Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias Físico-Matemáticas

Conceptos del Análisis de Supervivencia y una aplicación para pacientes con Diabetes tipo II.

Tesis presentada al

Colegio de Matemáticas

como requisito parcial para la obtención del grado de

Licenciado en Matemáticas Aplicadas

por

Karen Gabriela Tamayo Pérez.

asesorada por

Dr. Bulmaro Juárez Hernández.
Dr. Víctor Hugo Vázquez Guevara.

Puebla Pue.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción.

Se sabe que la diabetes es una de las principales causas de muerte en México y el número de personas con esta enfermedad sigue aumentando; hasta junio de 2014 se reportaron 193 026 casos nuevos de diabetes en todo México y 8216 en Puebla. Esto es un problema grave, ya que afecta no sólo al sector salud sino también afecta a la economía de México, dado que año tras año se invierten millones de pesos para el tratamiento de personas con diabetes [4] y [20].

La diabetes es una enfermedad crónica, es decir, una enfermedad de larga duración y lento progreso, en la cual el cuerpo es incapaz de producir insulina suficiente o es incapaz de utilizarla efectivamente. La insulina es una hormona producida y secretada por las células beta de los islotes pancreáticos; esta hormona es la encargada de facilitar que la glucosa que circula en la sangre penetre en las células y se convierta en energía [6].

El análisis de supervivencia es la rama de la estadística cuyas variables principales de estudio son los tiempos de vida de los individuos, es por ello que en principio se utiliza en la medicina.

En el presente trabajo se explican los conceptos básicos del análisis de supervivencia como son: tipos de censura, curva de supervivencia de Kaplan-Meier, función de riesgo y las principales distribuciones que se utilizan para el estudio de supervivencia. Después se aplican estos conceptos a un conjunto de datos obtenidos de pacientes de la comunidad de Zacapoaxtla que presentan diabetes tipo II. Primero se realiza un análisis descriptivo, es decir se obtuvo la curva de Kaplan-Meier la cual explica la probabilidad que los pacientes sobrevivan a esta enfermedad y su función de riesgo.
Posteriormente, se describen algunos métodos para estimar los parámetros de una distribución Weibull cuando existen datos completos y datos con censura, además de presentar dos pruebas de bondad de ajuste y ejemplificar estos métodos con la base de datos antes mencionada cuya censura fue aleatoria debido a la forma en que fueron entrando los pacientes al estudio. Se obtuvo que los pacientes siguen una distribución Weibull con parámetros: $\hat{\beta} = 1.82892$ y $\hat{\alpha} = 14.7051$ como

parámetros de escala y forma, respectivamente. Esto es importante puesto que ahora se pueden calcular los tiempos de vida promedio para pacientes que tienen diabetes en la comunidad de Zacapoaxtla.

El interés principal de este trabajo es comprender y desarrollar los conceptos básicos del análisis de supervivencia, desarrollar algunos de estos conceptos con programas como R y Mathematica, y mostrar una aplicación de estos conceptos en la vida real, que en este caso fue a datos de pacientes que tienen diabetes tipo II.

1.0.1. Materiales y procedimiento.

Como se mencionó anteriormente la diabetes es una enfermedad crónica, que puede prevenirse y que a pesar de ser una enfermedad controlable, puede traer fuertes complicaciones en la salud [3].

Los 4 síntomas de la diabetes son: Aumento de sed (Polidipsia), orinar frecuentemente (Poliuria), tener mucha hambre (Polifagia) y la pérdida de peso sin razón aparente [20].

Existen varios tipos de diabetes:

- Diabetes tipo I.
- Diabetes tipo II.
- Diabetes MODY.
- Diabetes Relacionada con Fibrosis Quística.
- Diabetes Gestacional.
- Diabetes Secundaria a Medicamentos.

Diabetes tipo I. Este tipo de diabetes se presenta usualmente en niños y jóvenes, no se observa producción de insulina debido a la destrucción autoinmune de las células β, es decir el sistema inmunológico ataca a estas células, por lo que las personas que tienen este tipo de diabetes no producen los niveles necesarios de insulina y necesitan inyectarse insulina diariamente.

Diabetes tipo II. Este tipo de diabetes es más común y se presenta en adultos de (40 - 70 años), aunque cada vez existen más casos de niños y adolescentes, en este tipo el organismo produce insulina pero no es suficiente o no la puede aprovechar provocando una acumulación de glucosa en la sangre, y el organismo se va deteriorando debido al exceso de azúcar en el cuerpo. Aún cuando las razones para desarrollar diabetes tipo II no se conocen los factores de riesgo son:

- Obesidad.

- Mala alimentación.
- Falta de ejercicio.
- Antecedentes hereditarios.
- Origen étnico.

Diabetes MODY. Se produce por defectos genéticos de las células beta y es hereditaria.
Diabetes Relacionada con Fibrosis Quística. La fibrosis afecta a varios órganos entre ellos al páncreas, por lo que se puede desarrollar diabetes.
Diabetes secundaria a medicamentos. Algunos medicamentos pueden alterar la secreción o la acción de la insulina.
Diabetes Gestacional. Se dice que una mujer tiene diabetes mellitus gestacional (DMG) cuando se le diagnostica diabetes por primera vez durante el embarazo.

Se tomó una muestra de 46 individuos del municipio de Zacapoaxtla ubicado al norte del estado de Puebla cuyas coordenadas son: $19^0$52'11" N $97^0$35'17" O, su población total del municipio de acuerdo al Censo de Población y Vivienda del INEGI es de 53,295 habitantes [7], de los cuales 25,534 son hombres y 27,761 son mujeres (2010). El municipio de Zacapoaxtla colinda al norte con los municipios de Cuetzalan del Progreso, al este con Tlatlauquitepec y Zaragoza, al sur con el de Zautla y al oeste con los de Xochiapulco.

La muestra consiste de 41 mujeres y 5 hombres de edad entre 26 y 83 años, los cuales entraron al estudio al presentar los síntomas antes mencionados, por lo que el doctor primero realiza un destroxtis el cual es una prueba en la que se extrae una gota de sangre y se pone en una tirilla reactiva, se inserta en un aparato y regresa el nivel de azúcar en la sangre, si éste es superior a 200 mg/dl entonces se sigue con otras pruebas para el diagnóstico de diabetes tipo II, ya que se han establecido que los niveles óptimos para las personas normales de glucosa son de 70 a 100 mg/dl en ayunas, es decir, sin haber consumido alimento; mientras, la cantidad de glucosa normal después de dos horas de comer es menor a 140 mg/dl.

Para confirmar el diagnóstico, se deben hacer uno o más de los siguientes exámenes [6]:

Exámenes de sangre para la diabetes:

Nivel de glucemia en ayunas: Se diagnostica diabetes si el resultado es mayor a 126 mg/dl en dos momentos diferentes.
Examen de hemoglobina A1c: Se diagnostica diabetes si el resultado del examen es 6.5 % o superior.
Prueba de tolerancia a la glucosa oral: Se diagnostica diabetes si el nivel de glucosa es superior a 200 mg/dl 2 horas después de ingerir una bebida azucarada especial, y al hacerles análisis clínicos se diagnostican con diabetes tipo II.

La variable de interés es el tiempo que los individuos tienen con la enfermedad desde que fueron diagnosticados, el estudio duró hasta octubre del 2014, por lo que en la muestra se observan 15 eventos presentados, de los demás pacientes sólo se sabe que hasta que terminó el estudio estaban vivos, es decir no aportan la información completa.

En la figura 1.1 se presenta una ilustración de datos censurados de 8 individuos que fueron sometidos a un estudio por un tiempo delimitado, el individuo 1 representa datos completos, es decir, presenta el evento y es observado. El individuo 2 tiene una censura por la derecha ya que hasta que terminó el estudio no había presentado el evento de interés, el individuo 3 presenta una censura por intervalo ya que no se sabe exactamente en que momento ocurió el evento pero se tiene un intervalo donde se sabe que ocurrió. En el capítulo 5 se introducen formalmente los conceptos de censura.

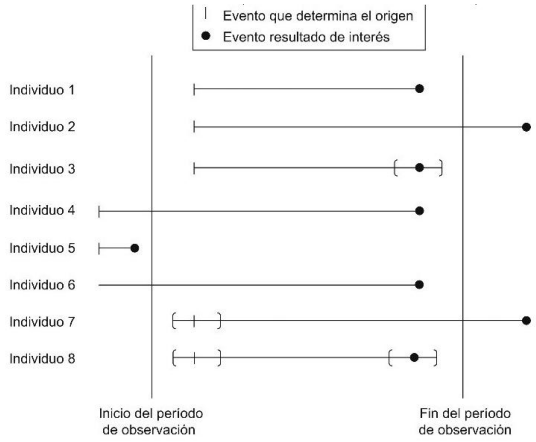


Figura 1.1: Datos con Censura

En este trabajo se utilizaron programas como R y Mathematica como apoyo para obtener gráficas, resolver ecuaciones, entre otras cosas [8] y [21].

Capítulo 2

Conceptos Básicos.

En este capítulo se estudian los conceptos básicos del análisis de supervivencia y se describen las distribuciones más utilizadas para su estudio. La variable de interés es el tiempo que transcurre hasta que se presenta un cierto evento, a estos eventos se les llamará fallos y, en general, se asume que la variable aleatoria tiempo T es una variable continua que puede tomar valores entre $[0, \infty)$ [14]. A la función de densidad de probabilidad (fdp) de la variable aleatoria T se le representa por $f_T(t)$ y la función de distribución de probabilidad $F_T(t)$ es la probabilidad acumulada de fallo hasta el tiempo t (Ver [10] y [15]).

$$F_T(t) = P[T \leq t] = \int_0^t f_T(x)dx,$$

Cumple que:

- Es monótona no decreciente.
- Es continua.
- $\lim_{x \to -\infty} F[t] = 0$ y $\lim_{x \to \infty} F[t] = 1$.

La probabilidad de que un individuo sobreviva después del tiempo t está dada por la función de supervivencia:

$$S(t) = P[t \leq T] = 1 - F_T[t] = \int_t^\infty f_T(x)dx.$$

Esta función es continua, monótona decreciente y cumple que $S(0) = 1$ y $S(\infty) = \lim_{t \to \infty} S(t) = 0$.

El p-cuantil de la distribución de T es el valor t_p tal que:

$$P[T \leq t_p] = p,$$

si el p-cuantil toma el valor de 0.5 es llamado la mediana de f_T.

Otro concepto importante para tiempos de vida es el de función de riesgo y se define como:

$$h(t) = \lim_{\Delta t \to 0} \frac{P[t \leq T \leq t + \Delta t | T \geq t]}{\Delta t}.$$

Es decir, la función de riesgo se define como el límite de la probabilidad condicionada de que un individuo falle antes del tiempo $t + \Delta$, si se sabe que dicho individuo no ha fallado al tiempo t. Entonces:

$$h(t) = \frac{f_T(t)}{S(t)}.$$

También se define la función de riesgo acumulado como:

$$H(t) = \int_0^t h(x)dx.$$

Además se verifica que:

$$f_T(t) = F_T'(t) = (1 - S(t))' = -S(t)',$$

entonces:

$$h(t) = \frac{f_T(t)}{S(t)} = \frac{-S(t)'}{S(t)} = -\frac{d}{dt} log S(t).$$

Más aún, integrando con respecto a t y usando que F(0)=0,

$$S(t) = \exp\left[-\int_0^t h(x)dx\right] = \exp(-H(t)),$$

y así:

$$H(t) = -\log S(t).$$

Una curva de riesgo típica de la vida de una persona o de alguna máquina es la llamada curva de bañera, la cual se comporta al principio de manera decreciente, esto es debido a que los primeros años el riesgo de falla es alto ya sea por la presencia de enfermedades tempranas o posibles defectos

de fábrica, al paso del tiempo el riesgo disminuye hasta que cae en una constante y después se vuelve creciente esto es por el envejecimiento tanto de las personas como de las máquinas, como se puede ver en la figura 2.1.

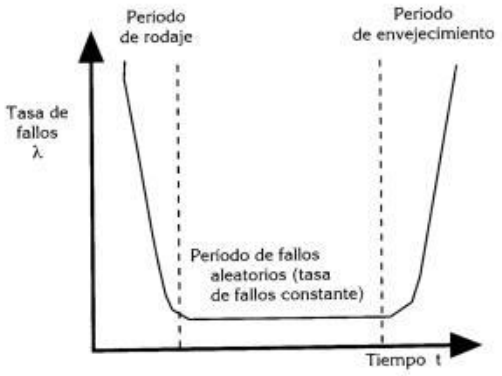


Figura 2.1: Curva de bañera

Algunas distribuciones importantes para el estudio de supervivencia se mencionan a continuación.

2.1. Distribución Exponencial.

La distribución exponencial históricamente ha sido utilizada para el estudio de tiempos de falla. Anteriormente fue utilizada por Clasius en relación con la teoría cinética de los gases en el siglo XIX y, recientemente en el estudio de manufactura por Epstein y Sobel [2].
Es menos utilizada en el estudio de supervivencia, debido a la propiedad de falta de memoria.

Se dice que una variable aleatoria T tiene una distribución exponencial si su función de densidad es de la forma:

$$f_T(t) = \lambda e^{-\lambda t} I_{[0,\infty)}(t) : \lambda > 0.$$

Donde se define a la función indicadora como

$$I_A(x) = \begin{cases} 1, & \text{si } x \in A. \\ 0, & \text{si } x \notin A. \end{cases} \tag{2.1}$$

Si una variable aleatoria T se distribuye exponencialmente con parámetro λ entonces se denotará por: $T \sim Exp(\lambda)$.

Su función de distribución está dada por:

$$F_T(t) = 1 - e^{-\lambda t} : \forall\ t \in \mathbb{R}^+ \ \ y \ \ \lambda > 0.$$

Además la función de supervivencia se puede obtener a partir de la función de probabilidad por lo tanto:

$$S(t) = 1 - F_T(t) = e^{-\lambda t} : \forall\ t \in \mathbb{R}^+ \ \ y \ \ \lambda > 0.$$

La función de riesgo de una distribución exponencial es una constante puesto que $h(t) = \frac{f_T(t)}{S(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda$, sobre todo el rango de T, por eso se dice que es una distribución sin memoria, ya que la falla será independiente al tiempo que lleva el individuo en el estudio.

La función de riesgo acumulado es:

$$H(t) = \lambda t \ ; \ \ \forall\, t > 0.$$

La esperanza y varianza están dadas por las siguientes expresiones:

$$E[T] = \frac{1}{\lambda} \quad y \quad Var[T] = \frac{1}{\lambda^2}.$$

La estimación de los cuantiles t_p en tiempos de vida es de mayor importancia que la media, debido a que siempre existen y en datos censurados son más fáciles de calcular.
De la definición de cuantil, se tiene:

$$F_T(t) = P[T \leq t] = 1 - e^{-\lambda t}$$

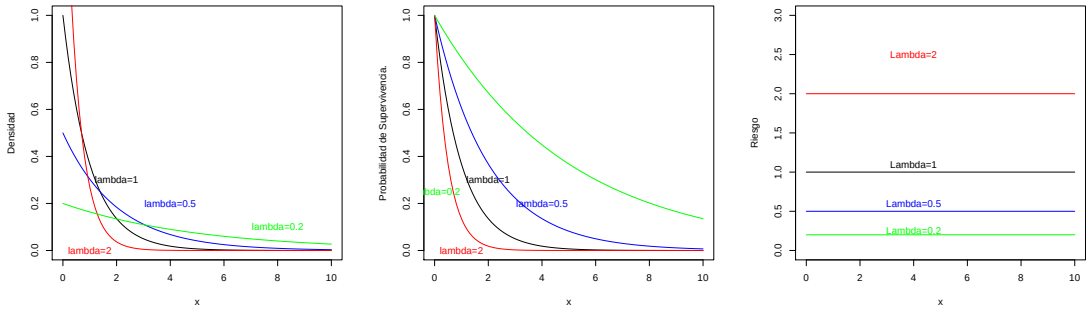
$$\Rightarrow e^{-\lambda t} = 1 - F_T(t)$$

$$\Rightarrow log(1 - F_T(t)) = -\lambda t$$

Por lo que el p-cuantil de la función exponencial es:

$$t_p = \frac{-1}{\lambda} \log(1 - p).$$

En la figura 2.2, se presenta la función de densidad, de supervivencia y de riesgo, para diferentes valores del parámetro λ.



(a) Función de Densidad. (b) Función de Supervivencia. (c) Función de Riesgo.

Figura 2.2: Distribución Exponencial.

2.2. Distribución Weibull.

Suponga que existen valores $\beta > 0$ y $\alpha > 0$ tales que la variable aleatoria $Y = \left(\frac{T}{\beta}\right)^{\alpha}$ se distribuye como una exponencial estándar, es decir su función de densidad es de la forma:

$$f_Y(y) = e^{-y} I_{(0,\infty)}(y).$$

Luego por el método de las transformaciones se tiene que la función de densidad para T está dada por:

$$f_T(t) = \frac{\alpha}{\beta}\left(\frac{t}{\beta}\right)^{\alpha-1} e^{-\left(\frac{t}{\beta}\right)^{\alpha}} I_{(0,\infty)}(t) : \alpha > 0 \quad y \quad \beta > 0. \tag{2.2}$$

Una variable aleatoria T tiene una distribución Weibull si tiene una función de densidad dada por la Ecuación (2.2), lo cual se representa como $T \sim W(\alpha, \beta)$, con β y α siendo los parámetros de escala y de forma respectivamente [9]; es fácil ver que si $\alpha = 1$ se obtiene una distribución exponencial con parámetro $\lambda = \frac{1}{\beta}$.

La distribución Weibull es una generalización de la distribución exponencial, esta distribución es utilizada con mayor frecuencia para estudios de tiempos de vida.
La función de distribución está dada por:

$$F_T(t) = 1 - e^{-\left(\frac{t}{\beta}\right)^{\alpha}} : \forall\, t \in \mathbb{R}^+, \lambda > 0 \quad y \quad \beta > 0,$$

por lo que la función de supervivencia es de la forma:

$$S(t) = 1 - F_T(t) = e^{-\left(\frac{t}{\beta}\right)^{\alpha}} : \forall\, t \in \mathbb{R}^+,\ \lambda > 0 \quad y \quad \beta > 0,$$

y su función de riesgo:

$$h(t) = \frac{\alpha}{\beta^{\alpha}} t^{\alpha-1}; \quad \forall\, t > 0.$$

La función de riesgo de una distribución Weibull tiene las siguientes propiedades:

- Es monótona decreciente si $\alpha < 1$, es decir la probabilidad de fallo disminuye al aumentar el tiempo.
- Es creciente si $\alpha > 1$, es decir la probabilidad de fallo aumenta cuando aumenta el tiempo.
- Se reduce a la constante de riesgo de una distribución exponencial cuando $\alpha = 1$, es decir el riesgo no depende del tiempo.

La función de riesgo acumulada es:

$$H(t) = \left(\frac{t}{\beta}\right)^{\alpha}.$$

El p-cuantil está dado por:

$$t_p = \beta[-\log(1-p)]^{\frac{1}{\alpha}}.$$

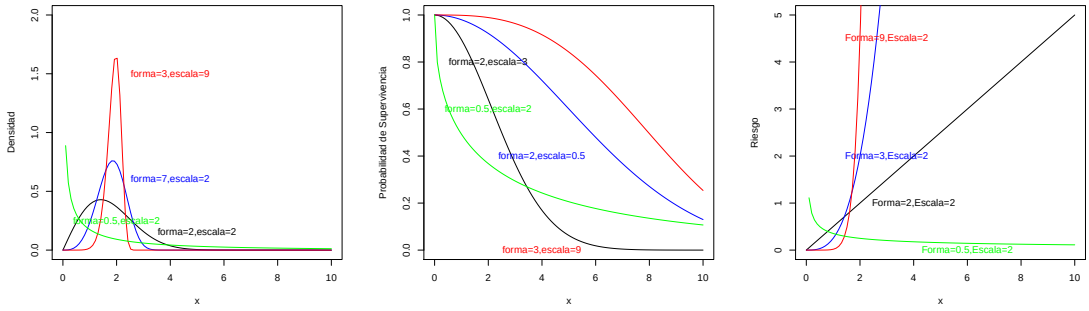
Su esperanza y varianza son respectivamente:

$$E[T] = \beta\Gamma\left(1 + \frac{1}{\alpha}\right) \quad y \quad V[T] = \beta^2\left[\Gamma\left(1 + \frac{2}{\alpha}\right) - \Gamma^2\left(1 + \frac{1}{\alpha}\right)\right].$$

Donde Γ es la función Gama [9].

La distribución Weibull es de gran importancia ya que se puede utilizar para modelar diferentes situaciones como son tiempos de falla para máquinas, terremotos, tamaño de las gotas de lluvia, análisis de supervivencia, entre otras. Esto es debido a las diferentes formas que puede tomar su función de riesgo.

En la figura 2.3 se muestra la función de densidad, de supervivencia y riesgo, para diferentes valores de los parámetros β y α.



(a) Función de Densidad. (b) Función de Supervivencia. (c) Función de Riesgo.

Figura 2.3: Distribución Weibull.

2.3. Distribución Normal.

La distribución normal es de gran importancia histórica debido a que es la base de muchos trabajos estadísticos, además de que se puede utilizar como una aproximación de muchas distribuciones. El soporte teórico para utilizar esta distribución está basado en la teoría del límite central, la cual dice que bajo ciertas condiciones la suma de variables aleatorias se distribuye de manera normal cuando el número de variables aleatorias sumadas, se incrementa [9].

No es tan utilizada en estudios de supervivencia debido a su carácter simétrico ya que los tiempos de supervivencia presentan asimetría [14].

Se dice que una variable aleatoria T se distribuye de forma normal si su función de densidad es de la forma:

$$f_T(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right] I_{(-\infty,\infty)}(t) : \sigma > 0 \quad y \quad \mu \in \mathbb{R},$$

donde μ y σ se denominan media y desviación estándar respectivamente, si una variable aleatoria se distribuye como una normal con media μ y desviación σ entonces se denota como $T \sim N(\mu, \sigma)$.

Su función de distribución y supervivencia respectivamente son:

$$F_T(t) = \int_{-\infty}^{t} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] dx,$$

y

$$S(t) = 1 - F(t) = \int_{t}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] dx.$$

Las cuales se pueden aproximan numéricamente, por lo que la función de riesgo no puede cal-

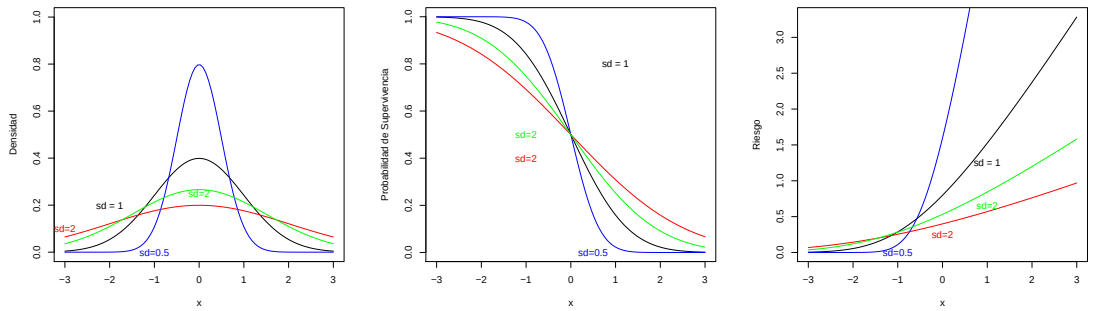
cularse analíticamente como la división entre la función de densidad y la función de supervivencia, pero se puede calcular de forma numérica.

Cuando $\mu = 0$ y $\sigma = 1$ se tiene la distribución normal estándar y se denota como $T \sim N(0,1)$ cuya función de densidad es de la forma:

$$f_T(t) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{t^2}{2}\right].$$

La esperanza y varianza de la distribución normal son: $E[T] = \mu$ y $Var[T] = \sigma^2$.

En la figura 2.4 se presenta la función de densidad, de supervivencia y riesgo para la distribución normal con diferentes valores de σ y $\mu = 0$.



(a) Función de Densidad. (b) Función de Supervivencia. (c) Función de Riesgo.

Figura 2.4: Distribución Normal.

2.4. Distribución Log-Normal.

En 1879, Galton señaló [9] que si se tienen $X_1, X_2, ..., X_n$ variables aleatorias positivas e independientes y si

$$T_n = \prod_{i=1}^{n} X_i,$$

entonces

$$\log T_n = \log \prod_{i=1}^{n} X_i = \sum_{i=1}^{n} \log X_i,$$

y como son variables aleatorias independientes, por el teorema de límite central la distribución de T_n es aproximadamente una normal si $n \to \infty$, de forma que la distribución límite de $\log T_n$ es una Log-Normal.

Se dice que una variable aleatoria T tiene una distribución Log-Normal, si la variable aleatoria $Z = \log T \sim N(\mu, \sigma)$, luego la notación $T \sim LN(\mu, \sigma)$ significa que T tiene una distribución Log-Normal con parámetros μ y σ.
Por lo tanto, si la función de densidad de la variable Z es de la forma:

$$f_Z(z) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2\right] I_{(-\infty,\infty)}(z) : \ \sigma > 0 \ \ y \ \ \mu \in \mathbb{R},$$

entonces la función de densidad para la variable T está dada por:

$$f_T(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\log t-\mu}{\sigma}\right)^2\right] \left(\frac{1}{t}\right) I_{(0,\infty)}(t) : \ \sigma > 0 \ \ y \ \ \mu \in \mathbb{R},$$

y la función de distribución está dada por:

$$F_T(t) = \int_{-\infty}^{\frac{\log t-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{u^2}{2}\right] du.$$

La función de supervivencia se obtiene como:

$$S(t) = 1 - F_T(t) = 1 - \int_{-\infty}^{\frac{\log t-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{u^2}{2}\right] du.$$

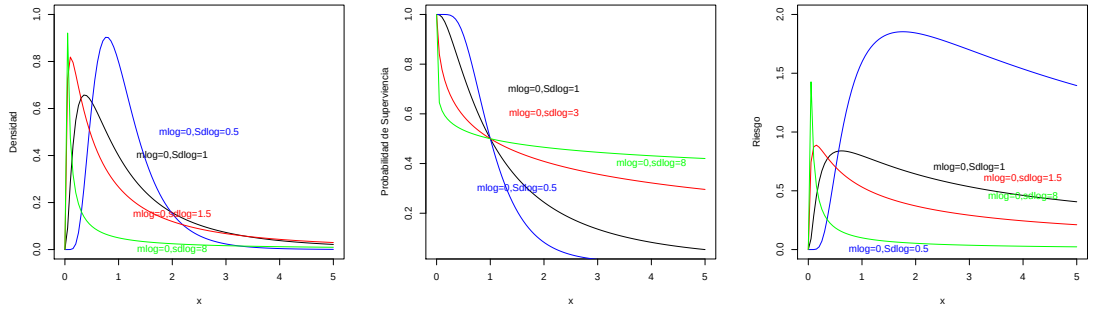
El valor esperado de la distribución Log-Normal, está dado por

$$E[T] = e^{(\mu+\sigma)},$$

y la varianza

$$Var[T] = e^{2\mu+\sigma}(1 - e(\sigma)).$$

La figura 2.5 presenta la función de densidad, supervivencia y riesgo de la distribución log-normal variando sus parámetros.



(a) Función de Densidad. (b) Función de Supervivencia. (c) Función de Riesgo.

Figura 2.5: Distribución Log-Normal.

2.5. Distribución Gama.

La distribución gama da representaciones útiles en situaciones físicas; se ha utilizado para hacer ajustes a distribuciones exponenciales que representan tiempos de vida [9].

Se dice que una variable T tiene una distribución gama con parámetros α y k, si su función de densidad es de la forma:

$$f_T(t) = \frac{1}{\alpha\Gamma(k)}\left(\frac{t}{\alpha}\right)^{k-1}\exp\left(\frac{-t}{\alpha}\right)I_{(0,\infty)}(t)\,;\quad \alpha>0\ \ y\ \ \ k>0,$$

donde α y k son parámetros de escala y forma, respectivamente. Si una variable T tiene distribución gama entonces se representa $T \sim \Gamma(\alpha, k)$. Cuando $k = 1$, al igual que en la distribución Weibull, se tiene a la distribución exponencial con parámetro $\lambda = \frac{1}{\alpha}$ [14].

Una función de interés en el análisis de supervivencia para poder definir la función de supervivencia, es la función gama incompleta, la cual se define como:

$$I(k,t) = \frac{1}{\Gamma(k)}\int_0^t x^{k-1}e^{-x}dx.$$

Y se define a la función de supervivencia en términos de la función gama incompleta como sigue:

$$S(t) = 1 - I(k,\frac{t}{\alpha}) = \frac{1}{\Gamma(k)}\int_{\frac{t}{\alpha}}^{\infty} x^{k-1}e^{-x}dx,$$

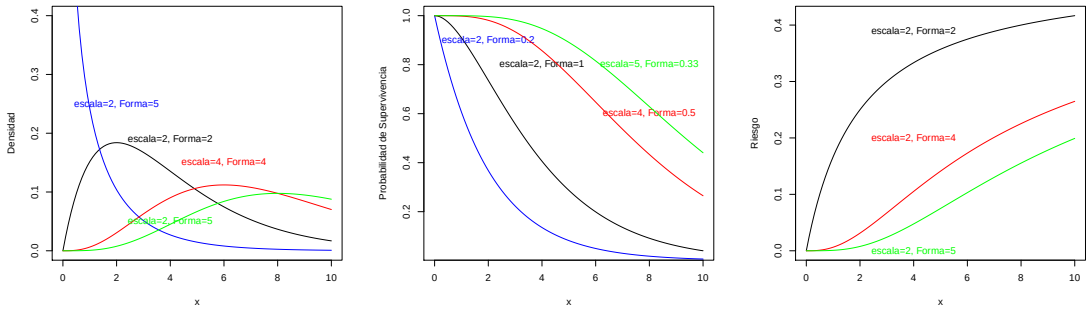
y la función de riesgo está dada por:

$$h(t) = \frac{t^{k-1}e^{\frac{-t}{\alpha}}}{\int_t^{\infty} x^{k-1}e^{-x}dx}.$$

la esperanza y varianza están dadas por:

$$E[T] = k\alpha \quad y \quad V[T] = k\alpha^2.$$

En la figura 2.6 se presenta la función de densidad, de supervivencia y de riesgo para la distribución Gama, variando sus parámetros.



(a) Función de Densidad. (b) Función de Supervivencia. (c) Función de Riesgo.

Figura 2.6: Distribución Gama.

2.6. Distribución Gumbell.

Se dice que una variable aleatoria T tiene una distribución Gumbell y se representará como $T \sim G(\tau, \omega)$ si tiene una función de densidad: ver [9]

$$f_T(t) = \omega^{-1} \exp\left[-\left(\frac{t-\tau}{\omega}\right) - \exp\left[-\left(\frac{t-\tau}{\omega}\right)\right]\right] I_{(\tau,\infty)}(t)\,; \;\; \tau \in \mathbb{R} \;\; y \;\; \omega > 0,$$

donde ω y τ son los parámetros de escala y localización, respectivamente.
Su función de distribución está dada por:

$$F_T(t) = \exp\left[-\exp\left[-\left(\frac{t-\tau}{\omega}\right)\right]\right]\;; \quad t > \tau, \;\; \tau \in \mathbb{R}, \;\; \omega > 0.$$

La función de supervivencia está dada por:

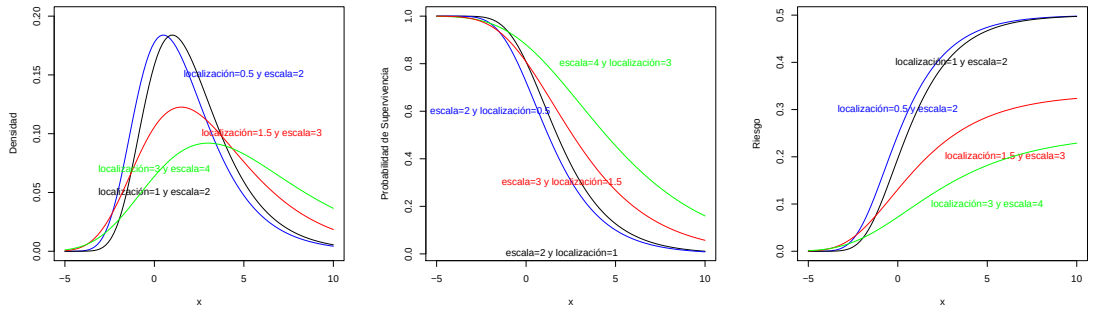
$$S(t) = 1 - F_T(t) = 1 - \exp\left[-\exp\left[-\left(\frac{t-\tau}{\omega}\right)\right]\right]\;;\; t > \tau, \;\; \tau \in \mathbb{R}, \;\; \omega > 0.$$

El p-cuantíl de esta distribución está dado por: [14]

$$t_p = \tau + \omega log[-log(1-p)].$$

La importancia de esta distribución en este estudio radica por la relación que tiene con la distribución Weibull; si una variable $T \sim W(\lambda, \beta)$, entonces la variable transformada $Y = LogT \sim G(\tau, \omega)$ donde $\tau = log\lambda$ y $\omega = \beta^{-1}$ [9].

La figura 2.7 muestra la función de densidad, de supervivencia y de riesgo para la distribución Gumbell para diferentes valores de los parámetros τ y ω.



(a) Función de Densidad. (b) Función de Supervivencia. (c) Función de Riesgo.

Figura 2.7: Distribución Gumbel.

Capítulo 3

Algunos procedimientos gráficos y no paramétricos.

En algunos experimentos para analizar datos y hacer resúmenes, entre otras cosas, es útil utilizar gráficos, pero en tiempos de vida la presencia de censura modifica los métodos estándar, por ejemplo en la fabricación de una tabla de frecuencia relativa, es importante la suposición de que los datos provienen de una muestra aleatoria de tamaño n completa, ya que la variable d_j, que se define cómo el número de observaciones que caen en cierto intervalo de tiempo y no se podría obtener cuando existe censura, ya que no se sabe exactamente en qué intervalo es que ocurrió el evento. En este capítulo se presentan algunas modificaciones que se hacen a los métodos estándar cuando existe censura en los datos [12].

3.1. Estimaciones no paramétricas de funciones de supervivencia y cuantiles.

3.1.1. Estimador producto-Límite.

En esta sección se presenta cómo graficar la función de supervivencia de una distribución de forma empírica, la cual es estimador de la función de supervivencia. Cuando no existe censura, para una muestra aleatoria de tamaño n, la función de supervivencia empírica es de la forma

$$\hat{S}(t) = \frac{\text{Número de observaciones} \geq t}{n} \; ; \; t \geq 0,$$

Es una función escalonada que en cada observación decrece $1/n$, cuando todas las observaciones son distintas, pero si hay d tiempos de vida igual a t entonces decrece d/n.

Para datos censurados se hace una modificación la cual se llama el estimador producto-límite (PL) de la función de supervivencia o el estimador de Kaplan-Meier, definido de la siguiente forma:

Definición 3.1 *Sea (t'_i, δ_i) una muestra aleatoria de tiempos de vida censurados, entonces:*

$$\hat{S}(t) = \prod_{t_j \leq t} \left(\frac{n_j - d_j}{n_j} \right),$$

es el estimador Producto-Límite, donde δ_i es el indicador de censura, $d_j = \sum I(t'_i = t_j, \delta_i = 1)$ representa el número de muertes en el tiempo t_j y $n_j = \sum I(t'_j \geq t_j)$ representa el número de individuos que están vivos y sin censura justo antes del tiempo t_j.

De igual forma la gráfica de la función es escalonada con saltos de tamaño $(n_j - d_j/n_j)$ para cada t_j.

El siguiente ejemplo muestra cómo se obtiene el estimador producto límite:

Se hace el análisis de los resultados de un tratamiento clínico, en el cual la droga 6-mercaptopurina (6-MP) fue comparada con el tratamiento de placebo, para pacientes con leucemia aguda, la tabla 3.1 da los tiempos de remisión para dos grupos de 21 pacientes cada uno [12].

$6-MP$	$Placebo$	$6-MP$	$Placebo$
6	1	16	8
6	1	17*	8
6	2	19*	8
6*	2	20*	11
7	3	22	11
9*	4	22	11
10	4	23	12
10*	5	25*	12
11*	5	32*	15
13	8	34*	22
		35*	23

Tabla 3.1: Longitudes de remisión(en semanas)para dos grupos de pacientes. Donde "*"significa que el dato es censurado.

Se escribe el estimador producto límite de forma recursiva como:

$$\hat{S}(t_j+) = \hat{S}(t_{j-1}+)\frac{n_j - d_j}{n_j}.$$

Por conveniencia se define a $\hat{S}(0) = 1$, entonces calculando de forma recursiva se obtiene el estimador producto límite, para el tratamiento de placebo, como se muestra en la tabla 3.2:

t_j	n_j	d_j	$\hat{S}(t_j)$
1	21	2	0.904762
2	19	2	0.809524
3	17	1	0.761905
4	16	2	0.666667
5	14	2	0.571429
8	12	4	0.380952
11	8	2	0.285714
12	6	2	0.190476
15	4	1	0.142857
17	3	1	0.0952381
22	2	1	0.047619
23	1	1	0.

Tabla 3.2: Estimador producto límite (PL) para pacientes que han recibido el tratamiento de placebo.

y para tratamiento de la droga 6-MP se muestra en la tabla 3.3:

t_j	n_j	d_j	$\hat{S}(t_j)$
6.	21.	3.	0.857143
7.	17.	1.	0.806723
10.	15.	1.	0.752941
13.	12.	1.	0.690196
16.	11.	1.	0.627451
22.	7.	1.	0.537815
23.	6.	1.	0.448179

Tabla 3.3: Estimador producto límite (PL) para pacientes que han recibido el tratamiento de la droga 6-MP.

La gráfica de la función de supervivencia para ambos tratamientos se muestra en la figura 3.1.

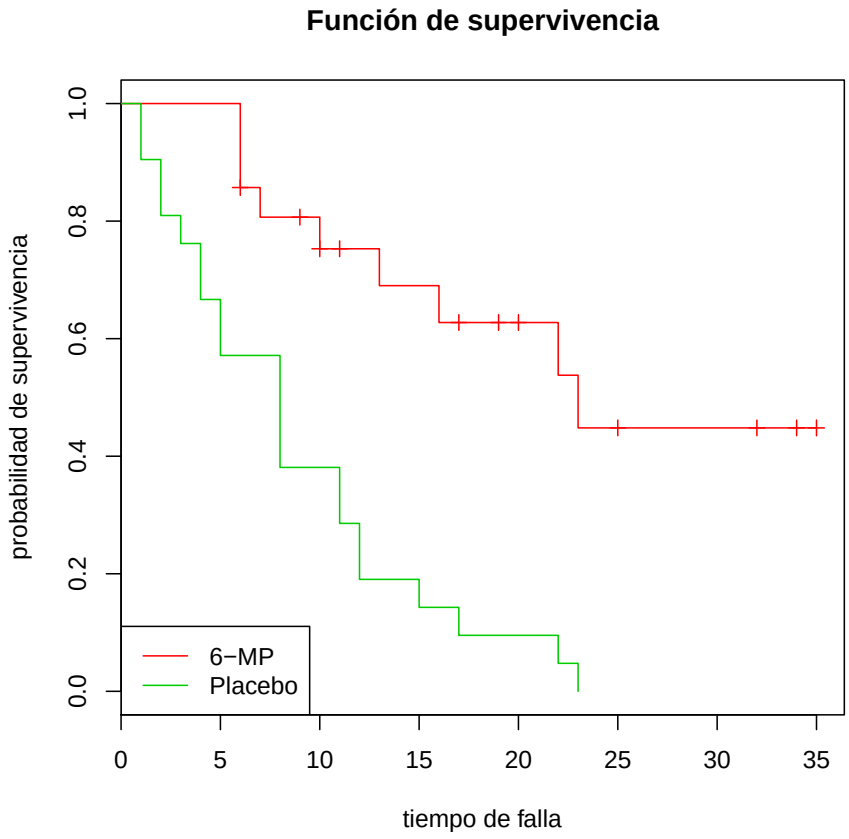


Figura 3.1: Función de supervivencia para la droga 6-MP y Placebo.

Donde se observa que la probabilidad de que sobrevivan los pacientes que están con el tratamiento de 6-MP es más alta que la de los pacientes que están bajo el tratamiento de placebo.

3.1.2. Estimación de la varianza del estimador de la función de supervivencia.

Dado el estimador de la función de supervivencia $\hat{S}(t_j+)$, es de interés conocer un estimador de su varianza.

Como anteriormente se dijo el estimador producto límite se define como

$$\hat{S}(t) = \prod_{j=1}^{n} \hat{p}_j,$$

donde $\hat{p_j} = \frac{n_j - d_j}{n_j}$, tomando el logaritmo se tiene

$$log(\hat{S}(t)) = log\left(\prod_{j=1}^{n} \hat{p}_j\right) = \sum_{j=1}^{n} log(\hat{p}_j).$$

Luego, tomando la varianza se tiene:

$$Var(log(\hat{S}(t))) = Var\left(\sum_{j=1}^{n} log(\hat{p}_j)\right).$$

Como la probabilidad de falla es independiente para todo t cuando está en el intervalo de tiempo $(t_j, t_j + 1)$ la expresión anterior queda de la siguiente forma:

$$Var(log(\hat{S(t)})) = \sum_{j=1}^{n} Var(log(\hat{p}_j)).$$

Dado que para cada intervalo de tiempo los individuos tienen sólo dos posibilidades, seguir vivos o morir, entonces tenemos un experimento de tipo Bernoulli y como todos los intervalos de tiempo son independientes entre sí, entonces el número de individuos que sobreviven a través de un intervalo de tiempo, sigue una distribución de tipo Binomial con probabilidad de supervivencia p_j, el número de individuos observados al tiempo t_{j+1} está dado por $n_j - d_j$ y utilizando el resultado de que la varianza de una distribución binomial con parámetros (n, p) es $np(1-p)$, entonces se tiene

$$Var(n_j - d_j) = n_j p_j (1 - p_j).$$

Además

$$Var(\hat{p_j}) = Var\left(\frac{n_j - d_j}{n_j}\right).$$

Como n_j es constante, utilizando las propiedades de la varianza, se tiene:

$$Var\left(\frac{n_j - d_j}{n_j}\right) = \frac{Var(n_j - d_j)}{n_j^2} = \frac{n_j \hat{p_j}(1 - \hat{p_j})}{n_j^2} = \frac{\hat{p_j}(1 - \hat{p_j})}{n_j}.$$

Para obtener la varianza del logaritmo, se utiliza el resultado de la aproximación de series de Taylor para la varianza de una función de una variable aleatoria, es decir, la varianza de una función $g(X)$, donde X es una variable aleatoria, está dada por:

$$Var(g(X)) \approx \left(\frac{dg(X)}{dX}\right)^2 Var(X),$$

entonces:

$$Var[log(\hat{p_j})] \approx \left(\frac{dlog(\hat{p_j})}{d\hat{p_j}}\right)^2 Var(\hat{p_j}) = \frac{1}{\hat{p_j}^2}\frac{\hat{p_j}(1-\hat{p_j})}{n_j},$$

haciendo la sustitución de $\hat{p_j} = \frac{(n_j - d_j)}{n_j}$, entonces la varianza estimada del logaritmo queda de la siguiente forma:

$$Var\left[log\left(\frac{n_j - d_j}{n_j}\right)\right] \approx \frac{d_j}{(n_j(n_j - d_j))},$$

de modo que

$$Var(log(\hat{S}(t))) = \sum_{j=1}^{n}(Var(log(\hat{p_j}))) \approx \sum_{j=1}^{n}\left(\frac{d_j}{n_j(n_j - d_j)}\right),$$

aplicando de nuevo la aproximación por series de Taylor

$$Var(log(\hat{S}(t))) \approx \left[\frac{dlog(\hat{S}(t))}{d\hat{S}(t)}\right]^2 Var[\hat{S}(t)] \approx \left[\frac{1}{\hat{S}(t)}\right]^2 Var[\hat{S}(t)],$$

de modo que despejando $Var[\hat{S(t)}]$ se tiene que:

$$Var[\hat{S(t)}] \approx \hat{S(t)}^2 \sum_{j=1}^{n}\left(\frac{d_j}{n_j(n_j - d_j)}\right).$$

La última expresión es conocida como la fórmula de Greenwood, y el error estándar se define como la raíz cuadrada de la varianza estimada [5].

3.2. El estimador de producto límite como un estimador de máxima verosimilitud.

El estimador de producto límite puede ser utilizado como un estimador de máxima verosimilitud no paramétrico asumiendo un conjunto de tiempos discretos, así suponiendo que $T_1, T_2, ..., T_n$ son tiempos de vida provenientes de una distribución discreta, con función de supervivencia $S(t)$ y función de riesgo $h(t)$ y suponiendo que existe censura, la función de verosimilitud es de la forma:

$$L = \prod_{i=1}^{n}\prod_{t=0}^{\infty} h(t)^{dN_i(t)}[1 - h(t)]^{Y_i(t)(1-dN_i(t))},$$

donde t_i son tiempos de vida, $Y_i(t) = I(t_i \geq t)$ y $dN_i(t) = I(t_i = t, \delta_i = 1)$,

si se hace $d_t = \sum_{i=1}^{n} dN_i(t)$ y $n_t = \sum_{i=1}^{n} Y_i(t)$, entonces puede reescribirse en la forma:

$$L = \prod_{t=0}^{\infty} h(t)^{d_t}[1 - h(t)]^{(n_t - d_t)}.$$

Si se considera el vector $h = (h(0), h(1), ...,)$ como el parámetro desconocido en la distribución de tiempos de vida, maximizando la función de verosimilitud, se obtiene el máximo en $\hat{h}(t) = d_t/n_t$. Además, como la función de supervivencia puede verse como $S(t) = \prod_{j:T_j<t}[1 - h(t_j)]$ entonces el estimador de máxima verosimilitud para $S(t)$ está dado por:

$$\hat{S}(t) = \prod_{s=0}^{t-1}[1 - h(\hat{t}_j)] = \prod_{s=0}^{t-1}\left(1 - \frac{d_s}{n_s}\right).$$

Además a partir de la ecuación de verosimilitud dada se puede ver a la matriz de información como una matriz diagonal cuyas entradas en la diagonal son $\frac{\partial^2}{\partial h(r)^2} logL = n_r/(h(r)(1 - h(r))$ [12].

3.3. Estimador de Nelson-Aalen.

Se define al estimador Nelson-Aalen como:

$$\sum_{j:t_j \leq t} \frac{d_j}{n_j},$$

donde $t_1, t_2, ..., t_k$ representan los tiempos sin censura.

La función $H(t)$ es útil para conocer la forma que tiene la función de riesgo, por ejemplo si $H(t)$ es una función lineal, entonces $h(t)$ es una constante [16].

Además un estimador de la varianza de la función acumulada de riesgo está dado por $\hat{var}[H(t)] = \sum_{j:t_j \leq t} \frac{d_j(n_j - d_j)}{n_j^3}$.

Otro estimador para $S(t)$ sugerido en casos continuos está dado en función del estimador de $H(t)$ y es de la forma $\hat{S}(t) = \exp[-\hat{H}(t)]$, de igual forma un estimador sugerido para $H(t)$ es $\hat{H}(t) = -log\hat{S}(t)$ [13].

En la tabla 3.4 se muestra el estimador Nelson-Aalen para el ejemplo antes mencionado [12].

t_j	$\hat{H}(t_j)$
1	0.095
2	0.201
3	0.259
4	0.384
5	0.527
8	0.860
11	1.110
12	1.444
15	1.694
17	2.027
22	2.527
23	3.527

Tabla 3.4: Estimador Nelson-Aalen

su gráfica de riesgo se muestra en la figura 3.2.

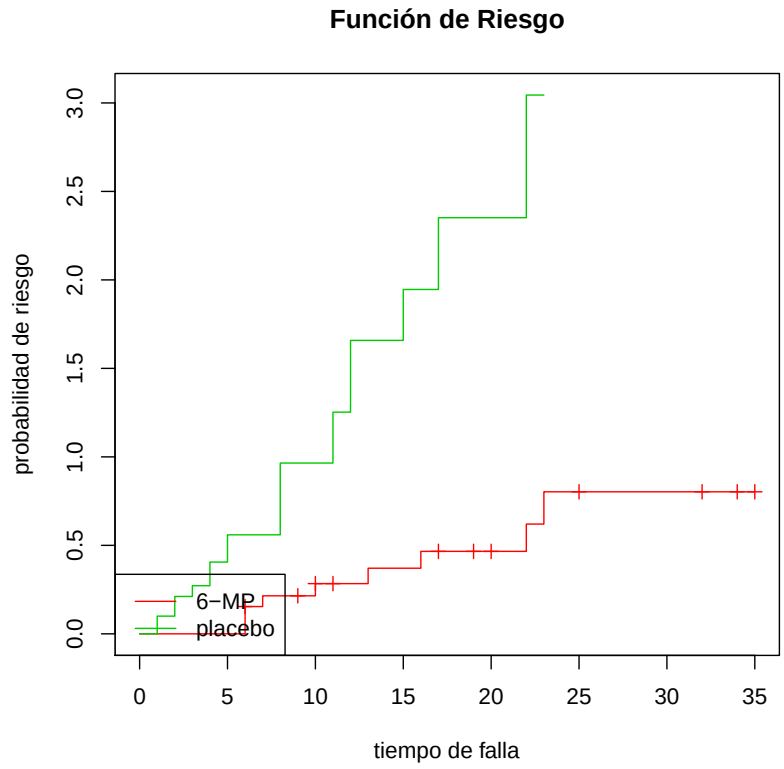


Figura 3.2: Gráfica de riesgo para la droga 6-Mp y Placebo.

3.4. Intervalos de estimación de probabilidades de supervivencia o cuantiles.

Aun cuando se tienen métodos no paramétricos, se pueden construir intervalos de confianza para tiempos de vida, de la misma forma que para métodos paramétricos, pero en la vida real son de mayor interés las probabilidades de supervivencia y los cuantiles, por lo que se empezará describiendo el método para obtener intervalos de confianza para probabilidades de supervivencia [12].

Dado que $\hat{S}(t)$ es el estimador producto límite para $S(t)$, entonces $\sqrt{n}(\hat{S}(t) - S(t))$ es asintóticamente normal, es decir, la variable

$$Z_1 = \frac{(\hat{S}(t) - S(t))}{\hat{\sigma_s}(t)},$$

se distribuye aproximadamente como una $N(0,1)$, en donde $\hat{\sigma_s}^2(t) = \hat{Var}(\hat{S}(t))$, así usando la aproximación pivotal se puede encontrar un intervalo de confianza de nivel α, entonces $P[a \leq Z_1 \leq b] = 1 - \alpha$, tomando a $a = -b$, donde, $b = z_{0.05}$, se obtiene el siguiente intervalo de confianza del 95 %,

$$\hat{S}(t) - z_{0.05}\hat{\sigma_s}(t) \leq S(t) \leq \hat{S}(t) + z_{0.05}\hat{\sigma_s}(t).$$

Cuando la muestra aleatoria tiene pocos datos censurados o $S(t)$ se acerca a 0 o a 1 entonces Z_1 puede no distribuirse como una N(0,1), por lo que para ajustar los intervalos se consideran funciones inyectivas $\phi(t) = g[S(t)]$ las cuales toman valores en $(-\infty, \infty)$ y el estimador de máxima verosimilitud es: $\hat{\phi} = g[(\hat{S(t)})]$ y la varianza estimada está dada por

$$\hat{\sigma_\phi}(t)^2 = (g'[\hat{S}(t)])^2 \hat{Var}[\hat{S}(t)].$$

Así la aproximación pivotal para una nueva variable definida como Z_2 es:

$$Z_2 = \frac{\hat{\phi}(t) - \phi(t)}{\hat{\sigma_\phi}(t)},$$

y esta variable está más cerca a la distribución normal estándar que la variable inicial Z_1 y esto es porque $\hat{S}(t)$ es un estimador no parámetrico de máxima verosimilitud y por las propiedades de los estimadores de máxima verosimilitud se justifica lo antes mencionado.

Algunas funciones ϕ usadas con frecuencia son la transformación logarítmica $\phi(s) = log\frac{(1-s)}{s}$ y la transformación log-log $\phi(s) = log(-log(s))$ [12].

Por ejemplo con la transformación $\phi(s) = log(-log(s))$, la transformación inversa es $S(t) =$

$e(-e^{\phi(t)})$ y el intervalo $\phi_L \leq \phi(t) \leq \phi_U$ se transforma en

$$e(-e^{\phi(u)}) \leq S(t) \leq e(-e^{\phi(L)}).$$

3.5. Intervalos de confianza por cuantiles.

Estimar la media es de interés en muchos experimentos pero en tiempos de vida es de mayor interés estimar cuantiles t_p de la distribución, como son la mediana o el cuantil $t_{0.5}$, la ventaja de estimar estos cuantiles es debido a que éstos siempre existen y porque son más fáciles de encontrar cuando hay datos con censura [16].

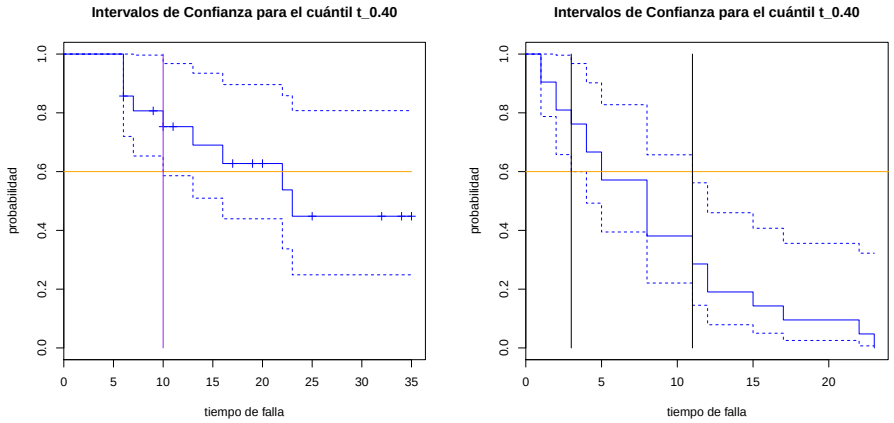
Se sabe que $\hat{S}(t)$ es una función escalonada, por lo que para algunos valores de p, existe un intervalo de t-valores que satisfacen $\hat{S}(t) = 1 - p$, es más para algún valor de p existe un t-valor y es común que se tome éste como el punto estimado para $\hat{t}_p$.

Estimar el cuantil t_p es de menor interés que encontrar un intervalo de confianza para ese cuantil, y es más fácil encontrar dicho intervalo si se invierte la relación entre la función de supervivencia y la distribución de los cuantiles.

Entonces si t_L es un límite de confianza inferior para t_p se observa que $P[t_L \leq t_p] = P[S(t_L) \geq 1 - p]$ por lo que si se quiere obtener un límite inferior para t_p, entonces se puede obtener encontrando el valor de t_L tal que $1 - p$ es un límite inferior α de confianza para $S(t)$, así el límite inferior para t_p es un valor t tal que $S(t) = 1 - p$.

Existe una manera gráfica para determinar un intervalo de confianza de nivel α para t_p y ésta es utilizando la gráfica de $\hat{S}(t)$, con sus respectivas bandas de confianza, las cuales también son funciones escalonadas, para encontrar límites de confianza de t_p, simplemente se encuentra la intersección entre las bandas de confianza y la línea $S(t) = 1 - p$ [12].

En el ejemplo que se mencionó anteriormente, si se quiere obtener un intervalo de confianza para el cuántil $t_{.40}$ para los tratamientos 6-MP y placebo, entonces $p = 0.40$ y se grafica la línea $S(t) = 1 - p = 1 - 0.40 = 0.60$ como se ve en la figura 3.3.



(a) Intervalo de confianza para el tratamiento 6-MP

(b) Intervalo de confianza para el tratamiento Placebo.

Figura 3.3: Intervalo de confianza para el cuántil $t_{0.40}$.

Por lo que el intervalo de confianza para el cuantil $t_{0.4}$ en el tratamiento 6-MP es $(10, \infty)$ y para el tratamiento de placebo es $(3, 11)$.

Capítulo 4

Estimación de los parámetros.

Suponga que se quiere estimar el parámetro $\underline{\hat{\Theta}}$ el cual es un vector k-dimensional, es decir, $\underline{\hat{\Theta}}= (\theta_1, \theta_2, ..., \theta_k)$, en este capítulo se describen algunos métodos que se utilizan para estimar este parámetro.
Existen dos maneras de estimar el parámetro $\underline{\theta}$, la estimación puntual y la estimación por intervalo, en la estimación puntual se calcula un valor numérico para $\underline{\theta}$ y en la estimación por intervalo se obtiene una región k-dimensional de tal forma que la probabilidad que la región contenga al parámetro $\underline{\theta}$ es un valor específico [15].

Definición 4.1 *Al conjunto de todos los posibles valores que puede tomar θ se le llama espacio paramétrico y se denota como Ω.*

Definición 4.2 *Sea $T_1, T_2, ..., T_n$ una muestra aleatoria de una distribución con densidad $f_T(t; \underline{\theta})$ con $\underline{\theta} \in \Omega$.*
Un estadístico es una función que depende de la muestra tal que no contiene parámetros desconocidos, es decir $\hat{\Theta} = g(T_1, ..., T_n)$.

Algunos estadísticos de interés son: el promedio muestral $\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n}$ y la varianza muestral $S^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n-1}$.

Definición 4.3 *Sea $T_1, T_2, ..., T_n$ una muestra aleatoria de una distribución con densidad $f_T(t; \underline{\theta})$ con $\underline{\theta} \in \Omega \subseteq \mathbb{R}^k$. $\underline{\hat{\Theta}}$ se dirá que es un estimador si $\hat{\Theta} = g(T_1, ..., T_n)$ es un estadístico cuyo rango es un subconjunto del espacio paramétrico de la población, es decir $\hat{\theta} \in \Omega$.*

4.1. Propiedades de los estadísticos.

Definición 4.4 *Se dice que un estadístico S es suficiente para $\underline{\theta}$ si se cumple que:*

$$f(t_1, t_2, ..., t_n | T = t) = \frac{f(t_1, t_2, ..., t_n, t)}{g(t)},$$

es decir que la probabilidad condicional de la muestra aleatoria dado que $T = t$ no depende de θ.

Definición 4.5 *Sea $F = \{f(T, \theta); \theta \in \Omega\}$ una familia de densidades de probabilidad, se dirá que es completa si para cualquier función continua g y para cada $\theta \in \Omega$ se cumple que $E[g(T)] = 0 \Leftrightarrow P[g(T) = 0] = 1$.*

Así, una estadística es completa si la familia de distribuciones de T es completa.

4.2. Propiedades de los estimadores.

Definición 4.6 *Un estimador $\hat{\Theta}$ de θ se dice que es insesgado si $E[\hat{\Theta}] = \theta$, de otra forma se dice que es un parámetro sesgado y el sesgo de este estimador se calcula como: $s[\hat{\Theta}] = E[\hat{\Theta}] - \theta$.*

Definición 4.7 *Un estimador $\hat{\Theta}$ de θ se dice que es asintóticamente insesgado si $E[\hat{\Theta}] \to \theta$ cuando $n \to \infty$.*

Definición 4.8 *Un estimador $\hat{\Theta}$ de θ se dice que es consistente, si $\forall\ \epsilon > 0\ P[|\hat{\Theta} - \theta|] > \epsilon \to 0$ cuando $n \to \infty$, es decir el valor del parámetro estimado converge al valor verdadero cuando la muestra tiende a infinito.*
Un estimador $\hat{\Theta}$ de θ es consistente si es insesgado y $lim_{n\to\infty} V[\hat{\Theta}] = 0$.

Definición 4.9 *Un estimador insesgado $\hat{\theta}$ de un parámetro θ es de varianza mínima si $V[\hat{\theta}] \leq V[\hat{\theta}^*]$ para todo estimador insesgado $\hat{\theta}^*$.*

Desigualdad de Crámer-Rao.
Suponga que $\hat{\varphi}(\theta) = g(T_1, T_2, ..., T_n)$ es un estimador insesgado de $\varphi(\theta)$ entonces:

$$V[\hat{\varphi}] \geq \frac{[\varphi'(\theta)]^2}{E\left[\frac{\partial ln[L(T_1, T_2, ..., T_n; \theta)]}{\partial \theta}\right]^2},$$

donde L es la función de verosimilitud. A la expresión del lado derecho de la desigualdad anterior se le llama cota inferior de Crámer-Rao.

Definición 4.10 *Un estimador $\hat{\Theta}$ es eficiente si es de varianza mínima.*

Note que si un estimador alcanza la cota de Crámer-Rao entonces es eficiente.

4.3. Métodos para obtener estimadores.

4.3.1. Método de Momentos.

El método de momentos se basa en la expresión de los k momentos en términos de los parámetros de la distribución, relacionando los momentos muestrales con los momentos poblacionales, obteniendo k ecuaciones que contienen los k parámetros desconocidos y cuyas soluciones resultan ser los estimadores de los parámetros [16].

Se define el k-ésimo momento muestral alrededor del cero como:

$$M'_k = \sum_{i=1}^{n} \frac{X_i^k}{n},$$

y el k-ésimo momento muestral alrededor de la media como:

$$M_k = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^k}{n}.$$

El k-ésimo momento alrededor del origen como $\mu'_k = E[T^k]$, por lo que para el caso discreto se define como: $E[T^k] = \sum t^k f(t)$ y para el caso continuo como: $E[T^k] = \int_{-\infty}^{\infty} t^k f(t)dt$.
El k-ésimo momento alrededor de la media como: $\mu_k = E[(T-\mu)^k]$, por lo que para el caso discreto $\mu_k = \sum (t-\mu)^k f(t)$ y para el caso continuo $\mu_k = \int_{-\infty}^{\infty} (t-\mu)^k f(t)dt$.

Sea $T_1, T_2, ..., T_n$ una muestra aleatoria, cuya función de densidad es de la forma $f(T, \underline{\theta})$, donde $\underline{\theta}$ es un parámetro desconocido, entonces para estimar los parámetros $(\theta_1, ..., \theta_n)$ por el método de momentos, se deben formar las k ecuaciones provenientes de los momentos muestrales igualados a los momentos poblacionales, es decir:

$$M'_k = \mu'_k,$$

por lo que las soluciones de este sistema $(\widetilde{\theta_1}, ..., \widetilde{\theta_n})$ son los estimadores de momentos para los parámetros.

Propiedades de los estimadores de momentos.

P_1 : Son consistentes e insesgados asintóticamente.

P_2 : Normalidad Asintótica.

A continuación se presenta el siguiente conjunto de datos obtenidos de una prueba que se hizo a 20 objetos hasta que éstos fallaron.
Los datos se presentan a continuación en la tabla 4.1.

11.24	1.92
12.74	22.48
9.60	11.50
8.86	7.75
5.73	9.37
30.42	9.17
10.20	5.52
5.85	38.14
2.99	16.58
18.92	13.36

Tabla 4.1: Tiempo de los 20 objetos.

Suponga que los tiempos de vida de los objetos siguen una distribución Weibull y se desea estimar los parámetros de la distribución por el método de momentos, como son 2 parámetros se toman los primeros dos momentos alrededor del origen, es decir las ecuaciones que se tienen que resolver son: $M_1' = \mu_1'$ y $M_2' = \mu_2'$, sustituyendo se tiene que $\mu_1' = \beta\Gamma(1+\frac{1}{\alpha}) = \bar{t}$ y $\mu_2' = \beta\Gamma(1+\frac{2}{\alpha})$, despejando a α y β se tiene que $\hat{\alpha}$ se obtiene resolviendo la ecuación:

$$\frac{s^2}{\bar{t}^2} = \frac{\Gamma(1+\frac{2}{\hat{\alpha}})}{\Gamma^2(1+\frac{1}{\hat{\alpha}})} - 1,$$

y $\hat{\beta}$ resolviendo:

$$\hat{\beta} = \frac{\bar{t}}{\Gamma(1+\frac{1}{\alpha})}.$$

Así, resolviendo las ecuaciones anteriores, los estimadores de $\hat{\alpha} = 1.41626$ y $\hat{\beta} = 13.8678$, por lo que los datos siguen una distribución Weibull con parámetros de escala y forma antes mencionados.

El código utilizado en Mathematica para resolver las ecuaciones anteriores se encuentra en el apéndice A.1.

4.3.2. Estimadores por Máxima Verosimilitud.

Se define la función de verosimilitud como la función de densidad conjunta de las n's variables, la cual está en función de $\underline{\theta}$, es decir si una muestra aleatoria $T_1, T_2, ..., T_n$ tiene una función de densidad $f(T; \underline{\theta})$ entonces la función de verosimilitud está dada por:

$$L(\underline{\theta}) = f(x_1, x_2, ..., x_3; \underline{\theta}) = \prod_{i=1}^{n} f(x_i; \underline{\theta}).$$

Definición 4.11 *El estimador de máxima verosimilitud es la variable aleatoria* $\hat{\Theta}$=$\hat{\Theta}(T_1, T_2, ..., T_n)$ *tal que para cada realización* $(t_1, t_2, ..., t_n)$ *de* $(T_1, T_2, ..., T_n)$ *el valor correspondiente de* $\hat{\theta}$ *de* $\hat{\Theta}$ *maximiza a la función de verosimilitud.*

Por lo que para obtener los estimadores de máxima verosimilitud primero se debe obtener la función de verosimilitud, maximizarla con respecto al parámetro $\underline{\theta}$, algunas veces es más fácil maximizar la función logaritmo de verosimilitud y se puede utilizar cualquier método conocido del cálculo.

Propiedades de los estimadores por máxima verosimilitud.

P_1 : Asintóticamente eficientes.

P_2 : Bajo ciertas condiciones son consistentes en error cuadrado medio.

P_3 : Asintóticamente insesgados.

P_4 : Es una función que depende de la estadística suficiente minimal.

En el ejemplo anterior se obtuvieron los estimadores por momentos para la distribución Weibull, ahora se obtienen por el método de máxima verosimilitud.

Como se supuso anteriormente los datos siguen una distribución Weibull por lo que la función de verosimilitud está dada por:

$$L(\alpha, \beta) = \prod_{i=1}^{n} \left(\frac{\alpha}{\beta^{\alpha}} \right) t_i^{\alpha - 1} e \left(\frac{t_i}{\beta} \right)^{\alpha},$$

utilizando el método de la segunda derivada se tiene que los estimadores de máxima verosimilitud se obtienen al resolver las siguientes ecuaciones:

$$\frac{\sum_{i=1}^{n} (t_i^{\hat{\alpha}} \log t_i)}{\sum_{i=1}^{n} t_i^{\hat{\alpha}}} - \frac{1}{\hat{\alpha}} - \frac{1}{n} \sum_{i=1}^{n} \log t_i = 0,$$

y

$$\hat{\beta} = \left(\frac{1}{n} \sum_{i=1}^{n} t_i^{\hat{\alpha}} \right)^{\frac{1}{\hat{\alpha}}}.$$

Así resolviendo las ecuaciones anteriores por el método de Newton-Raphson se tiene que los estimadores por máxima verosimilitud son: $\hat{\alpha} = 1.54873$ y $\hat{\beta} = 14.1216$, de forma que los datos pueden ajustarse a una distribución Weibull con los parámetros de escala y forma antes mencionados.

El código utilizado para resolver las ecuaciones anteriores se encuentra en el Apéndice A.2.

Definición 4.12 *Sea $T_1, T_2, .., T_n$ una muestra aleatoria proveniente de una función de densidad $f_T(t, \underline{\theta})$ con $\underline{\theta} \in \mathbb{R}^k$. Sean $Y_i = u_i(t_1, t_2, ..., t_n)$ para $i \in \{1, 2, ..., m\}$ con $m \leq n$ estadísticos basadas en la muestra aleatoria, entonces se dice que $(Y_1, Y_2, ..., Y_m)$ son estadísticos conjuntamente suficientes para $\underline{\theta}$ basados en la muestra aleatoria si la densidad condicional $f_{T_1,T_2,...,T_n|Y_1,Y_2,...,Y_m}$ no depende del parámetro $\underline{\theta}$.*

Teorema 4.1 *El estadístico $\underline{Y} = (Y_1, Y_2, ..., Y_m)$ es un estadístico conjuntamente suficiente para θ basado en la muestra aleatoria si:*

$$f_{\underline{T}}(\underline{t}; \underline{\theta}) = \phi_1(u_1(x), u_2(x), ..., u_m(x), \underline{\theta}) * \phi_2(t_1, t_2, ..., t_n),$$

donde ϕ_2 no depende de $\underline{\theta}$.

Note que para el caso de la función de distribución Weibull la función de densidad conjunta se puede escribir como:

$$f_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n; \alpha, \beta) = \frac{\alpha^n}{\beta^{n\alpha}} \prod_{i=1}^{n} x_i^{\alpha} e[\sum_{i=1}^{n}(-\beta^{-\alpha}) + (\alpha - 1) \sum_{i=1}^{n} \log[x_i]],$$

entonces tomando a $\phi_2 = 1$ y a ϕ_1 como la expresión completa en el lado derecho de la función anterior, por el teorema 4.1, se dice que la estadística $Y = \sum_{i=1}^{n} \log[x_i]$ es una estadística conjuntamente suficiente para $\theta = (\alpha, \beta)$ basadas en la muestra aleatoria.

4.3.3. Método de percentiles.

Se sabe que el percentil p es un valor tal que $0 < p < 1$ y cumple que $P[T < t_p] = F(T_p) = p$, así que para obtener un estimador de percentil primero se obtiene la expresión de los percentiles en términos de los parámetros del modelo, es decir $p = F(t_p, \theta)$.

Los estimadores de percentiles se obtienen de la función de distribución empírica de los datos, obteniendo k ecuaciones que contienen k parámetros desconocidos, solucionando estas ecuaciones se encuentran los estimadores [14].

La gráfica de la función de distribución empírica es una gráfica de la probabilidad acumulada, la cuál es una gráfica escalonada y es denotada por $\hat{F}(t)$.

En el ejemplo que se ha trabajado, se supuso que los datos se distribuyen como una distribución Weibull, entonces se sabe que el percentil está dado por

$$t_p = \beta[-\log(1-p)]^{\frac{1}{\alpha}}.$$

Para estimar β note que

$$t_{(1-e^{-1})} = \beta[-\log[1-(1-e^{-1})]^{\frac{1}{\alpha}}] = \beta[1^{\frac{1}{\alpha}}] = \beta,$$

por lo que se toma como estimador de $\hat{\beta} = t_{(1-e^{-1})} = t_{0.632}$.
Ahora despejando de la ecuación $t_p = \beta[-\log(1-p)]^{\frac{1}{\alpha}}$ al parámetro α se obtiene que el estimador es:

$$\hat{\alpha} = \frac{\log[-\log(1-p)]}{\log(\frac{t_p}{t_{0.632}})}.$$

Seki y Yokoyama sugieren utilizar a $p = 0.31$ para obtener el estimador de α [19].

Así resolviendo las ecuaciones se tiene que $\hat{\beta} = 11.1$ y $\hat{\alpha} = 2.6296$ y por lo tanto los datos pueden ajustarse a una distribución Weibull con los parámetros antes mencionados.

El código que se utilizó para resolver las ecuaciones antes mencionadas se encuentra en el Apéndice A.3.

4.3.4. WPP.

Bajo la transformación Weibull $y = \log(-\log[1-F(t)])$ y $x = \log(t)$, una gráfica de y contra x se le llama la gráfica de probabilidad Weibull.
Como $F(t) = 1 - e^{-\left(\frac{t}{\beta}\right)^{\alpha}}$, entonces

$$y = \log(-\log[1-(1-e^{-\left(\frac{t}{\beta}\right)^{\alpha}})]),$$

$$\Rightarrow y = \alpha\log(t) - \alpha\log(\beta),$$

y como $x = \log(t)$, entonces:

$$y = \alpha x - \alpha\log(\beta),$$

la cuál es una ecuación de una línea recta cuya pendiente es α y el punto de intersección con el eje y es $-\alpha \log(\beta)$, por lo que existe una relación lineal entre x y y.
Para hacer una estimación por aproximación primero se tiene que obtener la gráfica de probabilidades Weibull (WPP), después se tiene que ajustar una línea recta a la gráfica y estimar los parámetros de esta línea [16].

Siguiendo el ejemplo para datos completos se obtiene que $\hat{\alpha} = 1.61288$ y $\hat{\beta} = 12.9468$.

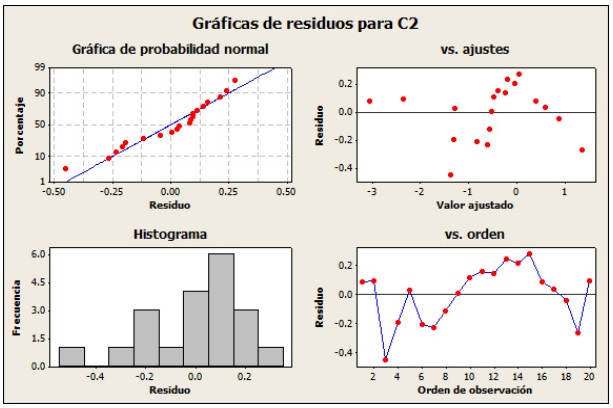


Figura 4.1: Gráfica de Residuales para el método WPP.

Con la gráfica 5.1 se observa que existe normalidad en los errores y dependencia entre los tiempos de vida y los errores.

El código que se utilizó para obtener los estimadores se encuentra en el Apéndice A.4.

Capítulo 5

Tipos de Censura.

Cuando se realiza un estudio sobre una muestra aleatoria de cierta población lo ideal sería que el resultado del estudio fuera observado para cada individuo de la muestra, por ejemplo, si se someten n individuos a un tratamiento que dura un periodo de tiempo $[0, t]$, para observar la eficacia de un nuevo medicamento que saldrá a la venta, es de interés que al término del estudio los n individuos de la muestra hayan cumplido con el tratamiento en tiempo y forma. Se sabe que esto no siempre es posible, debido a diferentes circunstancias que están fuera de las manos de la persona que realiza el experimento, a pesar de esta situación se quiere obtener la mayor información posible del individuo.

Dada una muestra aleatoria $T_1, T_2, ..., T_n$ que proviene de una función de distribución $F_T(t, \underline{\theta})$, con $\underline{\theta}$ un parámetro K-dimensional, T_i representa el tiempo de vida de algún individuo i con $0 \leq i \leq n$. Se denotará por $t_1, t_2, ..., t_n$ a la realización actual de los valores de T_i .

A continuación se presenta un ejemplo para ilustrar la diferencia entre datos completos y cuando existe censura [1].

En un estudio diseñado para evaluar el tiempo desde la primera infección de sitio de salida (en meses) en pacientes con insuficiencia renal, 43 pacientes utilizaron un catéter colocado quirúrgicamente (grupo 1) y 76 pacientes utilizaron la colocación percutánea de su catéter (grupo 2) en este caso se define a T_i como el tiempo de infección del i-ésimo paciente que está en el estudio.

Los resultados están dados en la siguiente tabla:

tiempos de infección	*Observaciones Censuradas*
1.5	2.5
3.5	2.5
4.5	3.5
4.5	3.5
5.5	3.5
8.5	4.5
8.5	5.5
9.5	6.5
10.5	6.5
11.5	7.5
15.5	7.5
16.5	7.5
18.5	7.5
23.5	8.5
26.5	9.5
	2.5
	2.5

Tabla 5.1: Tiempos de infección (meses) en pacientes con insuficiencia renal.

La tabla 5.1 muestra sólo los resultados de 32 pacientes, de los cuales sólo presentaron el evento 15 pacientes y los 17 restantes son datos censurados, es decir que hasta el termino del estudio los pacientes estaban vivos y no habían presentado el evento de interés.

5.1. Datos Completos.

Si para cada t_i con $0 \leq i \leq n$ los valores de la i-ésima realización son conocidos se dice que se tiene un conjunto de datos completos, es decir que para cada individuo se tendrá un resultado, no importa si es éxito o fracaso, el resultado es observable. Cuando se habla de tiempos de vida, se tendrá un éxito cuando el resultado es observable.
La función de verosimilitud asociada a los datos completos es:

$$L(\theta) = P[datos; \theta],$$

Ahora si se supone que $T_1, T_2, ..., T_n$ son tiempos de vida que tienen una función de densidad

$f_T(t)$ y un parámetro específico $\underline{\theta}$, entonces la función de verosimilitud está dada por:

$$L(\underline{\theta}) = \prod_{i=1}^{n} f_T(t_i, \underline{\theta}).$$

Esta función se puede maximizar con respecto a $\underline{\theta}$ para obtener un estimador, que se denomina estimador de máxima verosimilitud (EMV) y que se denotará por $\hat{\underline{\theta}}$.

5.2. Censura Tipo I.

Como se mencionó anteriormente lo ideal sería que si se tienen n individuos en un estudio se observarán n resultados, pero en la realidad esto difícilmente ocurre, es por ello que surgen los datos censurados.
Se define la variable

$$\delta_i = \begin{cases} 1, & \text{si} \quad t_i = T_i. \\ \\ 0, & \text{si} \quad T_i > t_i. \end{cases}$$

la cual indica si el i-ésimo individuo es un dato censurado o no observado ($\delta_i = 0$) o si es un dato observado ($\delta_i = 1$).

La censura por la derecha tipo I surge cuando se delimita el tiempo de un experimento, entonces el mecanismo de censura se aplica cuando cada individuo tiene un tiempo de censura fijo $C_i \geq 0$, entonces si $T_i \leq C_i$ se dice que T_i es un dato observado, de otro modo sólo se sabe que $T_i > C_i$, es decir sólo se sabe que sobrepaso el tiempo de censura, y por lo tanto es censurado [12].

Se observa que C_i, puede ser el mismo para todos los individuos o que C_i dependa de cada individuo, por ejemplo si se hace un estudio, se admiten n individuos a lo largo de un año, y la duración del estudio será por 2 años, entonces si un individuo entra al estudio a principios del año, digamos enero, su estudio terminará después de dos años en enero a diferencia del individuo que entra en junio ya que su estudio terminará dos años después en junio, aquí se observa que C_i va a variar dependiendo del individuo y de cuándo entró al estudio.

En esta notación se tiene que $t_i = min(T_i, C_i)$ que es el tiempo de censura para el individuo i [11].

Ahora dado que C_i son constantes fijas, tenemos las siguientes probabilidades

$$P[t_i = C_i, \delta_i = 0] = P[T_i > C_i] \quad \text{Dato censurado,}$$

$$P[t_i = C_i, \delta_i = 1] = f_T(t_i) \quad \text{Dato sin censura.}$$

De las probabilidades anteriores se obtiene la probabilidad para cualquier tipo de dato, ya sea con o sin censura, en términos de la función indicadora.

$$P[t_i = C_i, \delta_i] = f_T(t_i)^{\delta_i} P[T_i > C_i]^{1-\delta_i}.$$

Si se asume que las T_i son independientes entonces la función de verosimilitud es de la forma

$$L = \prod_{i=1}^{n} f_T(t_i)^{\delta_i} S(t_i+)^{1-\delta_i}.$$

En dónde $S(t_i+) = P[T_i > C_i]$ y además si $S(t_i)$ es continua en t_i entonces $S(t_i) = S(t_i+)$.

En la tabla 5.2 se muestran los tiempos de falla de 50 objetos que fueron probados, Ver [16], el estudio terminó después de ser observados por 12 horas.

0.80	1.26	1.29	1.85	2.41
2.47	2.76	3.35	3.68	4.46
4.65	4.83	5.21	5.26	5.36
5.39	5.53	5.64	5.80	6.08
6.38	7.02	7.18	7.60	8.13
8.46	8.69	10.52	11.25	11.90

Tabla 5.2: Tiempos de falla para datos con censura tipo I.

En los datos anteriores se dan los tiempos de falla de 30 objetos, de los 20 objetos restantes sólo se sabe que sobrepasaron las 12 horas, por lo que se tiene una censura de tipo I, suponiendo que se distribuyen como una distribución Weibull, se obtienen sus estimadores por el método de máxima verosimilitud.

La función de verosimilitud está dada por:

$$\frac{\alpha^k}{\beta^{\alpha k}} \left[\prod_{i=1}^{k} (t_i)^{\alpha-1}\right] \exp\left(-\frac{1}{\beta^\alpha} \sum_{i=1}^{k} t_i^\alpha + (n-k)v^\alpha\right),$$

donde se ordenan los datos de tal forma que los primeros k son datos sin censura, el resto censurados y v es el tiempo de censura [16].

El estimador de máxima verosimilitud de $\hat{\alpha}$ se obtiene resolviendo:

$$\frac{\sum_{i=1}^{50} t_i^{\hat{\alpha}} \log t_i}{\sum_{i=1}^{50} t_i^{\hat{\alpha}}} - \frac{1}{\hat{\alpha}} - \frac{1}{30}\sum_{i=1}^{30} \log t_i = 0,$$

y el estimador de $\hat{\beta}$ esta dado por:

$$\hat{\beta} = \left(\frac{1}{30}\left[\sum_{i=1}^{30} t_i^{\hat{\alpha}} + (50-30)v^{\hat{\alpha}}\right]\right)^{\frac{1}{\hat{\alpha}}}.$$

En este ejemplo el tiempo de censura es fijo para todos los datos, es decir $v = 12$ y resolviendo las ecuaciones anteriores se tiene que los estimadores por máxima verosimilitud son: $\hat{\alpha} = 1.33403$ y $\hat{\beta} = 8.43874$.

El código para resolver las ecuaciones anteriores se encuentra en el Apéndice B.1.

Para obtener los estimadores por el método WPP (Weibull Probability Plot) el procedimiento es similar al de datos completos, la diferencia radica en la gráfica de la función de distribución empírica ya que para cuando no existe censura la aproximación que se utiliza es: $\hat{F}(t) = \frac{1}{n}$, usando los estadísticos de orden $t_{(1)}, t_{(2)}, ..., t_{(n)}$.

Cuando existe censura, se utiliza la misma aproximación, sólo que en este caso no se grafican los datos censurados, solo se trabaja con los datos sin censura [17].

Por este método los estimadores obtenidos son: $\hat{\alpha} = 1.42764$ y $\hat{\beta} = 10.9433$.
El código para obtener los estimadores por el método WPP se encuentra en el Apéndice B.2.

5.3. Censura Tipo II.

Este tipo de censura surge cuando n individuos comienzan un estudio al mismo tiempo, y este termina cuando se han observado r fracasos. Por ejemplo si en un estudio donde n individuos han sido sometidos a un tratamiento en contra de cierta enfermedad y se puede observar que r individuos han fallecido después cierto tiempo, en ese momento se termina el estudio, y se dice que se tiene una censura de tipo II.

En este tipo de censura se escogen los primeros r valores más pequeños de los tiempos de vida de la muestra, luego para distribuciones continuas se toman los estadísticos de orden $T_{(1)}, T_{(2)}, \ldots, T_{(r)}$,

cuya función de densidad conjunta es

$$\frac{n!}{(n-r)!}\prod_{i=1}^{r} f(t_{(i)})S(t_{(r)})^{n-r}.$$

Despreciando la constante $\frac{n!}{(n-r)!}$ y en términos de (δ_i, t_i) la función de máxima verosimilitud queda de la forma

$$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}. \tag{5.1}$$

de donde nuevamente se observa que la función de verosimilitud está en términos de la función indicadora, por lo que todos los datos de la muestra, sean datos censurados o no, están indicados por la expresión 5.1 [12].

En el siguiente ejemplo que se muestra en la tabla 5.3, obtenido de [16], 30 objetos fueron probados, el estudio terminó al presentarse 20 fallos, por lo que se presenta una censura tipo II:

2.45	3.74	3.92	4.99	6.73
7.52	7.73	7.85	7.94	8.25
8.37	9.75	10.86	11.17	11.37
11.60	11.96	12.20	13.24	13.50

Tabla 5.3: Tiempos de falla para datos con censura tipo II.

La función de verosimilitud está dada por:

$$L(\beta,\alpha) = \frac{\alpha^r}{\beta^{\alpha r}}\left(\prod_{i=1}^{r} t_i^{\alpha-1}\right)\exp\left(-\frac{1}{\beta^\alpha}\left[\sum_{i=1}^{r} t_i^\alpha + (n-r)t_r^\alpha\right]\right),$$

donde los primero r datos son sin censura.

El estimador del parámetro de forma α se obtiene resolviendo:

$$\frac{\sum_{i=1}^{r} t_i^{\hat{\alpha}}\log[t_i] + (n-r)t_r^{\hat{\alpha}}\log[t_r]}{\sum_{i=1}^{r} t_i^{\hat{\alpha}} + (n-r)t_r^{\hat{\alpha}}} - \frac{1}{\hat{\alpha}} - \frac{1}{r}\sum_{i=1}^{r}\log[t_i] = 0,$$

y el estimador para el parámetro de escala se obtiene resolviendo:

$$\hat{\beta}^\alpha = \frac{1}{r}\left[\sum_{i=1}^{r} t_i^{\hat{\alpha}} + (n-r)t_r^{\hat{\alpha}}\right].$$

Por lo que para el ejemplo anterior los estimadores por máxima verosimilitud son : $\hat{\alpha} = 2.43625$ y $\hat{\beta} = 13.0843$.

El código para resolver las ecuaciones anteriores por el método de Newton-Rhapson se encuentra en el Apéndice C.1.

Por el método de WPP para datos censurados se obtuvo que los estimadores son: $\hat{\alpha} = 2.00868$ y $\hat{\beta} = 13.5951$.

El código para obtener los estimadores por el método WPP cuando existe censura tipo II, se encuentra en el apéndice C.2.

5.3.1. Censura tipo II progresiva.

Este proceso aunque no es viable en la práctica, es de importancia teórica conocerlo. En este proceso al igual que en la censura tipo II, se observan los primeros r_1 fracasos de la muestra que se tenía, se quitan los r_1 fracasos quedando una muestra de tamaño $n - r_1$, luego de los elementos que aún quedan en la muestra se van a remover n_1 datos, dejando una muestra de tamaño $n - r_1 - n_1$ datos, se vuelve a observar esta muestra hasta que se tengan r_2 nuevos fracasos que van a ser removidos otra vez, y de la muestra que es ahora de tamaño $n - r_1 - n_1 - r_2$, se quita de nuevo una muestra de n_2 elementos y así progresivamente se va reduciendo la muestra inicial, este procedimiento termina después de que se realiza varias series de repeticiones de este proceso.

Por simplicidad supongamos que en este proceso solo se tienen 2 repeticiones, es decir el procedimiento termina cuando se han observado r_2 fracasos en la muestra. En este punto se tendrían $n - n_1 - r_1$ elementos que aún no han fallado. Para hacer diferencia entre los primeros r_1 elementos que han fallado y los de la segunda etapa, los denotaremos como $T_{(1)}, ..., T_{(r_1)}$ y $T^*_{(1)}, ..., T^*_{(r_2)}$ respectivamente [12].

Entonces para los primeros r_1 fracasos la función de verosimilitud está dada por

$$\frac{n!}{(n-r_1)!} \prod_{i=1}^{r_1} f(t_{(i)}) S(t_{(r_1)})^{n-r_1}.$$

Luego se observa que los $t_{(i)}$ tienen una función de densidad truncada por la izquierda por lo

que la función de densidad y función de supervivencia normalizadas son respectivamente:

$$f_I(t) = \frac{f(t)}{S(t_{(r1)})} \quad S_I(t) = \frac{S(t)}{S(t_{(r1)})}.$$

Así que, la función de verosimilitud de los $T^*_{(i)}$ es de la forma

$$\frac{(n-r_1-n_1)}{(n-r_1-n_1-r_2)!)!}\prod_{i=1}^{r_2} f_I(t^*_{(i)})[S_I(t_{(r_2)})^{n-r_1-n_1-r}].$$

Ahora sí se combina la función de verosimilitud de las $T_{(i)}$ y $T^*_{(i)}$ se obtiene

$$\frac{n!(n-r_1-n_1)!}{(n-r_1)!(n-n_1-r_1-r_2)!}\prod_{i=1}^{r_1} f(t_{(i)})[S(t_{(r_1)})^{(n-r_1)}]\prod_{i=1}^{r_2} f(t^*_{t(i)})[S(t^*_{t(i)})^{(n-n_1-r_1-r)}].$$

Que de igual forma usando la función indicadora y despreciando la constante se tiene:

$$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}.$$

Del análisis anterior se concluye que la función de verosimilitud es de la forma

$$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i},$$

para los datos completos, censura tipo I, tipo II o tipo II progresiva.

5.4. Censura Aleatoria Independiente.

La censura aleatoria se tiene cuando se asume que cada individuo tiene un tiempo de vida T y un tiempo de censura C las cuales son variables aleatorias independientes y continuas [16], la función de verosimilitud es de la forma:

$$L(\underline{\theta}) = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i},$$

donde δ_i es como se definió anteriormente.

Suponga que se tiene el siguiente conjunto de datos mostrados en la tabla 5.4 con censura aleatoria independiente [16]:

9	14	18	25	27
31+	35	40	48	50+

Tabla 5.4: Tiempos de falla con censura aleatoria.

y suponga que los datos siguen una distribución Weibull.
Sea $u_i = min\{t_i, v_i\}$ donde t_i es el valor de la realización y v_i es el tiempo de censura para T_i, así los estimadores de máxima verosimilitud se obtienen resolviendo:

$$\hat{\beta}^{\alpha} = \frac{1}{n}\left(\sum_{i=1}^{n} u_i^{\hat{\alpha}}\right),$$

$$\frac{\sum_{i=1}^{n} u_i^{\hat{\alpha}} \log[u_i]}{\sum_{i=1}^{n} u_i^{\hat{\alpha}}} - \frac{1}{\hat{\alpha}} - \frac{1}{k}\sum_{i \in D} \log[u_i] = 0,$$

donde D es el conjunto de datos sin censura.

Luego, resolviendo las ecuaciones anteriores por el método de Newton-Raphson se tiene que $\hat{\alpha} = 2.10992$ y $\hat{\beta} = 32.7408$.

El código utilizado para resolver las ecuaciones anteriores se encuentra en el Apéndice D.1.
y por el método WPP se obtiene: $\hat{\alpha} = 1.60654$ y $\hat{\beta} = 35.7918$.

El código para obtener los estimadores por el método WPP se encuentra en el Apéndice D.2.

5.5. Inferencia por máxima verosimilitud con datos censurados.

Sea $\underline{\theta}$ un k-vector de parámetros, $L(\underline{\theta})$ una función de verosimilitud, $l(\underline{\theta})$ la función logaritmo de verosimilitud, se define como vector record al px1 vector

$$U(\underline{\theta}) = \frac{\partial}{\partial \underline{\theta}} l(\underline{\theta}).$$

La ecuación de verosimilitud como:

$$U(\underline{\theta}) = \frac{\partial}{\partial \underline{\theta}} l(\underline{\theta}) = 0.$$

Y a la matriz pxp

$$I(\underline{\theta}) = -\frac{\partial^2}{\partial \underline{\theta} \partial \underline{\theta}'} l(\underline{\theta}),$$

se le llama la matriz de información.

El parámetro estimado $\underline{\hat{\theta}}$ satisface a la ecuación record.
La matriz de Fisher se define como

$$I[\underline{\theta}] = E[I(\underline{\theta})].$$

Se sabe que en muestras aleatorias grandes $\underline{\hat{\theta}}$ se va a distribuir en forma normal con parámetros $\underline{\mu} = \underline{\theta}$ y $\underline{\sigma^2} = \frac{I[\underline{\theta}]}{n}$ y además el estadístico $\Delta(\underline{\theta}) = 2l(\underline{\hat{\theta}}) - 2l(\underline{\theta})$ se aproxima a una χ^2 [12].

5.6. Otros tipo de datos incompletos.

5.6.1. Observaciones intermitentes y Censuras por Intervalos.

Dado que al hacer un estudio no se puede tener bajo observación a la muestra todo el tiempo que dura el estudio, ya sean pacientes, máquinas, etc., surge un tipo de censura llamado censura por intervalos, el cual consiste en dividir el tiempo que durará el estudio en intervalos más pequeños de tiempo, es decir $0 = a_{i,0} \leq a_{i,1} \leq .... \leq a_{i,m_1}$ donde i es el individuo al que le pertenece ese intervalo y m_1 el tiempo que se estudia a dicho individuo [11].

Así las observaciones se hacen por intervalos, es decir se observa en el tiempo $a_{i,j-1}$ si no se ha presentado el evento, la siguiente observación se hace en el tiempo $a_{i,j}$, por lo tanto las observaciones se hacen en intervalos de la forma (U_i, V_i), es decir sólo se sabrá que $U_i \leq T_i \leq V_i$, se observa que si $V_i = \infty$ y $U_i = a_{i,m_1}$entonces el intervalo $(U_i, V_i) = (U_i, \infty)$ es una censura por la derecha [12].

Para obtener la función de verosimilitud se nota que las observaciones provienen de una función multinomial con probabilidad tal que

$$p_{i,j} = P[a_{i,j-1} \leq T_i \leq a_{i,j}] = F_i[a_{i,j}] - F_i[a_{i,j-1}].$$

Así la función de verosimilitud está dada por

$$L = \prod_{i=1}^{n} \left[F_i[U_i] - F_i[V_i]\right].$$

Se menciona un ejemplo donde se observa un intervalo por censura: suponga que un conjunto de n individuos va a ser observado durante un periodo de 3 meses para ver la eficacia de cierto medicamento que disminuye el colesterol en la sangre. El evento de interés es la disminución de colesterol en la sangre, pero éste sólo se puede medir mediante laboratorios, por lo que a los individuos se les asigna el tratamiento y se les programan las citas en donde se tomarán laboratorios. En este ejemplo puede observar una censura por intervalos, ya que los 3 meses serán

divididos en intervalos fijos de tiempo (citas programadas) y si se observa el evento, el investigador no sabrá la fecha exacta en que se presentó pero sabrá en qué intervalo de tiempo se presentó dicho evento.

Del ejemplo anterior se observa que se divide el tiempo del estudio en intervalos que son fijos, pero en la realidad no se puede sostener esta suposición ya que puede depender de los eventos previos para poder asignar la siguiente observación, por ejemplo, en un estudio clínico basado en la información del paciente se le asigna una fecha para su consulta.

Se observa que esta probabilidad ahora es condicional, ya sea que dependa de la información previa del individuo o de alguna covariable, por lo que se define a $H(a_{i,j-1})$ como la historia del tiempo de observación y ahora la elección de $a_{i,j}$ es condicional. En este caso se tiene

$$P[a_{i,j-1} \leq T_i \leq a_{i,j} | H(a_{i,j-1}), a_{i,j}] = \frac{F_i(a_{i,j}) - F_i(a_{i,j-1})}{1 - F(a_{i,j-1})},$$

donde se asume que $H(a_{i,j-1})$ incluye la información del individuo si está vivo y sin censura. Luego la función de verosimilitud es proporcional a:

$$\prod_{j=1}^{m_1-1} P(T_i \geq a_{i,j} | H(a_{i,j-1}), a_{i,j}) P(T_i \leq a_{i,m1} | H(a_{i,m-1})).$$

5.6.2. Censura Doble.

Este tipo de censura, si se observa como censura por intervalo surge de la observación de los tiempos de vida T_i cuando hay dos eventos distintos, por ejemplo cuando una persona adquiere el virus del VIH, y cuando se diagnostica SIDA. De igual forma, la censura es porque el tiempo T_i sólo se sabrá que fue observado en algún intervalo [12].

También se puede observar esta doble censura en un experimento cuando en éste se observa tanto censura por la izquierda como censura por la derecha, y se dice que existe una doble censura, por ejemplo cuando en un experimento se le pregunta a un cierto número de estudiantes si han probado la marihuana alguna vez, si algún individuo indica haber probado la marihuana pero no recuerda la fecha exacta en la que la probó, entonces se sabe que el evento, que en este caso es que sí han probado la marihuana, ha ocurrido pero no se sabe la fecha exacta, solo se tendría un cierto intervalo en el que ocurrió, por lo que tenemos una censura por la izquierda, ahora si otro individuo de la muestra niega haber probado la marihuana entonces tenemos una censura por la derecha, es decir hasta el tiempo t_i no se ha observado el evento por lo que en este caso $t_i < T_i$,

ya que en un futuro puede que el estudiante pruebe la marihuana [11].

5.6.3. Entrada Retrasada y truncación por la izquierda.

Cuando se realiza un estudio se puede dar el caso en que no todos los individuos de dicho estudio entren al mismo tiempo, por lo que se dice que el tiempo t será distinto a 0 y que cada individuo tendrá un tiempo $u_i > 0$ de entrada.

Seleccionar un individuo en un tiempo u_i en lugar de un tiempo $t = 0$ da origen a una truncación por la izquierda ya que se quiere que el evento sea observado es decir que $T_i \geq u_i$, por lo que el dato observado para el individuo i consiste de $(u_i, t_i, \delta_i, x_i)$, donde x_i representa la covariable.

Como se tiene una truncación por la izquierda definimos la función de supervivencia para esta variable truncada por la izquierda como $\frac{S(t|x)}{S(u|x)}$, por lo que la función de verosimilitud para n individuos con tiempos de vida independientes está dada por la siguiente expresión:

$$L = \prod_{i=1}^{n} \left(\frac{f_i(t_i)}{S_i(u_i)} \right)^{\delta_i} \left(\frac{S_i(t_i+)}{S_i(u_i)} \right)^{1-\delta_i}.$$

De donde se puede observar que $\frac{f_i(t_i)}{S_i(u_i)}$ es la función de densidad de la variable truncada y como se mencionó anteriormente la función de supervivencia es de la forma $\frac{S_i(t_i+)}{S_i(u_i).}$

Ahora se menciona un ejemplo para ilustrar un caso donde se da el ingreso tardío.

Un tiempo de vida se puede observar como el tiempo entre dos eventos por ejemplo el inicio de una enfermedad letal y la muerte por lo que en este caso el primer evento se tiene cuando se adquiere la enfermedad y el segundo evento es cuando se presenta la muerte y T entonces es el tiempo de supervivencia que tiene un individuo que ha adquirido dicha enfermedad. En este ejemplo se puede observar que no todos los individuos tendrán el mismo tiempo de entrada sino que cada uno tendrá su propio tiempo dependiendo de la fecha en que adquirió cierta enfermedad [11].

5.6.4. Observaciones Retrospectivas.

En un estudio retrospectivo se puede observar a los individuos desde su entrada en el tiempo u_i hasta el tiempo de falla o censura $u_i \leq t_i$

Para ilustrar esto se da el siguiente ejemplo: Se analizaron los datos de personas contagiados de

VIH por medio de transfusión sanguínea, los cuales después fueron diagnosticados con SIDA. La forma en que los datos fueron obtenidos es de forma retrospectiva ya que los datos fueron tomados en el año 1987 y pertenecen a esta muestra aquellas personas que han sido diagnosticadas con SIDA antes del 1 de julio de 1986 y que fueron infectados por transfusión sanguínea, se sabe que en este caso la fecha en que fueron infectados puede ser censurada por la izquierda. Los datos son usados para estimar la distribución del tiempo T entre la infección por VIH y el diagnóstico de VIH, por ello el conjunto de datos cumple que $T_i \geq v_i$ donde v_i es el tiempo entre la infección del individuo con VIH y el 1 de julio de 1986 [12]. Así la función de verosimilitud está dada por

$$\prod_{i=1}^{n} Pr(t_i|v_i, T_i \geq v_i) = \prod_{i=1}^{n} \frac{f_i(t_i)}{F_i(v_i)}.$$

Capítulo 6

Pruebas de bondad de Ajuste.

Bajo la suposición de que un conjunto de datos siguen una cierta distribución con parámetros obtenidos por los métodos descritos anteriormente, es importante verificar que esta suposición sea correcta [16],
Algunos métodos informales son los gráficos tales como las gráficas de probabilidad y las gráficas de residuos. Las pruebas de bondad de ajuste involucran pruebas de hipótesis, cuyo objetivo es probar la hipótesis nula: H_0 : Los datos provienen de una función de distribución acumulada $F(t;\underline{\theta})$ particular. Para esta prueba se pueden utilizar los estadísticos de Pearson's, ji-cuadrada y Kolmogorov-Smirmov.

6.1. Prueba de Kolmogorov- Smirnov.

El juego de hipótesis para esta prueba es: H_0 : Los datos analizados siguen una distribución específica. VS H_a : Los datos analizados no siguen dicha distribución.
El estadístico de prueba D_n se define como la distancia máxima entre la función de distribución empírica y la función de distribución acumulada, es decir:

$$D_n = Máx\{D_n^+, D_n^-\},$$

donde:

$$D_n^+ = máx_{i=1,2,\ldots,n}\left[\frac{i}{n} - F(t_{(i)})\right],$$

y

$$D_n^- = máx_{i=1,2,\ldots,n}\left[F(t_{(i)}) - \frac{(i-1)}{n})\right].$$

La regla de decisión es de la forma:
Si $D_n \leq D_\alpha \Rightarrow$ No se rechaza H_0 y sí $D_n \geq D_\alpha \Rightarrow$ se rechaza H_0 con un nivel de significancia α.

El p-valor se define como: $P[D_n \geq D_o | H_0 \; es \; cierta]$, donde D_o es el valor observado de la estadística. Si el p-valor es grande significa que siendo cierta la hipótesis nula era de esperarse el valor del estadístico D_n, por lo que no hay razón para rechazar H_0, de manera análoga si el p-valor fuera pequeño implicaría que siendo cierta la hipótesis nula, no se podría obtener el valor de D_n y por lo tanto pone en duda la veracidad de H_0, por lo que está se rechaza.

6.1.1. Prueba de Kolmogorov-Smirnov para datos completos.

Para el ejemplo donde se trabajaron datos completos, por el método de momentos se obtuvo que los estimadores son: $\hat{\alpha} = 1.41626$ y $\hat{\beta} = 13.8678$.

Realizando la prueba de Kolmogorov-Smirnov en R se tiene la figura 6.1:

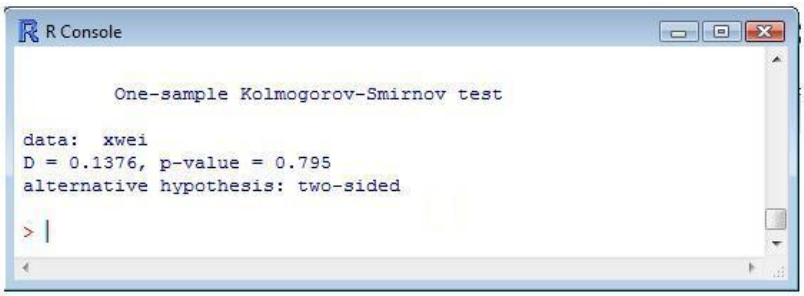


Figura 6.1: Prueba de Kolmogorov-Smirnov para los estimadores obtenidos por el método de momentos.

En este caso el valor del estadístico es $D_n = 0.1376$, si se toma un nivel significancia $\alpha = 0.05$, entonces $D_\alpha = 0.18841$, el cual se puede verificar en la tabla que se encuentra en el Apéndice K, entonces $D_n \leq D_\alpha$ y no se rechaza H_0.

También se puede observar que el p-valor es grande, entonces no se rechaza la hipótesis nula, es decir no se rechaza que los datos provengan de una función de distribución Weibull $F(t; \underline{\theta})$, donde $\underline{\theta} = (\hat{\beta}, \hat{\alpha}) = (13.8678, 1.41626)$.

Por el método de Máxima verosimilitud los estimadores que se obtuvieron son: $\hat{\alpha} = 1.54869$ y $\hat{\beta} = 14.1216$.

Realizando la prueba de Kolmogorov-Smirnov en R se tiene la figura 6.2:

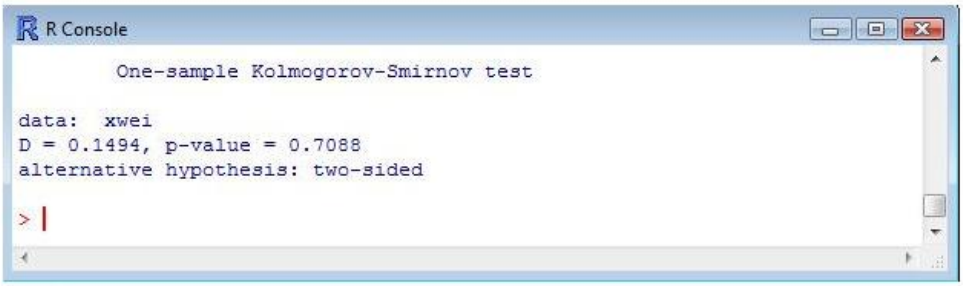


Figura 6.2: Prueba de Kolmogorov-Smirnov para los estimadores obtenidos por el método de máxima verosimilitud.

En este caso el valor del estadístico es $D_n = 0.1494$, si se toma un nivel significancia de $\alpha = 0.05$ entonces $D_\alpha = 0.18841$ por lo que $D_n \leq D_\alpha$ y no se rechaza H_0.

De igual manera debido a que el p-valor es grande no se rechaza la hipótesis nula, por lo que no se rechaza que los datos provengan de una función de distribución Weibull $F(t;\underline{\theta})$, donde $\underline{\theta} = (\hat{\beta}, \hat{\alpha}) = (14.1216, 1.54869)$.

Por el método de percentiles se obtuvo que los estimadores son: $\hat{\beta} = 11.1$ y $\hat{\alpha} = 2.6296$
Realizando la prueba de Kolmogorov-Smirnov en R se tiene la figura 6.3:

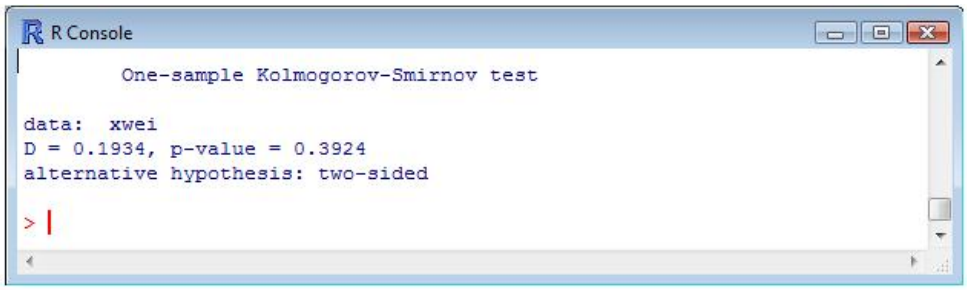


Figura 6.3: Prueba de Kolmogorov-Smirnov para los estimadores obtenidos por el método de percentiles.

El valor del estadístico es: $D_n = 0.1934$, si el nivel de significancia es $\alpha = 0.02$ entonces $D_\alpha = 0.21068$ por lo que $D_n \leq D_\alpha$ y no se rechaza H_0.
El p-valor es alto por lo que no se rechaza la hipótesis nula, es decir que no se rechaza que los datos provengan de una función de distribución Weibull $F(t;\underline{\theta})$, donde $\underline{\theta} = (\hat{\beta}, \hat{\alpha}) = (11.1, 2.6296)$. Note que si el nivel de significancia es $\alpha = 0.05$, es decir $D_\alpha = 0.18841$, entonces $D_\alpha < D_n$ por lo que se rechaza H_0. Es decir, aunque la aproximación es buena, los estimadores que se obtienen por los otros métodos son mejores, ésto es debido a que el método de percentiles es un método gráfico.

Por el método de WPP se obtuvo que los estimadores son: $\hat{\beta} = 12.9468$ y $\hat{\alpha} = 1.61288$.
Realizando la prueba de Kolmogorov-Smirnov en R se tiene la figura 6.4:

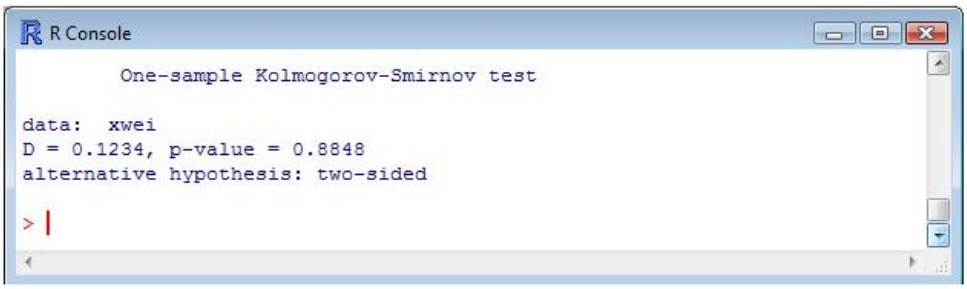


Figura 6.4: Prueba de Kolmogorov-Smirnov para los estimadores obtenidos por el método de WPP.

En este caso el valor del estadístico $D_n = 0.1234$ con un nivel significancia de $\alpha = 0.05$, $D_\alpha = 0.18841$ por lo que $D_n \leq D_\alpha$ y no se rechaza H_0.
De igual manera el p-valor es alto por lo que no se rechaza la hipótesis nula, es decir los datos provienen de una distribución $F(t;\underline{\theta})$, donde $\underline{\theta} = (\hat{\beta}, \hat{\alpha}) = (12.9468, 1.61288)$.

6.1.2. Prueba de Kolmogorov-Smirnov para datos con censura tipo I.

En la sección de datos con censura se presentó un ejemplo de datos con censura tipo I con la suposición que siguen una distribución Weibull, se obtuvieron los estimadores por máxima verosimilitud y por el método WPP. En esta sección se hará la prueba de Kolmogorov-Smirnov para verificar que los datos siguen una distribución Weibull.
Los estimadores obtenidos por el método de máxima verosimilitud son: $\hat{\alpha} = 1.33403$ y $\hat{\beta} = 8.43874$
Realizando la prueba de Kolmogorov-Smirnov en R se tiene la figura 6.5:

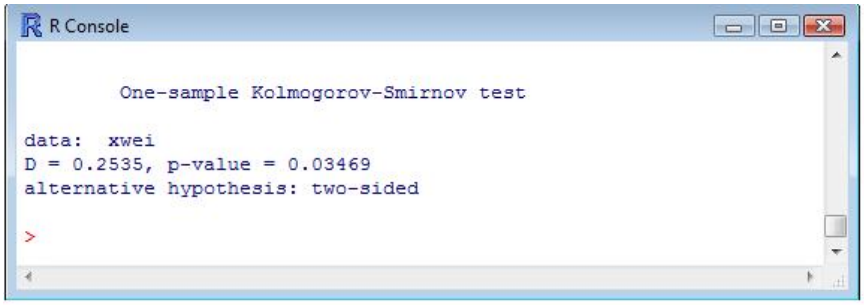


Figura 6.5: Prueba de Kolmogorov-Smirnov para los estimadores obtenidos por el método de máxima verosimilitud.

El estadístico obtenido es: $D_n = 0.2535$ y si se toma un nivel de significancia $\alpha = 0.02$, se observa que $D_\alpha = 0.27023$ y $D_n < D_\alpha$ por lo que no hay evidencia para rechazar H_0.

De igual modo se observa que el p-valor es alto, por lo que no se rechaza H_0, entonces se puede afirmar que los datos con censura tipo I se distribuyen como una Weibull con parámetros $\hat{\alpha} = 1.33403$ y $\hat{\beta} = 8.43874$.

Por el método de WPP, se obtuvieron los estimadores: $\hat{\alpha} = 1.42764$ y $\hat{\beta} = 10.9433$
Aplicando la prueba de Kolmogorov-Smirnov en R se obtiene la figura 6.6:

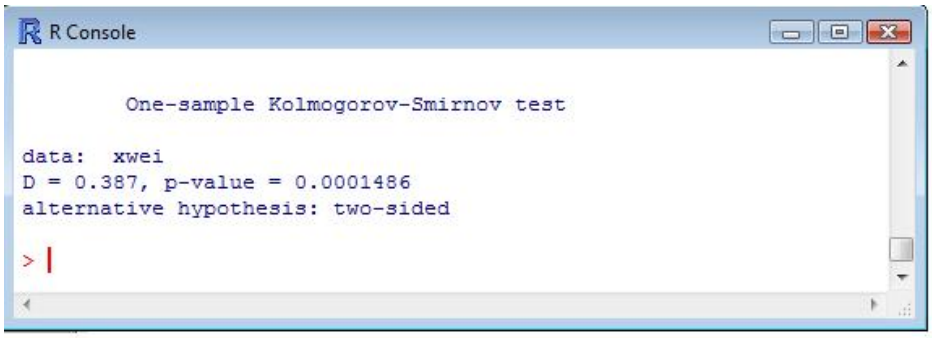


Figura 6.6: Prueba de Kolmogorov-Smirnov para los estimadores obtenidos por el método de WPP.

En este caso el estadístico de prueba es $D_n = 0.387$, si se toma un nivel de significancia $\alpha = 0.05$, se observa que $D_\alpha = 0.24170$ entonces $D_\alpha < D_n$ por lo que se rechaza H_0.

Además se observa que el p-valor es muy pequeño, por lo que se rechaza H_0, es decir los datos no se distribuyen como una Weibull con parámetros $\hat{\alpha} = 1.42764$ y $\hat{\beta} = 10.9433$.

6.1.3. Prueba de Kolmogorov-Smirnov para datos con censura tipo II.

Anteriormente se presentó un ejemplo de datos con censura tipo II bajo la suposición de que los datos siguen una distribución Weibull, se obtuvieron los estimadores por máxima verosimilitud y por el método WPP, en esta sección se presentan las pruebas de Kolmogorov-Smirnov para bondad de ajuste.

Los estimadores obtenidos por el método de máxima verosimilitud son: $\hat{\alpha} = 2.43625$ y $\hat{\beta} = 13.0843$.
Realizando la prueba de Kolmogorov-Smirnov en R se obtiene la figura 6.7:

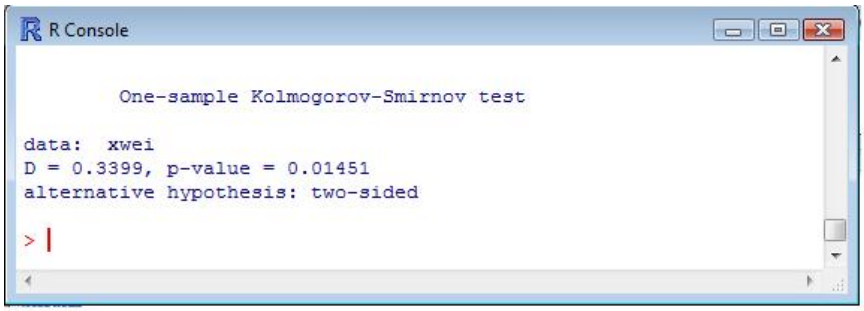


Figura 6.7: Prueba de Kolmogorov-Smirnov para los estimadores obtenidos por el método de máxima verosimilitud.

El estadístico de prueba es $D_n = 0.3399$ si se toma un nivel de significancia de $\alpha = 0.01$ se observa que $D_\alpha = 0.35241$ entonces $D_n < D_\alpha$, por lo tanto no se rechaza H_0, así se afirma que los datos provienen de una distribución Weibull con parámetros $\hat{\alpha} = 2.43625$ y $\hat{\beta} = 13.0843$.

Por el método WPP se obtuvo que los estimadores son: $\hat{\alpha} = 2.00868$ y $\hat{\beta} = 13.5951$.
Aplicando la prueba de Kolmogorov-Smirnov en R se tiene la figura 6.8

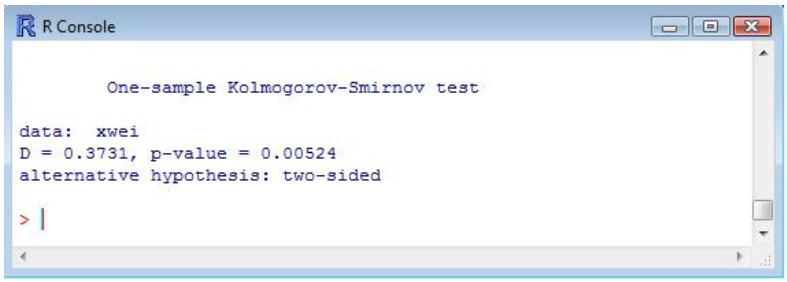


Figura 6.8: Prueba de Kolmogorov-Smirnov para los estimadores obtenidos por el método de WPP.

El estadístico es $D_n = 0.3731$ tomando un nivel de significancia de $\alpha = 0.05$ se tiene que $D_\alpha = 0.29408$ y $D_n > D_\alpha$ por lo que se rechaza H_0. Además el p-valor es muy pequeño, por lo que se rechaza H_0 es decir los datos no provienen de una distribución Weibull con parámetros: $\hat{\alpha} = 2.00868$ y $\hat{\beta} = 13.5951$.

6.1.4. Prueba de Kolmogorov-Smirnov para datos con censura aleatoria independiente.

En la sección de Censura se presentó un ejemplo donde los datos tienen una censura aleatoria independiente, bajo la suposición de que los datos se distribuyen como una Weibull, se obtuvieron los estimadores por máxima verosimilitud y por el método WPP.

Los estimadores por máxima verosimilitud son: $\hat{\alpha} = 2.10992$ y $\hat{\beta} = 32.7408$.

Aplicando la prueba de Kolmogorov-Smirnov se tiene la figura 6.9:

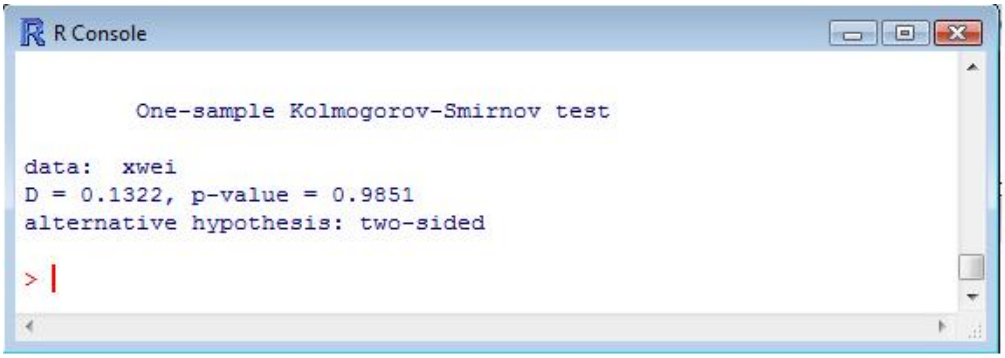


Figura 6.9: Prueba de Kolmogorov-Smirnov para los estimadores obtenidos por el método de máxima verosimilitud.

Para este caso el estadístico de prueba es $D_n = 0.1322$, así que sí se toma un nivel de significancia de $\alpha = 0.05$, $D_\alpha = 0.40925$ y $D_n < D_\alpha$ por lo tanto no se rechaza H_0.

Además como el p-valor es grande entonces no se rechaza H_0 por lo que no se rechaza que los datos provengan de una distribución Weibull con parámetros $\hat{\alpha} = 2.10992$ y $\hat{\beta} = 32.7408$.
Por el método WPP se obtuvo que los estimadores son: $\hat{\alpha} = 1.60654$ y $\hat{\beta} = 35.7918$.

Aplicando la prueba de Kolmogorov-Smirnov en R se tiene la figura 6.10:

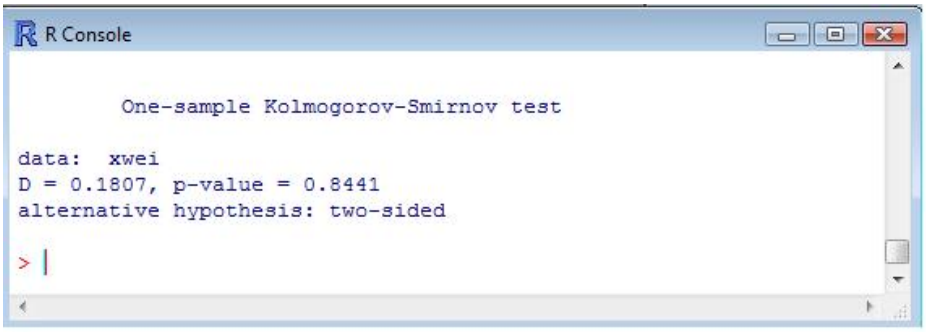


Figura 6.10: Prueba de Kolmogorov-Smirnov para los estimadores obtenidos por el método de WPP.

El estadístico de prueba es $D_n = 0.1807$, si se toma un nivel de significancia de $\alpha = 0.05$ entonces $D_\alpha = 0.40925$ y $D_n < D_\alpha$ por lo tanto no se rechaza H_0.
Además como el p-valor es grande entonces no se rechaza H_0 por lo que no se rechaza que los datos provengan de una distribución Weibull con parámetros $\hat{\alpha} = 1.60654$ y $\hat{\beta} = 35.7918$.

6.2. QQ-Plot.

Es un método gráfico para probar si un conjunto de datos proviene de cierta distribución propuesta, en este método se grafican los cuantiles o percentiles de la distribución propuesta contra los cuantiles de la distribución teórica.

Para el ejemplo que se presento en la tabla 4.1 donde se tiene datos completos, por el método de momentos se obtuvo que los estimadores son: $\hat{\alpha} = 1.41626$ y $\hat{\beta} = 13.8678$.
Realizando la prueba QQ-Plot se obtiene la gráfica 6.11:

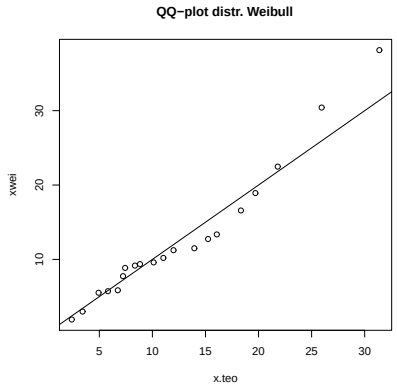


Figura 6.11: Prueba QQ-Plot para estimadores obtenidos por el método de momentos.

Se observa que los datos se distribuyen alrededor de la línea recta, por lo que se puede decir que los datos se distribuyen como una Weibull con parámetros $\hat{\alpha} = 1.41626$ y $\hat{\beta} = 13.8678$.

Los estimadores de máxima verosimilitud son: $\hat{\alpha} = 1.54869$ y $\hat{\beta} = 14.1216$.
Realizando la prueba QQ-Plot se obtiene la gráfica 6.12:

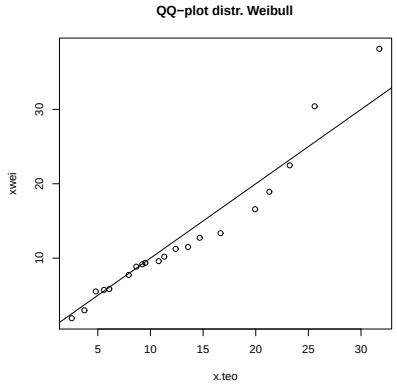


Figura 6.12: Prueba QQ-Plot para estimadores obtenidos por el método de máxima verosimilitud.

Se observa que los datos se distribuyen alrededor de la línea recta, por lo que se puede decir que los datos se distribuyen como una Weibull con parámetros $\hat{\alpha} = 1.54869$ y $\hat{\beta} = 14.1216$.
Por el método de percentiles se obtuvo que los estimadores son: $\hat{\alpha} = 2.6296$ y $\hat{\beta} = 11.1$.
Realizando la prueba QQ-Plot se obtiene la gráfica 6.13:

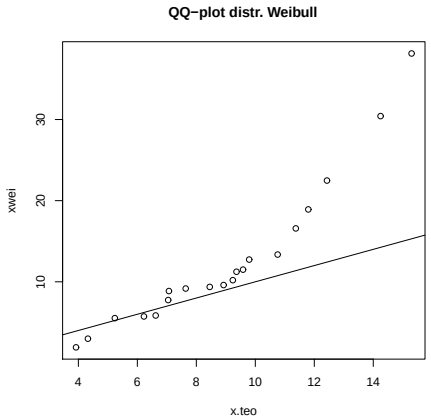


Figura 6.13: Prueba QQ-Plot para estimadores obtenidos por el método de percentiles.

Se observa que los datos no se distribuyen alrededor de la línea recta, se puede decir que los datos no se distribuyen como una Weibull con parámetros $\hat{\alpha} = 2.6296$ y $\hat{\beta} = 11.1$.
Por el método WPP se obtuvo que los estimadores son: $\hat{\alpha} = 1.61288$ y $\hat{\beta} = 12.9468$.
Realizando la prueba QQ-Plot se obtiene la gráfica 6.14:

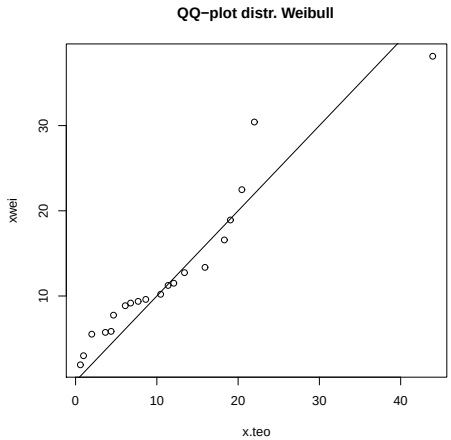


Figura 6.14: Prueba QQ-Plot para estimadores obtenidos por el método de WPP.

Se observa que los datos se distribuyen alrededor de la línea recta, por lo que se concluye que los datos se distribuyen como una Weibull con parámetros $\hat{\alpha} = 1.61288$ y $\hat{\beta} = 12.9468$.

6.2.1. Prueba QQ-Plot para datos con censura I.

Para el ejemplo donde se presentó datos con censura I mostrado en la tabla 5.2, por el método de máxima verosimilitud se obtuvo que los estimadores son: $\hat{\alpha} = 1.41626$ y $\hat{\beta} = 13.8678$.
Realizando la prueba QQ-Plot se obtiene la gráfica 6.15:

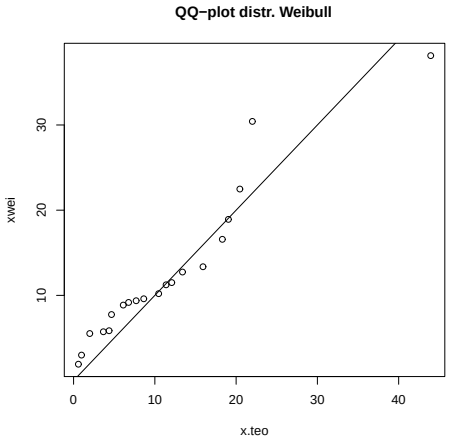


Figura 6.15: Prueba QQ-Plot para estimadores obtenidos por el máxima verosimilitud.

Los datos se distribuyen alrededor de la línea recta, así se asegura que los datos se distribuyen como una Weibull con parámetros $\hat{\alpha} = 1.41626$ y $\hat{\beta} = 13.8678$.

Por el método WPP se obtuvo que los estimadores son: $\hat{\alpha} = 1.42764$ y $\hat{\beta} = 10.9433$.
Realizando la prueba QQ-Plot se obtiene la gráfica 6.16:

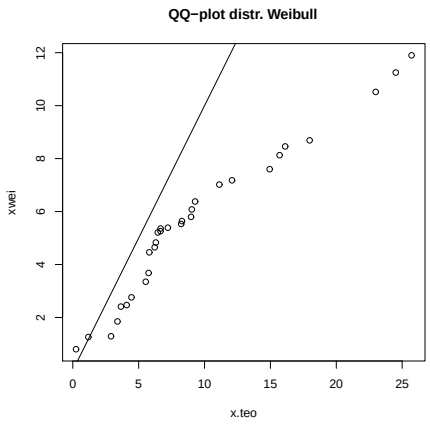


Figura 6.16: Prueba QQ-Plot para estimadores obtenidos por el método de WPP.

Los datos no se distribuyen alrededor de la línea recta, entonces se dice que los datos no se distribuyen como una Weibull con parámetros $\hat{\alpha} = 1.42764$ y $\hat{\beta} = 10.9433$, esto confirma lo que se

dijo en la prueba de Kolmogorov-Smirnov.

6.2.2. Prueba QQ-Plot para datos con censura tipo II.

Para el ejemplo donde se presentó datos con censura tipo II mostrado en la tabla 5.3, por el método de máxima verosimilitud se obtuvo que los estimadores son: $\hat{\alpha} = 2.43625$ y $\hat{\beta} = 13.0843$.

Realizando la prueba QQ-Plot se obtiene la gráfica 6.17:

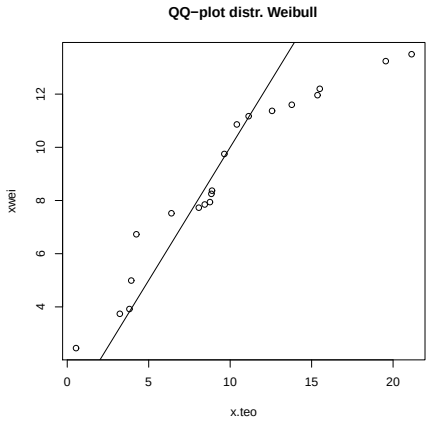


Figura 6.17: Prueba QQ-Plot para estimadores obtenidos por el máxima verosimilitud.

Se observa que los datos se distribuyen alrededor de la línea recta, entonces los datos se distribuyen como una Weibull con parámetros $\hat{\alpha} = 2.43625$ y $\hat{\beta} = 13.0843$.
Por el método WPP se obtuvo que los estimadores son: $\hat{\alpha} = 2.00868$ y $\hat{\beta} = 13.5951$.
Realizando la prueba QQ-Plot se obtiene la gráfica 6.18:

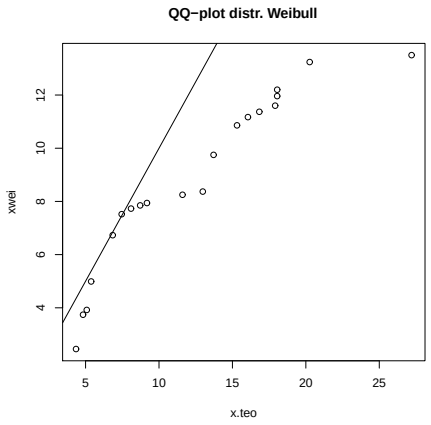


Figura 6.18: Prueba QQ-Plot para estimadores obtenidos por el método de WPP.

Se observa que los datos no se distribuyen alrededor de la línea recta, esto es por que los datos no se distribuyen como una Weibull con parámetros $\hat{\alpha} = 2.00868$ y $\hat{\beta} = 13.5951$, esto confirma lo que se dijo en la prueba de Kolmogorov-Smirnov.

6.2.3. Prueba QQ-Plot para datos con censura aleatoria.

Para el ejemplo donde se presentó datos con censura aleatoria mostrado en la tabla 5.4, por el método de máxima verosimilitud se obtuvo que los estimadores son: $\hat{\alpha} = 2.10992$ y $\hat{\beta} = 32.7408$

Realizando la prueba QQ-Plot se obtiene la gráfica 6.19:

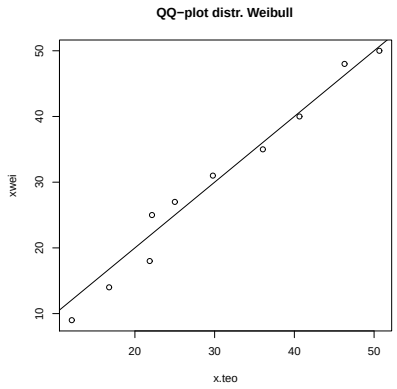


Figura 6.19: Prueba QQ-Plot para estimadores obtenidos por el máxima verosimilitud.

Se observa que los datos se distribuyen alrededor de la línea recta, por lo que se concluye que los datos se distribuyen como una Weibull con parámetros $\hat{\alpha} = 2.10992$ y $\hat{\beta} = 32.7408$.

Por el método WPP se obtuvo que los estimadores son: $\hat{\alpha} = 1.60654$ y $\hat{\beta} = 35.7918$
Realizando la prueba QQ-Plot se obtiene la gráfica 6.20:

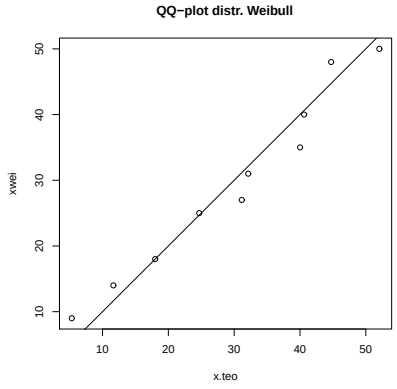


Figura 6.20: Prueba QQ-Plot para estimadores obtenidos por el método WPP.

Se observa que los datos se distribuyen alrededor de la línea recta, por lo tanto los datos se distribuyen como una Weibull con parámetros $\hat{\alpha} = 1.60654$ y $\hat{\beta} = 35.7918$

Capítulo 7

Caso de estudio, Diabetes Mellitus Tipo II.

En el Apéndice F se tienen los datos de pacientes con Diabetes, de los cuales los primeros 15 pacientes han fallecido y se sabe el tiempo promedio que vivieron con la enfermedad, los 31 pacientes que siguen son pacientes censurados, en este caso los pacientes están vivos y siguen bajo control.

Debido a que a los pacientes se les diagnosticó la enfermedad en diferentes fechas y que algunos pueden vivir más tiempo con la enfermedad, cada paciente tiene su propio tiempo de censura y tiempo de vida, así se tiene que los datos tienen una censura aleatoria.

Utilizando R se obtiene la función de supervivencia empírica [18] para cada individuo, la cual se muestra en la figura 7.1:

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	46	1	0.978	0.0215	0.9370	1.000
3	44	1	0.956	0.0304	0.8982	1.000
4	43	2	0.912	0.0422	0.8324	0.998
10	26	1	0.877	0.0532	0.7782	0.987
15	20	3	0.745	0.0833	0.5984	0.928
17	14	1	0.692	0.0928	0.5318	0.900
18	11	1	0.629	0.1035	0.4555	0.868
20	8	2	0.472	0.1237	0.2821	0.789
30	3	2	0.157	0.1348	0.0293	0.844
35	1	1	0.000	NaN	NA	NA

Figura 7.1: Función de supervivencia para datos de pacientes con diabetes tipo II.

por lo que se obtiene la curva de supervivencia empírica en la figura 7.2:

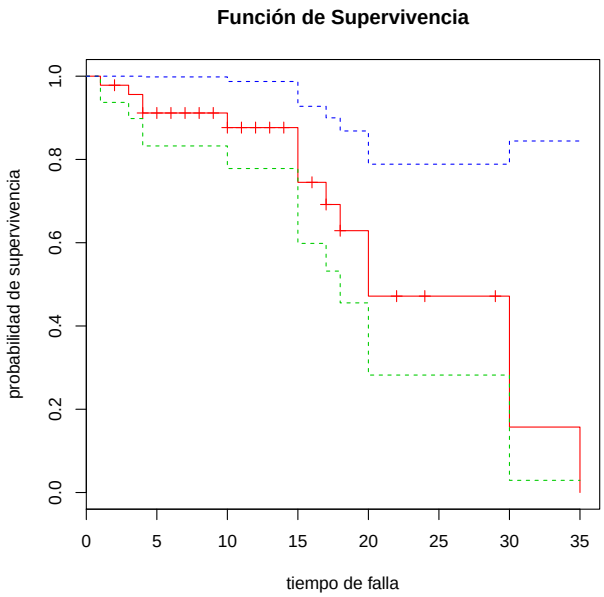


Figura 7.2: Curva de supervivencia con bandas de confianza para datos de pacientes con Diabetes tipo II. La banda de confianza la conforman los segmentos en color verde y los segmentos en color rojo.

Se observa que la probabilidad de sobrevivir a esta enfermedad es alta. Usando la transformación logaritmo logaritmo se obtiene la figura 7.3:

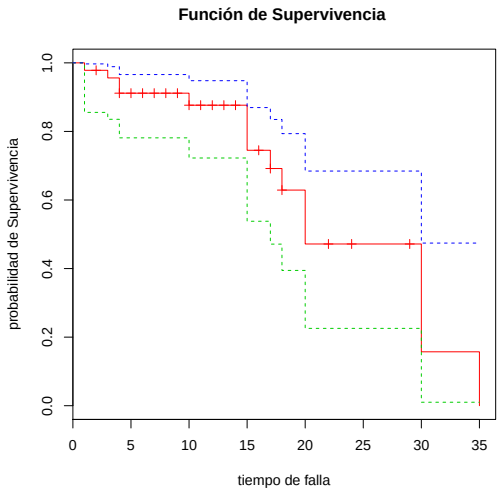


Figura 7.3: Curva de supervivencia con bandas de confianza para datos de pacientes con Diabetes tipo II, usando la transformación log-log.

La función de riesgo se muestra en la figura 7.4:

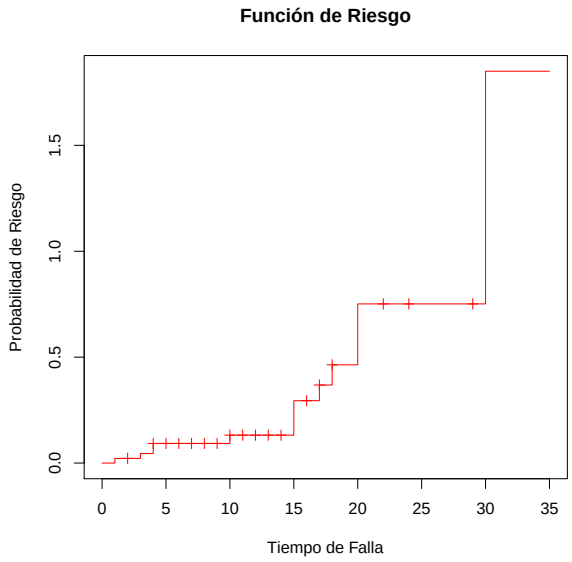


Figura 7.4: Función de riesgo para datos de pacientes con diabetes tipo II.

Se observa que en los primeros 15 años de vivir con diabetes el riesgo de morir es bajo, sin embargo al pasar los 15 años el riesgo aumenta de forma muy pronunciada.

Se obtienen los intervalos de confianza para el cuantil $t_{0.5}$ en la figura 7.5,

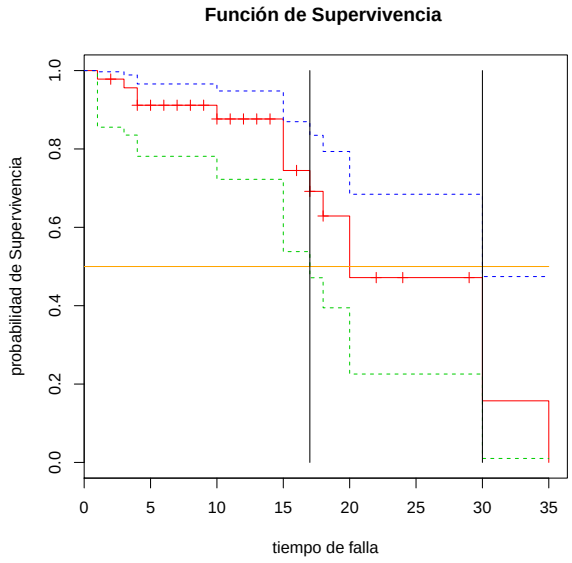


Figura 7.5: Intervalos de confianza para el cuantil $t_{0.5}$ para datos de pacientes con Diabetes tipo II.

Por lo que un intervalo de confianza para la mediana en este caso es: $(17, 30)$, como se muestra en la figura 7.5.

Si se toma el percentil $t_{0.4}$ se tiene que un intervalo de confianza para tal cuantíl es $(15, 30)$, como se muestra en la figura 7.6.

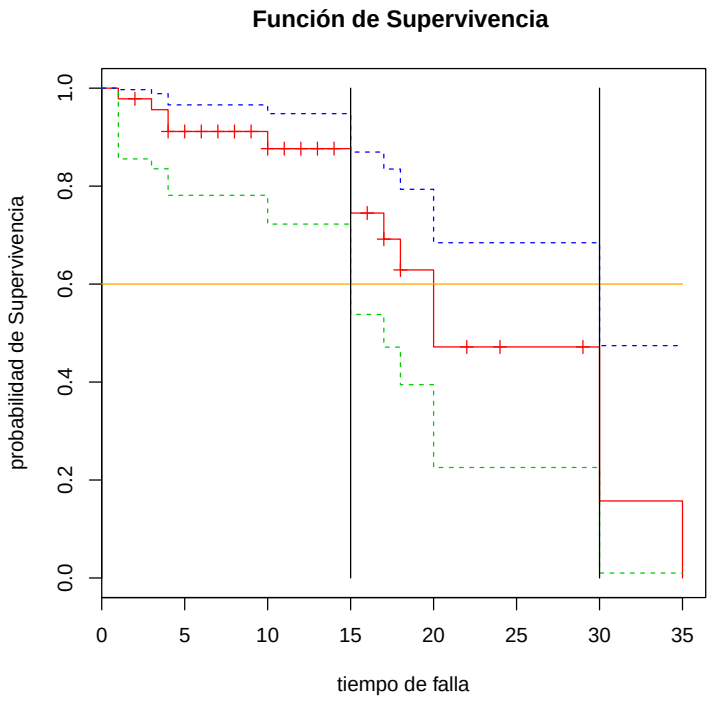


Figura 7.6: Intervalos de confianza para el percentil $t_{0.4}$ para datos de pacientes con Diabetes tipo II.

Como se explicó en capítulos anteriores, la distribución Weibull es una distribución utilizada en tiempos de vida, debido a las formas variadas que puede tomar su función de riesgo, por lo que para este caso se aproximan los datos a una distribución Weibull.

Para este caso los datos presentan una censura aleatoria, por lo que se estiman los parámetros de la distribución Weibull por el método de máxima verosimilitud y por el método de WPP.

Por el método de máxima verosimilitud se tiene que los parámetros estimados son: $\hat{\beta} = 1.82892$ parámetro de escala y $\hat{\alpha} = 14.7051$ parámetro de forma.

A continuación se presentan las pruebas de Kolmogorov-Smirnov y QQ-Plot, para verificar la bondad de ajuste del modelo propuesto al tomar los estimadores de Máxima Verosimilitud para los parámetros.

En R utilizando la prueba de Kolmogorov-Smirnov se tiene en la figura 7.7:

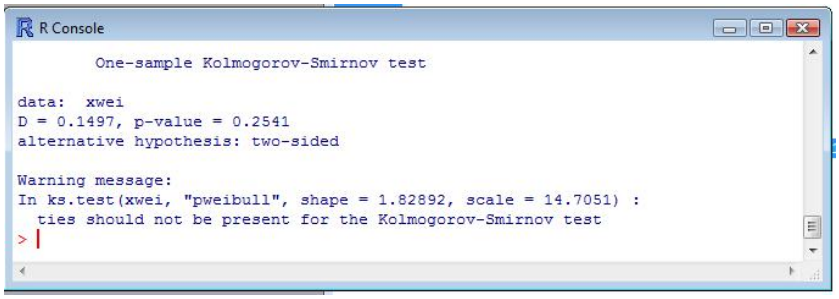


Figura 7.7: Prueba de Kolmogorov-Smirnov para datos de pacientes con Diabetes tipo II.

El valor del estadístico $D_n = 0.1497$ por lo que si se toma un nivel de significancia $\alpha = 0.05$ entonces $D_\alpha = 0.19625$ y $D_n < D_\alpha$, por lo que no se rechaza H_0, es decir que no se rechaza que los datos provengan de una distribución Weibull con parámetros: $\hat{\beta} = 1.82892$ y $\hat{\alpha} = 14.7051$.

Por el método QQ-Plot se tiene la figura 7.8:

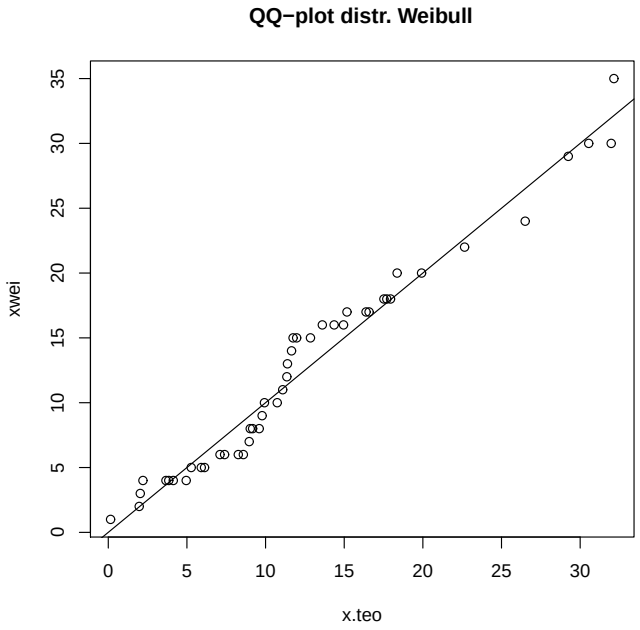


Figura 7.8: Prueba QQ-Plot para datos de pacientes con Diabetes tipo II.

Y se puede observar que el ajuste es bueno, por lo que se concluye que los datos siguen una distribución Weibull con parámetros: $\hat{\beta} = 1.82892$ y $\hat{\alpha} = 14.7051$.

Ahora, por el método WPP se tiene que los parámetros estimados son: $\hat{\beta} = 113.433$ y $\hat{\alpha} = 0.8116$.

De forma semejante a lo hecho para el modelo obtenido con estimaciones por Máxima Verosimilitud, a continuación se presentan las pruebas de Kolmogorov-Smirnov y QQ-Plot, para verificar la bondad de ajuste del modelo propuesto al tomar los estimadores para los parámetros, obtenidos por el método WPP.

En la figura 7.9 se muestra la prueba de Kolmogorov-Smirnov:

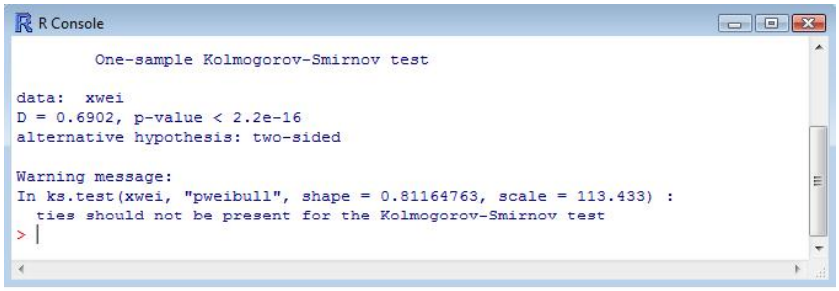


Figura 7.9: Prueba de Kolmogorov-Smirnov para datos de pacientes con Diabetes tipo II.

Se observa que el estadístico de prueba es $D_n = 0.6902$, por lo que para cualquier nivel de significancia, se rechaza H_0, es decir que los datos no provienen de una distribución Weibull con parámetros: $\hat{\beta} = 113.433$ y $\hat{\alpha} = 0.8116$.

Por el método QQ-Plot se tiene la figura 7.10:

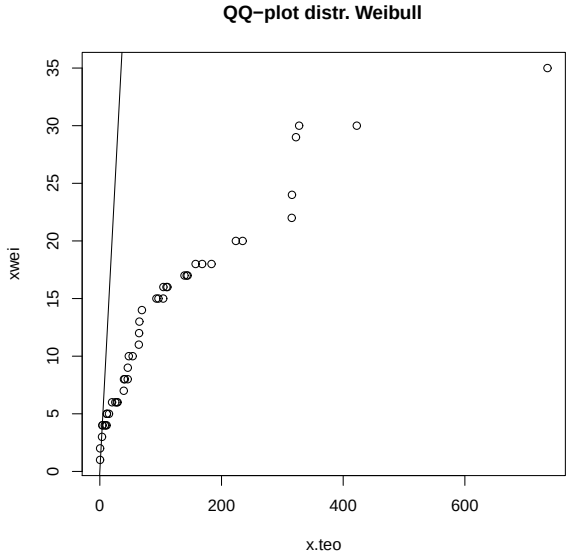


Figura 7.10: Prueba QQ-Plot para datos de pacientes con Diabetes tipo II.

Se observa que los datos no se ajustan a la línea, por lo que se afirma que los datos no siguen

una distribución Weibull con los parámetros antes mencionados.

Los datos de pacientes con Diabetes no siguen una distribución Weibull con los parámetros obtenidos por el método WPP para datos con censura aleatoria, esto es debido a que los datos presentan demasiados datos censurados y en estos casos el método WPP no es recomendable. Por lo que para el caso de estudio que se analiza, se utilizaran los estimadores de Máxima Verosimilitud.

Capítulo 8

Análisis de Resultados, Conclusiones e Investigaciones Futuras.

Se revisaron conceptos importantes del análisis de supervivencia, como son:

- Tipos de Censura: Dando ejemplos, que se pueden encontrar en diferentes experimentos, se explican los tipos de censura.
- Kaplan-Meier: Utilizando los ejemplos antes mencionados se calculó su respectiva curva de supervivencia y riesgo.
- Estimación: Bajo la suposición de que los datos siguen una distribución Weibull se estimaron los parámetros por diferentes métodos, tanto gráficos como analíticos, y considerando los diferentes tipos de censura.
- Pruebas de Bondad de ajuste: Se muestran dos pruebas de bondad de ajuste para los estimadores obtenidos por diferentes métodos y diferentes tipos de censura.

Se presentó como aplicación del análisis de supervivencia el caso de estudio de un problema que tiene gran impacto en nuestra sociedad que es la enfermedad de Diabetes Mellitus tipo II, de la población de Zacapoaxtla en el Estado de Puebla, se estudió el tiempo de vida de personas que presentan esta enfermedad desde que fue diagnosticada hasta octubre del 2014, considerando que se presenta una censura aleatoria independiente, se obtuvo su función de supervivencia empírica y de riesgo, las cuales nos indican que los primeros 15 años de un paciente con esta enfermedad la probabilidad de sobrevivir es alta, a diferencia de cuando llevan ya más de 20 años con la enfermedad.

Además se obtuvo, por dos diferentes métodos de estimación de parámetros, que los datos siguen una distribución Weibull con $\hat{\beta} = 1.82892$ como parámetro de escala y $\hat{\alpha} = 14.705$ parámetro

de forma y por dos pruebas de bondad de ajuste se comprobó lo antes mencionado, lo cual resulta importante ya que conociendo la distribución que siguen se puede calcular la probabilidad promedio de que los pacientes sobrevivan un tiempo específico. Para este ejemplo se concluyó que los estimadores obtenidos por el método de máxima verosimilitud fueron los que mejor ajustaron la distribución Weibull a los datos, ya que al realizar la prueba de bondad de ajuste para los estimadores obtenidos por el método gráfico WPP se rechazó la hipótesis nula, y la justificación de este hecho es la existencia de una gran cantidad de datos censurados.

Lo anterior nos lleva a afirmar que el análisis de supervivencia es una herramienta muy útil para el estudio de los tiempos de vida de pacientes con enfermedades como la diabetes tipo II. Además se hizo uso del software R, el cual es de acceso libre, facilita los cálculos y se pueden obtener las gráficas de supervivencia y riesgo de forma fácil comparado con otros programas.

Finalmente en este trabajo se proporcionan los programas en R y Mathematica que se implementaron en el análisis del problema estudiado.
Dado que la base de datos obtenidos para este trabajo contiene más información de la utilizada en la presente, se pretende en un futuro hacer el mismo estudio para personas con hipertensión arterial, ya que es otra enfermedad que está relacionada con el sobrepeso y que afecta a gran parte de la población en México. Además de hacer una extensión del modelo Weibull a un modelo de regresión sobre su parámetro de escala que involucre otras variables que están relacionadas con la Diabetes tipo II, como son: perímetro de la cintura, estatura, peso, índice de masa corporal (IMC), etc. Y hacer un estudio para comparar los diferentes tratamientos que son utillizados en el tratamiento de la Diabetes Mellitus Tipo II.

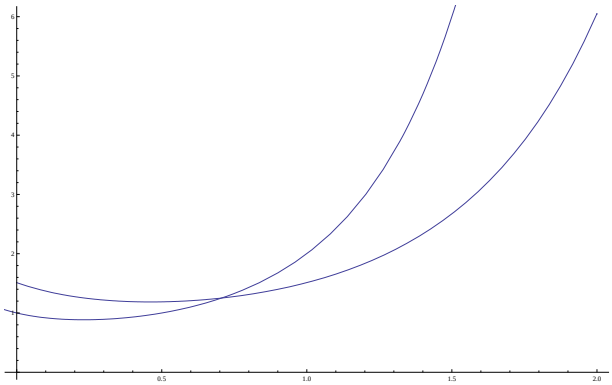
Investigar si es posible mejorar el ajuste de los datos considerando el modelo de la distribución Gama generalizada y comparando con el ajuste obtenido al usar como modelo la distribución Weibull.

Apéndice A

A.1. Programa I

Programa para obtener los estimadores para el ejemplo que se menciona en la sección de estimadores de momentos.

datos1 = Import[“EM.txt”, "Table"]
Do[t_i = datos1[i, 1], {i, 1, 20}]
$t = \sum_{i=0}^{n} \frac{t_i}{20}$
$s = \sum_{i=1}^{n} \frac{(t_i - t)^2}{n-1}$
$\frac{s}{t^2} = \frac{\Gamma[1+\frac{1}{\alpha}]}{\Gamma^2[1+\frac{2}{\alpha}]} - 1$
Sea $h = \frac{1}{\alpha}$ y se sustituye en la ecuación anterior.
Esta es la ecuación de momentos que se tiene que resolver para obtener el valor estimado de α
a = Plot[1.506181 * $Gamma^2$[1 +h], {h, -1, 4}]
b = Plot[Gamma[1 + 2h], {h, -1, 4}]
Show[a,b]



Con las gráficas de las funciones se desea observar en donde se encuentran las soluciones, en este caso se puede observar que la única solución es menor que 2.

FindRoot[1.560868*Gamma[1 + h] == Gamma[1 + 2h], {h, 2}]
Como se sabe que la solución es menor que 2, la instrucción anterior debe de acotarse por 2 para que se obtenga la raíz.
y se obtiene que {h ->0.706083}, por lo que el valor estimado es: $\hat{\alpha} = 1.41626$ Sustituyendo este valor en: $\beta = \frac{t}{\Gamma[1+\frac{1}{\hat{\alpha}}]}$
Se obtiene que el valor estimado es $\hat{\beta} = 13.8678$

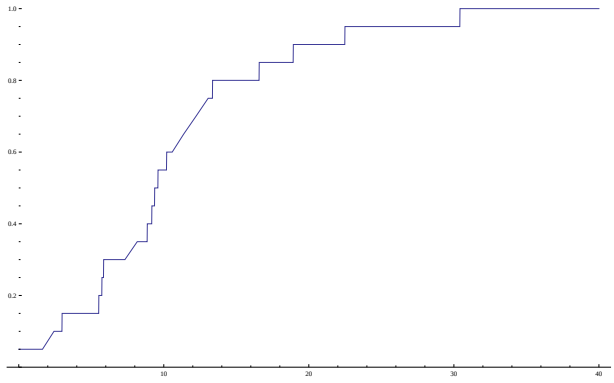
A.2. Programa II

Programa para obtener estimadores de máxima verosimilitud para el ejemplo que se menciona en la sección de estimadores de máxima verosimilitud, primero se tiene que obtener un intervalo en donde se encuentra la solución por medio de el método de la regla falsa, con el propósito de saber en qué punto comenzar las iteraciones del método de Newton-Raphson.

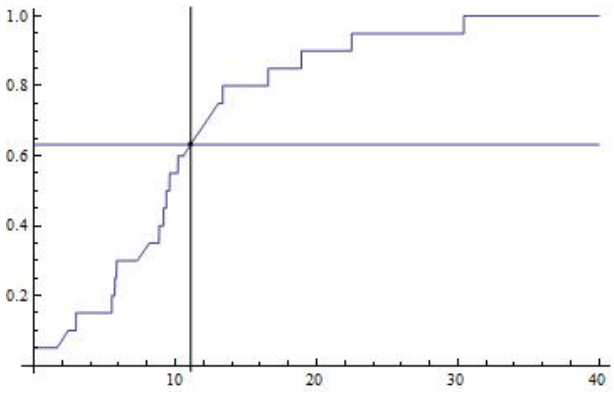
Una vez obtenido el punto inicial para empezar las iteraciones del método de Newton- Raphson, se aplica el programa para obtener las raíces de las ecuaciones.
Se importan los datos
datos1 = Import["EMD.txt", "Table"]
Do [t_i = datos1[[i, 1]], {i, 1, 20}]
Se escribe la ecuación a resolver y se calcula su derivada.
$l[\alpha] = \frac{\sum_{i=1}^{n}(t_i^{\alpha}\log t_i)}{\sum_{i=1}^{n} t_i^{\alpha}} - \frac{1}{\alpha} - \frac{1}{n}\sum_{i=1}^{n}\log t_i$
$l'[\alpha]$
Se da el punto inicial
$r_0 = 1$
Y con la siguiente instrucción se obtienen las iteraciones del método de Newton-Raphson.
Do$[r_{i+1} = r_i - \frac{l[r_i]}{l'[r_i]}, \{i, 0, 9\}]$
$Table[r_{i+1}, \{i, 0, 9\}]$
Así se obtiene los siguientes valores.
$\{1.40003, 1.53908, 1.54869, 1.54873, 1.54873, 1.54873, 1.54873, \quad 1.54873, 1.54873, 1.54873\}$ Por lo que $\hat{\alpha} = 1.54873$
y sustituyendo en $\hat{\beta} = \left(\frac{1}{n}\sum_{i=0}^{20} t_i^{\hat{\alpha}}\right)^{\frac{1}{\hat{\alpha}}}$ se obtiene que $\hat{\beta} = 14.1216$

A.3. Programa III

Programa para obtener los estimadores por el método de percentiles del ejemplo que se encuentra en la sección de percentiles.
Se importan los datos como estadísticos de orden.
datos1 = Import[“EMD1.txt", "Table"]
Do[t_i = datos1[i, 1], {i, 1, 20}]
Con la siguiente instrucción se calcula la función de distribución empírica. Do[F[t_i] = i/20, {i, 1, 20}]
Se define la función escalonada como sigue:
$f[x] = If[0 < x < 1.92, F[t_1], If[1.92 < x < 2.99, F[t_2], If[2.99 < x < 5.52, F[t_3], If[5.52 < x < 5.73, F[t_4], If[5.73 < x < 5.85, F[t_5], If[5.85 < x < 7.75, F[t_6], If[7.75 < x < 8.86, F[t_7], If[8.86 < x < 9.17, F[t_8], If[9.17 < x < 9.37, F[t_9], If[9.37 < x < 9.6, F[t_10], If[9.6 < x < 10.2, F[t_11], If[10.2 < x < 11.24, F[t_12], If[11.24 < x < 11.5, F[t_13], If[11.5 < x < 12.74, F[t_14], If[12.74 < x < 13.36, F[t_15], If[13.36 < x < 16.58, F[t_16], If[16.58 < x < 18.92, F[t_17], If[18.92 < x < 22.48, F[t_18], If[22.48 < x < 30.42, F[t_19], If[30.42 < x < 38.14, F[t_20], 1]]]]]]]]]]]]]]]]]]]]$
se grafica:
a = Plot[$f[x]$, {x, 0, 40}]



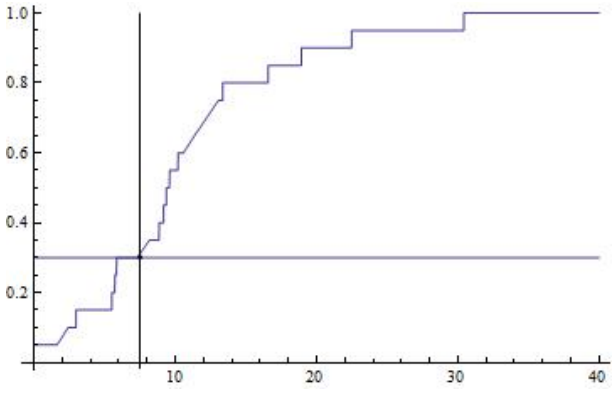
en la sección de percentiles se dijo que el estimador para $\hat{\beta} = t_{(1-e^{-1})}$ es decir se quiere encontrar el valor $t_{(1-e^{-1})}$ tal que $P[T \leq t_{(1-e^{-1})}] = 1 - \exp^{-1} = 0.632121$, por lo que graficamos la función $x = 0.632121$ y encontramos el punto donde se intersectan las gráficas, ese es el valor estimado de $t_{(1-e^{-1})} = \hat{\beta}$



Así el valor estimado es: $\hat{\beta} = 11.1$, y para α se tiene que resolver la ecuación:

$$\hat{\alpha} = \frac{\log[-\log(1-0.31)]}{\log(\frac{t_{0.31}}{t_{0.632}})}$$

Por lo que se tiene que estimar el valor de $t_{0.31}$



así el valor de $t_{0.31} = 7.5$ y sustituyendo en la ecuación anterior se tiene que el valor estimado para α es: $\hat{\alpha} = 2.6296$

A.4. Programa IV

Programa para obtener los estimadores por el método de WPP.

Se importan los datos acomodados de menor a mayor.

datos1 = Import["EMD2.txt", "Table"]

Do[t_i] = datos1[[i, 1]], {i, 1, 20}]

Se escribe la función de distribución empírica.

Do[F[t_i] = i/20, {i, 1, 20}]

Se hacen las transformaciones de los datos.

Do[y_i = Log[-Log[1 - F[t_i]]] // N, {i, 1, 19}]

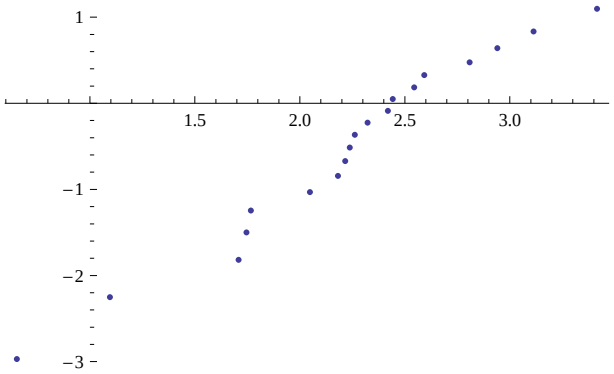
Do[x_i= Log[t_i], {i, 1, 19}]

Se grafican los datos

datos = Table[{x_i, y_i}, i, 1, 19]

ListPlot[datos]

Se obtiene una gráfica como la siguiente:



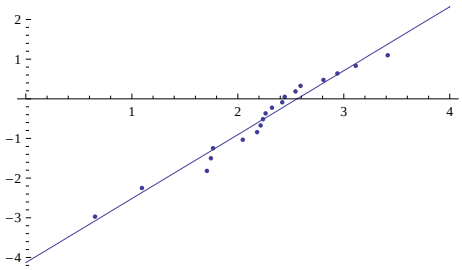
Se obtiene la regresión de los datos:

Regress[datos, x, x]

```
{ParameterTable → |          | Estimate   SE          TStat     PValue
                  | 1        | -4.13034   0.167002    -24.7322  9.07391×10^-15,
                  | x        | 1.61288    0.0717314   22.4849   4.37428×10^-14
 RSquared → 0.967469, AdjustedRSquared → 0.965555, EstimatedVariance → 0.0405253,
                          DF    SumOfSq    MeanSq       FRatio    PValue
 ANOVATable → Model       1     20.4884    20.4884      505.572   4.37428×10^-14
              Error       17    0.68893    0.0405253
              Total
```

y la ecuación de regresión es

$$-4.13034 + 1.61288x$$



La pendiente de la recta estima a α, por lo que $\hat{\alpha} = 1.61288$.
y resolviendo:
$\beta = Exp[-(y_0)/\hat{\alpha}]$
Así $\hat{\beta} = 12.9468$

Apéndice B

B.1. Programa V

Programa para obtener los estimadores por máxima verosimilitud para el caso de datos con censura tipo I.

datos1 = Import["DC1.txt", "Table"]
Do[t_i] = datos1[[i, 1]], {i, 1, 50}]
Se importan los datos, de los cuales los primeros 30 son datos sin censura y los otros 20 sólo se sabe que sobrepasaron las 12 horas de vida.

Se escribe la ecuación a resolver y se calcula su derivada.
$l[\alpha] = \frac{\sum_{i=1}^{50} t_i^{\hat{\alpha}} \log t_i}{\sum_{i=1}^{50} t_i^{\hat{\alpha}}} - \frac{1}{\hat{\alpha}} - \frac{1}{30} \sum_{i=1}^{30} \log t_i$
$l'[\alpha]$
Se da el punto inicial obtenido previamente por el método de la regla falsa.
$r_0 = 0.1$
Se hacen las iteraciones del método
Do[$r_{i+1} = r_i - \frac{l[r_i]}{l'[r_i]}, \{i, 0, 8\}$
y se obtiene:
{ 0.887841, 1.19497, 1.32074, 1.33391, 1.33403, 1.33403,1.33403, 1.33403, 1.33403}
Por lo que el estimador para $\hat{\alpha} = 1.33403$
y para beta se obtiene resolviendo:
$\hat{\beta} = \left(\frac{1}{30}\left[\sum_{i=1}^{30} t_i^{\hat{\alpha}} + (50-30)12^{\hat{\alpha}}\right]\right)^{\frac{1}{\hat{\alpha}}}$
$\hat{\beta} = 8.43874$

B.2. Programa VI

Programa para obtener los estimadores por el método WPP cuando existe censura tipo I.

datos1 = Import["DC1.txt", "Table"]

Do$[t_i]$ = datos1$[[i, 1]]$, {i, 1, 50}]

Se importan los datos, de los cuales los primeros 30 son datos sin censura y los otros 20 sólo se sabe que sobrepasaron las 12 horas de vida, por lo que para los últimos 20 datos $t_i = 12$

Se define la función de distribución empírica.

Do$[F[t_i]$ = i/50, {i, 1, 50}]

Se hace la transformación pero solo para los datos que no tienen censura, ya que para los datos censurados los tiempos de falla no son conocidos.

Do$[y_i = Log[-Log[1 - F[t_i]]]//N, \{i, 1, 30\}]$

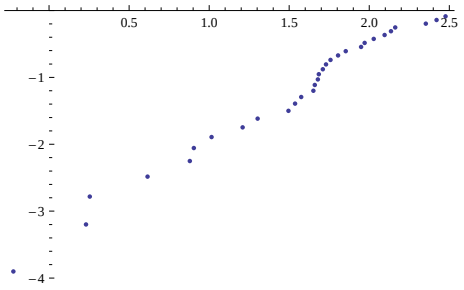
Do$[x_i = Log[t_i], \{i, 1, 30\}]$

Se grafican los datos

datos1 = Table$[\{x_i, y_i\}, \{i, 1, 30\}]$

b = ListPlot$[datos1]$

Y se obtiene la siguiente gráfica:



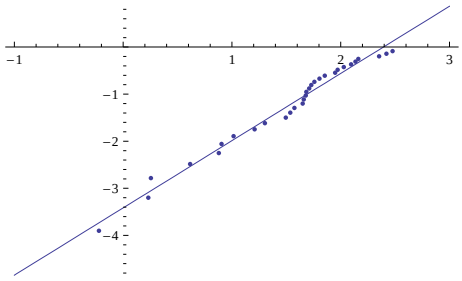
se obtiene la regresión de los datos:

Regress[datos1, x, x]

```
                          |  Estimate   SE          TStat      PValue
{ParameterTable → 1       | -3.41594   0.0627102  -54.4719   5.84343×10^-30,
                  x       |  1.42764   0.0376575   37.9111   0.

 RSquared → 0.980891, AdjustedRSquared → 0.980208, EstimatedVariance → 0.0183134,

                          DF     SumOfSq      MeanSq        FRatio      PValue
 ANOVATable → Model       1      26.3209      26.3209       1437.25     0.
              Error       28     0.512775     0.0183134
              Total       29     26.8337
```



De aquí se obtiene que: $\hat{\alpha} = 1.42764$ y el estimador para β se obtiene resolviendo:
$\beta = Exp[-(-3.4159)/1.4276364]$
Por lo que $\hat{\beta} = 10.943$

Apéndice C

C.1. Programa VII

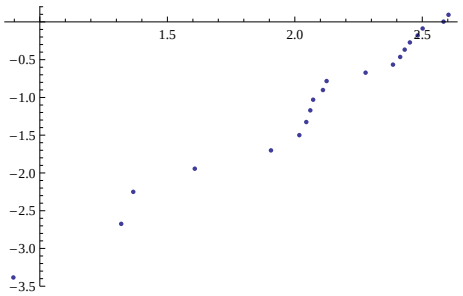
Programa para obtener los estimadores por máxima verosimilitud para el caso de datos con censura tipo II.
Se importan los datos, de los cuales los primeros 20 son datos sin censura y los otros 10 son datos con censura tipo II.
datos1 = Import["DC2.txt", "Table"]
Do$[t_i]$ = datos1$[[i, 1]]$, {i, 1, 20}]

Se escribe la ecuación a resolver y se calcula su derivada.
$l[\alpha] = \frac{\sum_{i=1}^{r} t_i^{\hat{\alpha}} \log[t_i] + (n-r) t_r^{\hat{\alpha}} \log[t_r]}{\sum_{i=1}^{r} t_i^{\hat{\alpha}} + (n-r) t_r^{\hat{\alpha}}} - \frac{1}{\hat{\alpha}} - \frac{1}{r} \sum_{i=1}^{r} \log[t_i]$
$l'[\alpha]$
Se da el punto inicial obtenido previamente por el método de la regla falsa.
$r_0 = 2$

Do$[r_{i+1} = r_i - \frac{l[r_i]}{l'[r_i]}, \{i, 0, 8\}$
y se obtiene:
{ 2.36293, 2.43419, 2.43625, 2.43625, 2.43625, 2.43625, 2.43625, 2.43625, 2.43625, 2.43625}
Por lo que el estimador para $\hat{\alpha} = 2.43625$
y para beta se obtiene:
$\hat{\beta} = \left(\frac{1}{r}\left[\sum_{i=1}^{r} t_i^{\hat{\alpha}} + (n-r) t_r^{\hat{\alpha}}\right]\right)^{\frac{1}{\hat{\alpha}}}$
$\hat{\beta} = 13.0843$

C.2. Programa VIII

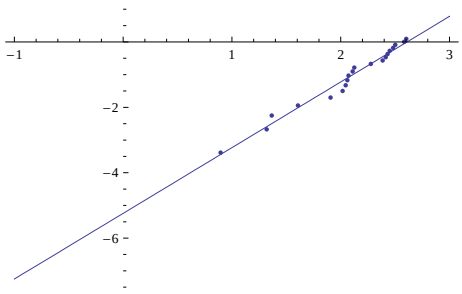
Programa para obtener los estimadores por el método WPP cuando existe censura tipo II.
Se importan los datos, de los cuales los primeros 20 objetos presentaron el evento y se terminó, por lo que los 10 objetos restantes son censurados.
datos1 = Import["DC2.txt", "Table"]
Do[t_i] = datos1[[i, 1]], {i, 1, 30}]
Se define la función de distribución empírica.
Do[$F[t_i]$ = i/30, {i, 1, 50}]
Se hace la transformación pero sólo para los datos que no tienen censura, ya que para los datos censurados los tiempos de falla no son conocidos.
Do[$y_i = Log[-Log[1 - F[t_i]]]//N, \{i, 1, 20\}$]
Do[$x_i = Log[t_i], \{i, 1, 20\}$]
Se grafican los datos
datos1 = Table[$\{x_i, y_i\}, \{i, 1, 20\}$]
b = ListPlot[$datos1$]
y se obtiene la siguiente gráfica:



se obtiene la regresión de los datos:

Regress[datos1, x, x]

```
                                  |  Estimate   SE          TStat      PValue
{ParameterTable → 1               | -5.24208   0.157423   -33.2994   1.26496×10^-17,
                  x               |  2.00868   0.0738559   27.1973   4.44089×10^-16
RSquared → 0.976244, AdjustedRSquared → 0.974924, EstimatedVariance → 0.02248,
                          DF     SumOfSq      MeanSq      FRatio      PValue
ANOVATable → Model        1      16.6283      16.6283     739.695     4.44089×10^-16
             Error        18     0.404639     0.02248
             Total        19
```



De aquí se obtiene que: $\hat{\alpha} = 2.00868$ y el estimador para β se obtiene resolviendo:
$\beta = Exp[-(-5.24208)/2.00868]$
Por lo que $\hat{\beta} = 13.5951$

Apéndice D

D.1. Programa IX

Programa para obtener los estimadores por máxima verosimilitud para el caso de datos con censura aleatoria.

Se importan los datos, de los cuales los primeros 8 son datos sin censura y los otros 2 son datos con censura aleatoria.

datos1 = Import["ca1.txt", "Table"]

Do$[t_i]$ = datos1$[[i, 1]]$, {i, 1, 10}]

Se escribe la ecuación a resolver y se calcula su derivada.

$l[\alpha] = \frac{\sum_{i=1}^{n} u_i^{\hat{\alpha}} \log[u_i]}{\sum_{i=1}^{n} u_i^{\hat{\alpha}}} - \frac{1}{\hat{\alpha}} - \frac{1}{k} \sum_{i \in D} \log[u_i]$

$l'[\alpha]$

Se da el punto inicial obtenido previamente por el método de la regla falsa.

$r_0 = 0.3$

Se hacen las iteraciones del método de Newton-Raphson.

Do$[r_{i+1} = r_i - \frac{l[r_i]}{l'[r_i]}, \{i, 0, 9\}$

y se obtiene:

{0.577776, 1.04097, 1.61443, 2.00812, 2.10571, 2.10992, 2.10992,2.10992, 2.10992, 2.10992}

Por lo que el estimador para $\hat{\alpha} = 2.10992$

y beta se obtiene resolviendo:

$\hat{\beta}^{\hat{\alpha}} = \frac{1}{n} \left(\sum_{i=1}^{n} u_i^{\hat{\alpha}} \right)$

Por lo tanto $\hat{\beta} = 32.7408$

D.2. Programa XI

Programa para obtener los estimadores por el método WPP cuando existe censura aleatoria independiente.

Se importan los datos, de los cuales los primeros 8 son datos sin censura y los otros 2 son datos censurados.

datos1 = Import["ca1.txt", "Table"]

Do[t_i] = datos1[[i, 1]], {i, 1, 10}]

para este caso se define

$u_i = min(t_i, c_i)$

Se escribe la función de distribución empírica.

Do[$F[t_i]$ = i/10, {i, 1, 10}]

Se hace la transformación pero solo para los datos que no tienen censura, ya que para los datos censurados los tiempos de falla no son conocidos.

Do[$y_i = Log[-Log[1 - F[t_i]]]//N, \{i, 1, 8\}$]

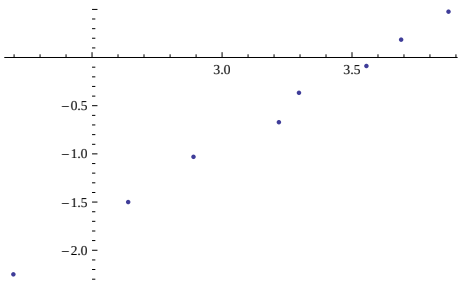
Do[$x_i = Log[t_i], \{i, 1, 8\}$]

Se grafican los datos

datos1 = Table[$\{x_i, y_i\}, \{i, 1, 8\}$]

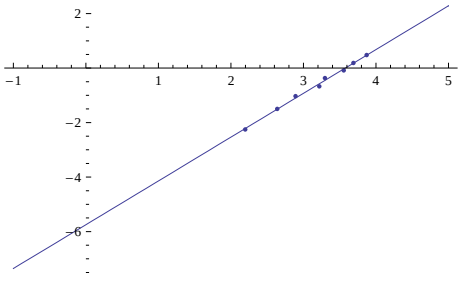
b = ListPlot[$datos1$]

y se obtiene la siguiente gráfica:



se obtiene la regresión de los datos:

Regress[datos1, x, x]



```
                                 | Estimate  SE          TStat     PValue
{ParameterTable → 1              | -5.74775  0.140822    -40.8156  1.44627×10^-8,
                  x              | 1.60654   0.0438242   36.6587   2.74894×10^-8
 RSquared → 0.995555, AdjustedRSquared → 0.994814, EstimatedVariance → 0.00429056,
                          DF    SumOfSq       MeanSq         FRatio     PValue
 ANOVATable → Model       1     5.7659        5.7659         1343.86    2.74894×10^-8
              Error       6     0.0257434     0.00429056
              Total       7     5.79164
```

De aquí se obtiene que: $\hat{\alpha} = 1.60654$ y el estimador para β se obtiene resolviendo:
$\beta = \exp[-(-5.74775)/1.60654]$
Por lo que $\hat{\beta} = 35.7918$

Apéndice E

Tabla del estadístico D_α

Tabla

Test de Kolmogorov-Smirnov sobre Bondad de Ajuste

	Nivel de significación α							
n	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
1	0.90000	0.95000	0.97500	0.99000	0.99500	0.99750	0.99900	0.99950
2	0.68337	0.77639	0.84189	0.90000	0.92929	0.95000	0.96838	0.97764
3	0.56481	0.63604	0.70760	0.78456	0.82900	0.86428	0.90000	0.92065
4	0.49265	0.56522	0.62394	0.68887	0.73424	0.77639	0.82217	0.85047
5	0.44698	0.50945	0.56328	0.62718	0.66853	0.70543	0.75000	0.78137
6	0.41037	0.46799	0.51926	0.57741	0.61661	0.65287	0.69571	0.72479
7	0.38148	0.43607	0.48342	0.53844	0.57581	0.60975	0.65071	0.67930
8	0.35831	0.40962	0.45427	0.50654	0.54179	0.57429	0.61368	0.64098
9	0.33910	0.38746	0.43001	0.47960	0.51332	0.54443	0.58210	0.60846
10	0.32260	0.36866	0.40925	0.45562	0.48893	0.51872	0.55500	0.58042
11	0.30829	0.35242	0.39122	0.43670	0.46770	0.49539	0.53135	0-55588
12	0.29577	0.33815	0.37543	0.41918	0.44905	0.47672	0.51047	0.53422
13	0.28470	0.32549	0.36143	0.40362	0.43247	0.45921	0.49189	0.51490
14	0.27481	0.31417	0.34890	0.38970	0.41762	0.44352	0.47520	0.49753
15	0.26589	0.30397	0.33750	0.37713	0.40420	0.42934	0.45611	0.48182
16	0.25778	0.29472	0.32733	0.36571	0.39201	0.41644	0.44637	0.46750
17	0.25039	0.28627	0.31796	0.35528	0.38086	0.40464	0.43380	0.45540
18	0.24360	0.27851	0.30936	0.34569	0.37062	0.39380	0.42224	0.44234
19	0.23735	0.27136	0..30143	0.33685	0.36117	0.38379	0.41156	0.43119
20	0.23156	0.26473	0.29408	0.32866	0.35241	0.37451	0.40165	0.42085
21	0.22517	0.25858	0.28724	0.32104	0.34426	0.36588	0.39243	0.41122
22	0.22115	0.25283	0.28087	0.31394	0.33666	0.35782	0.38382	0.40223
23	0.21646	0.24746	0.2749tl	0.30728	0.32954	0.35027	0.37575	0.39380
24	0.21205	0.24242	0.26931	0.30104	0.32286	0.34318	0.36787	0.38588
25	0.20790	0.23768	0.26404	0.29518	0.31657	0.33651	0.36104	0.37743

26	0.20399	0.23320	0.25908	0.28962	0.30963	0.33022	0.35431	0.37139
27	0.20030	0.22898	0.25438	0.28438	0.30502	0.32425	0.34794	0.36473
28	0.19680	0.22497	0.24993	0.27942	0.29971	0.31862	0.34190	0.35842
29	0.19348	0.22117	0.24571	0.27471	0.29466	0.31327	0.33617	0.35242
30	0.19032	0.21756	0.24170	0.27023	0.28986	0.30818	0.33072	0.34672
31	0.18732	0.21412	0.23788	0.26596	0.28529	0.30333	0.32553	0.34129
32	0.18445	0.21085	0.23424	0.26189	0.28094	0.29870	0.32058	0.33611
33	0.18171	0.20771	0.23076	0.25801	0.27577	0.29428	0.31584	0.33115
34	0.17909	0.21472	0.22743	0.25429	0.27271	0.29005	0.31131	0.32641
35	0.17659	0.20185	0.22425	0.25073	0.26897	0.28600	0.30597	0.32187
36	0.17418	0.19910	0.22119	0.24732	0.26532	0.28211	0.30281	0.31751
37	0.17188	0.19646	0.21826	0.24404	0.26180	0.27838	0.29882	0.31333
38	0.16966	0.19392	0.21544	0.24089	0.25843	0.27483	0.29498	0.30931
39	0.16753	0.19148	0.21273	0.23785	0.25518	0.27135	0.29125	0.30544
40	0.16547	0.18913	0.21012	0.23494	0.25205	0.26803	0.28772	0.30171
41	0.16349	0.18687	0.20760	0.23213	0.24904	0.26482	0.28429	0.29811
42	0.16158	0.18468	0.20517	0.22941	0.24613	0.26173	0.28097	0.29465
43	0.15974	0.18257	0.20283	0.22679	0.24332	0.25875	0.27778	0.29130
44	0.15795	0.18051	0.20056	0.22426	0.24060	0.25587	0.27468	0.28806
45	0.15623	0.17856	0.19837	0.22181	0.23798	0.25308	0.27169	0.28493
46	0.15457	0.17665	0.19625	0.21944	0.23544	0.25038	0.26880	0.28190
47	0.15295	0.17481	0.19420	0.21715	0.23298	0.24776	0.26600	0.27896
48	0.15139	0.17301	0.19221	0.21493	0.23059	0.24523	0.26328	0.27611
49	0.14987	0.17128	0.19028	0.21281	0.22832	0.24281	0.26069	0.27339
50	0.14840	0.16959	0.18841	0.21068	0.22604	0.24039	0.25809	0.27067
$n > 50$	$\frac{1.07}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.52}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$	$\frac{1.73}{\sqrt{n}}$	$\frac{1.85}{\sqrt{n}}$	$\frac{1.95}{\sqrt{n}}$

Apéndice F

Datos de pacientes con Diabetes

Individuo	*Sexo*	*Edad*	*Tiempo promedio con diabetes*	*Fecha de de función.*
1	F	81	30 años	06/09/2011
2	M	58	1 año	01/08/2011
3	F	57	15 años	05/05/2011
4	F	67	4 años	20/07/2014
5	F	81	15 años	09/01/2014
6	F	68	18 años	02/05/2014
7	M	59	10 años	04/05/2014
8	F	56	4 años	24/02/2014
9	F	72	20 años	26/12/2009
10	F	77	30 años	18/11/2009
11	F	75	20 años	18/09/2009
12	F	45	3 años	11/07/2009
13	F	83	35 años	01/05/2009
14	F	43	17 años	12/03/2009
15	F	57	15 años	05/05/2011
16	F	45	6 años	
17	F	40	6 años	
18	F	26	2 años	
19	F	30	6 años	
20	M	44	13 años	
21	F	47	8 años	
22	F	62	17 años	
23	F	41	8 años	
24	F	82	16 años	
25	F	74	16 años	
26	F	69	24 años	
27	F	77	22 años	
28	M	57	8 años	
29	F	42	14 años	
30	F	52	12 años	
31	F	40	6 años	
32	F	78	10 años	
33	F	60	18 años	
34	F	41	5 años	
35	F	52	17 años	
36	F	58	11 años	
37	F	67	29 años	
38	M	52	16 años	
39	F	56	7 años	
40	F	62	18 años	
41	F	50	4 años	
42	F	79	9 años	
43	F	65	5 años	
44	F	70	4 años	
45	M	66	5 años	
46	F	63	4 años	

Tabla F.1: Tabla de datos de pacientes diagnosticados con Diabetes.

Apéndice G

G.1. Programa utilizado para obtener los cálculos de los datos de pacientes con Diabetes en R.

Se cargan las librerías correspondientes al Análisis de Supervivencia
library(splines)
library(survival)
Se introducen los datos
datosc<-read.table("Diabetes.txt", header = TRUE, sep=' t')
Con la siguiente instrucción los datos introducidos van a clasificarse en datos con censura y datos completos, distinguidos por un signo (+) si es dato con censura
surv<-Surv(datosc $ X, datosc $ E)
La siguiente instrucción da un pequeño resumen de los datos, la mediana y su intervalo de confianza
survf<-survfit(surv T,datosc)
survf
Con la siguiente instrucción se da un resumen más amplio de los datos, incluyendo el valor de la función de supervivencia empírica
summary(survf)
Se obtiene la gráfica de la función de Supervivencia empírica
plot(survf, conf.int=F,col=2:4, main="Función de Supervivencia, xlab="Tiempo de Fallo", ylab="probabilidad de supervivencia")
La función de riesgo se obtiene con la siguiente instrucción
plot(survf, conf.int=F, fun="cumhaz", col=2:4, main="Función de Riesgo", xlab="Tiempo de Falla", ylab="Probabilidad de Riesgo")

Para las pruebas de bondad de ajuste primero se cargan los datos en una variable como se ve a continuación:

```
datosc
xwei<-datosc[,1]
```

Luego se da la distribución teórica

```
x.teo<-rweibull(n=46, shape=1.82892, scale=14.7051)
```

y con la siguiente instrucción R hace la prueba QQ-Plot

```
qqplot(x.teo, xwei, main="QQ-plot distr. Weibull")
abline(0,1)
```

Y con esta instrucción R hace la prueba de Kolmogorov-Smirnov

```
ks.test(xwei, "pweibull", shape=1.82892, scale=14.7051)
```

Bibliografía

[1] Dietz K.; Gail M.; Krickeberg K.; Samet J.; Tsiatis A. *Statistics for Biology and Health.*2003.

[2] Epstein; Sobel. *Life Testing. Journal of the American Statistical Association.* 1953

[3] García de los Ríos Manuel. *Guías Alad de diagnóstico, control y tratamiento de la diabetes mellitus tipo II.* 2013.

[4] *Federación Mexicana de diabetes, A.C. http : //www.fmdiabetes.org/fmd/pag/diabetes_numeros.php.*

[5] Godoy Aguilar Angel Manuel. *Introducción al Análisis de Supervivencia con R.* 2009.

[6] Instituto Mexicano del Seguro Social. *Diagnóstico y tratamiento de la Diabetes Mellitus Tipo 2.* 2010.

[7] Instituto Nacional de Estadística Geográfica e Informática. *XIII Censo General de Población y Vivienda 2010. http : //www.inegi.org.mxs/sistemas/mexicocifras/default.aspx?src =* 487&$e = 21$.

[8] R Development Core Team *Introducción a R, Notas sobre R: Un entorno de programación para Análisis de Datos y Gráficos.* 2000

[9] Johnson Norman L.; Kotz Samuel; Balakrishnan N. *Continuous Univariate Distributions.* Second edition. 1994.

[10] Kalbfleisch D John; Prentice Ross L. *The Statistical Analysis of Failure Time Data.* Second Edition. 2002.

[11] Klein John P.; Moeschberger Melvin L. *Survival Analysis Techniques for Censored and Truncated Data.* Second Edition.2003.

[12] Lawless Jerald F. *Statistical Models and Methods for life time.* Second edition. 2002.

[13] Lee Elisa T.; Wang John. *Statistical Methods for Survival Analysis.* Third edition. 2003.

[14] Martínez Fernández Laura. *Métodos de Inferencia para la distribución Weibull: Una Aplicación en Fiabilidad Industrial.* 2011.

[15] Mendenhall William; Wackerly Dennis D.; Scheaffer Richard L. *Estadística Matemática con aplicaciones.* Séptima Edición. 2008.

[16] Murthy Prabhakar; Xie Min; Jiang Renyan. *Weibull Models.* 2004.

[17] Nelson Wayne. *Applied Life Data Analysis.* 1982.

[18] Ricci Vito. *Fitting Distribution with R.* 2005.

[19] Seki T.; Yokoyama S. *Simple and Robust estimation of the Weibull parameters.* 1993.

[20] Tapia Zegarra Gino Guillermo; Chirinos Cáceres Jesús Luis; Tapia Zegarra Lenibet Miriam. *Características sociodemográficas y clínicas de los pacientes diabéticos tipo 2 con infecciones adquiridas en la comunidad admitidos en los servicios de Medicina del Hospital Nacional Cayetano Heredia.* 2000.

[21] Wolfram Mathematica ®Tutorial Collection. *MATHEMATICS AND ALGORITHMS.* 2008.