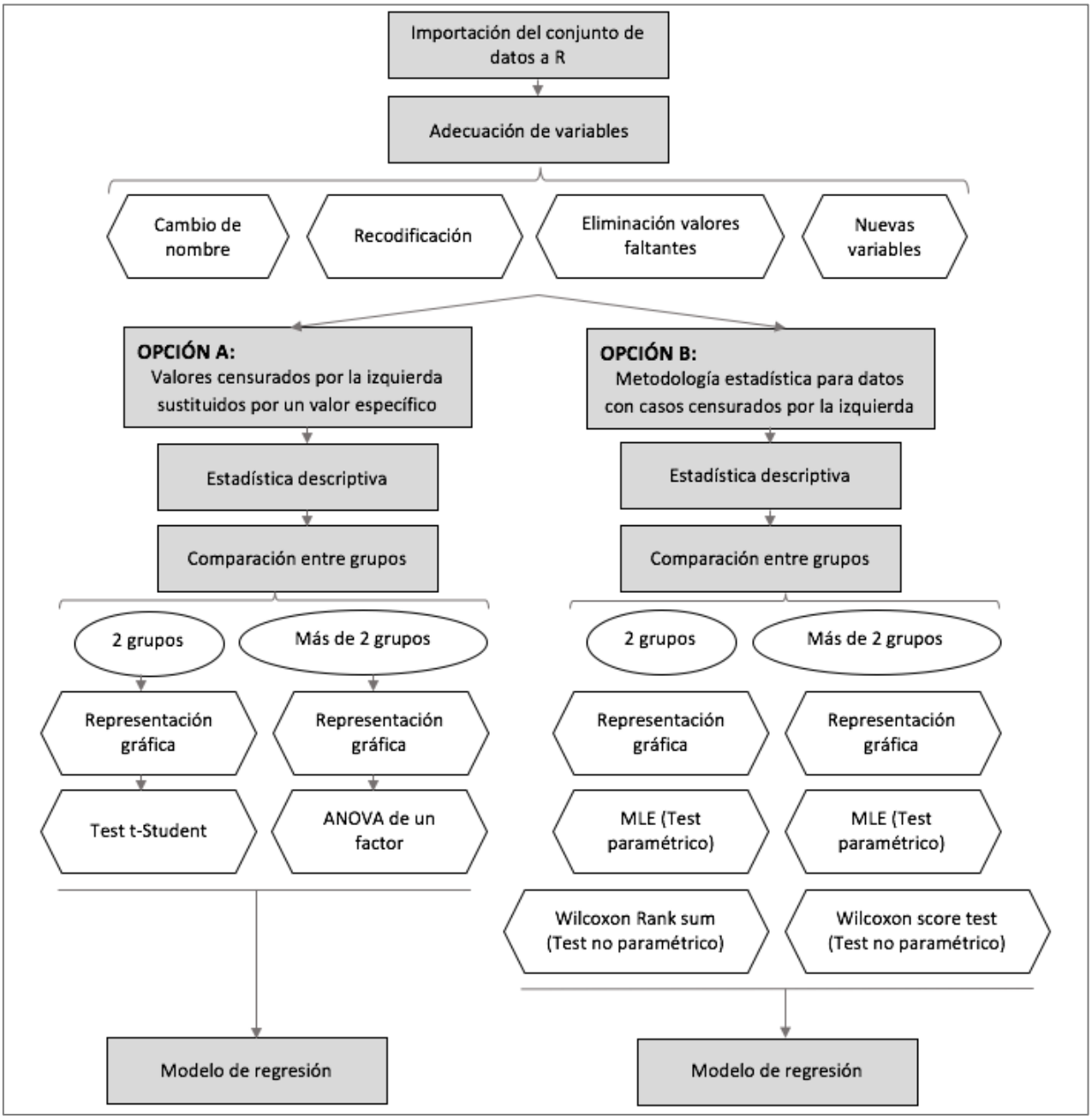


Figura 3. Diagrama de flujo del tratamiento de datos con casos censurados por la izquierda.

Seguidamente, se proporciona el código de R generado en cada una de las etapas incluidas en el diagrama de flujo de la Figura 3 y las correspondientes explicaciones exhaustivas sobre las funciones que se aplican. El documento ha sido generado a partir de la herramienta *RMarkdown* de R y la función Knitr, tal y como se muestra a continuación:

DOCUMENTO GENERADO EN R POR RMARKDOWN

1. Importación del conjunto de datos:

Los datos proporcionados se encuentran en un fichero en formato .csv, por este motivo para logar su importación en R se emplea la función "read.csv"; se indica que la primera fila del fichero es el encabezado (header=TRUE), que el elemento separador es la coma (sep=",") y que la primera columna es el número de fila (row.names=1):

```
datos<-read.csv("/Users/beatrizquintanillacasas/Desktop/BEATRIZ/MASTER BIOINF. Y BIOEST./TFM/DATOS/UHG_DEM.csv", header=TRUE, sep=",", row.names=1)

#Se muestran las primeras filas y la últimas del conjunto de datos importado para verificar que se ha realizado correctamente:
head(datos)
```

```
##    SEQN RIAGENDR RIDAGEYR DMDMARTL URXUHG URDUHGLC URXUCR
## 1 73560        1        9       NA   0.84        0     76
## 2 73564        2       61        2   0.69        0    242
## 3 73567        1       65        2   0.09        1    215
## 4 73583        2        7       NA   0.25        0    151
## 5 73585        1       28        1   0.09        1    100
## 6 73589        1       35        3   0.25        0    200
```

```
tail(datos)
```

```
##       SEQN RIAGENDR RIDAGEYR DMDMARTL URXUHG URDUHGLC URXUCR
## 2750 83715        1       58        5   0.09        1    115
## 2751 83718        2       60        2   0.31        0    300
## 2752 83720        1       36        5   0.17        0    399
## 2753 83721        1       52        1   0.61        0    122
## 2754 83729        2       42        3  83.03        0    117
## 2755 83731        1       11       NA   0.09        1    114
```

2. Adecuación de las variables:

2.1 Cambio de nombre de variables:

Para facilitar el tratamiento de datos, se renombran las variables que conforman el conjunto de datos dado que los nombres originales son demasiado largos y poco intuitivos:

```
colnames(datos)<-c("NUM", "GEN", "EDAD", "EC", "Hg", "CENS", "CREAT")
colnames(datos)
```

```
## [1] "NUM"   "GEN"   "EDAD"  "EC"    "Hg"    "CENS"  "CREAT"
```

2.2 Recodificación de variables:

Primero comprobamos mediante la función "str" el tipo de variables que conforman el conjunto de datos:

```
str(datos)
```

```
## 'data.frame':    2755 obs. of  7 variables:
##  $ NUM  : int  73560 73564 73567 73583 73585 73589 73592 73594 7359
5 73603 ...
##  $ GEN  : int  1 2 1 2 1 1 1 1 1 ...
##  $ EDAD : int  9 61 65 7 28 35 29 23 58 35 ...
##  $ EC   : int  NA 2 2 NA 1 3 5 5 1 1 ...
##  $ Hg   : num  0.84 0.69 0.09 0.25 0.09 0.25 0.16 0.09 0.09 0.22 ..
.
##  $ CENS : int  0 0 1 0 1 0 0 1 1 0 ...
##  $ CREAT: int  76 242 215 151 100 200 46 96 64 106 ...
```

Como vemos, es necesario recodificar alguna de ellas como factores con distintos niveles ya que la mayoría aparecen como variables de tipo *integer*:

```
#Variable GEN:
datos$GEN<-factor(datos$GEN, levels=c("1","2"), labels=c("M","F"))


#Variable EC:
datos$EC<-factor(datos$EC, levels=c("1","2","3","4","5","6","7","8"),
labels=c("Cas","Viud","Div", "Sep","Solt","VPar","R","NS/NC"))
```

En el caso de la variable CENS, es más adecuado transformarla en una variable de tipo *logical* que indique TRUE en los casos en que existe censura por la izquierda en la variable Hg y FALSE cuando no haya censura:

```
#Teniendo en cuenta que el LD es 0.13, los valores censurados serán aq
uellos que se encuentran por debajo del valor (los que son = o superio
res a 0.13 no estarán censurados):
datos$CENS<-datos$Hg<0.13
```

```
#Comprobamos las recodificaciones de las variables:
str(datos)
```

```
## 'data.frame':    2755 obs. of  7 variables:
##  $ NUM  : int  73560 73564 73567 73583 73585 73589 73592 73594 7359
5 73603 ...
##  $ GEN  : Factor w/ 2 levels "M","F": 1 2 1 2 1 1 1 1 1 1 ...
##  $ EDAD : int  9 61 65 7 28 35 29 23 58 35 ...
##  $ EC   : Factor w/ 8 levels "Cas","Viud","Div",..: NA 2 2 NA 1 3 5
5 1 1 ...
##  $ Hg   : num  0.84 0.69 0.09 0.25 0.09 0.25 0.16 0.09 0.09 0.22 ..
.
##  $ CENS : logi  FALSE FALSE TRUE FALSE TRUE FALSE ...
##  $ CREAT: int  76 242 215 151 100 200 46 96 64 106 ...
```

De este modo, observar la matriz de datos y comprender el significado de las mismas es más visual ya que se han añadido unas etiquetas descriptivas en lugar de números en las variables factoriales:

```
head(datos)
```

```
##     NUM GEN EDAD   EC   Hg  CENS CREAT
## 1 73560   M    9 <NA> 0.84 FALSE    76
## 2 73564   F   61 Viud 0.69 FALSE   242
## 3 73567   M   65 Viud 0.09  TRUE   215
## 4 73583   F    7 <NA> 0.25 FALSE   151
## 5 73585   M   28  Cas 0.09  TRUE   100
## 6 73589   M   35  Div 0.25 FALSE   200
```

2.3 Eliminación de casos con valores faltantes:

Si se explora el *dataset* creado (datos) se puede observar que existen datos faltantes (NA) para las variables Hg y CREAT, lo que puede significar que no se pudo realizar la analítica de orina de estos individuos:

```
summary(datos$Hg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0900  0.0900  0.2000  0.5065  0.4700 83.0300      89
```

Hay 89 valores ausentes en la variable Hg; como el objetivo de estudio son las mediciones realizadas en orina, se elabora un nuevo *dataset* (DATOS) eliminando las filas que contienen este tipo de datos faltantes:

```
#Creación del nuevo dataset:
DATOS<-subset(datos, !is.na(Hg))


#Dimensión del nuevo dataset:
dim(DATOS)
```

```
## [1] 2666    7
```

```
#Comprobación de ausencia de valores faltantes en la variable Hg del n
uevo dataset:
summary(DATOS$Hg)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0900  0.0900  0.2000  0.5065  0.4700 83.0300
```

Dado que es interesante expresar la concentración de Hg según la concentración de creatinina en orina, verificamos si existen datos faltantes en esta variable:

```
summary(DATOS$CREAT)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     5.0    55.0   100.0   114.7   156.0   524.0       1

#Hay 1 NA en esta variable, podemos inspeccionar de qué caso se trata:
VFD<-is.na(DATOS)
VFD<-data.frame(VFD)
which(VFD$CREAT==TRUE)

## [1] 548

DATOS[548,]

##       NUM GEN EDAD  EC   Hg CENS CREAT
## 566 75610   M   56 Cas 0.09 TRUE    NA
```

Considerando que únicamente hay un valor faltante y que el correspondiente nivel de mercurio está censurado, podemos eliminar este caso del conjunto de datos:

```
DATOS<-DATOS[-548,]
dim(DATOS)

## [1] 2665    7
```

2.4 Creación de nuevas variables:

La mayoría de los estudios de biomonitorización que incluyen el análisis de metabolitos en orina, suelen tomar también el valor de creatinina y expresan la concentración del metabolito de estudio en función de ésta. La creatinina en orina es un indicador de la función renal y puede ser variable entre individuos, por este motivo, es útil aplicar esta corrección a los valores crudos del metabolito con el objetivo de obtener resultados comparables entre sujetos.

Así pues, se crea una nueva variable llamada **HgCr** para expresar la concentración de mercurio en función de la concentración de creatinina, una variable cuantitativa continua cuyas unidades serán μg Hg/g creatinina:

```
DATOS$HgCr<-(DATOS$Hg*1000)/(10*DATOS$CREAT)
```

De ahora en adelante se utilizará esta variable para expresar la concentración de Hg en orina de los individuos y realizar tanto la estadística descriptiva como la comparación entre grupos.

Otra variable que merece la pena crear es una que indique la edad de los individuos agrupados según los grandes grupos de edad estándar, es decir, categorización de la variable EDAD:

```
EDAD_CAT<-vector()
EDAD_CAT[DATOS$EDAD<15]<-1
EDAD_CAT[DATOS$EDAD>=15 & DATOS$EDAD < 65]<-2
EDAD_CAT[DATOS$EDAD>=65]<-3
DATOS$EDAD_CAT<-as.factor(EDAD_CAT)
levels(DATOS$EDAD_CAT)<-c("6 a 14", "15 a 64", "65 o más")

head(DATOS)
```

```
##     NUM GEN EDAD   EC   Hg  CENS CREAT       HgCr EDAD_CAT
## 1 73560   M    9 <NA> 0.84 FALSE    76 1.10526316   6 a 14
## 2 73564   F   61 Viud 0.69 FALSE   242 0.28512397  15 a 64
## 3 73567   M   65 Viud 0.09  TRUE   215 0.04186047 65 o más
## 4 73583   F    7 <NA> 0.25 FALSE   151 0.16556291   6 a 14
## 5 73585   M   28  Cas 0.09  TRUE   100 0.09000000  15 a 64
## 6 73589   M   35  Div 0.25 FALSE   200 0.12500000  15 a 64
```

```
#Fijamos el dataframe con el que se va a trabajar para poder emplear l
as variables directamente sin tener que indicar a que conjunto de dato
s se refiere:
attach(DATOS)
```

3. Estadística descriptiva:

Tanto en el presente apartado de estadística descriptiva como en el resto de apartados que le siguen, se presentarán dos opciones de tratamiento de los datos censurados por la izquierda, que se definen a continuación:

1. OPCIÓN A. Valores censurados por la izquierda sustituidos por un valor específico: tal y como se ha comentado en la revisión bibliográfica, una de las prácticas más comunes en el tratamiento de datos medioambientales con casos censurados por la izquierda es la sustitución de valores, también llamada fabricación. En el presente conjunto de datos los valores censurados de las mediciones de mercurio en orina han sido sustituidos por el valor 0.09 ($LD/\sqrt{2}$, siendo LD=0.13).

2. OPCIÓN B. Metodología estadística para datos con casos censurados por la izquierda: otra opción más recomendable a la sustitución es el empleo de alguno de los métodos estadísticos mencionados en la revisión bibliográfica realizada en el presente proyecto, dependiendo de dos características del conjunto de datos sobre el que se trabaja: el tamaño de la muestra y la proporción de datos censurados.

Llevando a cabo estas dos opciones, es posible realizar una comparación de los resultados obtenidos y así poder comprobar con un mismo ejemplo las ventajas e inconvenientes que suponen cada uno de los métodos empleados. Sin embargo, es necesario recordar que la sustitución de valores censurados es una técnica no recomendada ya que puede introducir un error que no está presente en el conjunto de datos de forma natural y esto puede llevarnos a la obtención de resultados sesgados y conclusiones incorrectas.

OPCIÓN A

Se aplican los estadísticos descriptivos básicos sobre el conjunto de datos:

```
#Creación de una nueva función que proporciona estadísticos descriptiv
os básicos:
estad.basic<-function(x){
est<-cbind(mean(x), sd(x), t(quantile(x)))
colnames(est)<-c("Media","Desv.est", "Min.","Q1","Q2","Q3"," Máx.")
return(round(est,2))}


#Aplicación de la función estad.basic a la variable Hg:
estad.basic(HgCr)
```

```
##      Media Desv.est Min.   Q1   Q2   Q3  Máx.
## [1,]  0.53     1.73 0.02 0.13 0.26 0.55 70.97
```

OPCIÓN B

Teniendo en cuenta la tabla 1 y la figura 2 presentadas en la revisión bibliográfica, a partir del tamaño de muestra y la proporción de casos censurados es posible conocer qué metodología aplicar. Así, dado que el tamaño de muestra es conocido (N=2665), se procede a calcular la proporción de casos censurados a partir de la variable CENS que indica en qué casos la variable Hg está censurada por la izquierda por ser el valor inferior al LD (TRUE).

```
#Número de casos censurados:
ncens<-length(which(CENS==TRUE))


#Número total de casos (tamaño de muestra, N):
N<-dim(DATOS)[1]


#Proporción de casos censurados:
ncens/N
```

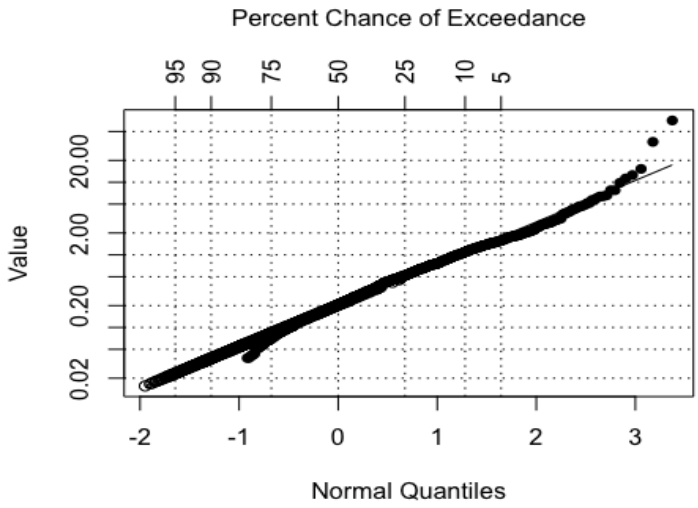
```
## [1] 0.3354597
```

La proporción de casos censurados por la izquierda es de 0.335 (33.5%). Atendiendo al diagrama de la figura 2, dado que la proporción de casos censurados se sitúa entre el 15

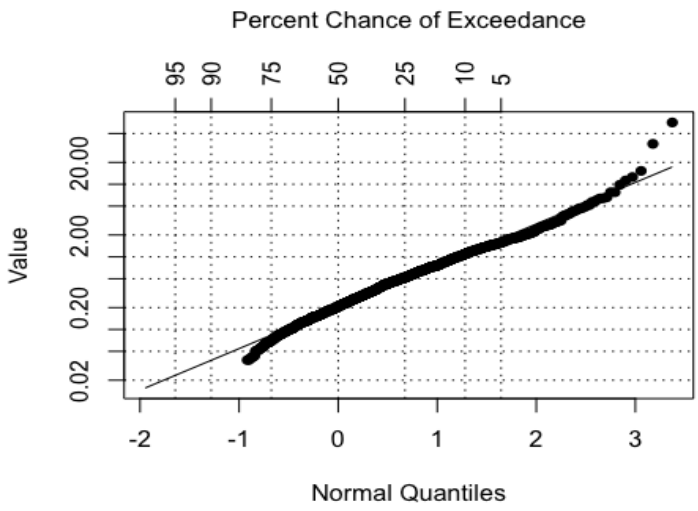
y el 50% y el tamaño de muestra es superior a 50 individuos, sería conveniente emplear el método MLE (*Maximum Likelihood Estimation*). Esta decisión coincide con las alternativas mostradas en la tabla 1, ya que con las características mencionadas y sabiendo que existe un único LD el método MLE sería el elegido. También sería posible emplear el método rROS (*Robust Regression on Order Statistics*) ya que funciona correctamente con cualquier tamaño de muestra, pero con muestras de gran tamaño el MLE es el más recomendado.

Como MLE es un método paramétrico, es necesario indicar la distribución con la que se trabaja. En datos de tipo medioambiental es muy común emplear la **distribución normal logarítmica (*lognormal*)** [19] y por ello en este caso se decide trabajar con esta distribución. Para llevar a cabo la aplicación de la metodología estadística se emplea el paquete NADA (*Nondetects and Data Analysis for Environmental Data*) [28].

```
library(NADA)

#Ajuste del modelo de regresión general siguiendo una distribución nor
mal logarítmica:
D_mle<-cenmle(HgCr, CENS, dist="lognormal")
D_mle

##            n        n.cen       median         mean           sd
## 2665.0000000  894.0000000    0.1939622    0.5065525    1.2220913

#Comprobación de que los datos se ajustan a una distribución logarítmi
ca:
D_ros<-ros(HgCr, CENS)
plot(D_ros, plot.censored=TRUE)#Incluyendo los datos censurados
```




```
plot(D_ros)#Sin incluir los datos censurados
```



Se observa que la suposición de que los datos siguen una distribución normal logarítmica es correcta, ya que las observaciones siguen la recta trazada. Por tanto, se procede a la obtención de los parámetros estimados:

```
summary(D_mle)

##              Value Std. Error     z        p
## (Intercept) -1.640     0.0293 -56.0 0.00e+00
## Log(scale)   0.326     0.0178  18.3 4.28e-75
##
## Scale = 1.39
##
## Log Normal distribution
## Loglik(model)= -2155.8   Loglik(intercept only)= -2155.8
## Loglik-r:  0
##
## Number of Newton-Raphson Iterations: 5
## n = 2665

mean(D_mle)

##      mean        se   0.95LCL   0.95UCL
## 0.5065525 0.0236731 0.4705893 0.5452640

sd(D_mle)

## [1] 1.222091

quantile(D_mle, conf.int=TRUE)

##   quantile      value    0.95LCL    0.95UCL
## 1     0.05 0.01985646 0.01781118 0.02213661
## 2     0.10 0.03284916 0.02990004 0.03608916
## 3     0.25 0.07617857 0.07084743 0.08191087
## 4     0.50 0.19396222 0.18313738 0.20542689
## 5     0.75 0.49385727 0.46566495 0.52375641
## 6     0.90 1.14527555 1.06404609 1.23270608
## 7     0.95 1.89466506 1.73843463 2.06493567
```

Si se comparan estos resultados con los obtenidos en la OPCIÓN A (sustitución de casos censurados):

```
estad.basic(HgCr)
```

```
##      Media Desv.est Min.   Q1   Q2   Q3  Máx.
## [1,]  0.53     1.73 0.02 0.13 0.26 0.55 70.97
```

Las medias obtenidas aplicando ambos métodos son muy similares (**MLE: 0.506** respecto a **Sustitución: 0.51**); mientras que la desviación estándar difiere en mayor medida en ambos casos (**MLE: 1.22** respecto a **Sustitución: 1.73**), como era de esperar dada la distribución de los datos en cada caso. Si se realiza la comparación entre los cuantiles, se observa que los valores obtenidos mediante MLE son menores que los obtenidos por sustitución de los casos censurados; en el Q1 y Q3 la diferencia es de 0.06 unidades y en la mediana (Q2) es de 0.07. Mientras que la diferencia existente entre los Q1 es esperable, dado que el porcentaje de datos censurados por la izquierda es del 33%, sorprende que los resultados por encima del percentil 33 sean tan distintos. Este efecto puede deberse a la bondad de ajuste del método MLE, la cual se reflejada en el valor logik-r proporcionado en la función "summary" del modelo ajustado y en este caso es 0 (bajo ajuste).

Aunque el objetivo de este trabajo no es la extracción de conclusiones relativas al estado de salud de los individuos en función de los niveles de mercurio que presentan, sí es posible comparar el valor medio de mercurio en orina corregido por creatinina que se ha obtenido con los valores de referencia de la tabla 3. Según las referencias que proporcionan los valores corregidos [24,25], se considera un valor normal por debajo de 5 μg Hg/g creatinina [24] y 7.09 μg Hg/g creatinina [25]; de modo que la media hallada (tanto por MLE como por sustitución de valores censurados) sería igual o inferior a los umbrales establecidos.

4. Comparación entre grupos:

Teniendo en cuenta las variables que se incluyen en el presente conjunto de datos, se cree conveniente realizar comparaciones entre grupos por lo que respecta a la concentración de mercurio en orina (corregida por la creatinina) que presentan. Por un lado, la comparación por género (2 grupos: Masculino y Femenino) y, por el otro, la comparación entre las categorías de edad (3 grupos: de 6 a 14, de 15 a 46 y >65).

4.1 Comparación de la concentración de mercurio en orina por género (Comparación entre dos grupos)

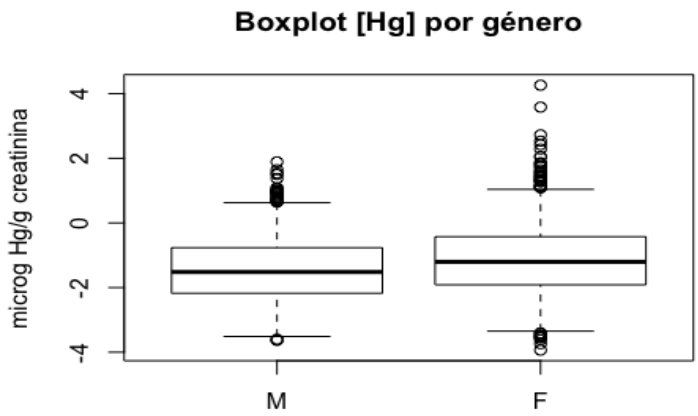
4.1.1 Representación gráfica:

Previo a la aplicación de tests estadísticos, es recomendable realizar una representación gráfica preliminar con el objetivo de obtener una idea de la distribución de los datos. Los diagramas de caja (*boxplots*) suelen resultar muy útiles para comparar los datos entre dos o más grupos a primera vista.

OPCIÓN A

Se realiza un *boxplot* con los valores del conjunto de datos, sabiendo que los valores censurados están sustituidos por la cifra 0.09, obteniendo el siguiente gráfico:

```
#En el apartado anterior se ha comprobado que los datos siguen una dis
tribución normal logarítmica, por tanto, el valor de la variable Hg se
expresa en este valor:
boxplot(log(DATOS$HgCr)~DATOS$GEN,
          main="Boxplot [Hg] por género",
          ylab="microg Hg/g creatinina", range=1)
```



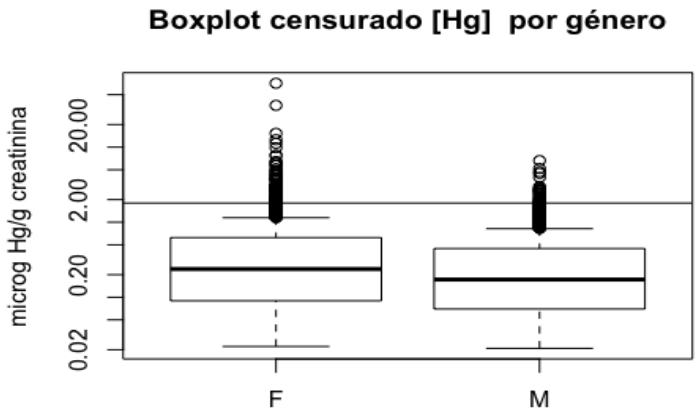
A simple vista, parece que la media de mercurio en orina del género femenino es superior a la del género masculino. Además, se observa que las varianzas de ambos grupos podrían ser homogéneas.

OPCIÓN B

El paquete NADA (*Nondetects and Data Analysis for Environmental Data*) posee la función "cenboxplot" que permite elaborar *boxplots* para observar de forma intuitiva la proporción de datos censurados mediante la elaboración de una caja a partir de los percentiles 25, 50 (línea horizontal) y 75. Por defecto, el *boxplot* se elabora en escala logarítmica, de modo que ya se ajusta a la distribución de nuestros valores:

```
library("NADA")
#Boxplot censurado en escala logarítmica:
```

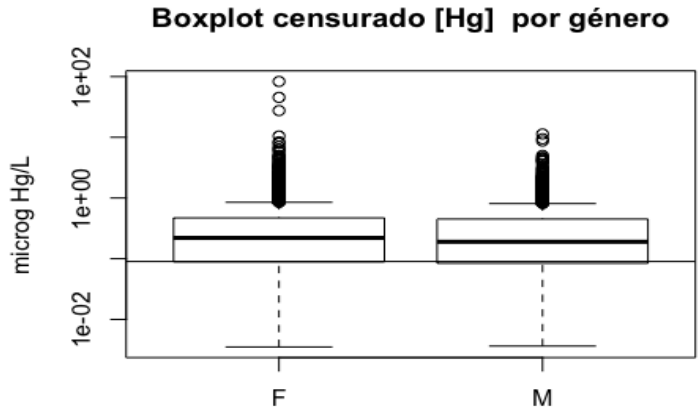
```
cenboxplot(DATOS$HgCr, DATOS$CENS, group=DATOS$GEN,
           main="Boxplot censurado [Hg]  por género",
           ylab="microg Hg/g creatinina", range=1)
```



En este caso, observamos que la línea trazada no coincide con el valor del LD (0.13) y la proporción de valores censurados no corresponde al 33% que se ha calculado anteriormente. Este hecho se debe a que se está empleando una variable expresada en función del valor de otra variable (mercurio en función de creatinina en orina) y, por tanto, el valor de ésta es distinto al original. Pese a que la variable auxiliar CENS sigue guardado la información sobre la censura, es posible que un caso no censurado tenga un valor de variable transformada menor al de una variable censurada; por este motivo la línea de censura no está formada por un único valor. No obstante, las cajas representadas son correctas y, a simple vista, parece que el grupo F (género femenino) tiene una media de Hg en orina mayor que el grupo M (género masculino). Además, se observa que las varianzas de los dos grupos son homogéneas, ya que la amplitud de las cajas es similar.

Si deseáramos obtener un gráfico de cajas con la línea de censura en el valor de censura real, se debería emplear la variable Hg en orina (sin corregir):

```
#Boxplot censurado en escala logarítmica de la variable Hg:
cenboxplot(DATOS$Hg, DATOS$CENS, group=DATOS$GEN,
           main="Boxplot censurado [Hg]  por género",
           ylab="microg Hg/L", range=1)
```



En este caso, sí se observa que la línea trazada coincide con el valor asignado a los casos censurados (0.09) y la proporción de datos que quedan por debajo de la línea de censura son los valores que no han sido detectados (censurados por la izquierda).

4.1.2 Tests estadísticos:

OPCIÓN A

En conjuntos de datos donde no hay presencia de casos censurados, es muy común emplear los t-test para la comparación de **medias** de una variable en dos grupos independientes. Sabemos que este tipo de test estadístico no es recomendable cuando se tienen casos censurados dado que estos valores producen afectan a la distribución de los datos y a su dispersión; pero en este caso se aplicará para observar qué resultados proporciona y compararlos con el estadístico aplicado en la Opción B. Se formulan las siguientes hipótesis, que dan lugar a la realización de un test t de Student de dos colas:

$$H0: media[Hg]mujeres = media[Hg]hombres$$

$$H1: media[Hg]mujeres \neq media[Hg]hombres$$

En realidad, las hipótesis se refieren a la igualdad del logaritmo de las medias de ambos grupos.

```
#Test de t de Student con dos colas, empleando el logaritmo de la conc
entración de Hg:
Hg.t.test<-t.test(log(DATOS$HgCr)~DATOS$GEN, alternative="two.sided",
conf.level=0.95, var.equal=TRUE)
Hg.t.test

##
##  Two Sample t-test
##
## data:  log(DATOS$HgCr) by DATOS$GEN
## t = -7.8429, df = 2663, p-value = 6.318e-15
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##  -0.3935213 -0.2361042
## sample estimates:
## mean in group M mean in group F
##       -1.456940       -1.142128
```

En el *boxplot* anterior se puede observar que las varianzas de las muestras son homogéneas (homocedasticidad de varianzas), así lo indicamos en la fórmula (var.equal=TRUE) para poder efectuar el test t Student. El resultado proporcionado indica que existen diferencias significativas entre las medias de ambos grupos con un nivel de significación del 5% (pvalor <0.05), por lo que se rechaza la hipótesis nula (H0).

OPCIÓN B

Tal y como se hizo para la estadística descriptiva, una alternativa para realizar un test de comparación de valores centrales entre 2 grupos puede ser un método de tipo MLE. En este caso, las hipótesis a testar son diferentes de la opción anterior ya que la información que obtendremos ajustando el modelo MLE es la razón de las medias:

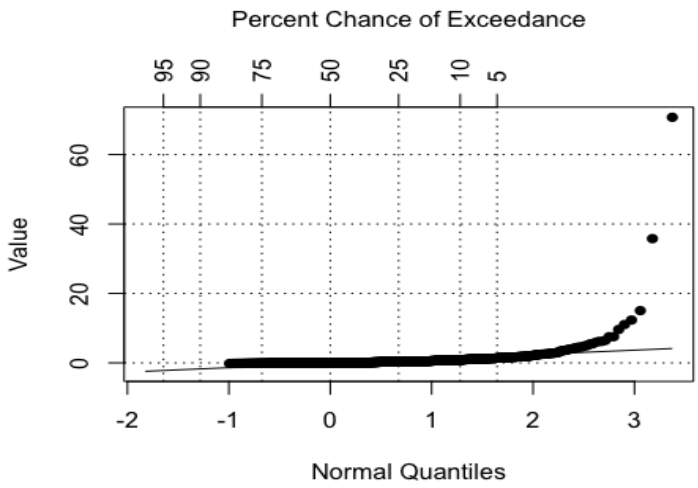
$$H0: \frac{media[Hg]Género\ Femenino}{media[Hg]Género\ Masculino} = 1$$

$$H1: \frac{media[Hg]Género\ Femenino}{media[Hg]Género\ Masculino} \neq 1$$

Como se trata de un método paramétrico, es necesario verificar que se cumplen dos suposiciones:

- Homocedasticidad de varianzas: las varianzas de los dos grupos no deben ser estadísticamente distintas. Esto puede comprobarse mediante los diagramas de cajas, vemos que la amplitud de las cajas relativas a cada grupo es muy similar.
- Siguen una distribución específica: en este caso, ya se ha visto que el modelo MLE general (D_mle) se ajustaba bien a la distribución normal logarítmica. Ahora es necesario comprobarlo en el nuevo modelo que se ajusta introduciendo los grupos:

```
#Ajuste del modelo MLE en base a los dos grupos, en una escala normal
logarítmica. Por defecto, la función ya realiza el test de dos colas.
Gen_mle<-cenmle(DATOS$HgCr, DATOS$CENS, DATOS$GEN, dist="lognormal")

#Comprobación del cumplimiento de la suposición de seguimiento de la d
istribución normal-logarítmica a través de un gráfico de probabilidad:
res<-residuals(Gen_mle)
plot(ros(res, CENS, forwardT=NULL))
```



La mayoría de puntos se ajustan a la recta teórica, excepto dos puntos que pueden ser posibles *outliers* ya que se separan mucho del resto; esto puede ser un problema ya que el método MLE es susceptible a valores *outliers*. Pasamos a observar los resultados proporcionados:

```
Gen_mle

##              Value Std. Error      z        p
## (Intercept) -1.813     0.0408 -44.45 0.00e+00
## DATOS$GENF   0.350     0.0562   6.22 5.07e-10
## Log(scale)   0.315     0.0178  17.71 3.57e-70
##
## Scale = 1.37
##
## Log Normal distribution
## Loglik(model)= -2136.7   Loglik(intercept only)= -2155.8
## Loglik-r:  0.1194096
##
## Chisq= 38.27 on 1 degrees of freedom, p= 6.2e-10
## Number of Newton-Raphson Iterations: 3
## n = 2665
```

La estimación de la diferencia entre las medias de los grupos es 0.350 y va asociada a un pvalor <0.05, por lo que se indica que existen diferencias significativas entre las medias de la concentración de Hg en orina de ambos grupos y se rechazaría la hipótesis nula (H0). Atendiendo a las hipótesis, las cuales se expresan en ratios, es necesario transformar este valor de escala logarítmica a normal:

```
exp(0.350)

## [1] 1.419068
```

Esto significa que la media de la [Hg] en el género femenino es aproximadamente 1.42 veces mayor que la media de la [Hg] en el género masculino, y esta diferencia es estadísticamente significativa; resultado que coincide con lo que se ha podido observar

en los *boxplots*. Además, es posible calcular los intervalos de confianza empleando el error estándar estimado del modelo MLE:

```
#Función para calcular los intervalos de confianza [19]:
library("NADA")
IC_mle<-function(cenobj, conf.level=0.95, ngroups=2){
  results<-summary(cenobj)
  variance<-results[8]
  v2<-unlist(variance)
  v3<-matrix(data<-v2, ncol=ngroups+1, nrow=ngroups+1)
  coef<-unlist(results[7])
  diffs<-coef[2:ngroups]
  vars1<-diag(v3)
  ses<-sqrt(vars1[2:ngroups])
  critical1<-(1-conf.level)/2
  critical2<-abs(qnorm(critical1))
  width<-critical2*ses
  lls<-diffs-width
  uls<-diffs+width
  interval<-cbind(lls,uls)
  colnames(interval)<-c("lower","upper")
  return(interval)}

IC_mle(cenobj=Gen_mle, conf.level = 0.95, ngroups=2)

##                               lower     upper
## coefficients.DATOS$GENF 0.2394402 0.4599181
```

Realizando la transformación de escala, estos corresponderían a:

```
exp(0.239)

## [1] 1.269979

exp(0.459)

## [1] 1.582491
```

El IC no incluye el 1, por lo que se rechaza la H0 (sí hay diferencias significativas entre las medias de los dos grupos por lo que respecta a la concentración de Hg).

También es posible aplicar **métodos no paramétricos** para realizar un contraste de valores centrales entre grupos, los cuales son más robustos en caso de la presencia de valores *outliers*. Aunque estos métodos pueden aplicarse a muestras tanto grande como pequeñas, con tamaños de muestra muy grandes y si los valores siguen una distribución concreta (como es el caso del presente conjunto de datos) pueden no resultar tan potentes a la hora de detectar diferencias entre grupos. De todos modos, se procede a su aplicación con el objetivo de visualizar las diferencias respecto al método paramétrico MLE. Las hipótesis planteadas son diferentes al método paramétrico, ya que en este caso se trata de la razón de medianas en lugar de medias:

$$H0: \frac{mediana[Hg]Género\ Femenino}{mediana[Hg]Género\ Masculino} = 1$$

$$H1: \frac{mediana[Hg]Género\ Femenino}{mediana[Hg]Género\ Masculino} \neq 1$$

```
#Dado que sólo se tiene un LD, se aplica el test de "Wilcoxon rank sum
" (los valores se ordenan y se comparan los rangos relativos de los do
s grupos, por eso no es necesario relizar la transformación normal log
arítmica de los datos):
Hg.rank.test<-wilcox.test(DATOS$HgCr~DATOS$GEN, alternative="two.sided
", conf.int=TRUE, conf.level=0.95)
Hg.rank.test

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  DATOS$HgCr by DATOS$GEN
## W = 739920, p-value = 1.008e-13
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -0.07945620 -0.04465413
## sample estimates:
## difference in location
##            -0.06160747
```

Del mismo modo que el resultado obtenido en el test paramétrico MLE, el pvalor proporcionado es inferior a 0.05 por lo que existen diferencias significativas entre las medianas (medida de tendencia central) entre los dos grupos con un nivel de significación del 5%.

4.2 Comparación de la concentración de mercurio en orina por grupos de edad (Comparación entre más de dos grupos)

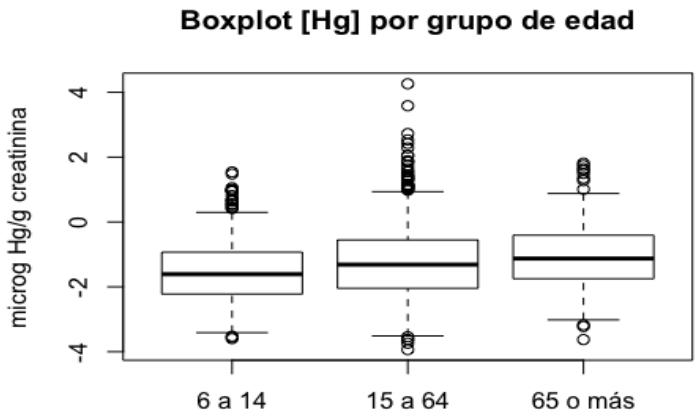
4.2.1 Representación gráfica:

Tal y como se ha efectuado en el anterior apartado, se realiza una representación gráfica preliminar en forma de diagramas de caja (*boxplots*).

OPCIÓN A

Se realiza un *boxplot* con los valores del conjunto de datos, sabiendo que los valores censurados están sustituidos por la cifra 0.09, obteniendo el siguiente gráfico:

```
#En el apartado anterior se ha comprobado que los datos siguen una dis
tribución normal logarítmica, por tanto, el valor de la variable Hg se
expresa en este valor:
boxplot(log(DATOS$HgCr)~DATOS$EDAD_CAT,
          main="Boxplot [Hg] por grupo de edad",
          ylab="microg Hg/g creatinina", range=1)
```

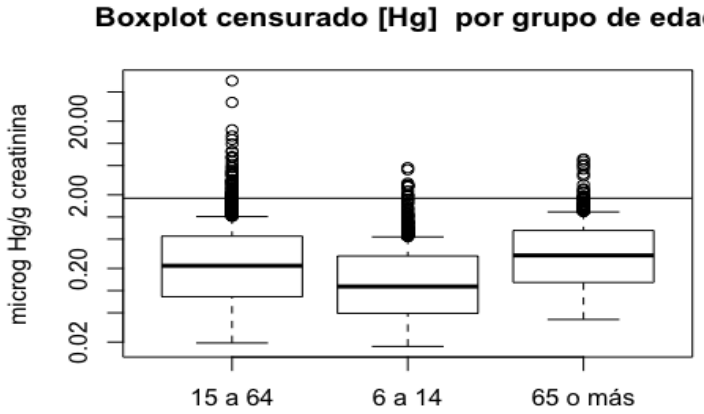


A simple vista parece que la media de mercurio en orina va aumentando a medida que se incrementa la edad de grupo; esto puede ser razonable ya que el mercurio es un metal pesado bioacumulable y es esperable que se encuentre en mayor proporción en individuos de mayor edad.

OPCIÓN B

El paquete NADA (*Nondetects and Data Analysis for Environmental Data*) posee la función "cenboxplot" que permite elaborar *boxplots* para observar de forma intuitiva la proporción de datos censurados mediante la elaboración de una caja a partir de los percentiles 25, 50 (línea horizontal) y 75. Por defecto, el *boxplot* se elabora en escala logarítmica, de modo que ya se ajusta a la distribución de nuestros valores:

```
library("NADA")
#Boxplot censurado en escala logarítmica:
cenboxplot(DATOS$HgCr, DATOS$CENS, group=DATOS$EDAD_CAT,
           main="Boxplot censurado [Hg]  por grupo de edad",
           ylab="microg Hg/g creatinina", range=1)
```

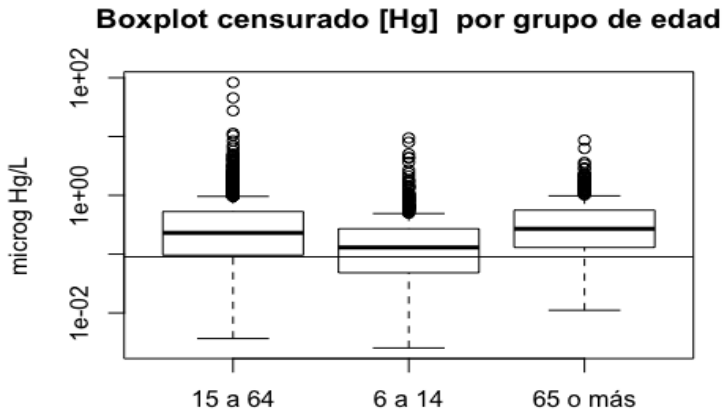


En este caso, ocurre el mismo fenómeno observado en el *boxplot* del apartado anterior con la línea trazada a causa del empleo de la variable Hg en orina corregida. No obstante, ya se ha comentado que las cajas representadas sí son correctas y se puede observar que la media más baja de Hg en orina corresponde al grupo de 6 a 14 años.

Entre los otros dos grupos de edad más avanzada no se observan unas diferencias claras a simple vista; es necesario aplicar tests estadísticos para obtener resultados fiables.

Como en la comparación entre género, si se emplea la variable Hg en orina (sin corregir) la línea de censura que aparece en el *boxplot* sí es la correcta:

```
#Boxplot censurado en escala logarítmica de la variable Hg:
cenboxplot(DATOS$Hg, DATOS$CENS, group=DATOS$EDAD_CAT,
           main="Boxplot censurado [Hg]  por grupo de edad",
           ylab="microg Hg/L", range=1)
```



La proporción de datos que quedan por debajo de la línea de censura son los valores que no han sido detectados (censurados por la izquierda).

4.2.2 Tests estadísticos:

OPCIÓN A

El método más común para realizar comparaciones entre grupos es el análisis de la varianza (ANOVA), aunque del mismo modo que ocurría con los t-test, este no es el método más adecuado a emplear en conjuntos de datos que contienen casos censurados. La razón es que para aplicar una ANOVA sería necesario sustituir los valores y eso, como ya se sabe, introduce error en los datos. A continuación, se aplica una ANOVA al presente conjunto de datos censurado con valores sustituidos, para observar qué ocurre. Se formulan las siguientes hipótesis:

$$H0: media[Hg]6-14 años = media[Hg]15-65 años = media[Hg] > 65 años$$

$$H1: media[Hg]i \neq media[Hg]j,$$
$$i \neq j \ para \ \ al \ \ menos \ \ un \ \ i,j \ (i \ es \ el \ nivel \ de \ la \ variable \ y \ j \ es \ la \ observación)$$

```
#Primero se ajusta el modelo empleando los logaritmos de la variable H
gCr:
modelo<-aov(log(DATOS$HgCr)~DATOS$EDAD_CAT)
```

```
#Aplicación del ANOVA sobre el modelo:
anova(modelo)
```

```
## Analysis of Variance Table
##
## Response: log(DATOS$HgCr)
##                  Df  Sum Sq Mean Sq F value    Pr(>F)
## DATOS$EDAD_CAT    2   60.03  30.014  27.894 1.025e-12 ***
## Residuals      2662 2864.30   1.076
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El resultado proporcionado indica que existen diferencias significativas entre las medias de los grupos con un nivel de significación del 5% (pvalor<0.05), por lo que se rechaza la hipótesis nula (H0).

El siguiente paso es realizar un contraste múltiple para conocer qué grupos son distintos entre ellos:

```
library(agricolae)
LSD.test(modelo,"DATOS$EDAD_CAT", group=T, console = T)
```

```
##
## Study: modelo ~ "DATOS$EDAD_CAT"
##
## LSD t Test for log(DATOS$HgCr)
##
## Mean Square Error:  1.075997
##
## DATOS$EDAD_CAT,  means and individual ( 95 %) CI
##
##          log.DATOS.HgCr.       std    r       LCL        UCL
Min
## 15 a 64        -1.267967 1.0820576 1698 -1.317328 -1.2186059 -3.936
173
## 6 a 14         -1.548261 0.9372130  576 -1.633011 -1.4635112 -3.595
789
## 65 o más       -1.058535 0.9770608  391 -1.161399 -0.9556713 -3.625
821
##                Max
## 15 a 64  4.262198
## 6 a 14   1.542427
## 65 o más 1.812779
##
## alpha: 0.05 ; Df Error: 2662
## Critical Value of t: 1.960856
##
## t-Student: 1.960856
## Alpha    : 0.05
## Minimum difference changes for each comparison
##
## Means with the same letter are not significantly different
##
```

```
##
## Groups, Treatments and means
## a     65 o más    -1.058535
## b     15 a 64     -1.267967
## c     6 a 14      -1.548261
```

Al tratarse de una comparación múltiple en la que se realizan contrastes de hipótesis simultáneas, podrían aparecer resultados significativos en alguno de los contrastes únicamente por azar. En este caso, tendríamos 3 contrastes de hipótesis, por lo que la probabilidad de que al menos un resultado sea significativo es:

P(un resultado significativo)= 1 - P = 1 - (1-0.05)^{3} = 0.1426

Como vemos, tenemos un 14.26% de probabilidad de aceptar un resultado significativo aun no siéndolo por efecto del azar. Para remediar este efecto empleamos la corrección de Bonferroni en el test LSD para ajustar el nivel de significación:

```
library(agricolae)
LSD.test(modelo,"DATOS$EDAD_CAT", group=T, p.adj="bonferroni", console
= T)

##
## Study: modelo ~ "DATOS$EDAD_CAT"
##
## LSD t Test for log(DATOS$HgCr)
## P value adjustment method: bonferroni
##
## Mean Square Error:  1.075997
##
## DATOS$EDAD_CAT,  means and individual ( 95 %) CI
##
##          log.DATOS.HgCr.       std    r       LCL        UCL
Min
## 15 a 64        -1.267967 1.0820576 1698 -1.317328 -1.2186059 -3.936
173
## 6 a 14         -1.548261 0.9372130  576 -1.633011 -1.4635112 -3.595
789
## 65 o más       -1.058535 0.9770608  391 -1.161399 -0.9556713 -3.625
821
##               Max
## 15 a 64  4.262198
## 6 a 14   1.542427
## 65 o más 1.812779
##
## alpha: 0.05 ; Df Error: 2662
## Critical Value of t: 2.395494
##
## t-Student: 2.395494
## Alpha    : 0.05
## Minimum difference changes for each comparison
##
## Means with the same letter are not significantly different
```

```
##
##
## Groups, Treatments and means
## a     65 o más    -1.058535
## b     15 a 64     -1.267967
## c     6 a 14      -1.548261
```

Si observamos la información que aparece al final, cada uno de los niveles de la variable conforman un grupo independiente (a: 65 o más, b: 15 a 64, c: 6 a 14), lo que indica que todos los grupos son distintos entre sí.

OPCIÓN B

En este apartado se realiza un test paramétrico (MLE) que tiene en cuenta que el conjunto de datos contiene datos censurados y, por tanto, es más adecuado que el método representado en la opción A. Las hipótesis planteadas son las siguientes:

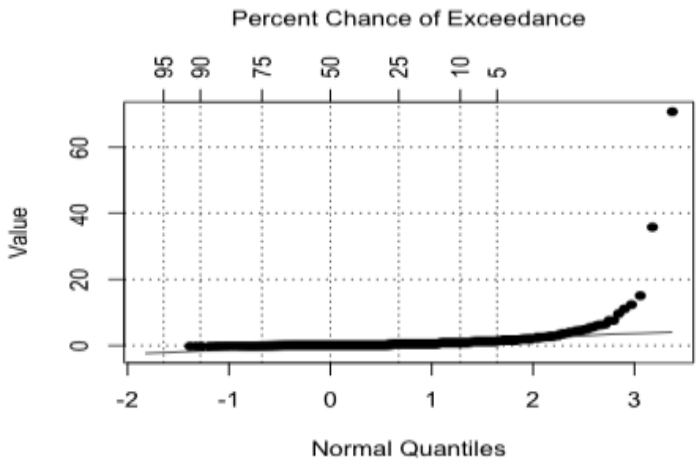
$$H0: \frac{media[Hg]\ poblacional}{media[Hg]\ diferentes\ niveles\ EDAD\_CAT} = 1$$

$$H1: \frac{media[Hg]\ poblacional}{media[Hg]\ diferentes\ niveles\ EDAD\_CAT} \neq 1$$

Como se trata de un método paramétrico, es necesario verificar que se cumplen dos suposiciones:

- Homocedasticidad de varianzas: las varianzas de los dos grupos no deben ser estadísticamente distintas. Esto puede comprobarse mediante los diagramas de cajas, vemos que la amplitud de las cajas relativas a cada grupo es muy similar.
- Siguen una distribución específica: en este caso, ya se ha visto que el modelo MLE general (D_mle) se ajustaba bien a la distribución normal logarítmica. Ahora es necesario comprobarlo en el nuevo modelo que se ajusta introduciendo los grupos:

```
#Ajuste del modelo lognormal:
Edad_mle<-cenmle(DATOS$HgCr, DATOS$CENS, DATOS$EDAD_CAT, dist="lognorm
al")


#Comprobación del cumplimiento de la suposición de seguimiento de la d
istribución normal-logarítmica a través de un gráfico de probabilidad:
res<-residuals(Edad_mle)
plot(ros(res, CENS, forwardT=NULL))
```



La mayoría de puntos se ajustan a la recta teórica, excepto dos puntos que pueden ser posibles *outliers* ya que se separan mucho del resto; esto puede ser un problema ya que el método MLE es susceptible a valores *outliers*. Pasamos a observar los resultados proporcionados:

```
Edad_mle

##                          Value Std. Error      z         p
## (Intercept)             -2.108     0.0641 -32.90 2.03e-237
## DATOS$EDAD_CAT15 a 64    0.541     0.0722   7.48  7.25e-14
## DATOS$EDAD_CAT65 o más   0.816     0.0958   8.52  1.57e-17
## Log(scale)               0.311     0.0178  17.52  9.61e-69
##
## Scale = 1.36
##
## Log Normal distribution
## Loglik(model)= -2114.5   Loglik(intercept only)= -2155.8
## Loglik-r:  0.1748519
##
## Chisq= 82.75 on 2 degrees of freedom, p= 0
## Number of Newton-Raphson Iterations: 3
## n = 2665
```

La categoría de referencia seleccionada por defecto es "6 a 14", lo cual ya es adecuado dado que en el boxplot se observaba que era la categoría más distinta al resto. El test de significación realizado ha sido un Chi-square con 2 grados de libertad y el pvalor obtenido ha sido 0, lo cual indica que hay diferencias entre los grupos de edad con un nivel de significación del 5%. Con este resultado, el siguiente paso es obtener los intervalos de confianza para conocer qué grupos son diferentes entre sí:

```
#Generamos los intervalos de confianza con la función creada anteriorm
ente:
IC_mle(cenobj=Edad_mle, conf.level = 0.95, ngroups=3)

##                                          lower     upper
## coefficients.DATOS$EDAD_CAT15 a 64  0.3990353 0.6822328
## coefficients.DATOS$EDAD_CAT65 o más 0.6282808 1.0036144
```

El primer IC compara la categoría "6 a 14" con la "15-64" no contiene el 1, por lo que podemos afirmar que existen diferencias significativas entre estas dos categorías. Por el contrario, el segundo IC que compara la categoría "6 a 14" con la ">65", sí que incluye el 1 y por este motivo decimos que no existen diferencias entre estos dos niveles.

En este caso también sería necesario tener en cuenta la aplicación de la corrección de Bonferroni debido a la ejecución simultánea de contrastes de hipótesis, pero como se ha observado en el ANOVA que no había diferencias entre los resultados obtenidos aplicando la corrección y sin aplicarla, se decide no aplicarla.

También es posible aplicar **métodos no paramétricos** para realizar un contraste de valores centrales entre grupos, como se ha realizado para la comparación por género. los cuales son más robustos en caso de la presencia de valores *outliers*. Como se ha explicado en el apartado anterior, estos métodos pueden aplicarse a muestras tanto grande como pequeñas, aunque con tamaños de muestras muy grandes y si los valores siguen una distribución concreta (como es el caso del presente conjunto de datos) pueden no resultar tan potentes a la hora de detectar diferencias entre grupos. De todos modos, se procede a su aplicación con el objetivo de visualizar las diferencias respecto al método paramétrico MLE:

```
#Wilcoxon sore test:
Hg.score.test<-cendiff(obs=DATOS$HgCr, censored=DATOS$CENS, groups=DAT
OS$EDAD_CAT)
Hg.score.test

##                             N Observed Expected (O-E)^2/E (O-E)^2/V
## DATOS$EDAD_CAT=6 a 14     576      182      282     35.58      66.3
## DATOS$EDAD_CAT=15 a 64   1698      769      723      2.92      10.9
## DATOS$EDAD_CAT=65 o más   391      201      147     20.05      31.0
##
##  Chisq= 81.4  on 2 degrees of freedom, p= 0

#En vista de las diferencias entre grupos, pasamos a observar diferenc
ias entre dos grupos. Se eligen los grupos que anteriormente NO han pr
esentado diferencias para verificar si este método es adecuado:
DATOS.EDAD<-DATOS[DATOS$EDAD_CAT=="6 a 14" | DATOS$EDAD_CAT=="65 o más
",]
Hg.score.DATOS.EDAD<-cendiff(obs=DATOS.EDAD$HgCr, censored=DATOS.EDAD$
CENS, groups=DATOS.EDAD$EDAD_CAT)
Hg.score.DATOS.EDAD

##                                N Observed Expected (O-E)^2/E (O-E)^
2/V
## DATOS.EDAD$EDAD_CAT=6 a 14   576      191      265      20.6      8
2.4
## DATOS.EDAD$EDAD_CAT=65 o más 391      210      136      40.3      8
2.4
##
##  Chisq= 82.4  on 1 degrees of freedom, p= 0
```

En este caso los resultados obtenidos difieren respecto a los proporcionados por MLE, ya que los IC confirmaban que no había diferencias entre las medianas de los grupos "6 a 14" y ">65" ya que incluía el 1; en este caso el test paramétrico indica que si existen diferencias con un pavlor=0. Este resultado tiene más lógica que el obtenido mediante el test paramétrico, dado que el mercurio es un metal pesado que se bioacumula en el organismo y tiene más sentido que en individuos con una edad más avanzada presenten una concentración en orina más elevada (si sólo se tiene en cuenta la edad y el resto de variables se consideran iguales).

5. Regresión:

El objetivo de este apartado en realizar modelos de regresión con el objetivo de estimar cómo varía la variable respuesta, en este caso HgCr, en función de una variable continua. Teniendo en cuenta el conjunto de datos con el que está trabajando, la variable regresora es la edad de los individuos (EDAD).

El modelo lineal a ajustar sigue la siguiente ecuación:

$$y = \beta_0 + \beta_1 x$$

siendo y la variable de interés, β_0 equivale al incercepto y es el valor estimado de la variable de interés cuando x (covariable) es igual a 0, β_1 es la pendiente estimada de la regresión lineal.

OPCIÓN A

Se ajusta el modelo de regresión donde la variable respuesta es el logaritmo de "HgCr" con una variable regresora "EDAD", tomando los valores censurados como valores fijos (0.09):

```
r.mod<-aov(log(HgCr)~EDAD, DATOS)
r.mod
```

```
## Call:
##    aov(formula = log(HgCr) ~ EDAD, data = DATOS)
##
## Terms:
##                       EDAD Residuals
## Sum of Squares   175.4271 2748.9054
## Deg. of Freedom         1      2663
##
## Residual standard error: 1.016001
## Estimated effects may be unbalanced
```

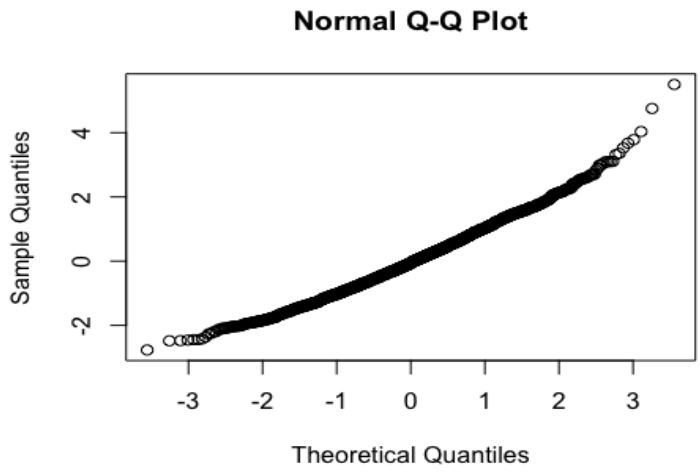
```
#Se calcula el anova para saber si la regresión es significativa:
anova(r.mod)
```

```
## Analysis of Variance Table
##
## Response: log(HgCr)
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## EDAD          1  175.43 175.427  169.94 < 2.2e-16 ***
## Residuals 2663 2748.91   1.032
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

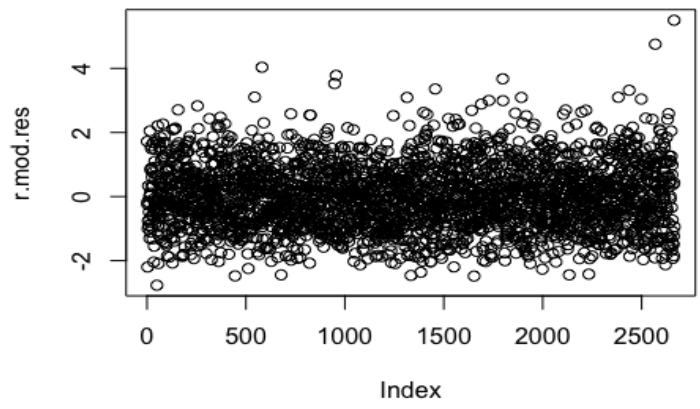
En el modelo creado, la edad de los individuos es significativa con un nivel de significación del 5% sobre el valor de Hg en orina de los individuos (pvalor<0.05).

A continuación, se comprueban las suposiciones del modelo: ajuste a la distribución normal logarítmica y homocedasticidad de varianzas, a partir del estudio de los residuos del modelo.

```
r.mod.res<-residuals(r.mod)
#Distribución:
qqnorm(r.mod.res)
```




```
#Homocedasticidad de varianzas:
plot(r.mod.res)
```



Observamos que los residuos del modelo se ajustan a la distribución normal logarítmica y no se observan diferencias de variancias.

Por otro lado, también se considera la opción de realizar una **regresión con interacción** de dos factores aleatorios principales: los grupos de edad (EDAD_CAT) y el género de los individuos de estudio (GEN).

El modelo lineal a ajustar sigue la siguiente ecuación:

$$y_{ijk} = \alpha_i + \beta_j + \alpha\beta_{ij}$$

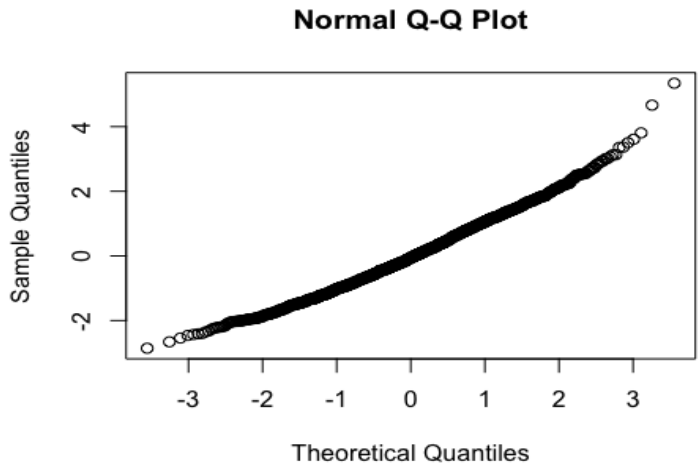
donde i=1,2,3 y j=1,2, k=1,...,n, α_i es el efecto del nivel i del factor principal EDAD_CAT, β_j es el efecto del nivel j del factor principal GEN, $\alpha\beta_{ij}$ es la interacción entre el nivel i del factor EDAD_CAT y el nivel j del factor GEN; y_{ijk} es la respuesta k en el i-ésimo nivel de EDAD_CAT y el j-ésimo nivel de GEN.

```
r.mod.int<-aov(log(HgCr)~EDAD_CAT+GEN+EDAD_CAT*GEN, DATOS)
r.mod.int

## Call:
##    aov(formula = log(HgCr) ~ EDAD_CAT + GEN + EDAD_CAT * GEN, data
= DATOS)
##
## Terms:
##                   EDAD_CAT       GEN EDAD_CAT:GEN Residuals
## Sum of Squares     60.0281   61.1068      13.6496 2789.5480
## Deg. of Freedom          2         1            2      2659
##
## Residual standard error: 1.024254
## Estimated effects may be unbalanced

anova(r.mod.int)

## Analysis of Variance Table
##
## Response: log(HgCr)
##                Df  Sum Sq Mean Sq F value    Pr(>F)
## EDAD_CAT        2   60.03  30.014 28.6094 5.092e-13 ***
## GEN             1   61.11  61.107 58.2471 3.200e-14 ***
## EDAD_CAT:GEN    2   13.65   6.825  6.5054  0.001519 **
## Residuals    2659 2789.55   1.049
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

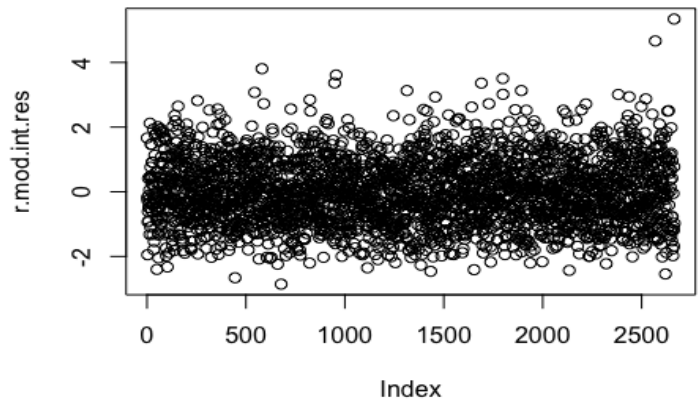
Según esta aproximación, la interacción de las variables categóricas edad y género es significativa sobre la concentración de Hg en orina (pvalor<0.05); así como ambos factores por separado también lo son si observamos los pvalores proporcionados en cada caso.

A continuación, y como se ha realizado para el anterior modelo de regresión, se comprueban las suposiciones del modelo: ajuste a la distribución normal logarítmica y homocedasticidad de varianzas, a partir del estudio de los residuos del modelo con interacción.

```
r.mod.int.res<-residuals(r.mod.int)
#Distribución:
qqnorm(r.mod.int.res)
```




```
#Homocedasticidad de varianzas:
plot(r.mod.int.res)
```



OPCIÓN B

Se ajusta un modelo de regresión MLE (método paramétrico) donde la variable respuesta es la variable censurada "HgCr" (junto con la variable indicadora de la censura) y la variable regresora es la edad de los individuos (variable cuantitativa continua):

```
#Ajuste del modelo con escala normal logarítmica:
r.mle<-cenreg(Cen(obs=DATOS$HgCr, censored=DATOS$CENS)~DATOS$EDAD, dis
t="lognormal")
r.mle
```

```
##                Value Std. Error     z        p
## (Intercept) -2.2643    0.05543 -40.8 0.00e+00
## DATOS$EDAD   0.0169    0.00124  13.6 2.16e-42
## Log(scale)   0.2869    0.01773  16.2 6.60e-59
##
## Scale = 1.33
##
## Log Normal distribution
## Loglik(model)= -2064.4   Loglik(intercept only)= -2155.8
## Loglik-r:  0.2575407
##
## Chisq= 182.9 on 1 degrees of freedom, p= 0
## Number of Newton-Raphson Iterations: 3
## n = 2665
```

```
#Generamos los intervalos de confianza con la función creada anteriorm
ente, indicamos 2 grupos porqué es el número de variables incluidas en
la regresión:
IC_mle(cenobj=r.mle, conf.level = 0.95, ngroups=2)
```

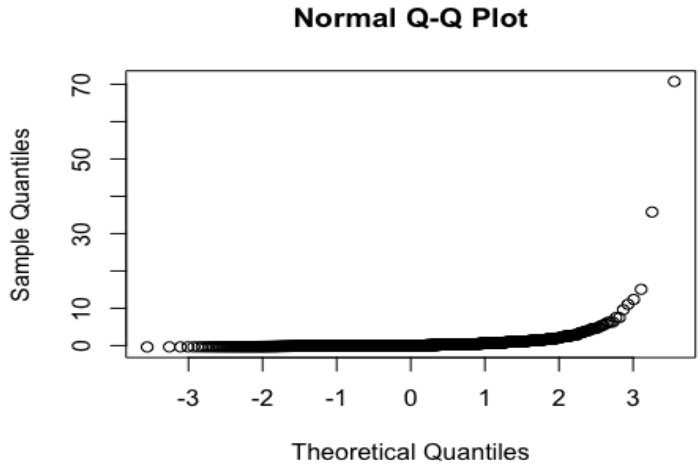
```
##                                lower      upper
## coefficients.DATOS$EDAD 0.01448781 0.01934789
```

A partir de los resultados proporcionados podemos conocer que la pendiente estimada para el efecto de la edad sobre la variable HgCr es 0.0169, esto quiere decir que por cada incremento de unidad en la variable EDAD la variable HgCr se incrementa en 0.0169. Sin embargo, el valor del coeficiente de correlación (*loglik-r*) obtenido para la regresión es bastante bajo (0.258); cuando este valor es cercano a cero indica que la correlación entre ambas variables es baja. Por último, con un pvalor= 0 (<0.05) proporcionado según una Chi-square con 1 grado de libertad, sabemos que el modelo de regresión elaborado es significativo por lo que respecta a la explicación de la variabilidad.

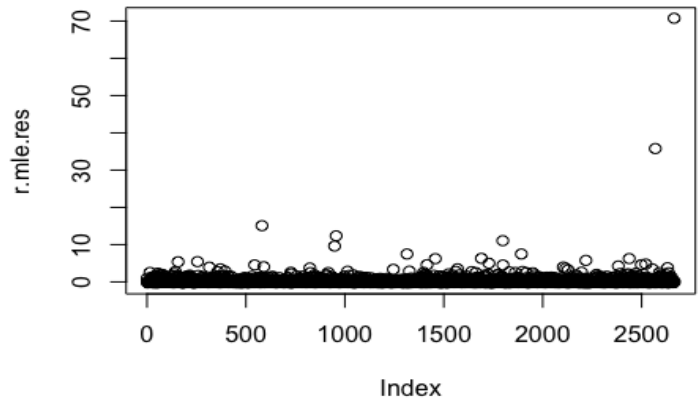
Los IC obtenidos, junto con la estimación de la pendiente, afirman que se trata de una relación positiva entre las variables ya que en este no se incluye el 0 y son valores positivos.

A continuación, se comprueban las suposiciones del modelo: ajuste a la distribución normal logarítmica y homocedasticidad de varianzas, a partir del estudio de los residuos del modelo.

```
r.mle.res<-residuals(r.mle)
#Distribución:
qqnorm(r.mle.res)
```




```
#Homocedasticidad de varianzas:
plot(r.mle.res)
```



El gráfico de distribución verifica el cumplimiento de la suposición de normalidad de forma fácil y visual, la comprobación de varianzas homogéneas es más difícil de observar dados los casos censurados. Aun así, podemos intuir que la dispersión de los datos se mantiene a lo largo de la variable y no hay variaciones considerables.

Por otro lado, al igual que en la opción A, también se considera la opción de realizar una **regresión con interacción** de dos factores aleatorios principales: los grupos de edad (EDAD_CAT) y el género de los individuos de estudio (GEN).

El modelo lineal a ajustar sigue la siguiente ecuación:

$$y_{ijk} = \alpha_i + \beta_j + \alpha\beta_{ij}$$

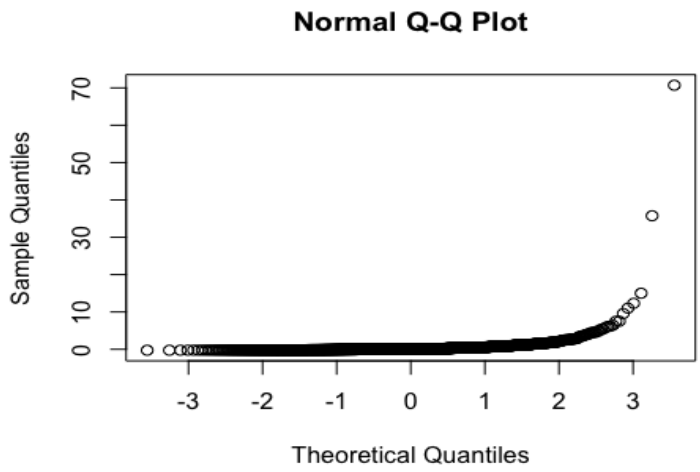
donde i=1,2,3 y j=1,2, k=1,...,n, α_i es el efecto del nivel i del factor principal EDAD_CAT, β_j es el efecto del nivel j del factor principal GEN, $\alpha\beta_{ij}$ es la interacción entre el nivel i del factor EDAD_CAT y el nivel j del factor GEN; y_{ijk} es la respuesta k en el i-ésimo nivel de EDAD_CAT y el j-ésimo nivel de GEN.

```
r.mle.int<-cenreg(Cen(obs=DATOS$HgCr, censored=DATOS$CENS)~DATOS$EDAD_
CAT*DATOS$GEN, dist="lognormal")
r.mle.int

##                                      Value Std. Error       z
p
## (Intercept)                        -2.1312     0.0860 -24.788 1.21e-
135
## DATOS$EDAD_CAT15 a 64               0.3395     0.0988   3.436  5.89e
-04
## DATOS$EDAD_CAT65 o más              0.7319     0.1324   5.528  3.24e
-08
## DATOS$GENF                          0.0672     0.1250   0.538  5.90e
-01
## DATOS$EDAD_CAT15 a 64:DATOS$GENF    0.3736     0.1427   2.619  8.83e
-03
## DATOS$EDAD_CAT65 o más:DATOS$GENF   0.1513     0.1889   0.801  4.23e
-01
## Log(scale)                          0.2983     0.0178  16.797  2.56e
-63
##
## Scale = 1.35
##
## Log Normal distribution
## Loglik(model)= -2092.9   Loglik(intercept only)= -2155.8
## Loglik-r:  0.2148365
##
## Chisq= 125.93 on 5 degrees of freedom, p= 0
## Number of Newton-Raphson Iterations: 3
## n = 2665
```

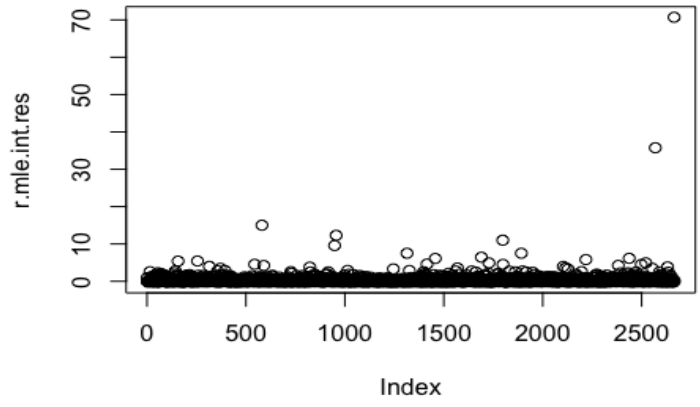
Se observa que el género de los individuos para el nivel de referencia de la variable EDAD (6 a 14 años) no es significativo (pvalor = 0.59), por lo que se puede afirmar que no hay diferencias entre ambos sexos hasta la adolescencia. Por lo que respecta al nivel de referencia de la variable GEN (Masculino) existen diferencias significativas entre los individuos jóvenes y adultos y los jóvenes y ancianos; mientras que para el género femenino en el grupo de edad más avanzado no es tan significativa. Tal y como se ha visto en apartados anteriores, la diferencia entre el género masculino y femenino incluso incluyendo la variable edad en el modelo resulta significativa; sin embargo, sería necesario elaborar un modelo con variables adicionales que informaran sobre el estilo de vida de los individuos que pueden afectar a la variable respuesta para obtener un modelo significativo y poder extraer conclusiones más fiables.

A continuación, se comprueban las suposiciones del modelo: ajuste a la distribución normal logarítmica y homocedasticidad de varianzas, a partir del estudio de los residuos del modelo.

```
r.mle.int.res<-residuals(r.mle.int)
#Distribución:
qqnorm(r.mle.int.res)
```




```
#Homocedasticidad de varianzas:
plot(r.mle.int.res)
```



Del mismo modo que para la regresión anterior, el gráfico de distribución verifica el cumplimiento de la suposición de normalidad; la comprobación de varianzas homogéneas es más difícil de observar dados los casos censurados, pero aun así podemos intuir que la dispersión de los datos se mantiene a lo largo de la variable y no hay variaciones considerables.

4
Conclusiones

La realización del presente proyecto final de máster ha permitido llevar a cabo una primera inmersión en el fenómeno de la censura estadística en su totalidad, primero des de un punto de vista teórico y posteriormente práctico. Teniendo en cuenta todos aquellos ámbitos donde la censura puede suponer un problema a la hora de realizar el tratamiento de datos y, de un modo más específico, de qué modo afecta a aquellos estudios de temática medioambiental. Especialmente se ha estudiado su incidencia en la evaluación de la exposición al mercurio ambiental a través de la medición de mercurio en orina.

A lo largo de las tareas realizadas, se ha comprobado que la censura estadística es mucho más común de lo que se podía esperar por parte de un investigador novel; sobre todo el tipo de censura por la izquierda. Este tipo de censura puede afectar a cualquier tipología de datos analíticos, ya que en muchas ocasiones el resultado obtenido es inferior al límite de detección o cuantificación del instrumento. Este hecho puede afectar de forma negativa a la interpretación de los resultados si no se emplean los métodos estadísticos adecuados, en función de la muestra de estudio y la proporción de datos censurados obtenidos. Se ha podido comprobar que, aunque la sustitución de los valores censurados resulta más fácil de aplicar y no requiere un conocimiento de herramientas estadísticas específicas, en ocasiones puede proporcionarnos resultados poco fiables. Sin embargo, también hay que tener en cuenta que el empleo de métodos más sofisticados, como el MLE, en ocasiones puede comportar otro tipo de inconvenientes como la baja bondad de ajuste dependiendo de los datos con los que se trabaje. Por tanto, es importante tener presente el objetivo del tratamiento de los datos con casos censurados y la naturaleza de los mismos para, en base a esto, decidir qué método estadístico merece la pena aplicar.

Los objetivos generales planteados al inicio del proyecto, así como los respectivos objetivos específicos, han podido cumplirse con éxito. Esto ha sido debido a una correcta organización y planificación de las tareas fijadas. En alguna ocasión ha habido un ligero desvío en la planificación temporal, pero este ha sido mitigado según las acciones correctoras previstas en el plan de trabajo y no ha causado otros inconvenientes mayores en el desarrollo del proyecto.

Como futuras líneas de trabajo sobre el tema planteado, una de ellas podría ser la aplicación de metodología estadística sobre datos con múltiples límites de detección. Pese a que el concepto es similar, es necesario considerar las dificultades adicionales que esto puede suponer; como el hecho de que la aplicación de estadísticos de orden puede perder el rigor necesario debido a que puede haber datos cuantificados con valores inferiores a otros casos censurados con un LD mayor.

Por otro lado, si se pretende profundizar en la evaluación de la exposición al mercurio medioambiental sería necesario disponer de más variables relacionadas, como la frecuencia de consumo de pescado, lugar de residencia (rural o urbano), presencia de amalgamas dentales, profesión (si tiene una exposición ocupacional al mercurio o no), entre otras. De este modo, podrá realizarse un modelo de regresión teniendo en cuenta variables que pueden afectar a la exposición del individuo al mercurio y las posibles interacciones entre ellas.

5
Glosario

ACRÓNIMOS EMPLEADOS EN LA MEMORIA:

CDC: *Centers for Disease Control and Prevention*

GerES: *German Environmental Survey*

Hg: Mercurio

ICM: *Iterative Convex Minorant*

ICP-MS: *Inductively Coupled Plasma – Mass Spectrometry*

KM: *Kaplan-Meier* method

LD: Límite de detección

MLE: *Maximum Likelihood Estimation*

N: Tamaño de muestra

NA: dato faltante (*missing value*)

NADA: *Nondetects and Data Analysis for Environmental Data*

NCHS: *National Center of Health Statistics*

NHANES: *National health and Nutrition Examination Survey*

ROS: *Regression on Order Statistics*

rROS: *robust Regression on Order Statistics*

USEPA: *United States Environmental protection Agency*

SIMBOLOGÍA Y UNIDADES:

[Hg]: Concentración de mercurio

L: Litro

μg: microgramo

nmol: nanomol

pm: peso molecular

pa: peso atómico

6
Bibliografía

[1] Helsel DR. (2012) Statistics for censored environmental data using minitab® and R (2nd ed.). Denver, CO: Wiley.

[2] Arrizabalaga EB. (2007) Interpretación de curvas de supervivencia. Revista Chilena de Cirugía, 59(1): 75-83.

[3] Zhang Z. (2010) Interval censoring. Statistical Methods in Medical Research, 19: 53-70. **DOI:** 10.1177/0962280209105023

[4] Banta-Green CJ, Brewer AJ, Ort C, Helsel DR, Williams JR, Field JA. (2016) Using waterwaster-based epidemiology to estimate drug consumption – Statistical analyses and data presentation. Science of Total Environment, 568: 856-863.
DOI: 10.1016/j.scitotenv.2016.06.052

[5] Shoari N, Dubé JS, Chenouri S. (2015) Estimating the mean and standard deviation of environmental data with below detection limit observations: Considering highly skewed data and model misspecification. Chemosphere, 138: 599-608.
DOI: 10.1016/j.chemosphere.2015.07.009

[6] Cdc.gov. (2017) CDC - NBP - Biomonitoring Summaries – Mercury. Disponible en: https://www.cdc.gov/biomonitoring/Mercury_BiomonitoringSummary.html
[Fecha acceso 13 Mar. 2017].

[7] Becker K, Kaus S, Krause C, Lepom P, Schulz C, Seiwert M, et al. (2002) German Environmental Survey 1998 (GerES III): environmental pollutants in blood of the German population. International Journal of Hygiene and Environmental Health, 205 (4): 297-308. **DOI:** 10.1078/1438-4639-00155

[8] Benes B, Spevackova V, Smid J, Cejchanova M, Cerna M, Subrt P, et al. (2000) The concentration levels of Cd, Pb, Hg, Cu, Zn, and Se in blood of the population in the Czech Republic. Central European Journal of Public Health, 8(2): 117-119.

[9] Caldwell KL, Mortensen ME, Jones RL, Caudill SP, Osterloh JD. (2009) Total blood mercury concentrations in the U.S. population: 1999-2006. International Journal of Hygiene and Environmental Health, 212(6):588-598. **DOI:** 10.1016/j.ijheh.2009.04.004

[10] Becker K, Schulz C, Kaus S, Seiwert M, Seifert B. (2003) German environmental survey 1998 (GerES III): environmental pollutants in the urine of the German population. International Journal of Hygiene and Environmental Health, 206: 15-24.
DOI: 10.1078/1438-4639-00188

[11] Apostoli P, Cortesi I, Mangili A, Elia G, Drago I, Gagliardi T, et al. (2002) Assessment of reference values for mercury in urine: the results of an Italian polycentric study. Science of the Total Environment, 289(1-3): 13-24.
DOI: 10.1016/S0048-9697(01)01013-0

[12] McPherson G. (1990) Statistics in scientific investigation: Its basis, Application and Interpretation. New York: Springer.

[13] García-Berthou E, Alcaraz C. (2004) Incongruence between test statistics and P values in medical papers. BMC Medical Research Methodology, 4,13.
DOI: 10.1186/1471-2288-4-13

[14] Faraway JJ. (2015) Linear models with R. (2nd ed.). Boca Raton, FL: CRC Press (Taylor & Francis Group)

[15] Gijbels I. (2010) Censored data. WIREs Computational Statistics, 2(2):178-188.
DOI: 10.1002/wics.80

[16] Klein JP, Moeschberger ML (2003). Survival analysis: techniques for Censoring and Truncated Data. (2nd ed). New York: Springer.

[17] Hough G. (2010) Sensory shelf life estimation of food products. Taylor & Francis Group,LLC. Chapter 4: Survival analysis applied to sensory shelf life.

[18] Robertson GL. (2010) Food Packaging and Shelf Life. Taylor and Francis group, LCC. Chapter 3: Shelf Life Testing Methodology and Data Analysis; Guillet M & Rodrigue N.

[19] Huston C & Juarez-Colunga E. (2009) Guidelines for computing summary statistics for data-sets containing non-detects. Department of Statistics and Actual Science, Simon Fraser University (Burnaby, BC).

[20] CDC (2001). National Report on Human Exposure to Environmental Chemicals. Atlanta, Georgia.

[21] Batariova A, Spevackova V, Benes B, Cejchanova M, Smid J, Cerna M (2006). Blood and urine levels of Pb, Cd and Hg in the general population of the Czech Republic and proposed reference values. International Journal of Hygiene and Environmental Health, 209: 359-366. **DOI:** 10.1016/j.ijheh.2006.02.005

[22] Bleda-Hernández MJ, Tobías-Garcés A. (2002) Aplicación de los modelos de regression de tobit en la modelización de variables epidemiológicas censuradas. Gaceta Santitaria, 16(2): 188-195. **DOI:** 10.1016/S0213-9111(02)71651-8

[23] Ramírez AV. (2008) Intoxicación ocupacional por mercurio. Anales de la Facultad de Medicina, 69(1):46-51.

[24] INSHT. NTP 184 (1986): Mercurio. Control ambiental y biológico.

[25] Brodkin E, Copes R, Mattman A, Kennedy J, Kling R, Yassi A. (2007) Canadian Medical Association Journal, 175(1): 59-63. **DOI:** 10.1503/cmaj.060790

[26] Cdc.gov. (2017). National Health and Nutrition Examination Survey. Disponible en: https://wwwn.cdc.gov/Nchs/Nhanes/2013-2014/UHG_H.htm [Fecha acceso 20 abril 2017].

[27] R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

[28] Lopaka Lee (2017). NADA: Nondetects and Data Analysis for Environmental Data. R package version 1.6-1. https://CRAN.R-project.org/package=NADA