UNIVERSIDAD DE EL SALVADOR

FACULTAD MULTIDISCIPLINARIA ORIENTAL

DEPARTAMENTO DE CIENCIAS NATURALES Y MATEMÁTICA

SECCIÓN DE MATEMÁTICA

TRABAJO DE GRADO:

"MODELOS EN LAS FUNCIONES DE SUPERVIVENCIA Y ALGUNAS APLICACIONES"

PRESENTADO POR:

HERIBERTO SORTO

HILDA MARINA MARTÍNEZ HENRÍQUEZ

PARA OPTAR AL GRADO DE:

LICENCIADO/A EN ESTADÍSTICA

DOCENTE DIRECTOR:

MS.EST. MARÍA DEL TRANSITO GUTIEREZ REYES

CIUDAD UNIVERSITARIA ORIENTAL, FEBRERO DE 2014

SAN MIGUEL EL SALVADOR CENTROAMÉRICA

UNIVERSIDAD DE EL SALVADOR

AUTORIDADES

ING. MARIO ROBERTO NIETO LOVO
RECTOR

MS.D ANA MARÍA GLOWER DE ALVARADO
VICE-RECTORA ACADEMICA

DRA. ANA LETICIA ZAVALETA DE AMAYA
SECRETARIA GENERAL

LIC. FRANCISCO CRUZ LETONA
FISCAL GENERAL

FACULTAD MULTIDISCIPLINARIA ORIENTAL

AUTORIDADES

LIC. CRISTOBAL HERNAN RÍOS BENÍTEZ
DECANO

LIC. CARLOS ALEXANDER DÍAZ
VICE-DECANO

LIC. JORGE ALBERTO ORTEZ HERNÁNDEZ
SECRETARIO

LIC. EDWIND JEOVANNY TREJOS CABRERA
ADMINISTRADOR ACADÉMICO

M.EST. JOSÉ ENRY GARCIA
JEFE DEL DEPARTAMENTO DE CIENCIAS NATURALES Y MATEMÁTICA

ING. DOLORES BENEDICTO SARAVIA
COORDINADOR DE LA SECCIÓN DE MATEMÁTICA

TRABAJO DE GRADUACIÓN APROBADO POR:

_____

MSC. OSCAR ULISES LIZAMA VIGIL.

COORDINADOR DE PROCESOS DE GRADUACIÓN

DEPTO. DE CIENCIAS NATURALES Y MATEMÁTICA

_____

MS.EST. MARÍA DEL TRANSITO GUTIÉRREZ REYES

ASESOR DIRECTOR

_____

LICDA. MARÍA OLGA QUINTANILLA DE LOVO

ASESOR METODOLÓGICO.

AGRADECIMIENTOS Y DEDICATORIA

A DIOS TODO PODEROSO:

Quiero darle gracias a Dios, por haberme permitido culminar mi carrera, además de brindarme la Sabiduría y Bendición en este proceso.

A MIS PADRES:

Juan Francisco Martínez Rodríguez y María de la Paz Henríquez de Martínez

Por brindarme su amor, dedicación, entrega y por toda la ayuda que me han brindado siempre, y porque son un ejemplo de que cuando se quiere algo en la vida se puede lograr.

A MIS HERMANOS:

Juan Carlos Martínez Henríquez, Melquis Alejandro Martínez Henríquez y Rafael Antonio Rodríguez por el apoyo que me dieron durante la carrera.

A MIS TIOS/AS:

A todos mis tíos especialmente a **Juana Noemy Martínez de Granados y José Eduardo Granados.** Por sus palabras, y por toda la ayuda que me brindaron.

A MIS AMIGOS:

A todos mis amig@s, especialmente a **Ilsia Dalia Gómez y Dina Bernarda Vanegas,** por apoyarme siempre y haberme ayudado en los momentos más difíciles.

Hilda Marina Martínez Henríquez

A nuestro **SEÑOR JESUCRISTO** todo poderoso, por haberme permitido culminar con éxito el esfuerzo de todos estos años de estudio, además de cuidarme, brindarme la Sabiduría y bendición en este proceso. Para el mis agradecimientos infinitos.

DEDICAD

A mi padres especialmente a mi madre **Hilda de Jesús Sorto Portillo** por su constantes oraciones, sacrificios y esfuerzos, lo que hizo posible el triunfo profesional alcanzado. Para ella mi amor, respeto y obediencia.

A Papá **Gilberto Romero Campos**, mi hermano **Carlos Mario Sorto Romero** y mis amigos **Pastor Eduardo Lizama** y **Lic**. **Esmelí Enomeí Romero** que desde el cielo están con migo, me iluminan, bendicen y que siempre recordare, amare y llevare en mi corazón.

A mis hermanos/as **José Magno Sorto Romero y Familia, Mariana de Jesús Sorto Romero sus hijas y sus dos nietos y María Adela Sorto de Lizama su esposo y sus dos niñas.** A **Mónica Dabeiva Navarrete** y a sus tres hijos/as especialmente **Julia Nathaly Rodríguez Navarrete Alvarado.** A **Tca. Bibl. Rosa Marroquín Quijano y su niño**. A mis tíos/as especialmente a **Rodolfo Sorto Portillo**. Mis primos/as y a toda mi familia. Que de una u otra forma me ayudaron y participaron para que lograra el presente éxito profesional. Gracias por sus palabras de aliento y fe en mí.

A COMPAÑEROS/AS Y AMIGOS/AS:

Quienes de una u otra forma han contribuido y participado para alcanzar la meta trazada, ya que con su ayuda esta se hizo más fácil. Al grupo los **"Toños".** A **Prof. Fredy Rojas y Obed Granados** por apoyarme siempre y haberme motivado en los momentos más difíciles a seguir adelante.

A LA UNIVERSIDAD Y A MIS CATEDRÁTICOS/AS.

Por haber permitido adquirir los conocimientos necesarios y la experiencia necesaria para poderla aplicar en la práctica. A **Ms. Ets. María del Transito Gutiérrez Reyes, Licda. María Olga Quintanilla de Lovo, Msc. Jorge Alberto Martínez Gutiérrez y Prof. Francisco Stanley Madrid Pérez** con afecto, respeto y admiración.

Heriberto Sorto

ÍNDICE **PÁGINA**

INTRODUCCIÓN

Posiblemente el origen del nombre "Análisis de supervivencia" ocurrió en Inglaterra a principios del siglo XVI cuando se realizó el registro de nacimientos y defunciones con el cual en 1662 apareció el primer estudio de datos poblacionales, titulado "Observations on the London Bills of Mortality" (Observaciones sobre las partidas de defunción en Londres).

Un estudio similar sobre la tasa de mortalidad en la ciudad de Breslau, en Alemania, realizado en 1691, hace énfasis a la importancia del conocimiento estadístico sobre los sucesos relacionados con la expiración no sólo de vidas humanas sino, como se hace actualmente, en la vida de componentes.

Hoy en día, la Probabilidad y la Estadística están íntimamente unidas entre sí, desempeñan un papel fundamental en prácticamente todos los campos del saber, tanto en las Ciencias Naturales (Física, Química, Biología, etcétera) como en las Ciencias Humanas (Economía, Psicología, Sociología, etcétera), papel que va cobrando cada vez mayor importancia. Con respecto a la Medicina, tiene pocos años que se han estado utilizando los resultados estadísticos, dando un gran impulso al desarrollo de la Medicina y la Estadística, llegando a fusionar algunas de ellas para dar inicio a áreas como la Bioestadística, Biotecnología, Biomedicina, entre otras.

El análisis estadístico, de tiempos de vida, tiempos de supervivencia, o datos de tiempo de fallas es un tema reciente de gran importancia en muchas áreas (la Biomedicina, Ingeniería, Ciencias Sociales entre otras). La aplicación de la metodología de distribución en

tiempos de vida va desde la investigación hasta la durabilidad de los productos manufacturados a los estudios de enfermedades humanas y su tratamiento.

El presente trabajo se dividirá en 3 capítulos.

En el capítulo uno se presentan e ilustran las distribuciones más utilizadas en el análisis de datos de tiempos de vida. Se presenta la función de densidad, de probabilidad, de supervivencia y la función de riesgo ó Hazard, su definición, y el efecto de sus parámetros.

En el capítulo dos se estudian los diferentes tipos de datos de tiempo de vida que pueden presentarse al realizar experimentos. En este capítulo se estudian los fenómenos que pueden encontrar datos incompletos o que incluyen en incertidumbre respecto al tiempo en que ocurre la falla, estos datos se llaman datos censurados; por lo que se definen los diferentes tipos de censura (censura por la derecha, tipo I, tipo II, aleatoria, por la izquierda y censura por intervalos). En gran parte de los estudios se presentan variables o covariables explicativas por lo que se muestra a los modelos de regresión de tiempo de falla acelerado y el modelo de regresión de riesgo proporcional.

Por último en el capítulo tres se trabajará con algunas aplicaciones de los modelos aplicados al análisis de supervivencia.

PLANTEAMIENTO DEL PROBLEMA

El análisis de supervivencia es una técnica adecuada para el análisis de estudios longitudinales caracterizado por una duración variable del seguimiento para los sujetos que se incorporan en momentos distintos o al existir observaciones incompletas, conocidas como datos censurados y una función importante del análisis de supervivencia es incluirlos porque aportan información muy útil.

En los estudios de supervivencia solo se necesitan un par de valores: el tiempo del seguimiento de un individuo u objeto y una variable que indique si el tiempo de estudio es completo o censurado. Los modelos en las funciones de supervivencia juegan un papel importante en muchos campos de aplicación, como la Medicina, Psicología, Análisis de Conglomerados, Pruebas de Vida, entre otros.

OBJETIVOS DE LA INVESTIGACION.

Los objetivos del presente trabajo son:

- Conocer la base teórica de las principales distribuciones de supervivencia aplicadas en las funciones de supervivencia.

- Disponer de un conjunto de distribuciones lo suficientemente flexible para facilitar su análisis.

- Desarrollar algunas aplicaciones en los modelos de las funciones de supervivencia.

CAPÍTULO 1. MODELOS Y FUNCIONES DE SUPERVIVENCIA

INTRODUCCIÓN

El análisis estadístico que es referido en varias ocasiones a tiempos de vida, tiempos de supervivencia, o datos de tiempo de fallas es un tema importante en muchas áreas, como son la Biomédica, Ingeniería y Ciencias Sociales. La aplicación de la metodología de distribución en tiempos de vida va desde la investigación de la durabilidad de los productos manufacturados a los estudios de enfermedades humanas y su tratamiento.

En este capítulo se presenta e ilustra los métodos estadísticos y el análisis de datos del tiempo de vida. La metodología de la distribución del tiempo de vida se utiliza ampliamente en la Biomedicina y Ciencias de la Ingeniería, de esta forma la mayoría de los ejemplos provienen de estas áreas.

Los distintos tipos de datos se denominarán por conveniencia, datos del "tiempo de vida", considerando situaciones en las que el tiempo de ocurrencia de algún evento de interés se refiere a los individuos en una población determinada. Por ejemplo, en algunas ocasiones los eventos son muertes reales de personas y el "tiempo de vida" es la longitud de la vida de éstas, medida a partir de algún punto de partida en particular. En otros casos el "tiempo de vida" y las palabras "muerte" o "falla", denotarán al evento de interés, utilizadas en sentido figurado. Al discutir las aplicaciones, se tiene que otros términos tales como el "tiempo supervivencia" y el "tiempo de falla" son usados frecuentemente.

Los siguientes ejemplos ilustran algunas formas en que los datos del tiempo de vida pueden ocurrir.

EJEMPLO 1.1. Los artículos fabricados con componentes mecánicos o electrónicos suelen ser sometidos a pruebas de tiempos de vida a fin de obtener información sobre su durabilidad. Esto implica la puesta en operación de los elementos, a menudo en un entorno de laboratorio y son observados hasta que fallen.

EJEMPLO 1.2. Los demógrafos y las ciencias sociales están interesados en la duración de ciertos "estados" de vida para los seres humanos, considérese por ejemplo el matrimonio, específicamente los matrimonios formados durante el año 2012 en un país en particular. Entonces el tiempo de vida de un matrimonio sería su duración, un matrimonio puede terminar debido a la anulación, al divorcio o a la muerte.

EJEMPLO 1.3. En los estudios médicos que tratan con enfermedades potencialmente mortales se está interesado en el tiempo de supervivencia de los individuos con la enfermedad medida desde la fecha del diagnóstico o algún otro punto de partida, es común comparar tratamientos para una enfermedad al menos en parte en términos de la distribución del tiempo de supervivencia de los pacientes que recibieron los distintos tratamientos.

EJEMPLO 1.4. Un experimento estándar en la investigación de sustancias cancerígenas en la que los animales de laboratorios son sometidos a una dosis de la sustancia y luego son observados para ver si desarrollan tumores. Una variable de interés es el tiempo de aparición de un tumor, medida a partir de cuando la dosis se administra.

La definición del tiempo de vida incluye una escala de tiempo así como su origen y una especificación del evento (por ejemplo, falla o muerte) que determina el tiempo de vida, en algunos casos es difícil decir exactamente cuándo se produce el evento: por ejemplo, la aparición de un tumor en el ejemplo 1.4.

La escala de tiempo no siempre es real o el tiempo es cronológico, especialmente cuando las máquinas o equipos son de interés. Por ejemplo, los kilómetros conducidos podrían ser utilizados como una escala de tiempo para los vehículos de motor, el número de páginas de salida para una impresora de una computadora o fotocopiadora.

Los principales problemas en los tiempos de supervivencia consisten en especificar los modelos que representen las distribuciones del tiempo de vida y hacer inferencias basadas en estos modelos. El objetivo de la modelización y el análisis estadístico consiste en incluir una descripción o estimación de las distribuciones, la comparación de las distribuciones, el fomento de la comprensión científica, el proceso o sistema de mejora, predicción y decisión.

Covariables o variables explicativas que puedan ser relacionadas con los tiempos de vida, usualmente características prominentes en estas actividades. En algunos casos puede haber más de una variable asociada al tiempo de vida de un individuo o una persona, puesto que la persona puede morir de diferentes maneras. Los tipos de modelos utilizados en una gama de análisis del tiempo de vida se extiende totalmente de las formas clásicas de paramétricas a no paramétricas; son comunes en estos casos los modelos semi-paramétricos que tienen características tanto paramétricas y no paramétricas.

El tiempo cronológico necesita observar que los tiempos de vidas de todos los individuos en un estudio puedan ser lo suficientemente grandes de tal forma que existen limitaciones prácticas para impedir la observación real. Esto nos lleva a lo que se denomina "censurado", en el que se conoce el tiempo de vida de un individuo sólo cuando excede un valor determinado. En el ejemplo 1.1, se tiene una prueba de vida que podría ser terminada después de digamos 28 días, si un artículo no había fallado en ese momento, podríamos saber que su vida útil supera los 28 días y se refieren a ese valor como un "tiempo de censura". En general, no es posible determinar con exactitud cuándo una falla o muerte ocurre, porque los individuos solo viven en ciertos tiempos. En ese caso sólo es posible saber que el tiempo de vida está en algún intervalo (L, R) refiriéndose a éstos como "intervalo censurado".

En este capítulo se revisan algunos de los modelos más conocidos en el análisis de tiempos de vida, para esto se da inicio a las definiciones de los diferentes tipos de modelos, tanto continuos como discretos, se trabajan la función de densidad de probabilidad, la función de supervivencia (o confiabilidad) y la función de riesgo (Hazard), para cada uno de estos modelos. Después se mencionarán algunas observaciones de las funciones de Hazard, además se ilustrarán algunas funciones observando el comportamiento de la función de Hazard, según los diferentes tipos de parámetros.

Para finalizar el capítulo se muestran las familias de loc-escala, es decir familias de distribuciones que son invariantes con respecto a los parámetros de localidad y escala además se muestran algunos ejemplos de estas familias y la relación entre sus parámetros.

DISTRIBUCIONES DE TIEMPOS DE VIDA

1.1. Modelos Continuos

El estudio inicia considerando el caso de tiempo de vida de una variable aleatoria continua ***T***. Específicamente, sea ***T*** una variable aleatoria no negativa que representa los tiempos de vida de personas en una población. Luego, en esta parte se establece que todas las funciones, a menos que se indique lo contrario, se definen en el intervalo $[0, +\infty[$.

Función de Densidad de Probabilidad (*f.d.p*). Si ***T*** es una variable aleatoria continua, entonces la *f.d.p* de ***T*** es una función $f(t)$ que cumple no ser negativa y tal que para dos números cualesquiera, *a* y *b*, con $a \leq b$

$$Prob(a \leq \boldsymbol{T} \leq b) = \int_a^b f(t)dt$$

$$= \lim_{a \to -\infty} \int_a^0 f(t)dt + \lim_{b \to +\infty} \int_0^b f(t)dt = 1$$

Así, la probabilidad de que ***T*** tome valores en el intervalo $[a, b]$ es calculada integrando la *f.d.p* sobre el intervalo de tiempo deseado.

Función de Distribución Acumulada (*f.d.a*) ó Función de Distribución $F(t)$ de una variable aleatoria $\boldsymbol{T}$ se define como:

$$F(t) = Prob(\boldsymbol{T} \leq t).$$

$F(t)$ denota la probabilidad de que la variable aleatoria $\boldsymbol{T}$ tome valores menores o iguales a *t*. $F(t)$ es una función continua, no negativa y monótona creciente en todos los reales, tal que

$$\lim_{t \to -\infty} F(t) \to 0$$

$$\lim_{t \to +\infty} F(t) \to 1$$

Funciones Monótonas. Una función f se dice que es creciente en un conjunto *S* si $f(t_1) \leq f(t_2)$ para cada par de puntos t_1 y t_2 de *S* con $t_1 < t_2$. Si se verifica la desigualdad estricta $f(t_1) < f(t_2)$ para todo $t_1 < t_2$ en s se dice que la función es creciente en sentido estricto en *S*. Análogamente, una función se dice que es decreciente en *S* si $f(t_1) \geq f(t_2)$ para todo $t_1 < t_2$ en *S*. Si $f(t_1) > f(t_2)$ para todo $t_1 < t_2$ en *S* la función se denomina decreciente en sentido estricto en *S*. Una función se denomina monótona en *S* si es creciente ó decreciente. Monótona en sentido estricto significa que f, es estrictamente creciente ó es estrictamente decreciente. En general, el conjunto *S* se considera, un intervalo abierto ó un intervalo cerrado.

Por otra parte, sea $f(t)$ la *f.d.p* de $\boldsymbol{T}$, entonces su *f.d.a* estará dada por:

$$F(t) = Prob(\boldsymbol{T} \leq t) = \int_0^t f(t)dt.$$

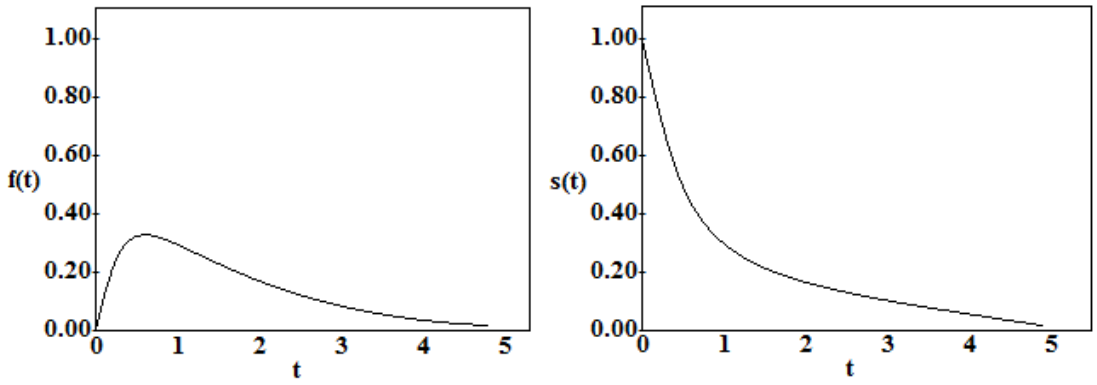


Figura 1.1. Función de Densidad de Probabilidad y Función de Supervivencia.

Función de supervivencia. La función básica empleada para describir los fenómenos de tiempo-evento es la función de supervivencia denotada por $s(t)$, también llamada tasa de supervivencia acumulativa o función de fiabilidad (figura 1.1). Esta función es la probabilidad de que un sujeto en estudio no experimente el evento de interés (sobreviva) antes de un momento dado, por tanto, sea $\boldsymbol{T}$ una variable aleatoria no negativa (de tiempo de falla) con Función de Distribución $F(t)$ y Función de Densidad de Probabilidad $f(t)$.

$$s(t) = Prob(\boldsymbol{T} > t)$$

O visto de otra forma

$$s(t) = 1 - F(t)$$

$$s(t) = 1 - Prob(\boldsymbol{T} \leq t)$$

$Se\ lee: uno\ menos\ la\ Probabilidad(un\ individuo\ falle\ antes\ del\ tiempo\ t)$

Por tales caracteristicas, $s(t)$ es una función monótona decreciente y tiene las siguientes propiedades:

1. $s(t)$ es una función continua monótona no creciente.

2. $s(0) = 1$

3. $\lim_{t \to +\infty} s(t) = 0$

Esto es, la probabilidad de sobrevivir al menos al tiempo cero es uno, obviamente, ha de alcanzar en todo caso una edad, por pequeña que sea, mayor a la de su nacimiento, porque el concepto de cohorte inicial, implica de nacidos vivos, no se acepta el concepto de nacidos muertos (mortinatos). Y la de sobrevivir un tiempo infinito es cero porque no hay probabilidad de que un individuo supere la edad extrema.

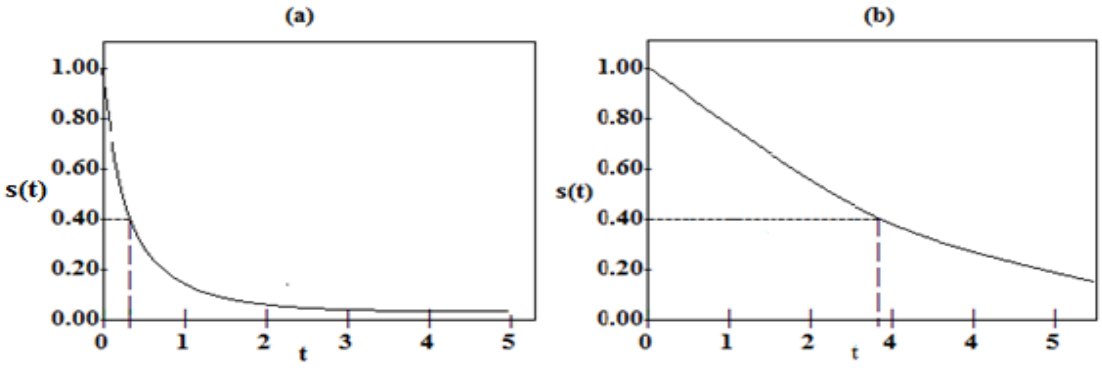
Cuando $\boldsymbol{T}$ es una variable aleatoria continua, la función de supervivencia es la integral de la función de densidad de probabilidad, esto es,

$$s(t) = Prob(\boldsymbol{T} > t) = 1 - F(t) = \int_t^{+\infty} f(t)dt \qquad (1.1.1)$$

En algunos textos, la participación de sistemas de tiempos de vida o artículos manufacturados, es referida como una función de rentabilidad.

Para describir el recorrido de la supervivencia, se hace la representación grafica de $s(t)$. Esta grafica es llamada curva de supervivencia. Muchos tipos de curvas de supervivencia pueden presentarse y analizarse de manera particular, pero es importante notar que todas tienen las mismas propiedades basicas, son monótonas no crecientes, igual a uno en cero y a cero cuando el tiempo tiende a infinito. La tasa de decrecimiento, varía de acuerdo al riesgo que experimente el evento al tiempo t, pero es difícil determinar en esencia la modelación de la falla solamente observando la curva de supervivencia. No obstante, el uso de esta curva representa un análisis importante en la práctica, y es usual comparar dos o más curvas de supervivencia para comprender el comportamiento que tienen entre ellas a lo largo del tiempo.

En la representación grafica, una curva de supervivencia empinada, como la que se muestra en la figura 1.2-(a) representa baja tasa de supervivencia o corto tiempo de supervivencia. Una curva de supervivencia plana o gradual como la que se muestra en la figura 1.2-(b) representa alta tasa de supervivencia o mayor supervivencia.



Figuras 1.2. Curvas de supervivencia.

Función de riesgo. La función de riesgo del tiempo de supervivencia $\boldsymbol{T}$ da la tasa de falla condicional. Esta se define como la probabilidad de falla durante un intervalo de tiempo muy pequeño, suponiendo que el sujeto de estudio ha sobrevivido hasta el inicio del intervalo, o como el límite de la probabilidad de que un sujeto falle en un intervalo muy corto, t a $t + \Delta t$ dado que el individuo ha sobrevivido hasta el tiempo t. La función de riesgo queda definida como:

$$h(t) = \lim_{\Delta t \to 0^+} \frac{Prob(t \leq \boldsymbol{T} < t + \Delta t \,/\, \boldsymbol{T} \geq t)}{\Delta t}$$

$$h(t) = \lim_{\Delta t \to 0^+} \frac{Prob(\{t \leq \boldsymbol{T} < t + \Delta t\} \cap \{\boldsymbol{T} \geq t\}}{Prob(\boldsymbol{T} \geq t)\Delta t}$$

$$h(t) = \lim_{\Delta t \to 0^+} \frac{Prob(t \leq \boldsymbol{T} < t + \Delta t)}{Prob(\boldsymbol{T} \geq t)\Delta t}$$

$$h(t) = \lim_{\Delta t \to 0^+} \frac{F(t + \Delta t) - P(\boldsymbol{T} \leq t)}{Prob(\boldsymbol{T} \geq t)\Delta t}$$

$$h(t) = \frac{1}{Prob(\boldsymbol{T} \geq t)} \lim_{\Delta t \to 0^+} \frac{F(t + \Delta t) - F(t)}{\Delta t}$$

$$h(t) = \frac{1}{Prob(\boldsymbol{T} \geq t)} \frac{d}{dt} F(t)$$

$$h(t) = \frac{f(t)}{s(t)}$$

$$\therefore h(t) = \frac{f(t)}{s(t)} \qquad (1.1.2)$$

La función de riesgo a veces tiene otros nombres, entre ellos **tasa de riesgo** y **fuerza de mortalidad.**

NOTA

Cuando se trabaja en cuestiones de confiabilidad, en ingeniería, se tienen las siguientes notaciones y definiciones equivalentes.

- $R(t) = P(\boldsymbol{T} > t) = 1 - F(t)$, llamada función de confiabilidad

- $h(t)$, se le conoce como función de Hazard.

Las funciones $f(t), F(t), s(t)\ y\ h(t)$ son matemáticamente equivalentes de la distribución de $\boldsymbol{T}$.

Se pueden deducir expresiones para $s(t)$ y $f(t)$ en términos de $h(t)$.

$$f(t) = \frac{d}{dt}\big(F(t)\big)$$

$$f(t) = \frac{d}{dt}\big(1 - s(t)\big)$$

$$f(t) = -\frac{d}{dt}s(t)$$

$$f(t) = -s'(t)$$

$$\therefore f(t) = -s'(t) \qquad (1.1.3)$$

Partiendo de la ecuación (1.1.2) tenemos:

$$h(t) = \frac{f(t)}{s(t)}$$

Entonces sustituyendo la ecuación (1.1.3) en (1.1.2) se tiene:

$$h(t) = \frac{-s'(t)}{s(t)}$$

$$h(t) = -\frac{s'(t)}{s(t)}$$

Entonces

$$h(t) = -\frac{d}{dt}ln\big(s(t)\big) \qquad (1.1.4)$$

Luego

$$ln\big(s(t)\big) = -\int_0^t h(t)dt.$$

$$e^{ln(s(t))} = e^{-\int_0^t h(t)dt}$$

Entonces

$$s(t) = e^{\left(-\int_0^t h(t)dt\right)} \qquad (1.1.5)$$

También es importante definir la función de riesgo acumulada, denotada por $H(t)$. Donde

$$H(t) = \int_o^t h(t)dt \qquad (1.1.6)$$

Usando la ecuación (1.1.5) se obtiene que la función de supervivencia está dada por

$$s(t) = e^{[-H(t)]}$$

$$si \lim_{t\to+\infty} s(t) = 0\,, \text{entonces} \quad \lim_{t\to+\infty} h(t) \to +\infty$$

Finalmente despejando $f(t)$ de la siguiente ecuación $h(t) = \frac{f(t)}{s(t)}$ se tiene $f(t) = h(t)s(t)$ y sustituyendo la ecuación (1.1.5) se obtiene el siguiente expresion:

$$f(t) = h(t)e^{\left(-\int_0^t h(t)dt\right)}. \qquad (1.1.7)$$

1.2. Modelos Discretos

En el caso discreto. Sea $\boldsymbol{T}$ una v.a. discreta que toma valores t_j con j = 1, 2,..., *n.* La función de riesgo se define para los valores t_j y proporciona la probabilidad condicional de falla al tiempo $t = t_j$, dado que el individuo estaba vivo antes de t_j, por lo tanto, $\boldsymbol{T}$ puede ser tratada como una variable aleatoria discreta.

Función de probabilidad (f.p). Se asume que $\boldsymbol{T}$ puede tomar los valores $t_1, t_2, \dots, n$ con $0 \leq t_1 < t_2 < \cdots$, la *f.p* es:

$$f(t_j) = P(\boldsymbol{T} = t_j), \qquad j = 1, 2, \dots, n$$

Con función de densidad acumulada (*f.d.a*)

$$F(t_j) = P(\boldsymbol{T} \leq t) = \sum_{\{j|t_j>t\}} f(t_j).$$

Mientras que su función de supervivencia

$$s(t_j) = P(\boldsymbol{T} > t) = \sum_{\{j|t_j>t\}} f(t_j). \qquad (1.2.1)$$

Cuando es considerada como una función para toda $t \geq 0$, $s(t)$ es una función continua por la izquierda escalonada no creciente,

$$s(0) = 1$$

$$\lim_{t \to +\infty} s(t) = 0$$

La función de riesgo discreta está definida como:

$$h(t_j) = P(\boldsymbol{T} = t_j | \boldsymbol{T} > t_j)$$

$$h(t_j) = \frac{P(\boldsymbol{T} = t_j \cap \boldsymbol{T} > t_j)}{P(\boldsymbol{T} > t_j)}$$

$$h(t_j) = \frac{P(\boldsymbol{T} = t_j)}{P(\boldsymbol{T} > t_j)}$$

$$h(t_j) = \frac{f(t_j)}{s(t_j)}, \quad j = 1, 2, \ldots, n \qquad (1.2.2)$$

Pero

$$f(t_j) = P(\boldsymbol{T} = t_j)$$

$$f(t_j) = P(\boldsymbol{T} = t_j) + P(\boldsymbol{T} = t_{j+1}) + \cdots + P(\boldsymbol{T} = t_k)$$

$$-\left[P(\boldsymbol{T} = t_{j+1}) + P(\boldsymbol{T} = t_{j+2}) + \cdots + P(\boldsymbol{T} = t_k)\right]$$

$$f(t_j) = s(t_j) - s(t_{j+1})$$

Así de la expresión anterior y sustituyendo en la ecuación (1.2.2) se tiene:

$$h(t_j) = \frac{s(t_j) - s(t_{j+1})}{s(t_j)}$$

$$h(t_j) = 1 - \frac{s(t_{j+1})}{s(t_j)} \qquad (1.2.3)$$

Pero de la ecuación (1.1.5) $s(t) = e^{\left(-\int_0^t h(x)dx\right)} = H(t)$, similarmente de la ecuación (1.2.3).

$$\frac{s(t_{j+1})}{s(t_j)} = 1 - h(t_j)$$

Aplicando logaritmo natural se tiene $ln\left(s(t_{j+1})\right) - ln\left(s(t_j)\right) = ln\left(\left(1 - h(t_j)\right)\right)$, sumando todos los términos desde 0 hasta t, resulta

$$ln(s(t)) = ln\left(\prod_{j:t_j<t} \left(1 - h(t_j)\right)\right)$$

Finalmente, se obtiene:

$$s(t) = \prod_{j:t_j<t} [1 - h(t_j)]. \qquad (1.2.4)$$

1.3. Algunas observaciones sobre la función de riesgo

La función de riesgo es una característica importante de la distribución del tiempo de vida. Indica la forma en que el riesgo de falla varía con la edad o con el tiempo y esto es de interés en la mayoría de las aplicaciones. Información previa sobre la forma de la función de riesgo puede ayudar a guiar la selección del modelo.

Por último, si los factores que afectan al tiempo de vida de un individuo varían con el tiempo, a menudo es fundamental para el enfoque del modelo a través de la función de riesgo.

La figura 1.3 muestra las funciones de densidad continuas y la figura 1.4 las funciones de riesgo y *f.d.p* para cuatro distribuciones continuas. Las formas de las gráficas de las funciones de riesgo son cualitativamente diferentes:

a) tiene una función de riesgo monótona creciente,

b) tiene una función de riesgo monótona decreciente,

c) tiene una función de riesgo llamada forma de bañera o en forma de U y

d) muestra una función de riesgo en forma de bañera inversa.

Los modelos con éstas y otras formas son útiles en la práctica. Por ejemplo, la tasa de muerte de los individuos de una población que se siguen desde el nacimiento hasta su muerte, tienen una función de riesgo en forma de bañera. En general estamos familiarizados con este patrón en las poblaciones humanas, puesto que después de un período inicial en el cual las muertes pueden ocurrir, principalmente por defectos de nacimiento o enfermedades infantiles, después de esta etapa la tasa de mortalidad es relativamente baja y constante hasta la edad de 30 años ó menos, posteriormente se incrementa con la edad. Este patrón también se manifiesta en poblaciones biológicas y en las poblaciones de artículos manufacturados, algunos de los cuales contienen defectos.

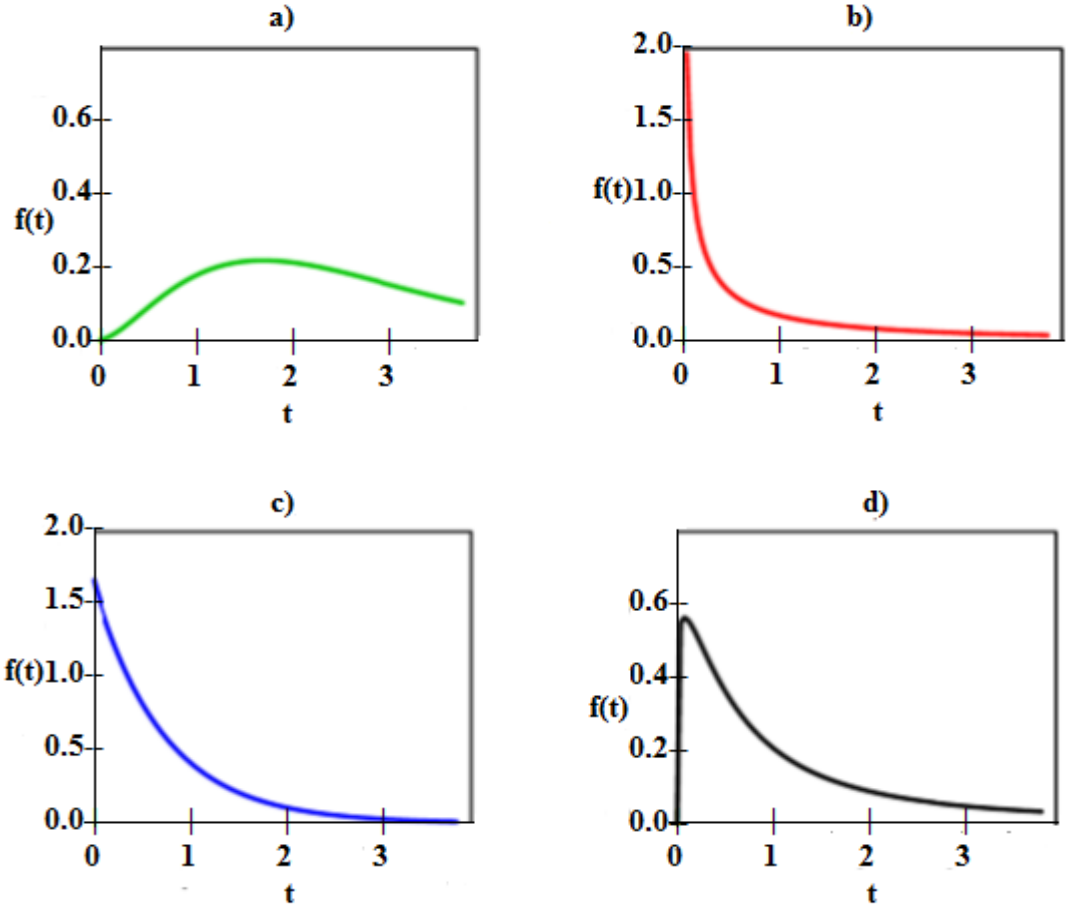


Figura 1.3. Algunas funciones continuas de densidad de probabilidad

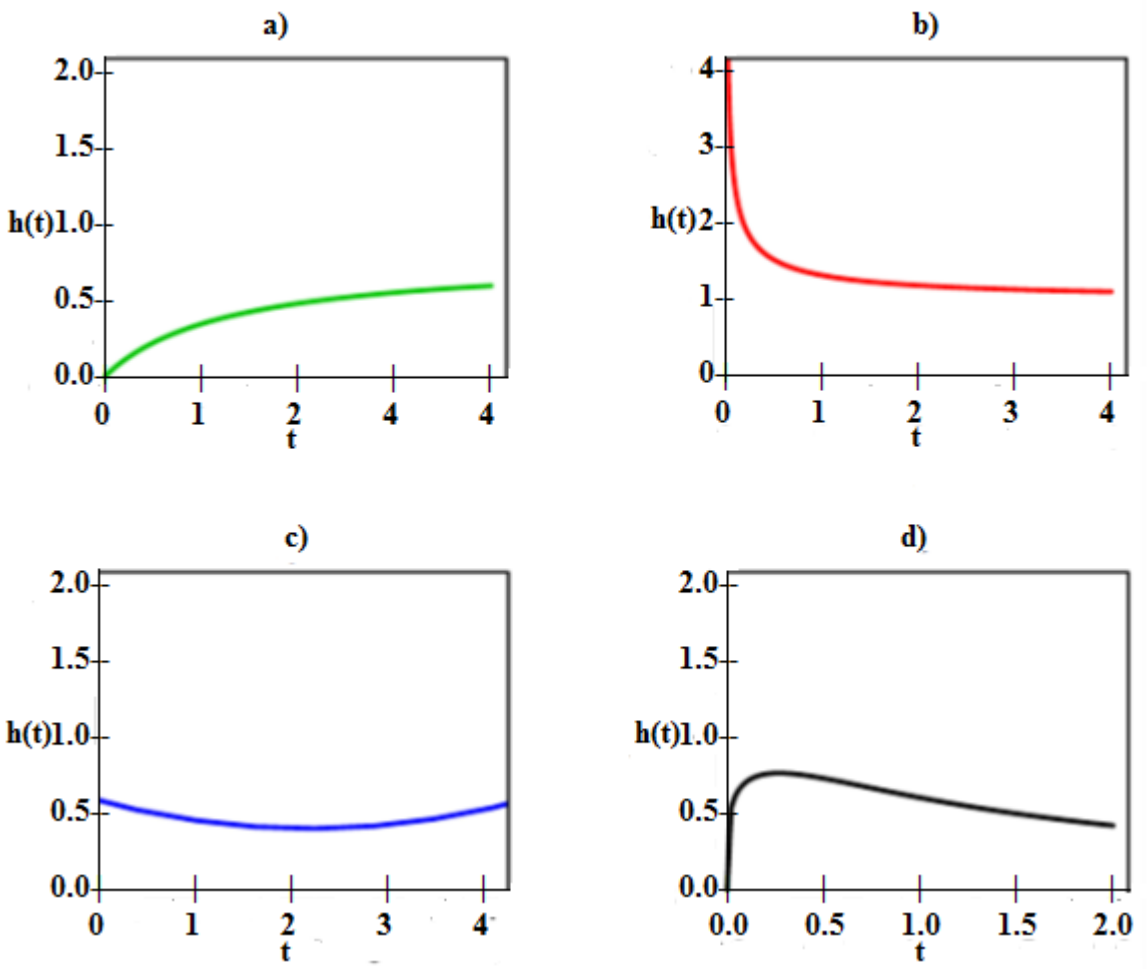


Figura 1.4. Algunas funciones continuas de riesgo.

Algunas distribuciones con incremento en las funciones de riesgo son vistas en personas para las que algún tipo de envejecimiento o deterioro se lleva a cabo. Además, las poblaciones que muestran una función de riesgo en forma de bañera a veces afectan a las personas débiles, dejando una población reducida con una función de riesgo creciente, por ejemplo, los fabricantes pueden utilizar una inspección en el proceso, en el que los artículos son sometidos a un breve período de operación antes de ser enviados a los clientes. De esta manera los artículos defectuosos o de mala calidad que podrían fallar se removerían de la población, lo que con frecuencia deja una población residual que presenta una función

creciente de riesgo. Ciertos tipos de dispositivos electrónicos con visualización de riesgo decreciente, como cuando los artículos con defectos fallan son removidos de la población. Alrededor de las funciones de riesgo constantes tienden a ocurrir en un contexto estable donde la falla o muerte se debe a fenómenos aleatorios tales como choques o accidentes. La forma (d) de la figura 1.4, donde $h(t)$ aumenta primero a un máximo y luego decrece, se encuentra en muchas aplicaciones, por ejemplo, en el caso de la supervivencia después del tratamiento para cáncer, donde algunos individuos se curan.

Los factores o covariables que afectan al tiempo de vida de los individuos, se les conocen como, "variables en el tiempo" o "variables dependiente del tiempo", por ejemplo, en estudios de la edad en que los fumadores desarrollan enfermedades crónicas, el tipo y el nivel de tabaquismo en cada individuo puede variar con el tiempo. La duración de un matrimonio (ejemplo 1.2) puede verse afectada por la presencia de niños o la situación laboral de la pareja, la cual puede cambiar con el tiempo. Cuando hay covariables o variables que cambian en el tiempo, por lo general es indispensable reflexionar sobre los modelos en términos de sus funciones de riesgo. No se puede discutir la relación del tiempo de vida y las covariables sin considerar la covariable "historia", es decir, los valores de las covariables tomando el tiempo como un enfoque de utilidad general es considerar la función de riesgo en el tiempo condicional t sobre los valores anteriores de la variable o covariable.

1.4. Algunos modelos paramétricos importantes

Algunos tiempos de falla pueden ser caracterizados por familias de distribuciones especificas que solo dependen de uno o varios parámetros desconocidos, los cuales proporcionan las características especificas del modelo en estudio. La selección de un modelo paramétrico es usualmente mediante la función de riesgo, pues de acuerdo a la información que el investigador tenga del fenómeno que causa la falla, puede determinar las características que el modelo debe seguir en la forma de la tasa de riesgo conforme avanza el tiempo. Por ejemplo, puede ser que el riesgo de muerte de un paciente después de someterse a alguna cirugía sea creciente las primeras horas y después (si sobrevive), su salud se estabilice hasta lograr su recuperación. En este caso, una función de riesgo creciente en valores pequeños del tiempo, que alcance un máximo y luego sea decreciente puede ser conveniente para modelar este fenómeno.

Utilizar un modelo paramétrico es restrictivo en el sentido de que se pueden exigir formas específicas del riesgo en el tiempo. Por ejemplo, el modelo exponencial que presenta riesgo constante, resultaría inadecuado para modelar el tiempo que tarda un individuo en morir cuando se le ha detectado una enfermedad terminal, pues en este caso, el riesgo debe ser claramente creciente. No obstante, puede haber situaciones donde se tenga evidencia para suponer que el riesgo puede ser constante en el tiempo, si fuera de interés modelar el tiempo que tarda en romperse la cuerda del violín de un concertista, puede ser que éste dependa de la dificultad de las piezas que el concertista tenga que tocar y el tiempo que invierta en practicar para perfeccionar el sonido, de modo que podría pensarse que la falla de la cuerda

puede suceder en cualquier momento, independiente del tiempo que lleve colocada en el instrumento.

Parámetros de las distribuciones. En el estudio de la distribución de una muestra los parámetros juegan un papel importante para determinar la distribución poblacional, puesto que la población puede tener la distribución supuesta pero con otros parámetros que no sean los propuestos. De esta forma crece el interés en conocer y estudiar los parámetros de la distribución que se supone tiene la población.

Los parámetros generalmente son del siguiente tipo:

- **Localización:** Este tipo de parámetro generalmente se relaciona con la media ó alguna medida de tipo central, se caracteriza por realizar un desplazamiento en una distribución de referencia, que comúnmente se llama distribución estándar.

- **Escala:** Este tipo de parámetro generalmente se relaciona con la desviación estándar o alguna medida de variación, se caracteriza por mostrar la amplitud de la gráfica sobre el eje de las ordenadas o dicho de otra manera, la misma forma que toma la gráfica pero en escala diferente sobre el eje de las ordenadas. Es decir, los parámetros de escala representan una transformación de una distribución de referencia, que comúnmente se llama distribución estándar.

- **Forma o asimetría:** Este tipo de parámetro como su nombre lo indica se relaciona con la forma de la distribución, ya que para algunos valores la función puede ser creciente y para otros decreciente la distribución en un mismo segmento de estudio.

1.4.1. Distribución Exponencial

Sea $\boldsymbol{T}$ una variable aleatoria (v.a.) que representa el tiempo de vida de un componente. Se dice que $\boldsymbol{T}$ tiene distribución exponencial con parámetro $\lambda > 0$ (y se denota $\boldsymbol{T} \sim e^{\lambda}$, si su función de densidad está dada por:

$$f(t) = \lambda e^{-\lambda t} \quad t \geq 0 \tag{1.4.1.1}$$

Donde: λ es un parámetro de escala, la media y varianza correspondiente a la Distribución Exponencial son: $\mu = E(\boldsymbol{T}) = \frac{1}{\lambda}$ y $V(\boldsymbol{T}) = \frac{1}{\lambda^2}$.

La función de supervivencia correspondiente es:

$$s(t) = e^{-\lambda t}. \quad \lambda > 0 \tag{1.4.1.2}$$

Considerando la ecuación (1.1.2) su función de riesgo es:

$$h(t) = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda \tag{1.4.1.3}$$

La distribución a menudo se expresa usando la reparametrización $\theta = \lambda^{-1}$, en este caso la *f.d.p* es:

$$f(t) = \theta^{-1}\mathrm{e}^{\frac{-t}{\theta}} \qquad (1.4.1.4)$$

usando la notación $\boldsymbol{T}\sim\exp(\theta)$ para indicar que la v.a $\boldsymbol{T}$ tiene distribución definida en (1.4.1.4). La distribución donde $\theta = 1$ es llamada distribución exponencial estándar, su *f.d.p* es mostrada en la Figura 1.5.

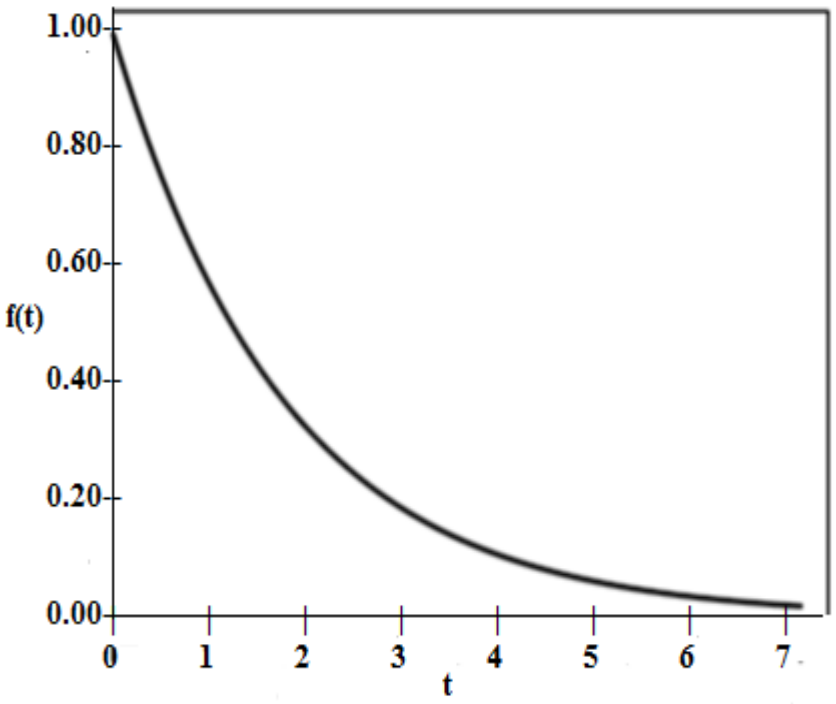


Figura 1.5. *f.d.p* de la Distribución Exponencial Estándar

Históricamente, la distribución exponencial fue la primera distribución que se utilizó para los modelos de tiempo de vida. Esto fue en parte por la disponibilidad de métodos estadísticos simples. Pero actualmente la suposición de una función de riesgo constante es muy restrictiva y actualmente tienen una aplicación bastante limitada.

1.4.2. Distribución Weibull

Otra de las distribuciones que se usa con gran frecuencia para modelar tiempos de vida es la Distribución Weibull. La aplicación de esta distribución a los tiempos de vida o durabilidad de artículos manufacturados es común. La Distribución Weibull es usada para modelar los tiempos de vida de diversos tipos de artículos, tal como: componentes de un automóvil y el aislante eléctrico. Además es usada en aplicaciones biológicas y médicas, por ejemplo, en el estudio del tiempo de ocurrencia de tumores en poblaciones humanas o en animales de laboratorio.

Se dice que una variable aleatoria continua no negativa $\boldsymbol{T}$ tiene distribución Weibull con metros $\lambda > 0$ (parámetro de escala) y $\beta > 0$ (parámetro de forma) y se denota $\boldsymbol{T} \sim$ Weibull (λ, β) si su función de densidad está dada por:

$$f(t) = \lambda\beta(\lambda t)^{\beta-1} e^{[-(\lambda t)^{\beta}]} \qquad (1.4.2.1)$$

La media y varianza correspondiente a la Distribución Weibull son:

$$\mu = E(\boldsymbol{T}) = \frac{\Gamma\left(1+\frac{1}{\beta}\right)}{\lambda} \;\; y \;\; V(\boldsymbol{T}) = \frac{\Gamma\left(1+\frac{2}{\beta}\right) - \Gamma^2\left(1+\frac{1}{\beta}\right)}{\lambda^2}$$

La figura 1.6 muestra algunos tipos de *f.d.p* para diferentes valores del parámetro β

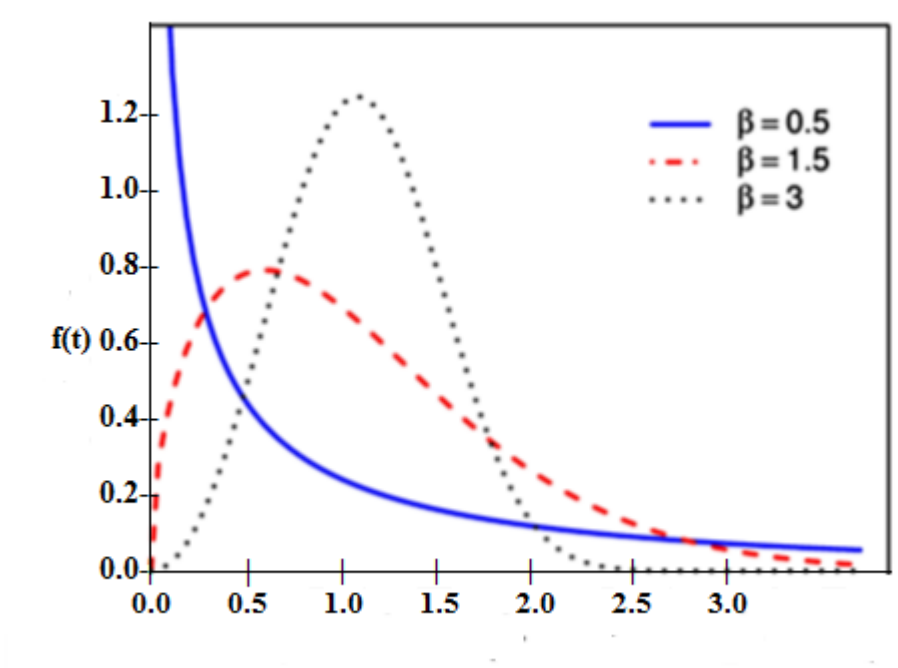


Figura 1. 6. *f.d.p* de la Distribución Weibull para $\lambda = 1$ y $\beta = 0.5, 1.5$ y 3

Además la función de supervivencia está dada por:

$$s(t) = e^{(-(\lambda t)^{\beta})} \tag{1.4.2.2}$$

Mientras que su función de riesgo es:

$$h(t) = \lambda\beta(\lambda t)^{\beta-1} \tag{1.4.2.3}$$

La Distribución Weibull es una generalización de la distribución exponencial, su tasa de falla puede ser constante, creciente o decreciente dependiendo de los parámetros, por lo que su aplicación es más amplia. Su función de riesgo es monótona creciente si $\beta > 1$, decreciente si $\beta < 1$ y constante si $\beta = 1$; el modelo es bastante flexible (Figura 1. 7), se ha

encontrado que proporciona una buena descripción de muchos tipos de datos de tiempos de vida, además este modelo tiene expresiones sencillas para las funciones de supervivencia y riesgo lo cual explica parte de su popularidad.

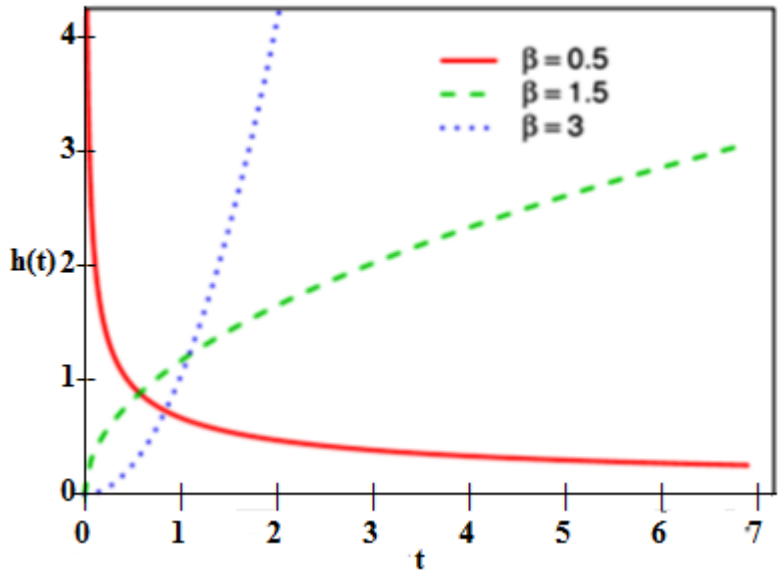


Figura 1. 7. Funciones de riesgo de la Distribución Weibull para $\lambda = 1\ y\ \beta = 0.5, 1.5\ y\ 3$

Típicamente, los valores de β varían de aplicación a aplicación, pero en algunas situaciones las distribuciones con β en el rango de 0.5 a 3 son apropiadas. La Figura 1. 7 muestra algunas funciones de riesgo para $\lambda = 1$ y diferentes valores de β. Note que el efecto para diferentes valores de λ en la Figura 1.6 serían sólo cambios de la escala en el eje horizontal (t) y no en la forma básica de la gráfica.

1.4.3. Distribución Gumbel o de valores extremos

Es conveniente introducir una distribución que está estrechamente relacionada con la Distribución Weibull. Esta es la primera Distribución Asintótica Gumbel o de valores extremos, en lo sucesivo se hará referencia a ella como Distribución Gumbel (también se le llama Distribución de Valores Extremos).

Se dice que una variable aleatoria continua *Y* tiene Distribución Gumbel, denotada por $Y \sim \text{Gumbel}(u, b)$ su *f.d.p* está dada por:

$$f(y) = b^{-1} e^{\left[\frac{y-u}{b} - e^{\left(\frac{y-u}{b}\right)}\right]}; \quad -\infty < y < +\infty \qquad (1.4.3.1)$$

Donde, $b > 0$ (parámetro de escala) y $u \in \mathbb{R}$, (parámetro de localización); además $u = -ln(\lambda)$ $y\, b = \frac{1}{\beta}$.

Una relación importante entre la Distribución Weibull y la Gumbel es la siguiente: Si $\boldsymbol{T}$ tiene una Distribución Weibull, entonces tiene una Distribución Gumbel.

Partiendo de la siguiente ecuación $s(t) = Prob(\boldsymbol{T} > t) = 1 - F(t) = \int_t^{+\infty} f(t)dt$ se obtiene:

$$s(y) = P[ln(\boldsymbol{T}) > y] = P\left[e^{ln(\boldsymbol{T})} > e^{y}\right] = P[\boldsymbol{T} > e^{y}]$$

$$s(y) = s_{\boldsymbol{T}}(e^{y}) = e^{[-(\lambda e^{y})^{\beta}]}$$

$$s(y) = e^{[-(\lambda^{\beta} e^{y\beta})]} = e^{\left[-e^{y\beta} e^{ln\left(\lambda^{\beta}\right)}\right]}$$

$$s(y) = e^{[-e^{y\beta} e^{\beta ln(\lambda)}]} = e^{\left[-e^{\beta(y+ln(\lambda))}\right]}$$

$$s(y) = e^{\left[-e^{\left(\frac{y-u}{b}\right)}\right]}$$

Donde $u = -ln(\lambda)$ y $b = \frac{1}{\beta}$. por tanto:

$$s(y) = e^{\left[-e^{\left(\frac{y-u}{b}\right)}\right]}; \quad -\infty < y < +\infty \qquad (1.4.3.2)$$

En el análisis de datos suele ser conveniente trabajar con el logaritmo de los tiempos de vida, por lo que en la Distribución Gumbel frecuentemente presenta este tipo de problemas.

La función de riesgo de la variable Gumbel está dada por:

$$h(y) = b^{-1} e^{\left[\frac{y-u}{b}\right]}, \quad -\infty < y < +\infty.$$

La Distribución Gumbel, Gumbel$(0,1)$ con $u = 0$ y $b = 1$ se denomina Distribución Gumbel Estándar y no afectan la forma de $f(y)$, únicamente la localización y escala.

Es decir, si $Y \sim \text{Gumbel}(u, b)$ entonces, $\frac{Y-u}{b} \sim \text{Gumbel}\,(0,1)$ la Figura 1. 8 muestra una gráfica de la *f.d.p*.

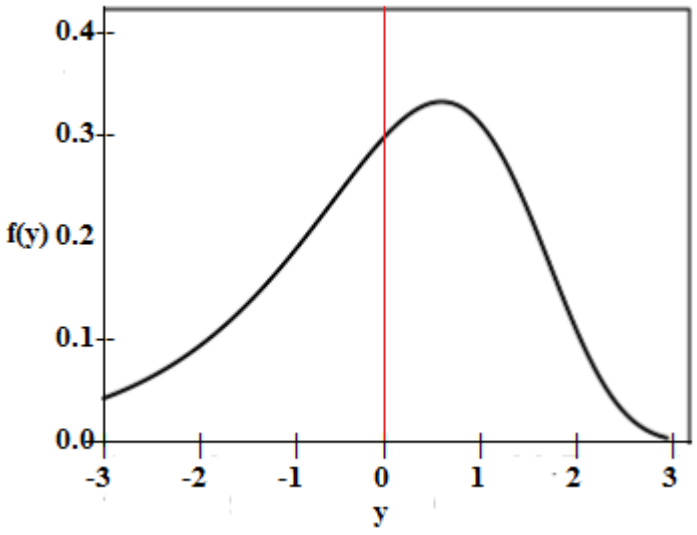


Figura 1. 8. *f.d.p* Estándar de la Distribución Gumbel

Los momentos para la Distribución Gumbel Estándar son $M(\theta) = \Gamma(1+\theta)$ y en forma general, la media y varianza correspondiente a la Distribución Gumbel son:

$\mu = E(Y) = u - \gamma b$ y $V(Y) = \frac{\pi^2 b^2}{6}$; en donde $\gamma \cong 0.5772$ ... es la constante de Euler.

1.4.4. Distribución Log-Normal

El uso de la distribución normal en Estadística es muy común; sin embargo, para el análisis del tiempo de vida es útil sólo para ciertos datos, cuando μ > 0 y su coeficiente de variación es pequeño $\left(\frac{\mu}{\sigma}\right)$, se usa frecuentemente para modelar el logaritmo del tiempo de vida.

La Distribución Log-Normal ha sido usada como un modelo en diversas aplicaciones tanto en la ingeniería como en medicina y en otras áreas.

Se dice que una variable aleatoria $\boldsymbol{T}$ tiene Distribución Log-Normal, si $Y = ln(\boldsymbol{T})$ tiene una Distribución Normal con media μ y varianza σ^2, con *f.d.p* de la forma:

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right]}; \quad -\infty < y < +\infty$$

Partiendo de que $t = e^y$ se puede encontrar la *f.d.p* para la Log-Normal.

$t = e^y \Rightarrow y = ln(t)$ y $\frac{dy}{dt} = \frac{1}{t}$, entonces

$$f_{\boldsymbol{T}} = f_y(ln(t)).\left|\frac{dy}{dt}\right| = \frac{1}{\sigma(2\pi)^{\frac{1}{2}}} e^{\left[-\frac{1}{2}\left(\frac{ln(t)\ -\mu}{\sigma}\right)^2\right]}\frac{1}{t}$$

Por lo tanto,

$$f(t) = \frac{1}{\sigma t\sqrt{2\pi}} e^{\left[-\frac{1}{2}\left(\frac{ln(t)\ -\mu}{\sigma}\right)^2\right]}; \quad t > 0 \qquad (1.4.4.1)$$

donde μ es el parámetro de localización y σ parámetro de escala.

Una variable aleatoria $\boldsymbol{T}$ tiene Distribución Log-Normal con parámetros $-\infty < \mu < \infty$ y $\sigma > 0$, y se denota por $\boldsymbol{T} \sim logN(\mu, \sigma^2)$ si su *f.d.p* es definida como en la ecuación (1.4.4.1).

Así entonces,

$$F(t) = P[\boldsymbol{T} \leq t]$$

$$F(t) = P[e^{Y} \leq t] = P[Y \leq ln(t)]$$

$$F(t) = \left[\frac{Y-\mu}{\sigma} \leq \frac{ln(t)-\mu}{\sigma}\right] = P\left[Z \leq \frac{ln(t)-\mu}{\sigma}\right]$$

$$F(t) = \varphi\left(\frac{ln(t)-\mu}{\sigma}\right)$$

Por tanto la función de supervivencia de la Distribución Log- Normal está dada por:

$$s(t) = 1 - \varphi\left(\frac{ln(t)-\mu}{\sigma}\right) \qquad (1.4.4.2)$$

Donde:

$$\varphi(x) = \int_{-\infty}^{x} \frac{1}{(2\pi)^{\frac{1}{2}}} e^{-u^2} du$$

De la ecuación (1.1.2) la función de riesgo es:

$$h(t) = \frac{f(t)}{s(t)} = \frac{\frac{1}{\sigma t(2\pi)^{\frac{1}{2}}} e^{\left[-\frac{1}{2}\left(\frac{ln(t)-\mu}{\sigma}\right)^2\right]}}{1 - \varphi\left(\frac{ln(t)-\mu}{\sigma}\right)}$$

Donde $h(0) = 0$ y crece hasta alcanzar un máximo y decrece aproximándose a 0 cuando $t \to \infty$. De esta forma se plantean muchas situaciones, por ejemplo cuando una población consiste en una mezcla de individuos que tienen cortos y largos tiempos de vida. Los ejemplos incluyen la supervivencia después del tratamiento para algunas formas de cáncer, donde las personas que son curadas favorecen su supervivencia por un plazo largo.

La Figura 1.9 muestra algunas *f.d.p* de la Distribución Log-Normal y la figura 1.10 las funciones de riesgo para $\mu = 0$ y diferentes valores de σ de la Distribución Log-Normal.

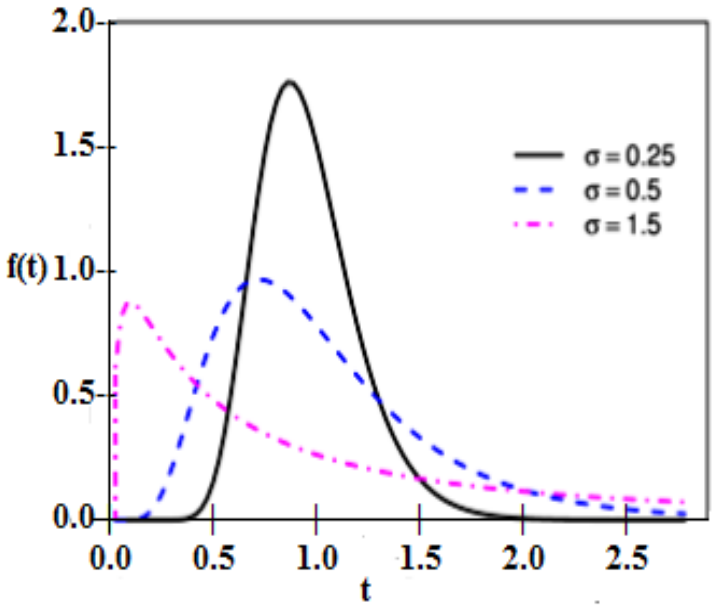


Figura 1.9 *f.d.p* de la Distribución Log-Normal con $\mu = 0$ y $\sigma = 0.25, 0.5,$ y 1.5.

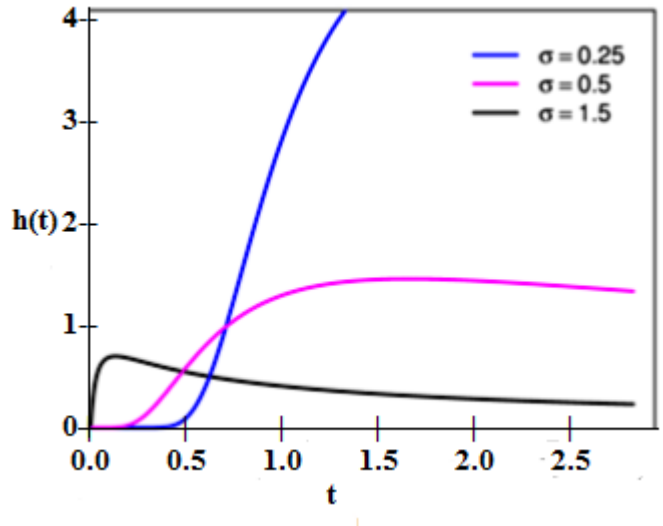


Figura 1. 10 Funciones de riesgo de la Distribución Log-Normal con $\mu = 0$ y $\sigma = 0.25, 0.5,$ y 1.5

La notación $Y \sim N(\mu, \sigma^2)$ se usa para denotar que Y es normal con media μ y varianza σ^2 y $\boldsymbol{T} \sim logN(\mu, \sigma^2)$ es usado para denotar que $\boldsymbol{T}$ tiene la *f.d.p* definida en (1.4.4.1)

1.4.5. Distribución Log-Logística

Una variable aleatoria $\boldsymbol{T}$ tiene Distribución Log-Logística con parámetros $\alpha > 0$ y $\beta > 0$, y se denota por $Y \sim LLogist(\alpha, \beta)$ si tiene *f.d.p* definida por:

$$f(t) = \frac{\left(\frac{\beta}{\alpha}\right)\left(\frac{t}{\alpha}\right)^{\beta-1}}{\left[1+\left(\frac{t}{\alpha}\right)^{\beta}\right]^2}; \qquad t > 0. \tag{1.4.5.1}$$

Donde, α es parámetro de escala y β es parámetro de forma, su forma estándar se obtiene cuando $\alpha = 1$, es decir $LLogist(1, \beta)$ y ésta dependerá del parámetro β. Además, los momentos están dados por $E(\boldsymbol{T}^r) = \alpha^r \Gamma\left(1 + \frac{r}{\beta}\right) \Gamma\left(1 - \frac{r}{\beta}\right)$, para $\beta > r$ en otros casos no existe, luego $E(\boldsymbol{T}) = \alpha\Gamma\left(1 + \frac{1}{\beta}\right) \Gamma\left(1 - \frac{1}{\beta}\right)$, para $\beta > 1$.

La función de supervivencia está dada por:

$$s(t) = \left[1 + \left(\frac{t}{\alpha}\right)^{\beta}\right]^{-1} \tag{1.4.5.2}$$

y su función de riesgo es definida como:

$$h(t) = \frac{f(t)}{s(t)} = \frac{\dfrac{\left(\frac{\beta}{\alpha}\right)\left(\frac{t}{\alpha}\right)^{\beta-1}}{\left[1 + \left(\frac{t}{\alpha}\right)^{\beta}\right]^2}}{\left[1 + \left(\frac{t}{\alpha}\right)^{\beta}\right]^{-1}} = \frac{\dfrac{\left(\frac{\beta}{\alpha}\right)\left(\frac{t}{\alpha}\right)^{\beta}}{\left(\frac{t}{\alpha}\right)}}{\left[1 + \left(\frac{t}{\alpha}\right)^{\beta}\right]^2}}{\dfrac{1}{\left[1 + \left(\frac{t}{\alpha}\right)^{\beta}\right]}} = \frac{\dfrac{\left(\frac{\beta}{\alpha}\right)\left(\frac{t}{\alpha}\right)^{\beta}}{\left(\frac{t}{\alpha}\right)\left[1 + \left(\frac{t}{\alpha}\right)^{\beta}\right]^2}}{\dfrac{1}{\left[1 + \left(\frac{t}{\alpha}\right)^{\beta}\right]}} = \frac{\left(\frac{\beta}{\alpha}\right)\left(\frac{t}{\alpha}\right)^{\beta}\left[1 + \left(\frac{t}{\alpha}\right)^{\beta}\right]}{\left(\frac{t}{\alpha}\right)\left[1 + \left(\frac{t}{\alpha}\right)^{\beta}\right]^2}$$

$$h(t) = \frac{\left(\frac{\beta}{\alpha}\right)\left(\frac{t}{\alpha}\right)^{\beta}}{\left(\frac{t}{\alpha}\right)\left[1 + \left(\frac{t}{\alpha}\right)^{\beta}\right]} = \frac{\left(\frac{\beta}{\alpha}\right)\left(\frac{t}{\alpha}\right)^{\beta-1}}{\left[1 + \left(\frac{t}{\alpha}\right)^{\beta}\right]}.$$

Por lo que:

$$h(t) = \frac{\left(\frac{\beta}{\alpha}\right)\left(\frac{t}{\alpha}\right)^{\beta-1}}{\left[1+\left(\frac{t}{\alpha}\right)^{\beta}\right]} \qquad (1.4.5.3)$$

La Distribución Log-Logística recibe el nombre del hecho de que $Y = ln(\boldsymbol{T})$ tienen una Distribución Logística con *f.d.p.*

$$f(y) = \frac{b^{-1}\mathrm{e}^{\left[\frac{(y-u)}{b}\right]}}{\left[1+\mathrm{e}^{\left[\frac{(y-u)}{b}\right]}\right]^2} \; ; \; -\infty < y < +\infty \qquad (1.4.5.4)$$

donde $u = ln(\alpha)$ parámetro de localización y $b = \beta^{-1}$ parámetro de escala, donde $-\infty < u < +\infty$ y $b > 0$.

La notación $Y \sim Logist(u, b)$ indica que *Y* tiene *f.d.p* definida en (1.4.5.4), se puede observar que su forma estándar se obtiene cuando $u = 0$ y $b = 1$; es decir, $Logist(0, 1)$ y su forma no depende de los parámetros. La media y varianza correspondiente a la Distribución Log-Logística son: $E(Y) = u$ y $V(Y) = \frac{\pi^2 b^2}{3}$.

La Figura 1.11 muestra la *f.d.p* para b = 0.14, 0.28 y 0.83, estos valores son elegidos de modo que la varianza de *Y* sea más o menos la misma que la varianza de una Distribución Normal con $\sigma =$ 0.25, 0.5 y 1.5 respectivamente, note la similitud de la Figura 1.9 con la Figura 1.11. La Distribución Logística y Normal tienen formas similares y se puede ver que

para $\beta > 1$ la función de riesgo tiene la misma característica de forma que la Log-Normal; además $h(0)$, creciente hasta alcanzar un máximo y luego se aproxima a cero en forma monótona cuando $t \to +\infty$, para $\beta \leq 1$ la función de riesgo es monótona decreciente.

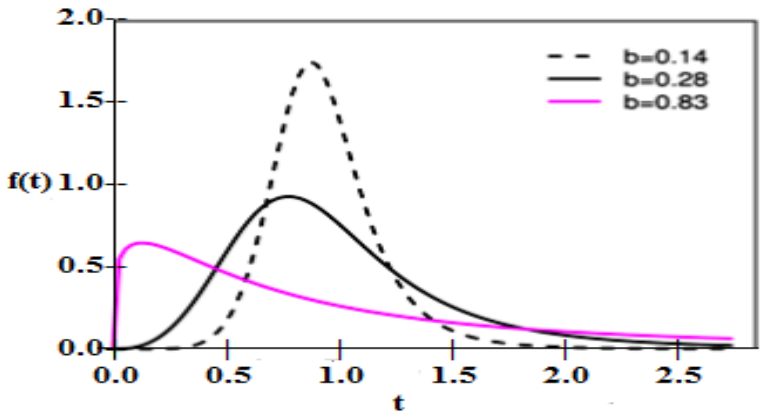


Figura 1. 11. *f.d.p* de la Distribución Log-Logística con $u = 0$ y $b = 0.14, 0.28,$ y 0.83

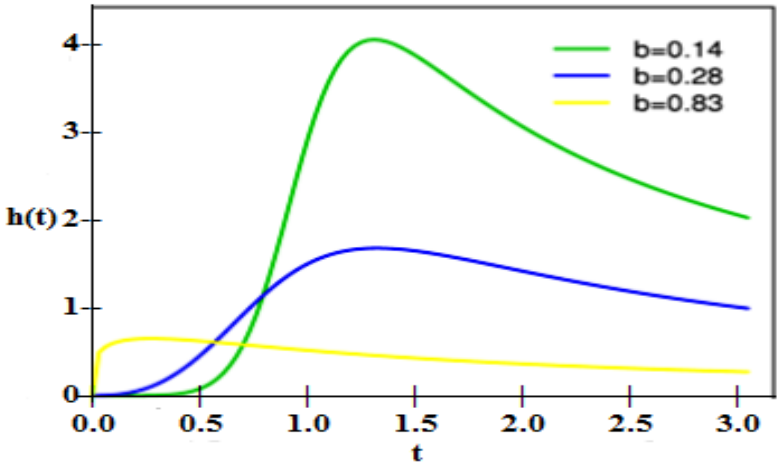


Figura 1. 12. Funciones de riesgo de la Distribución Log-Logística con $u = 0$ y $b = 0.14, 0.28,$ y 0.83

También es conveniente señalar la similitud entre las gráficas de las figuras 1.10 y 1.12.

1.4.6. Distribución Gamma

Es una distribución adecuada para modelizar el comportamiento de variables aleatorias continuas con asimetría positiva. Es decir, variables que presentan una mayor densidad de sucesos a la izquierda de la media que a la derecha. En su expresión se encuentran dos parámetros, siempre positivos, (k) y (λ) de los que depende su forma y alcance por la derecha, y también la función Gamma $\Gamma(k)$, responsable de la convergencia de la distribución.

Una variable aleatoria $\boldsymbol{T}$ tiene Distribución Gamma con parámetros $k > 0$ y $\lambda > 0$, si su *f.d.p* es de la forma:

$$f(t) = \frac{\lambda(\lambda t)^{k-1}\mathrm{e}^{-\lambda t}}{\Gamma(k)}; \qquad t > 0 \tag{1.4.6.1}$$

donde λ es el parámetro de escala y *k* parámetro de forma, de manera que si denotamos a la Distribución $\mathrm{Gamma}(\lambda, t)$, la Distribución Gamma Estándar se obtiene cuando $\lambda = 1$, $\mathrm{Gamma}(1, t)$ y su forma de la distribución depende del valor del parámetro *k*. Además, en general el valor esperado viene dado por $E(\boldsymbol{T}^r) = \lambda^r k(k+1)(k+2)\ldots(1+r-1)$, luego y $E(\boldsymbol{T}) = \lambda k$ y la varianza por $V(\boldsymbol{T}) = \lambda^2 k$.

La *f.d.p* de la Distribución Gamma Estándar es:

$$f(t) = \frac{t^{k-1}e^{-t}}{\Gamma(k)}; \qquad t > 0 \tag{1.4.6.2}$$

Esta Distribución, como la Distribución Weibull incluye la Distribución Exponencial como un caso particular (k = 1), las Funciones de Supervivencia y de Riesgo involucran a la Función Gamma incompleta dada por:

$$I = (k,t) = \frac{1}{\Gamma(k)}\int_0^t u^{k-1}e^{-u}\,du \tag{1.4.6.3}$$

Integrando la ecuación (1.4.6.2), se puede encontrar la Función de Supervivencia $s(t)$.

$$F(t) = \int_0^t f(t)dt = \int_0^t \frac{\lambda(\lambda t)^{k-1}e^{-\lambda t}}{\Gamma(k)}dt$$

$$F(t) = \frac{1}{\Gamma(k)}\int_0^t \lambda(\lambda t)^{k-1}e^{-\lambda t}dt$$

Sea $u = \lambda t \Rightarrow du = \lambda dt$

$$F(t) = \frac{1}{\Gamma(k)}\int_0^t \lambda(u)^{k-1}e^{-u}du$$

$$F(t) = I(k,\lambda t)$$

Por lo que la Función de Supervivencia es:

$$s(t) = 1 - I(k, \lambda t).$$

La Función de Riesgo viene dada por $h(t) = \frac{\lambda(\lambda t)^{k-1}e^{-\lambda t}}{\Gamma(k)(1-I(k,\lambda t))}$; para $0 < k < 1$, $h(t)$ es monótona decreciente con:

$$\lim_{t \to +\infty} h(t) = \lambda$$

$$\lim_{t \to 0} h(t) = +\infty$$

Caracteristicas de la Función de Riesgo de la Distribución Gamma:

- $h(t)$ es creciente si $k > 0$.
- Constante si $k = 1$.
- Decreciente si $k < 1$.

La notación $Y \sim \text{Gamma}(1, k)$ será usada para indicar que la variable aleatoria *Y* tiene *f.d.p.* definida en (1.4.6.2). Note que si $\boldsymbol{T}$ tiene *f.d.p* definida en (1.4.6.1) entonces $\lambda \boldsymbol{T} \sim \text{Gamma}(1, k)$, la Distribución Gamma con un parámetro está estrechamente relacionada

con la Distribución Ji-Cuadrada (χ^2). Si $Y \sim \text{Gamma}(1, k)$ entonces $2Y$ tiene distribución χ^2 con $2k$ grados de libertad, siendo referida como $\chi^2_{(2k)}$.

.

La Figura 1.13 muestra las *f.d.p* y la Figura 1.14 las funciones de riesgo para algunas Distribuciones Gamma. La Distribución Gamma no es tan usada como la Weibull, Log-Normal y Log-Logistica que ajustan una variedad de datos de tiempo de vida adecuadamente; sin embargo, también se plantea porque en algunas situaciones afectan a la Distribución Exponencial, esto porque se sabe que el resultado de la suma de variables aleatorias exponenciales independiente e idénticamente distribuidas (*i.i.d*) tienen una Distribución Gamma. Específicamente, si $\boldsymbol{T}_1, \ldots \boldsymbol{T}_n$ son independientes, cada una con Distribución Exponencial dada en (1.4.1.1), entonces $\boldsymbol{T}_1 + \cdots + \boldsymbol{T}_n$ tiene Distribución Gamma con parámetros λ y $k = n$. .

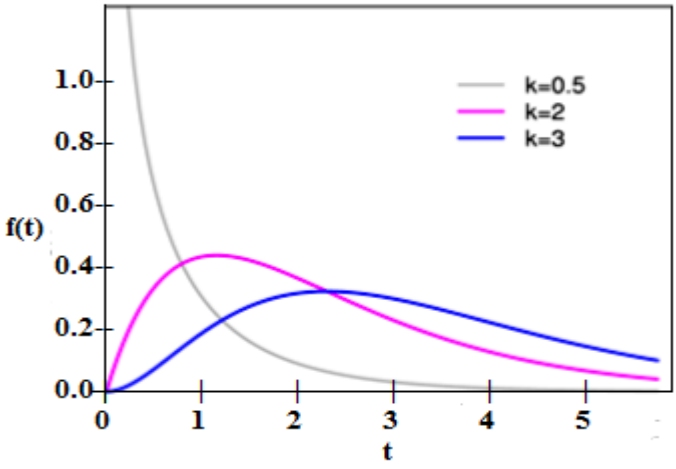


Figura 1. 13. *f.d.p* de la Distribución Gamma para $\lambda = 1$ y $k = 0.5, 2, 3$

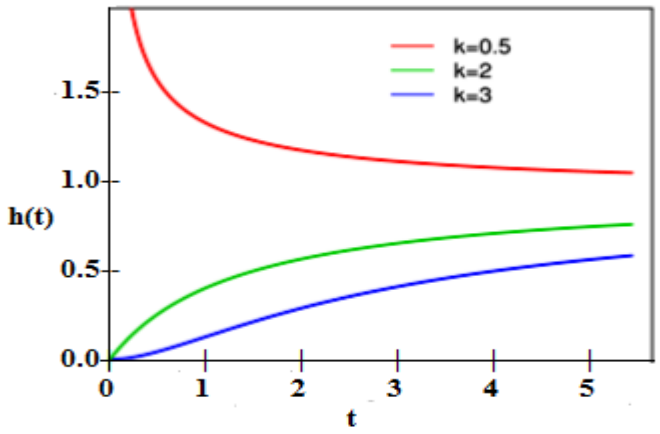


Figura 1. 14. Funciones de riesgo de la Distribución Gamma para $\lambda = 1$ y $k = 0.5, 2, 3$

1.5. Modelos Log-Loc-Escala

Por todo lo expuesto anteriormente se nota que existen modelos que se obtienen por medio de la transformación logaritmo de otro modelo conocido, que es la representación estándar de este último, y la forma de la distribución no depende de ningún parámetro de forma. A estos modelos los llamaremos **Logarítmicos de Localización y Escala**, Log-Loc-Escala.

1.5.1. Distribución Weibull (Log-Gumbel) con la Distribución Gumbel

Si su *f.d.p* es definida como (1.4.2.1), donde $t > 0,\ \lambda > 0$ (parámetro de escala) y $\beta > 0$ (parámetro de forma). La distribución Weibull Estándar, Weibull $(1, \beta)$, depende del parámetro de forma β.

Sea $\boldsymbol{T}$ una variable aleatoria con Distribución Weibull, con parámetro de escala λ y de forma β, entonces $Y = ln(\boldsymbol{T})$ tienen una Distribución Gumbel, $\text{Gumbel}(u, b)$, con $u = -ln(\lambda)$ y $b = \frac{1}{\beta}$ parámetros de localización y escala, respectivamente. Luego,

$$f(y) = \frac{1}{b} e^{\left[\frac{y-u}{b} - e^{\left(\frac{y-u}{b}\right)}\right]}; \ para\ y \in \mathbb{R}_0^+, u \in \mathbb{R}_0^+\ y\ b > 0.$$

La Distribución Gumbel Estándar es $\frac{(Y-u)}{b} \sim \text{Gumbel}\ (0, 1)$.

1.5.2. Distribución Log-Normal con la Distribución Normal

La *f.d.p* dada en la ecuación (1.4.4.1) con parámetros de localización $exp(u)$ y forma σ^2. La Distribución Log-Normal Estándar $LogN(0, \sigma^2)$ depende de la varianza.

Sea $\boldsymbol{T}$ una variable aleatoria con Distribución Log-Normal, con parámetros de localización $exp(\mu)$ y forma σ^2, entonces $Y = ln(\boldsymbol{T})$ tiene una Distribución Normal, $N(\mu, \sigma^2)$, con parámetro de localización μ y escala σ^2. Luego,

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right]}; \qquad -\infty < y < +\infty.$$

La Distribución Normal Estándar es $\frac{(Y-u)}{\sigma} \sim N(0, 1)$.

1.5.3. Distribución Log-Logística con la Distribución Logística

La *f.d.p* definida en la ecuación (1.4.5.1), con $\alpha > 0$ (parámetro de escala) y $\beta > 0$ (parámetro de forma). La Distribución Log-Logística Estándar $Y \sim LLogist(1, \beta)$ depende del parámetro β.

Sea $\boldsymbol{T}$ una variable aleatoria con Distribución Log-Logística, con parámetro de escala α y forma β; entonces $Y = ln(\boldsymbol{T})$ tiene una distribución, Logística(u, b), con parámetro de localización $u = ln(\alpha)$ y escala $b = \beta^{-1}$.

Luego,

$$f(y) = \frac{\frac{1}{b} e^{\left[\frac{(y-u)}{b}\right]}}{\left[1 + e^{\left[\frac{(y-u)}{b}\right]}\right]^2}; \quad para\ y \in \mathbb{R}_0^+.$$

La Distribución Logística Estándar es $\frac{(Y-u)}{b} \sim Logist\ (0, 1)$.

Una variable aleatoria Y definida en $]-\infty, +\infty[$ que provenga de un modelo Log-Loc-Escala tiene una *f.d.p* de forma:

$$f(y) = \frac{1}{b} f_0 \left(\frac{y-u}{b}\right); \quad -\infty < y < +\infty \qquad (1.5.3.1)$$

donde u es el parámetro de localización y $b > 0$ es parámetro de escala; además, $f_0(z)$ es una función de densidad específica en $]-\infty, +\infty[$. Para estas variables aleatorias las funciones de supervivencia y distribución son $s_0\left[\frac{(y-u)}{b}\right]$ y $F_0\left[\frac{(y-u)}{b}\right]$, respectivamente; donde:

$$F_0 = \int_{-\infty}^{z} f_0(w)dw = 1 - s_0(z).$$

De los tres casos analizados, note que la estandarización de la variable aleatoria $Z = \frac{(y-u)}{b}$ tiene las funciones anteriores $f_0(z)$ y $s_0(z)$ en su forma estándar determinada en (1.5.3.1) cuando $u = 0, b = 1$. Así en los casos vistos se tiene que $-\infty < z < +\infty$:

$f(z) = e^{[z-e^z]}$; de donde $s_0 = e^{(-e^z)}$ Gumbel

$f(z) = \frac{1}{\sqrt{2\pi}} e^{\left[-\frac{1}{2}z^2\right]}$; de donde $s_0 = 1 - \Phi(z)$ Normal

$f(z) = \frac{e^z}{(1+e^z))^2}$; de donde $s_0 = \frac{1}{1-e^z}$ Logística.

Como se ha revisado en los modelos de Localización-Escala la distribución para los tiempos de vida se obtiene a través de la transformación $\boldsymbol{T} = e^Y$. Así, en forma general la función de supervivencia para $\boldsymbol{T}$ puede ser expresada como:

$$P(\boldsymbol{T} \geq t) = s_0\left(\frac{ln(t) - u}{b}\right) = s_0^*\left[\left(\frac{t}{\alpha}\right)^{\beta}\right] \qquad (1.5.3.2)$$

donde $\alpha = e^{u}, \beta = b^{-1}$ y $s_0^*(x)$ la función de supervivencia definida en $(0, +\infty)$ por la relación $s_0^*(x) = s_0^*(ln(x))$.

Las familias de distribuciones con 3 ó más parámetros pueden ser obtenidas por la generalización (1.5.3.1), donde $f_0 = (z) = F_0(z) = s_0(z)$ para que incluyan uno o más parámetros de forma.

1.6. Modelos Log-Gamma y Gamma Generalizado

Se analizaron algunos ejemplos de los Modelos Log-Loc-Escala, ahora se revisará un ejemplo de un modelo generalizado con tres parámetros, el Modelo Log-Loc-Escala llamado Log-Gamma Generalizado y su correspondiente Modelo Loc-Escala llamado Gamma Generalizado. Para el Modelo Log-Loc-Escala Generalizado se tiene una versión original con la Distribución Gamma para $\left(\frac{\boldsymbol{T}}{\alpha}\right)^{\beta}$ con $k > 0$, es decir, si en la expresión para Gamma $(1, k) \sim \frac{y^{k-1}e^{-y}}{\Gamma(k)}$, se hace un cambio de variable, $f_{\boldsymbol{T}} = f_Y(\varphi(t)\varphi'(t))$, con $y = \varphi(t) = \left(\frac{t}{\alpha}\right)^{\beta}$, al derivarse se tiene:

$$\frac{dy}{dt} = \frac{d}{dt}\left(\frac{t}{\alpha}\right)^{\beta} = \frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1}.$$

Luego, sustituyendo este valor en la función de densidad para la Distribución Gamma $(1, k)$ definida en la ecuación (1.4.6.1)

$$f(t) = \frac{1}{\Gamma(k)}\left[\left(\frac{t}{\alpha}\right)^{\beta}\right]^{k-1} e^{-\left(\frac{t}{\alpha}\right)^{\beta}} \left(\frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1}\right) = \frac{\beta}{\alpha}\frac{1}{\Gamma(k)}\left(\frac{t}{\alpha}\right)^{\beta k-1} e^{-\left(\frac{t}{\alpha}\right)^{\beta}}.$$

Es, decir, la función de densidad que se consideró inicialmente estará dada por:

$$f(t) = \frac{\beta}{\alpha\Gamma(k)}\left[\left(\frac{t}{\alpha}\right)\right]^{\beta k-1} e^{-\left(\frac{t}{\alpha}\right)^{\beta}};\ \text{donde } t, k, \beta > 0 \qquad (1.6.1)$$

Aquí, α es un parámetro de escala y k, β parámetros de forma.

Un problema con esta definición consiste en este caso que la media y la varianza correspondiente a esta Distribución viene dada por $\mu = E(\boldsymbol{T}) = V(\boldsymbol{T}) = k$, dependen del parámetro de forma. Para evitar este problema se puede considerar $W = \frac{(Y-u_1)}{b_1}$, donde $Y = ln(\boldsymbol{T})$, $u_1 = ln(\alpha)$ y $b_1 = \beta^{-1}$, de tal forma que se tiene:

$$W = \frac{Y - u_1}{b_1} = \beta\big(ln(\boldsymbol{T}) - ln(\alpha)\big) = \beta\left[ln\left(\frac{\boldsymbol{T}}{\alpha}\right)\right] \Rightarrow \frac{\boldsymbol{T}}{\alpha} = e^{\frac{W}{\beta}}.$$

Realizando el cambio de variable para (1.6.1), tomando en cuenta que:

$$W = \frac{ln(T) - ln(\alpha)}{\beta^{-1}} = \beta\left[ln\left(\frac{\beta}{\alpha}\right)\right]$$

$$\frac{dt}{dW} = \frac{d}{dW}\alpha e^{\frac{W}{\beta}} = \frac{\alpha}{\beta}e^{\frac{W}{\beta}}.$$

Se tiene:

$$f(W) = \frac{\beta}{\alpha}\frac{1}{\Gamma(k)}\left(e^{\frac{W}{\beta}}\right)^{\beta k-1} e^{-\left(e^{\frac{W}{\beta}}\right)^{\beta}}\left[\frac{\alpha}{\beta}e^{\frac{W}{\beta}}\right]$$

$$f(W) = \frac{1}{\Gamma(k)}\left(e^{Wk}e^{\frac{W}{\beta}}\right)e^{-e^{W}}\left[e^{\frac{W}{\beta}}\right]$$

$$f(W) = \frac{1}{\Gamma(k)}(e^{Wk})e^{-e^{W}}.$$

Finalmente, se tiene:

$$f(t) = \frac{1}{\Gamma(k)}e^{(Wk-e^{W})};\ -\infty < W < +\infty,\ k > 0 \qquad (1.6.2)$$

Siendo la Distribución Log-Gamma con valor esperado $E(W) = \frac{d}{dk}ln\big(\Gamma(k)\big)$ y varianza $V(W) = \frac{d^2}{dk^2}ln\big(\Gamma(k)\big)$. Además, se sabe que los comportamientos para ambas funciones cuando $k \to +\infty$ son $ln(k)$ y k^{-1}, respectivamente.

Por tanto, actualmente se define la transformación Log-Gamma para la variable $Z = \sqrt{k}\ \big(W - ln(k)\big)$. Para determinar su función de densidad se tiene:

$$Z = \sqrt{k}\ \big(W - ln(k)\big) \Rightarrow W = \frac{Z}{\sqrt{k}} + ln(k) \Rightarrow \frac{dW}{dz} = \frac{1}{\sqrt{k}}$$

Realizando el cambio de variable en la ecuación (1.6.2)

$$f(z) = \frac{1}{\Gamma(k)}\left[e^{\left[\frac{z}{\sqrt{k}}+ln(k)\right]k - e^{\left[\frac{z}{\sqrt{k}}+ln(k)\right]}}\right]\frac{1}{\sqrt{k}}$$

$$f(z) = \frac{1}{\sqrt{k}}\ \frac{1}{\Gamma(k)}e^{\left[z\sqrt{k}+ln\left(k^k\right)- e^{\frac{z}{\sqrt{k}}}e^{ln(k)}\right]}$$

Así, se obtiene la función de densidad correspondiente es:

$$f_0(z,k) = \frac{k^{k-\frac{1}{2}}}{\Gamma(k)}e^{\left[z\sqrt{k}-ke^{\frac{z}{\sqrt{k}}}\right]};\ -\infty < z < +\infty. \qquad (1.6.3)$$

Con función de supervivencia, en la cual la variable estandarizada $\frac{(Y-u)}{b}$ es la forma:

$$s_0(z,k) = \left(1+\frac{1}{k}e^{z}\right)^{-k};\ -\infty < z < +\infty. \qquad (1.6.4)$$

donde $k > 0$ es el tercer parámetro, el caso especial cuando $k = 1$ es igual a la Distribución Logística Estándar y el límite cuando $k \to +\infty$ da el valor extremo de la distribución.

La Figura 1.15 muestra diferentes *f.d.p* definidas en la ecuación (1.6.3) para $k = 0.5, 1, 10$.

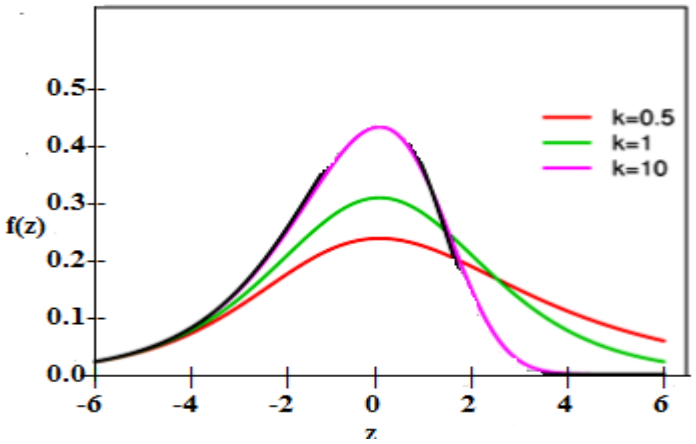


Figura 1.15 *p.d.f* de la Distribución Log-Gamma para $k = 0.5, 1, 10$.

Luego, se obtiene el modelo Log-Gamma Generalizado cuando sus otros dos parámetros de la familia de distribuciones de la ecuación (1.6.3) con $Z = \frac{(Y-u)}{b}$. Obteniendo los parámetros de localización u, escala b y forma k.

$$f(y;u,b,k) = \frac{1}{b} f_0(z;k) = \frac{k^{k-\frac{1}{2}}}{b\Gamma(k)} e^{\left(\frac{y-u}{b}\right)\sqrt{k} - k e^{\frac{\left(\frac{y-u}{b}\right)}{\sqrt{k}}}} \tag{1.6.5}$$

Propiedades de los Modelos Log-Generalizado

a) Cuando $k = 1$ coincide con la Distribución Gumbel.

b) Cuando $k \to +\infty$ converge a la Distribución Normal Estándar.

c) El valor esperado $E(Z)$ y la varianza $V(Z)$ dependen de k.

De forma inversa se obtiene la Distribución Gamma Generalizada para $\boldsymbol{T} = e^{Y}$ ó $Y = ln(\boldsymbol{T})$ con $u = ln(\alpha)$ y $b = \beta^{-1}$, para el cambio de variable $\frac{dy}{dt} = \frac{d}{dt} ln(t) = \frac{1}{t}$ sustituyendo en la ecuación (1.6.5)

$$f(t;\ \alpha, \beta, k) = \frac{\beta k^{k-\frac{1}{2}}}{\Gamma(k)} e^{\left[\beta(ln(t)-ln(\alpha))\sqrt{k}-k e^{\left[\frac{\beta(ln(t)-ln(\alpha))}{\sqrt{k}}\right]}\right]} \frac{1}{t}$$

$$f(t;\ \alpha, \beta, k) = \frac{\beta k^{k-\frac{1}{2}}}{\alpha\Gamma(k)} e^{\left[\beta\sqrt{k} ln\left(\frac{t}{\alpha}\right)-k e^{\frac{\beta}{\sqrt{k}} ln\left(\frac{t}{\alpha}\right)}\right]}$$

$$f(t;\ \alpha, \beta, k) = \frac{\beta k^{k-\frac{1}{2}} \left(\frac{t}{\alpha}\right)^{\beta\sqrt{k}}}{\alpha\Gamma(k)} e^{\left[-k\left(\frac{t}{\alpha}\right)^{\beta\sqrt{k}}\right]}.$$

Finalmente, se obtiene el Modelo Gamma Generalizado

$$f(t;\ \alpha, \beta, k) = \frac{\beta}{\alpha\Gamma(k)} k^{k-\frac{1}{2}} \left(\frac{t}{\alpha}\right)^{\beta\sqrt{k}-1} e^{\left[-k\left(\frac{t}{\alpha}\right)^{\frac{\beta}{\sqrt{k}}}\right]};\ \text{donde } t, \alpha, \beta, k > 0 \qquad (1.6.6)$$

Aquí α es un parámetro de escala y β, k parámetros de forma.

La figura 1.16 muestra las *f.d.p* para el Modelo Gamma Generalizado con $k = 0.5, 1, 10$.

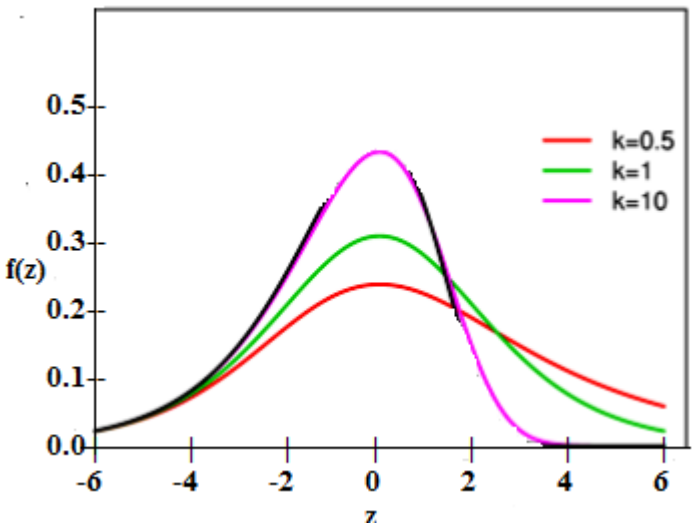


Figura 1.16 *f.d.p* del Modelo Gamma Generalizado para $k = 0.5, 1, 10$.

Propiedades del Modelo Gamma Generalizado.

a) Cuando $k = 1$ coincide con la Distribución Weibull (Log-Gumbel).

b) Cuando $k \to +\infty$ converge a la Distribución Log-Normal Estándar.

c) El valor esperado y la varianza del Modelo Gamma Generalizado dependen de k. Esto se obtiene del r-ésimo momento.

$$E(\boldsymbol{T}^r) = \int_{-\infty}^{+\infty} t^r f(t;\, \alpha, \beta, k)dt = \frac{\beta}{\alpha\Gamma(k)} k^{k-\frac{1}{2}} \int_0^{+\infty} t^r \left(\frac{t}{\alpha}\right)^{\beta\sqrt{k}-1} e^{\left[-k\left(\frac{t}{\alpha}\right)^{\frac{\beta}{\sqrt{k}}}\right]} dt$$

Con $x = \left(\frac{t}{\alpha}\right)^{\frac{\beta}{\sqrt{k}}} \Rightarrow \frac{t}{\alpha} = \left(\frac{x}{k}\right)^{\frac{\sqrt{k}}{\beta}} \Rightarrow dt = \frac{\alpha}{\beta\sqrt{k}} \left(\frac{x}{k}\right)^{\frac{\sqrt{k}}{\beta}-1} dx$

$$E(\boldsymbol{T}^r) = \frac{\beta}{\alpha\Gamma(k)} k^{k-\frac{1}{2}} \int_0^{+\infty} \left[\alpha \left(\frac{x}{k}\right)^{\frac{\sqrt{k}}{\beta}}\right]^r \left[\left(\frac{x}{k}\right)^{\frac{\sqrt{k}}{\beta}}\right]^{\beta\sqrt{k}-1} \mathrm{e}^{[-x]} \left(\frac{\alpha}{\beta\sqrt{k}} \left(\frac{x}{k}\right)^{\frac{\sqrt{k}}{\beta-1}}\right) dx$$

$$E(\boldsymbol{T}^r) = \frac{\alpha^r}{\alpha\Gamma(k)} k^{k-\frac{1}{2}} \int_0^{+\infty} \left(\frac{x}{k}\right)^{k+\frac{r\sqrt{k}}{\beta-1}} \mathrm{e}^{-x} dx$$

$$E(\boldsymbol{T}^r) = \frac{\alpha^r}{\Gamma(k)} k^{-\frac{1}{2}-\frac{r\sqrt{k}}{\beta-1}} \int_0^{+\infty} \mathrm{e}^{-x} (x)^{k+\frac{r\sqrt{k}}{\beta-1}} dx$$

Obteniendo los momentos,

$$E(\boldsymbol{T}^r) = \left(\frac{\alpha}{k^{\frac{\sqrt{k}}{\beta}}}\right)^r \frac{\Gamma\left(k + \frac{r\sqrt{k}}{\beta}\right)}{\Gamma(k)}.$$

Con la formula anterior se obtiene fácilmente la esperanza y la varianza de $\boldsymbol{T}$

$$E(\boldsymbol{T}) = \frac{\alpha}{k^{\frac{\sqrt{k}}{\beta}}} \frac{\Gamma\left(k + r\frac{\sqrt{k}}{\beta}\right)}{\Gamma(k)}.$$

$$V(\boldsymbol{T}) = \left(\frac{\alpha}{k^{\frac{\sqrt{k}}{\beta}}}\right)^2 \left[\frac{\Gamma\left(k + 2\frac{\sqrt{k}}{\beta}\right)}{\Gamma(k)} - \left(\frac{\Gamma\left(k + \frac{\sqrt{k}}{\beta}\right)}{\Gamma(k)}\right)^2\right].$$

1.7. Distribución Inversa Gaussiana

Surge en un proceso Wiener de tiempo continuo con parámetro $\gamma > 0$ y parámetro de dispersión σ^2 primero cruza un origen determinado $d > 0$. El proceso Wiener es un proceso Gaussiano Estocástico $\{X(t), t > 0\}$ con $X(0) = 0$ y una de sus propiedades es que $X(t) \sim N(\gamma t, \sigma^2 t)$ para $t > 0$. La variable aleatoria $\boldsymbol{T} = (t: X(t) = d)$ tiene *f.d.p*

$$f(t) = \frac{d}{\sigma(2\pi t^3)^{\frac{1}{2}}} e^{\left[\frac{-(d-\gamma t)^2}{2\sigma^2 t}\right]}; \quad t > 0.$$

Esta *f.d.p* depende solo $\frac{d}{\gamma}$ y $\frac{d}{\sigma}$; además es común reparametrizar definiendo $\mu = \frac{d}{\gamma}$ y $\lambda = \frac{d^2}{\sigma^2}$ Con esta reparametrización la función queda como:

$$f(t) = \frac{\lambda^{\frac{1}{2}}}{(2\pi t^3)^{\frac{1}{2}}} e^{\left[\frac{-\lambda(t-\mu)^2}{2t\mu^2}\right]}; \quad t > 0 \tag{1.7.1}$$

su *f.d.a* es:

$$F(t) = \Phi\left[\left(\frac{t}{\mu} - 1\right)\left(\frac{\lambda}{t}\right)^{\frac{1}{2}}\right] + e^{\frac{2\lambda}{\mu}} \Phi\left[-\left(\frac{t}{\mu} + 1\right)\left(\frac{\lambda}{t}\right)^{\frac{1}{2}}\right]. \tag{1.7.2}$$

Donde Φ es la *f.d.a* de la Distribución Normal Estándar. Denotaremos este modelo por $IG(\mu, \lambda)$.

La Distribución Inversa Gaussiana algunas veces es plausible en ajustes de modelos donde la falla ocurre cuando un proceso de deterioro alcanza un cierto nivel. En general se trata de una cierta flexibilidad de los modelos de dos parámetros con propiedades que son similares a los de la Distribución Log-Normal. La figura 1.17 muestra las funciones de densidad para $\mu = 1$ y varios valores para λ, mientras que la figura 1.18 muestra las correspondientes funciones de riesgo.

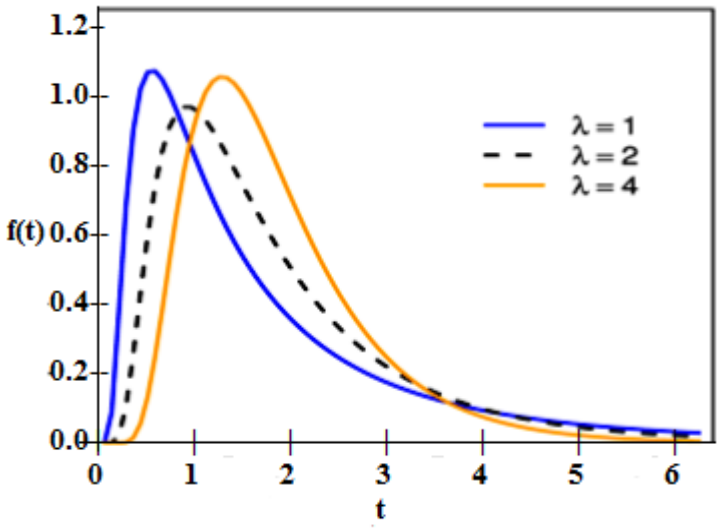


Figura 1.17. *f.d.p* de la Distribución Inversa Gaussiana para $\mu = 1$ y $\lambda = 1, 2$ y 4

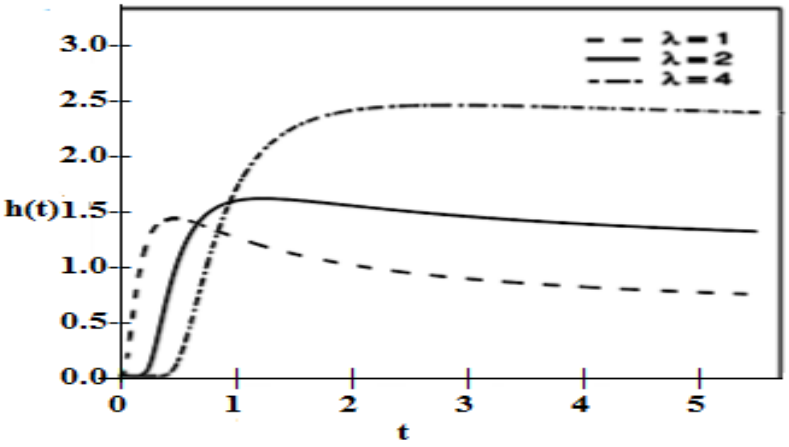


Figura 1.18. Funciones de Riesgo de la Distribución Inversa Gaussiana para $\mu = 1$ y $\lambda = 1, 2\ y\ 4$

CAPÍTULO 2. CENSURA Y MÁXIMA VEROSIMILITUD

INTRODUCCIÓN

En el capítulo anterior se mostró que los datos de tiempo de vida a menudo tienen la característica de terminar después de realizar las observaciones; por lo tanto, representan un tipo de censura en las observaciones. Sin embargo, la información disponible sobre un conjunto de datos del tiempo de vida puede presentar otras restricciones, como el problema de encontrar datos incompletos o que incluyan incertidumbre respecto al tiempo en que ocurre la falla. Un desafío importante para el análisis de este tipo de datos de vida consiste en desarrollar una metodología que se ocupe de la censura.

Por otro lado, en los procedimientos de inferencia estadística son bien usadas las funciones de verosimilitud basadas en los datos observados. En este capítulo se establecerá la forma de verosimilitud bajo censura y otras condiciones asociadas con la selección y observación de los individuos del estudio.

Supóngase que la distribución de probabilidad de posibles datos observables en un estudio se especifica hasta el vector de parámetros $\boldsymbol{\theta}$. La Función de Verosimilitud para $\boldsymbol{\theta}$, es proporcional a la probabilidad de los datos que se observaron, es decir:

$$L(\boldsymbol{\theta}) \propto P(Datos\,;\,\boldsymbol{\theta}).$$

Donde $Datos$ denota a los datos observados y P a la densidad de probabilidad o la función masa de la cual se supone que los datos surgen. Una notación más formal para $L(\boldsymbol{\theta})$, podría ser $L(\boldsymbol{\theta}: Datos)$**.**

Una pregunta que surge con frecuencia en el análisis de supervivencia es ¿Cómo los individuos son seleccionados y observados en el estudio de las distribuciones de tiempos de vida?. La motivación a esta pregunta se debe a que en el análisis de supervivencia esto tiene una variedad de formas dependiendo de los factores tales como el tiempo (cronológico) necesario para observar los eventos que definan el tiempo de vida, la viabilidad de los individuos sobre el paso del tiempo y el mecanismo para el registro del tiempo de vida y el valor de las covariables.

Luego, muchos estudios individuales de seguimiento longitudinal a través del tiempo son referidos como prospectos de estudio, algunos ejemplos incluyen pruebas de vida, exámenes clínicos y otro tipo de seguimiento de estudio (ver ejemplos 1.1 y 1.3). El grupo o cohorte de individuos en tales estudios es frecuente, pero no necesario, son seleccionados aleatoriamente de una población de individuos que se encuentran en el tiempo de origen $(t = 0)$ para el tiempo de vida de la variable $\boldsymbol{T}$. Con frecuencia en este tipo de estudio las limitaciones en la información colectada pueden ser impuestas por el tiempo disponible para su recolección, por el costo u otras limitaciones. En estos casos en que el seguimiento termine antes de que el individuo falle son causa que su tiempo de vida sea censurado por la derecha. En algunos ajustes sólo para determinar si un individuo falla o no como una

sucesión de puntos $a_1 < a_2 < \cdots < a_m$ en este caso, el tiempo de vida se encuentra en el intervalo $[a_{j-1}, a_j[$, conocido como intervalo censurado.

Algunas veces, los individuos no pueden ser seleccionados aleatoriamente y por consecuente $t = 0$. Una posibilidad es que sean seleccionados aleatoriamente de una población de individuos con vida; si u es un individuo y t el valor del tiempo de selección, entonces la condición inicial sobre los datos en que $t > u$ debe reflejarse en la función de máxima verosimilitud. Otra posibilidad es que la observación de datos para los individuos es por lo menos en la parte retrospectiva, esto significa que al menos en algunos de los datos utilizados en la función de máxima verosimilitud se presenta en forma cronológica antes del tiempo en que los individuos son seleccionados, en este caso puede haber condiciones no ignorables que se aplican a los tiempos de vida para los individuos que son seleccionados.

De esta forma en esté capítulo se revisan los siguientes temas: el primero relacionado con los experimentos de censura, seguido de los tipos de censura, revisaremos el método de máxima verosimilitud para la estimación de parámetros introducido por Fisher, este método será utilizado en la propuesta del trabajo para estimar los parámetros en el siguiente capítulo y se refiere a los modelos de regresión, modelos de riesgo proporcional de Cox.

2.1. Experimentos de censura

Debido a factores que escapan del control experimental se presenta pérdida de información que se produce en estudios de supervivencia y confiabilidad cuando los tiempos de vida de las unidades muestrales presentan algún tipo de censura. En supervivencia algunos pacientes pueden sobrevivir al final del ensayo clínico, o pasar a otro tratamiento, o morir por otra causa ajena a la enfermedad en estudio, produciéndose en cualquier caso una observación incompleta para algunas unidades muestrales, admitiendo que dichas unidades están censuradas. A continuación se muestran experimentos con datos censurados.

2.1.1. Experimento con censura tipo I

El mecanismo de censura tipo I se aplica cuando se observa a cada individuo en un tiempo $C_i > 0$ fijo de antemano, tal que $\boldsymbol{T}_i$ es observado en el i-ésimo tiempo, si $\boldsymbol{T}_i < C_i$, de otra manera sólo se conoce que $\boldsymbol{T}_i > C_i$. La censura tipo I surge a menudo como un estudio de especificaciones en un período de tiempo. Este tipo de censura también es denominada temporal y es la más habitual en estudios médicos.

Luego, al realizar un estudio sobre "tiempo de falla o muerte" en un paciente se tiene un tiempo limitado, por ende no es posible esperar a que todos los pacientes en estudio mueran. Esto indica que parece razonable detener el estudio en un tiempo de falla prefijado C_i, obteniendo solamente los tiempos de falla inferiores a C_i.

Una notación general para la censura tipo I, es:

$$t_i = min(\boldsymbol{T}_i, C_i) \quad \delta_i = I(\boldsymbol{T}_i < C_i) \tag{2.1.1.1}$$

EJEMPLO 2.1. Los resultados de un examen clínico, en los que la droga 6-mercaptopurina (6- MP) se comparó con un placebo con respecto a la capacidad de mantener la remisión en pacientes con leucemia aguda. La tabla 2.1 muestra el tiempo de remisión para dos grupos de 21 pacientes cada uno. Un grupo recibió el placebo y en el otro la droga 6-MP. Las observaciones son tiempos censurados; para estos pacientes, la enfermedad está en un estado de remisión al final del estudio.

Tabla 2. 1. Tiempos de remisión (semanas) para dos grupos de pacientes

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

* Observaciones censuradas

La censura es común en pruebas clínicas, por lo que a menudo la prueba termina antes de que todas las personas han “fallado”. Además, las personas pueden entrar en estudio varias ocasiones, por lo tanto puede estar bajo observación en diferentes periodos de tiempo.

El ejemplo 2.1 discute una prueba clínica concerniente a la duración de remisión de pacientes con leucemia, la cual fue planeada para llevarse a cabo durante un año con pacientes que estuvieron a prueba durante ese período. El tiempo de vida de la variable para

$\boldsymbol{T}_i$ un paciente fue la duración de la remisión a partir del momento de estudio y C_i puede ser el tiempo entre su fecha de inicio y el fin del estudio.

2.1.2. Experimento con censura aleatoria

Un proceso simple aleatorio de censura que es a menudo uno de los objetivos en el cual se asume que cada individuo tiene tiempo de vida $\boldsymbol{T}$ y tiempo de censura C, donde $\boldsymbol{T}$ y C son variables aleatorias continuas e independientes, con función de supervivencia $s(t)$ y $G(t)$ respectivamente, además todos los tiempos de vida y tiempos de censura se asumen mutuamente independientes, además se asume que $G(t)$ no depende de alguno de los parámetros de $s(t)$.

2.1.3. Experimento con censura tipo II

El término de censura tipo II se refiere cuando se inicia el estudio de n individuos durante un tiempo, el estudio se termina una vez que la r-ésima falla ocurra. Aunque algunas pruebas de vida son formuladas con censura tipo II, tienen la desventaja que el tiempo total $r(t)$ en que la prueba se ejecutará es aleatorio, luego se desconoce al principio de la prueba.

Por lo anterior se puede decir que la censura tipo I es más común en experimentos planeados. Las propiedades exactas del muestreo de los procedimientos estadísticos son basadas en la censura tipo II.

2.2. Datos censurados

Al realizar un análisis en general y en tiempos de vida se espera utilizar todos los datos que se encuentren disponibles, pero en el análisis de tiempos de vida a menudo ocurre que todos o algunos de los datos presentan el problema de estar incompletos o incluyen incertidumbre respecto a cuándo ocurrió la falla.

Para esto, los datos de vida serán clasificados en dos categorías:

- Completos o no censurados (toda la información esta disponible)

- Censurados (en donde un poco de información esta perdida)

Datos completos. Un dato (u observación) al tiempo t se dice que es completo si representa el tiempo "exacto" en el que ocurrió la falla de la unidad o individuo (tiempo de vida). Por ejemplo si los tiempos exactos en que fallan cinco unidades puestas a prueba son: 25, 50, 52, 53 y 60 horas, respectivamente, entonces se tienen datos completos.

Note que para tener datos completos es necesaria una observación constante de las n unidades o individuos puestos a prueba, además, el experimento finaliza cuando falla la última unidad en funcionamiento, siendo esto a veces una de las desventajas. Pero cabe mencionar que son los más fáciles de analizar, porque pueden utilizarse los métodos estadísticos directamente para realizar inferencias.

Datos censurados. Son datos donde no se conoce el tiempo total hasta la aparición del fracaso/éxito, bien porque el individuo/componente se retiro del estudio, ó bien porque se acabo el estudio.

En muchos casos cuando los datos de vida son analizados, todas las unidades o individuos en la muestra podrían no haber fallado (el evento de interés no fue observado) o los tiempos exactos de vida (tiempos en que ocurrieron las fallas) de las unidades no son todos conocidos.

Se introducirá una notación para los datos censurados, se supone que *n* individuos tienen tiempos de vida representados por las variables aleatorias $\boldsymbol{T}_1, \dots, \boldsymbol{T}_n$, luego t_i es el tiempo de vida o tiempo censurado. Por otro lado, sea la variable

$$\delta_i = I = \begin{cases} 1, & \boldsymbol{T}_i = t_i \\ 0, & \boldsymbol{T}_i > t_i \end{cases}$$

Llamada indicador de censura o indicador de estado para t_i, ya que indica si t_i es observado en el tiempo de vida $\delta_i = 1$ o si presenta censura de tiempo $\delta_i = 0$.

Los principales tipos de censura que suelen considerarse son: censura por la derecha, por la izquierda y por intervalo.

2.2.1. Datos censurados por la derecha

Censura por la derecha: se define así porque el límite inferior en el tiempo de vida es válido para algunos individuos. Los datos censurados por la derecha pueden ocurrir por varias razones:

- Cuando las pruebas se planifican, como cuando se decide terminar la prueba antes de que todos los artículos tengan fallas.

- Cuando las pruebas no se planifican; por ejemplo, cuando una persona en estudio presenta "perdida durante el seguimiento" porque, se alejan de la región donde se lleva a cabo el estudio. Es decir, este tipo de censura ocurre cuando el tiempo exacto de vida t de un individuo no es observado, pero se conoce que excedió un cierto tiempo, por decir t^+, ver (Figura 2.1)

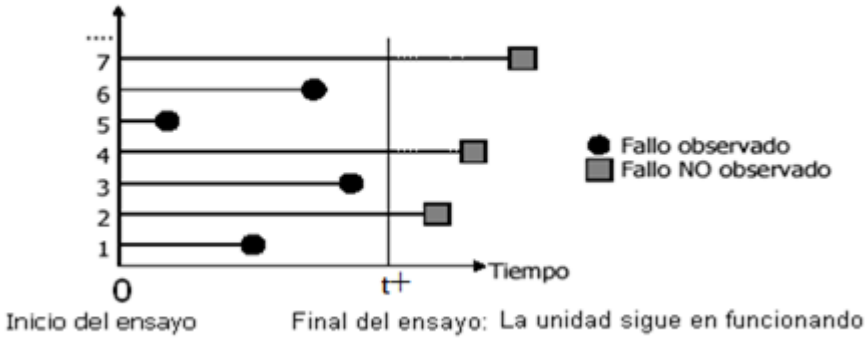


Figura 2.1. Censura por la derecha

El proceso de censura a menudo no es ninguno de los tipos discutidos hasta ahora, algunas veces puede ser complicado e imposible hacer un modelo. Por ejemplo, la decisión de terminar una prueba de vida o un análisis clínico al tiempo t, o retirar ciertos individuos puede estar basada en información de fallas antes del tiempo t. Afortunadamente se puede mostrar que bajo ciertas condiciones, en general la verosimilitud es de la forma:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i)^{\delta_i} s(t_i^+)^{1-\delta_i}$$

Puede ser empleada de la forma usual al hacer inferencia acerca de la distribución de tiempo de vida bajo estudio.

La idea clave, para un enfoque en general, es considerar la falla y el proceso de censura para un grupo de individuos en el paso del tiempo.

Suponga que n individuos son observados a partir de $t = 0$ hasta que fallen o sean censurados. Asumiendo que el tiempo de vida y el tiempo de censura son casos discretos, por conveniencia y sin pérdida de generalidad asuma que los valores para cada uno son $t = 0, 1, 2, \dots, n$. Suponga ahora que las covariables no varían en el tiempo, y sea $h_i(t)$ y $s_i(t)$ la función de riesgo y supervivencia (ver las ecuaciones (1.2.3) y (1.2.4) para el individuo, condicionada a los valores observados en las covariables.

Introduciendo una notación adicional dirigida a la evolución de las fallas y el proceso de censura sobre el tiempo, para $t = 0, 1, 2, \dots, n$; sea

$$Y_i(t) = I(\boldsymbol{T}_i \geq t_i, El\ individuo\ i\ no\ es\ censurado\ antes\ de\ t)$$
$$dN_i(t) = Y_i(t)I(\boldsymbol{T}_i = t)$$
$$dC_i(t) = Y_i(t)I(El\ individuo\ i\ es\ censurado\ en\ t).$$

La variable $Y_i(t)$ es a menudo llamada la indicadora de riesgo, toma el valor 1 si el individuo i está vivo y no censurado antes del tiempo t. Las variables $dN_i(t)$ y $dC_i(t)$ indican cuando se observó una falla y el evento de censura en el tiempo t, respectivamente. Entre todos los valores $\{dN_i(t),\ dC_i(t), t \geq 0\}$ sólo uno es diferente a cero para cualquier individuo.

Además se definen los vectores

$$d\boldsymbol{N}_i(t) = \big(dN_1(t), \dots, dN_n(t)\big), \qquad d\boldsymbol{C}_i(t) = \big(dC_1(t), \dots, dC_n(t)\big).$$

$$\mathcal{H}(t) = \{(d\boldsymbol{N}(s), d\boldsymbol{C}(s)), \qquad s = 0, 1, 2, \dots, t-1\}.$$

Refiriéndonos a $\mathcal{H}(t)$ como el histórico de fallas y el proceso de censura al tiempo t. Consta de la información acerca de todas las fallas y los eventos de censura que han ocurrido hasta el tiempo $t-1$.

El punto importante en este momento es que los datos que se observan pueden ser representados como:

$$Datos = (d\boldsymbol{N}(t), d\boldsymbol{C}(t); t = 0, 1, 2, \dots, n).$$

Además, la probabilidad de $Datos$ se puede descomponer como:

$$P(Datos) = \prod_{t=0}^{\infty} P(d\boldsymbol{N}(t)|\mathcal{H}(t))P(d\boldsymbol{C}(t)|d\boldsymbol{N}(t), \mathcal{H}(t)) \qquad (2.2.1.1)$$

Donde $\mathcal{H}(0)$ es nula. En la ecuación (2.2.1) todas las probabilidades son condicionales sobre el valor de las covariables, pero por simplicidad se suprimió en la ecuación.

2.2.2. Datos censurados por la izquierda

Censura por la izquierda. El tercer tipo de censura es similar a la censura por intervalos y es llamado censura por la izquierda. Este tipo de censura resulta cuando se sabe

que el tiempo exacto t a la falla de una unidad ocurrió antes de un cierto tiempo, digamos t (ver Figura 2. 2). Así, en datos censurados por la izquierda, por ejemplo, se puede conocer que cierta unidad falló antes de las 100 horas pero no se conoce exactamente cuándo. En otras palabras, tal unidad podría haber fallado en algún tiempo entre 0 y 100 horas. Esto es idéntico a datos censurados por intervalos en los cuales el tiempo de inicio del intervalo es cero.

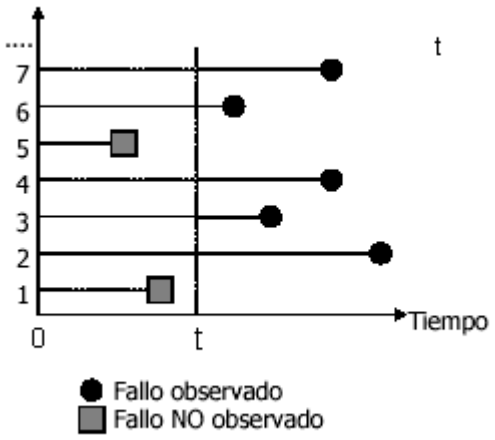


Figura 2.2. Censura por la izquierda

2.2.3. Datos censurados por intervalos

Una ocurrencia común para individuos en un estudio para observar de forma intermitente en puntos discretos del tiempo. Empezaremos por considerar un sistema donde cada individuo $i = 0, 1, 2, \dots, n$ son observados en un conjunto de tiempo pre-especificado $0 = a_{i0} < a_{i1} \dots < a_{im} < +\infty$, si los individuos no han presentado fallas en el tiempo $a_{i,j-1} (j = 1, \dots, m_i)$ se observa al tiempo a_{ij}. En este caso se desconoce el tiempo exacto en

que ocurre la falla de una unidad, la única información que se tiene es que la falla se presenta en un cierto intervalo de tiempo $]a_{ij-1}, a_{ij}]$.

Los datos observados se encuentran en el intervalo $]U_i, V_i]$ para cada individuo, donde $U_i < \boldsymbol{T}_i < V_i$, así se dice que los tiempos son censurados por intervalos (ver figura 2.3)

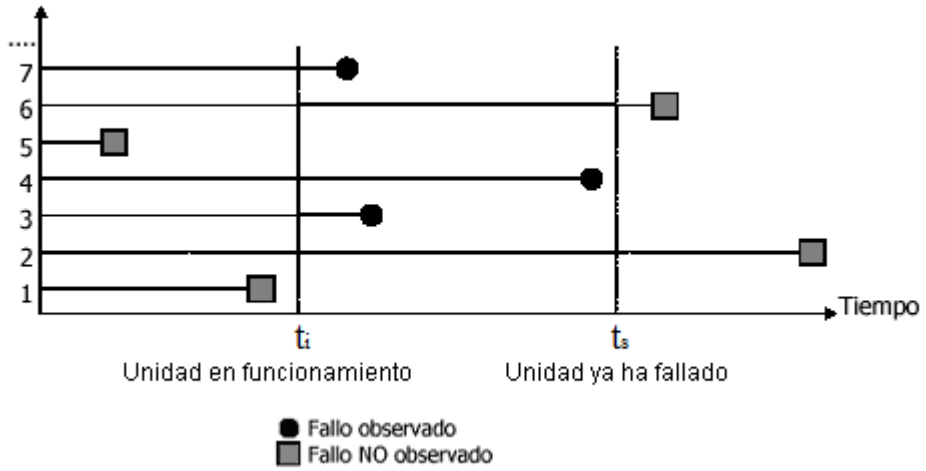


Figura 2.3. Censura Arbitraria o por Intervalos

En este tipo de datos censurados los supuestos de los tiempos de observación se fijan de antemano.

A veces la falla de una pieza de un material o equipo sólo puede determinarse por inspección.

EJEMPLO 2.2. Un tiempo de vida asociado con componentes metálicos tales como cuerpos de un avión, tubos de presión en los reactores nucleares o el tiempo en que las vías del ferrocarril presentan algún tipo de error.

En algunos estudios longitudinales en humanos es posible observar únicamente en intervalos, en donde sólo se pueden hacer observaciones durante uno o dos años. El tiempo de algunos tipos de eventos puede ser determinado, pero otros no.

EJEMPLO 2.3. La determinación en que un niño ha alcanzado la pubertad, la prueba se llevará a cabo en un periodo, por lo que la edad de inicio de la pubertad se encuentra en un intervalo censurado.

EJEMPLO 2.4. Si se corre una prueba sobre cinco unidades y se inspecciona cada 100 horas, solamente se sabrá si una unidad falló o no falló entre las inspecciones. Más específicamente, si se inspecciona una cierta unidad a las 100 horas y encontramos que está operando y luego al realizar otra inspección a las 200 horas encontramos que ya no se encuentra funcionando, entonces solamente se sabe que tal unidad falló en el intervalo 100 y 200 horas.

2.3. Máxima Verosimilitud

El método de Máxima Verosimilitud es uno de los métodos estadísticos más utilizados en la estimación de parámetros, consiste fundamentalmente en maximizar la función de

densidad conjunta de una muestra aleatoria $\boldsymbol{T}_1, \dots, \boldsymbol{T}_n$, pero con respecto a los parámetros de la función de densidad de la variable aleatoria cuando se proporciona una realización de la muestra aleatoria.

.

Supóngase que se tiene una muestra de observaciones independientes $t_1, \dots, t_n$ de la población de interés, en donde ninguna de las observaciones es censurada y t_i son tiempos de vida. Además su función de densidad es $f(t; \boldsymbol{\theta})$; $(\boldsymbol{\theta} = \theta_1, \theta_2, \dots, \theta_m)$, donde la función f es conocida con parámetros $\theta_1, \theta_2, \dots, \theta_m$. Entonces la función de verosimilitud esta definida por:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i; \boldsymbol{\theta})$$

En las propiedades del procedimiento estadístico basadas en datos censurados es necesario considerar el proceso por el que los tiempos de vida parecen tiempos censurados. Para hacer esto, se requiere de un modelo de probabilidad de tal forma que pueda establecerse un mecanismo de censura. Curiosamente, se torna a observar la función de verosimilitud para que los parámetros de tiempo de vida tomen la misma forma bajo una amplia variedad de mecanismos. Luego, se requiere considerar algunos tipos específicos de censura en los siguientes ejemplos se dará una formulación general.

Entonces los datos observados consisten en $(t_i,\ \delta_i),\ i = 1, 2, \dots, n$ con esta notación ocasionalmente t_i representará a una variable aleatoria ó el valor de una realización, por lo que hay que tener cuidado de no confundir en esta parte con la realización de la variable.

La función de Máxima Verosimilitud toma la forma:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i|\boldsymbol{\theta})^{\delta_i} s(t_i^+|\boldsymbol{\theta})^{1-\delta_i} \qquad (2.3.1)$$

El estudio de la función de Máxima Verosimilitud se lleva a cabo para obtener los Estimadores de Máxima Verosimilitud (EMV) $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_m$ de los parámetros $\theta_1, \theta_2, \dots, \theta_m$ y son los valores que maximizan la verosimilitud $L(\boldsymbol{\theta})$, o equivalentemente, maximizan el logaritmo de la verosimilitud $\ell(\boldsymbol{\theta})$.

En el caso de funciones diferenciables no monótonas los estimadores de máxima verosimilitud pueden encontrarse de las formas siguientes:

1. Resolviendo el sistema de ecuaciones:

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j} = 0 \; para \; j = 1, 2, \dots, m.$$

El sistema de ecuaciones puede ser demasiado complejo y en general casi nunca se puede resolver analíticamente.

Para la solución al siguiente sistema $\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j} = 0$ suelen utilizarse métodos numéricos, como el de Newton-Raphson.

2. Se calcula la matriz Hessiana.

3. Los valores estacionarios encontrados en el paso 1, se sustituyen en la matriz Hessiana y se aplica el criterio de decisión para determinar si se trata de puntos máximos de la función.

Para saber si se tiene un máximo se forma la Matriz Hessiana, y los valores críticos se sustituyen en las variables. Posteriormente, se calculan los determinantes de los menores de los elementos de la diagonal principal.

a) Si todos los determinantes son positivos se tiene un mínimo.

b) Si los determinantes cambian de signo alternativamente, iniciando con menos, entonces se tiene un máximo.

De los cursos de cálculo se sabe que la matriz Hessiana se forma con las segundas derivadas parciales de la función en estudio. Por ejemplo, sea $f(x,y)$ una función en dos variables la matriz Hessiana estará dada por:

$$J = \begin{bmatrix} \frac{\partial^2}{\partial x^2} f(x,y) & \frac{\partial^2}{\partial x \partial y} f(x,y) \\ \frac{\partial^2}{\partial y \partial x} f(x,y) & \frac{\partial^2}{\partial y^2} f(x,y) \end{bmatrix}$$

De tal forma que en este caso los determinantes de sus valores serán:

$$det \begin{bmatrix} \frac{\partial^2}{\partial x^2} f(x,y) & \frac{\partial^2}{\partial x \partial y} f(x,y) \\ \frac{\partial^2}{\partial y \partial x} f(x,y) & \frac{\partial^2}{\partial y^2} f(x,y) \end{bmatrix}$$

El problema aumenta en complejidad conforme aumenta la cantidad de parámetros, por tales condiciones es recomendable utilizar algún paquete matemático para resolver el problema.

EJEMPLO 2.5. Sea $X_1, X_2, \dots, X_n$ una muestra aleatoria de densidad normales con parámetros μ y σ^2. Entonces los estimadores de máxima verosimilitud de μ y σ^2.

En este caso la Función de Verosimilitud está dada por:

$$L(\mu,\sigma^2)=\prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-\left[\frac{(x_i-\mu)^2}{2\sigma^2}\right]}$$

$$L(\mu,\sigma^2)=\left(\frac{1}{2\pi}\right)^{\frac{n}{2}}\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}}\mathrm{e}^{W^*};\quad W^*=-\left(\sum_{i=1}^{n}(x_i-\mu)^2\right)^{-2\sigma^2}$$

Aplicando el logaritmo natural:

$$\ell(\mu,\sigma^2)=-\frac{n}{2}ln(2\pi)-\frac{n}{2}ln(\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Derivando parcialmente, con respecto a cada parámetro

$$\frac{\partial}{\partial\mu}[\ell(\mu,\sigma^2)]=-0-0-\frac{1}{2\sigma^2}(-2)\sum_{i=1}^{n}(x_i-\mu)=\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)$$

$$\frac{\partial}{\partial\sigma^2}[\ell(\mu,\sigma^2)]=-0-\frac{n}{2}\left(\frac{1}{\sigma^2}\right)+\frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2 \qquad .$$

$$\frac{\partial}{\partial\sigma^2}[\ell(\mu,\sigma^2)]=\frac{1}{2(\sigma^2)^2}\left[\sum_{i=1}^{n}(x_i-\mu)^2-n\sigma^2\right].$$

Igualando a cero cada ecuación se obtiene el sistema de ecuaciones:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)=0$$

$$\frac{1}{2(\sigma^2)^2}\left[\sum_{i=1}^{n}(x_i-\mu)^2-n\sigma^2\right]=0$$

Resolviendo se obtiene los valores críticos:

$$\begin{cases}\hat{\mu}=\overline{X}\\ \hat{\sigma}^2=\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i-\hat{\mu})^2\end{cases}$$

Para saber si es un máximo se calculan sus segundas derivadas parciales:

$$\frac{\partial^2}{\partial\mu^2}[\ell(\mu,\sigma^2)]_{(\hat{\mu},\hat{\sigma}^2)}=-\frac{n}{\sigma^2}\Big|_{(\hat{\mu},\hat{\sigma}^2)}=-\frac{n}{\hat{\sigma}^2}$$

$$\frac{\partial}{\partial\sigma^2}[\ell(\mu,\sigma^2)]_{(\hat{\mu},\hat{\sigma}^2)}=\left[\frac{n}{2}\left(\frac{1}{(\sigma^2)^2}\right)-\frac{1}{(\sigma^2)^3}\sum_{i=1}^{n}(x_i-\mu)^2\right]_{(\hat{\mu},\hat{\sigma}^2)}=-\frac{n}{2(\hat{\sigma}^2)^2}$$

$$\frac{\partial^2}{\partial\mu\partial\sigma^2}[\ell(\mu,\sigma^2)]_{(\hat{\mu},\hat{\sigma}^2)}=\left[\left(\frac{1}{(\sigma^2)^2}\right)\sum_{i=1}^{n}(x_i-\mu)\right]_{(\hat{\mu},\hat{\sigma}^2)}=0$$

Se forma el determinante de la Matriz Hessiana:

$$\begin{bmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & -\frac{n}{2(\hat{\sigma}^2)^2} \end{bmatrix}$$

Por otro lado, los determinantes de los menores son:

$$-\frac{n}{\hat{\sigma}^2} < 0 \text{ y } \frac{n^2}{2(\hat{\sigma}^2)^3} > 0.$$

Según la regla, se tiene un máximo, ya que los determinantes de los menores cambian su signo de menos a más. Por lo tanto, los estimadores de Máxima Verosimilitud para la media y varianza poblacional (μ, σ^2) son, respectivamente:

$$\hat{\mu} = \overline{X} \text{ y } \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2$$

2.3.1. Máxima Verosimilitud para datos con censura tipo I

La función de verosimilitud para la censura tipo I está basada en la distribución de probabilidad de (t_i, δ_i) para $i = 1, 2, \ldots, n$, donde t_i y δ_i son variables aleatorias en la ecuación (2.1.1.1), C_i son constantes fijas y t_i pueden ser valores menores que C_i con:

$$P(t_i = C_i, \delta_i = 0) = P(\boldsymbol{T}_i > C_i)$$

$$P(t_i, \delta_i = 1) = f(t_i); \quad t_i < C_i.$$

Donde la probabilidad de la segunda expresión denota la *f.d.p* o función masa de acuerdo si $\boldsymbol{T}_i$ tiene una distribución continua o discreta como t_i, por lo que, sus puntos tienen una función de distribución de probabilidad.

$$f(t_i)^{\delta_i}\left(P\big((\boldsymbol{T}_i > C_i)\big)^{1-\delta_i}\right) \tag{2.3.1.1}$$

Asumiendo que los tiempos de vida $\boldsymbol{T}_1, \dots, \boldsymbol{T}_n$ son estadísticamente independientes, se obtiene la función de Verosimilitud de la ecuación (2.3.1.1) como:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i|\boldsymbol{\theta})^{\delta_i} s(t_i^+|\boldsymbol{\theta})^{1-\delta_i} \tag{2.3.1.2}$$

Note que el termino $s(t_i^+)$ es igual a $P(\boldsymbol{T}_i > t_i)$, en general si $s(t_i)$ es continuo en t_i, entonces $s(t_i^+) = s(t_i)$. El ajuste para la ecuación (2.3.1.2) cuando las covariables x_i están presentes en el modelo se hace reemplazando a $s(t)$ por $f(t)$ con $s_i(t) = s(t) = \mathrm{e}^{-\lambda t}$ y $f_i(t) = f(t|x_i)$, respectivamente.

EJEMPLO 2.6. Supóngase que los tiempos de vida $\boldsymbol{T_i}$ son independientes y tienen Distribución Exponencial con *f.d.p* $f(t) = \lambda e^{(-\lambda t)}$ y $s(t) = e^{-\lambda t}$, entonces su Función de Verosimilitud (2.3.1.2) está dada por:

$$L(\lambda) = \prod_{i=1}^{n} \left(\lambda e^{-\lambda t_i}\right)^{\delta_i} \left(e^{-\lambda t_i}\right)^{1-\delta_i}$$

$$L(\lambda) = \lambda^r e^{W^{**}}; \text{ donde } \quad W^{**} = -\lambda \sum_{i=1}^{n} t_i \quad \text{y} \quad r = \sum_{i=1}^{n} \delta_i$$

r es el número de observaciones de los tiempos de vida que están censurados, por lo que la Función de Log-Verosimilitud $\ell(\lambda) = ln\left(L(\lambda)\right)$

$$\ell(\lambda) = r ln(\lambda) - \lambda \sum_{i=1}^{n} t_i.$$

La estimación de Máxima Verosimilitud está dada por la solución de $\frac{d\ell(\lambda)}{d\lambda} = 0$

$$\frac{d\ell(\lambda)}{d\lambda} = \frac{r}{\lambda} - \sum_{i=1}^{n} t_i = 0 \quad \Rightarrow \hat{\lambda} = r\left(\sum_{i=1}^{n} t_i\right)^{-1}.$$

Para verificar si la función de máxima verosimilitud alcanza su máximo se calcula la segunda derivad de $\ell(\lambda)$.

$$\frac{\partial^2}{\partial\lambda^2}\ell(\lambda) = -\frac{r}{\lambda^2} \Rightarrow \frac{\partial^2}{\partial\lambda^2}\bigg|_{\lambda=\hat{\lambda}} < 0$$

Por lo que el logaritmo de verosimilitud alcanza su máximo en $\hat{\lambda}$.

2.3.2. Máxima Verosimilitud para datos con censura aleatoria

Como en el caso de la censura tipo I, $t_i = min(\boldsymbol{T}_i, C_i)$ y $\delta_i = 1$; si $\boldsymbol{T}_i \leq C_i$ (no hay censura) y $\delta_i = 0$; si $\boldsymbol{T}_i > C_i$; (si hay censura). Los datos de los n individuos se asumen consistentes para los pares (t_i, δ_i), para $i = 1, 2, \dots, n$, así mismo los posibles resultados de C_i para todo $i = 1, 2, \dots, n$. Entonces, si $f(t)$ y $g(t)$ son $f.d.p$ para $\boldsymbol{T}_i$ y δ_i su función de densidad de probabilidad es la siguiente:

$$P(t_i = t, \delta_i = 0) = P(C_i = t, \boldsymbol{T}_i > C_i) = g(t)s(t).$$

$$P(t_i = t, \delta_i = 1) = P(\boldsymbol{T}_i = t, C_i \geq \boldsymbol{T}_i) = f(t)G(t).$$

Esto puede ser mezclado en la siguiente expresion:

$$P\left(t_i = t, \begin{cases}\delta_i = 0\\ \delta_i = 1\end{cases}\right) = [f(t)G(t)]^{\delta_i}[g(t)s(t)]^{1-\delta_i}.$$

Por lo tanto, la distribución de $(t_i, \delta_i), i = 1, 2, \dots, n$ es:

$$(t_i, \delta_i) = \prod_{i=1}^{n} [f(t_i)G(t_i)]^{\delta_i}[g(t_i)s(t_i)]^{1-\delta_i}$$

Entonces $G(\boldsymbol{T})$ y $\mathrm{g}(t)$ no involucran a algún parámetro de la función $f(t)$, los cuales pueden pasarse por alto, por lo tanto el núcleo de la función de máxima verosimilitud queda dado por:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i|\boldsymbol{\theta})^{\delta_i} s(t_i^+|\boldsymbol{\theta})^{1-\delta_i}.$$

La cual tiene la misma forma que la ecuación (2.3.1.2)

2.3.3. Máxima Verosimilitud para datos con censura tipo II

Con la censura tipo II, el valor de r se elige antes de que los datos sean colectados y los r tiempos de vida más pequeños en una muestra aleatoria $\boldsymbol{T}_1, \dots, \boldsymbol{T}_n$. Para las distribuciones continuas se puede ignorar la posibilidad de lazos y denotar a los r tiempos de vida más pequeños como $\boldsymbol{T}_{(1)} \leq \boldsymbol{T}_{(2)} \dots \leq \boldsymbol{T}_{(n)}$. Si $\boldsymbol{T}_i$ tiene función de densidad de probabilidad $f(t)$ y función de supervivencia $s(t)$, entonces toma la forma general de:

$$L = \frac{n!}{(n-r)!}\left\{\prod_{i=1}^{n} f\left(t_{(i)}\right)\right\} s\left(t_{(r)}\right)^{n-r} \qquad (2.3.3.1)$$

La función de Verosimilitud para los datos con censura tipo II está basada en (2.3.3.1). Note que en términos de la notación (δ_i, t_i), se tiene que $\delta_i = 0$ y $t_{(i)} = t_{(r)}$ para aquellos individuos cuyo tiempo de vida está censurado, la ecuación (2.3.3.1) da la función de Verosimilitud de la forma (2.3.1.2) como censura tipo I.

EJEMPLO 2.7. Considere la distribución exponencial como en el ejemplo 2.6, pero suponga que ahora los tiempos de vida tienen censura tipo II.

La Log-Verosimilitud sigue la forma del ejemplo 2.6.

$$\ell(\lambda) = rln(\lambda) - \lambda \sum_{i=1}^{n} t_i,$$

por lo que se puede escribir como sigue:

$$\ell(\lambda) = rln(\lambda) - \lambda[W].$$

Así, el estimador de Máxima Verosimilitud para el estimador de λ es $\hat{\lambda} = \frac{r}{W}$, donde

$$W = \left(\sum_{i=1}^{n} t_i \right) + (n - r)t_{(r)}$$

Puesto que r es fija, la estadística w es suficiente para λ. Se puede mostrar que con los datos considerados como variables aleatorias, $2\lambda w = \frac{2r\lambda}{\hat{\lambda}} \sim \chi^2_{(2r)}$ tiene una distribución ji-cuadrada con $2r$ grados de libertad.

2.3.4. Máxima Verosimilitud en datos censurados por la derecha

Hasta ahora no se han hecho suposiciones sobre los mecanismos de censura, sin embargo es necesario. Los supuestos que se han convertido en estándar en el análisis de datos de tiempo de vida requieren que:

$$P(d\boldsymbol{N}(t)|\mathcal{H}(t)) = \prod_{i=1}^{n} h_i(t)^{dN_i(t)}[1 - h_i(t)]^{Y_i(t)\left(1-dN_i(t)\right)} \qquad (2.3.4.1)$$

Efectivamente, se requiere la probabilidad dado $\mathcal{H}(t)$ y el valor de la covariable, el mecanismo de falla para los individuos como riesgo al tiempo t que opera independientemente, para $t = 0, 1, 2, \dots, n$.

$$P(dN(t) = 1|\mathcal{H}(t)) = Y_i(t)h_i(t). \qquad (2.3.4.2)$$

En la ecuación (2.3.4.1) el efecto de que si $Y_i(t) = 0$ es que, no hay información acerca del individuo i en el tiempo t, y el término en la verosimilitud es igual a uno. Note que el valor de $Y_i(t)$ es determinado por $\mathcal{H}(t)$.

La condición (2.3.4.2) representa una independencia condicional (en $\mathcal{H}(t)$ y en los valores de la covariable) entre la falla y la censura en el tiempo t.

En ecuación (2.3.4.2), la probabilidad de que un individuo que está vivo, no censurado, sólo antes del tiempo t es observado que falla en el tiempo t es $h_i(t)$.

Si el término $P(d\boldsymbol{C}(t)|d\boldsymbol{N}(t))$ en (2.2.1.1) no involucra ninguno de los parámetros que especifica $h_i(t)$, el plan de censura es llamado no informativo.

Estos términos pueden reducir la forma de Verosimilitud, usando (2.3.4.1) en (2.2.1.1) obteniendo:

$$\prod_{i=1}^{n}\prod_{t=0}^{+\infty} h_i(t)^{dN_i(t)}[1-h_i(t)]^{Y_i(t)\left(1-dN_i(t)\right)} \qquad (2.3.4.3)$$

Cada individuo es observado a su falla o a ser censurado en algún tiempo t. En este caso, la falla al tiempo t, $dN_i(t)=1$ y $Y_i(s)=I(s\leq t)$; en caso de censura en t, $dN_i(t)=0$ y $Y_i(s)=I(s\leq t)$. Entonces (ver la ecuación (1.2.2) y (1.2.4)).

$$s_i(t)=\prod_{i=1}^{n}\big(1-h_i(t)\big), \quad f_i(t)=h_i(t)s_i(t).$$

Por lo que la ecuación (2.3.4.3), en notación (t_i, δ_i) se puede escribir así:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f_i(t_i|\boldsymbol{\theta})^{\delta_i} s_i(t_i + 1|\boldsymbol{\theta})^{1-\delta_i} \qquad (2.3.4.4)$$

Entonces, $s_i(t_i + 1) = s_i(t^+)$, la Verosimilitud es exacta de forma (2.3.1.2) encontrada previamente para el tipo I y otras formas de censura.

2.3.5. Máxima Verosimilitud en datos censurados por intervalos

La función de verosimilitud para la muestra de *n* individuos independientes para datos censurados por intervalos esta dada por:

$$L = \prod_{i=1}^{n} [F_i(V_i) - F_i(U_i)] \qquad (2.3.5.1).$$

Donde $F_i(t)$ es la función de distribución de $\boldsymbol{T}_i$ asumiendo que $F_i(0) = 0$.

Así, los datos censurados por intervalos reflejan incertidumbre respecto al tiempo exacto en que las unidades fallaron dentro de un intervalo. Este tipo de datos frecuentemente viene de pruebas o situaciones donde los individuos de interés no son observados constantemente.

Como se puede apreciar el método de máxima verosimilitud es simple en su esencia, pero obviamente tiene todas las dificultades de la localización de máximos en una función.

2.4. Modelos de Regresión

En la mayoría de los estudios hay covariables o variables explicativas, tales como: tratamientos, indicadores de grupo, características individuales o condiciones ambientales, cuya relación con el tiempo de vida es de interés. Esto lleva a la consideración de los modelos de regresión. En el ejemplo 2.1 se describe una situación en la que se presentan covariables, el siguiente es un ejemplo adicional.

EJEMPLO 2.8. En un estudio realizado a 65 pacientes con problema de mieloma para determinar si el tiempo de supervivencia estaba relacionado con las variables explicativas fueron examinados en relación con 16 variables explicativas. Incluyendo medidas fisiológicas, tales como el recuento de glóbulos blancos de la persona en el momento del diagnóstico, los factores cualitativos, como la presencia o ausencia de infección al momento del diagnóstico y las características personales como el sexo y edad.

Las covariables pueden variar con el tiempo de tal forma que centraremos la atención en los métodos paramétricos, principalmente en covariables fijas (donde el tiempo no varía).

El análisis de regresión del tiempo de vida involucra las especificaciones para la distribución de un tiempo de vida ***T*,** dado un vector de covariables **x**. Por ejemplo, considerando la distribución de Weibull con *f.d.p* de la forma:

$$f(t) = \frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1} e^{\left[-\left(\frac{t}{\alpha}\right)^{\beta}\right]}. \qquad (2.4.1)$$

Con parámetro de escala α y parámetro de forma β, un modelo de regresión para el que α o β dependan de **x**, puede ser considerado. Como α y β son valores positivos, un par de especificaciones convenientes son como $\alpha(\mathbf{x}) = e^{(\delta'\mathbf{x})}$ y $\beta(\mathbf{x}) = e^{(\gamma'\mathbf{x})}$, donde δ y γ son vectores de coeficientes de la regresión de la misma longitud que **x**, en este caso, $\alpha(\mathbf{x}) > 0$ y $\beta(\mathbf{x}) > 0$, sin restricciones para δ ó γ.

Un modelo Weibull que en muchas situaciones resulta útil, cuando sólo α está en función de **x**, por lo que la función de supervivencia de ***T*** es:

$$s(t|\mathbf{x}) = e^{\left[-\left(\frac{t}{\alpha(\mathbf{x})}\right)^{\beta}\right]}; \quad t \geq 0. \qquad (2.4.2)$$

El log-tiempo de vida $Y = logT$ en este caso; tiene función de supervivencia dada por:

$$s(y|\mathbf{x}) = e^{\left[-e^{\left(\frac{y-u(\mathbf{x})}{b}\right)}\right]}; \ -\infty < y < +\infty. \qquad (2.4.3)$$

Donde $u(\mathbf{x}) = ln\big(\alpha(\mathbf{x})\big)$ y $b = \beta^{-1}$. Note que es una Distribución Gumbel (o de valor extremo) de forma:

$$f(y) = \frac{1}{b}\mathrm{e}^{\left[\frac{(y-u)}{b}\right]}\mathrm{e}^{\left[-\mathrm{e}^{\left[\frac{(y-u)}{b}\right]}\right]};\ -\infty < y < +\infty,\ \text{ con } u = u(\mathbf{x}).$$

En términos de ***T***, el modelo (2.4.2) se conoce como un modelo Log-Loc-Escala o modelo de tiempo de falla acelerado, éste es uno de los más utilizados en el tipo de regresión paramétrica.

2.4.1. Modelos de Regresión Log-Loc-Escala (Tiempo de falla acelerado)

Los modelos de regresión Loc-Escala que consideran la función de supervivencia de y dado **x** toman una forma similar a la ecuación siguiente:

$$s(y|\mathbf{x}) = s_0\left(\frac{y-u(\mathbf{x})}{b}\right);\ -\infty < y < +\infty. \qquad (2.4.1.1)$$

Donde, $s_0(z)$ es independiente de **x.** Otra manera de expresarlo es:

$$y = u(\mathbf{x}) + bz. \qquad (2.4.1.2)$$

Donde Z es una variable aleatoria con función de supervivencia $s_0(z)$.

La familia de modelos para los que Z tiene una distribución normal estándar es una base del análisis de regresión, es decir, con datos de tiempo de vida el uso de las Distribuciones Gumbel (valor extremo), Logística y otras distribuciones para Z son comunes.

La función de supervivencia para t dado $\mathbf{x}$ correspondiente a la ecuación (2.4.1.1) es de la forma:

$$s(t|\mathbf{x}) = s_0^*\left[\left(\frac{t}{\alpha(\mathbf{x})}\right)^{\beta}\right]; \quad t \geq 0. \tag{2.4.1.3}$$

Donde $\alpha(\mathbf{x}) = e^{[u(\mathbf{x})]}$, $\delta = \beta^{-1}$ y $s_0^*(t) = s_0\big(ln(t)\big)$.

Las covariables efectivamente modifican la escala de tiempo y la ecuación (2.4.1.3) es referida como un modelo de tiempo de falla acelerado. En particular, si $\alpha(\mathbf{x}) > 1$ el efecto del vector de covariables es reducir el tiempo de aceleración y si $\alpha(\mathbf{x}) < 1$ el efecto es acelerar a éste.

Las Figuras 2.4 y 2.5 muestran el efecto de las covariables en las funciones de densidad de probabilidad y de supervivencia del Log-Escala de vida "y". Diferentes vectores de covariables $\mathbf{x}_1$ y $\mathbf{x}_2$ dan funciones que son transformaciones de ellas, tienen la misma forma pero están separadas por una distancia $u(\mathbf{x}_1) - u(\mathbf{x}_2)$. Tales modelos son específicamente utilizados cuando los tiempos de vida para individuos diferentes pueden variar por órdenes de importancia por ejemplo las fallas de un fluido aislante eléctrico.

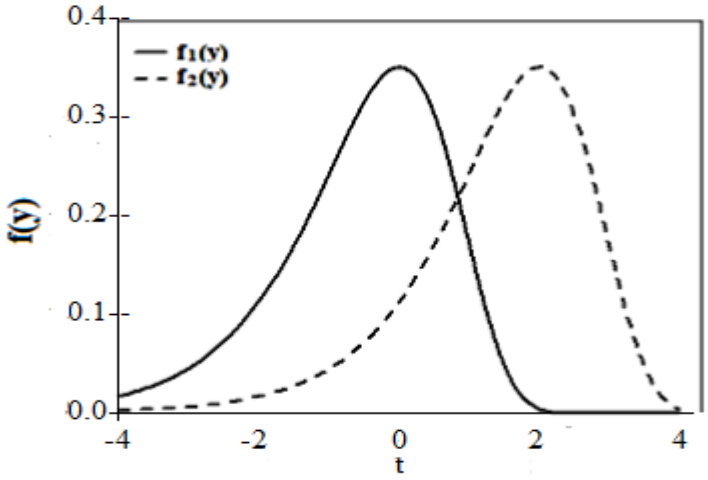


Figura 2.4. Funciones de Densidad para Modelos de Regresión Loc-Escala

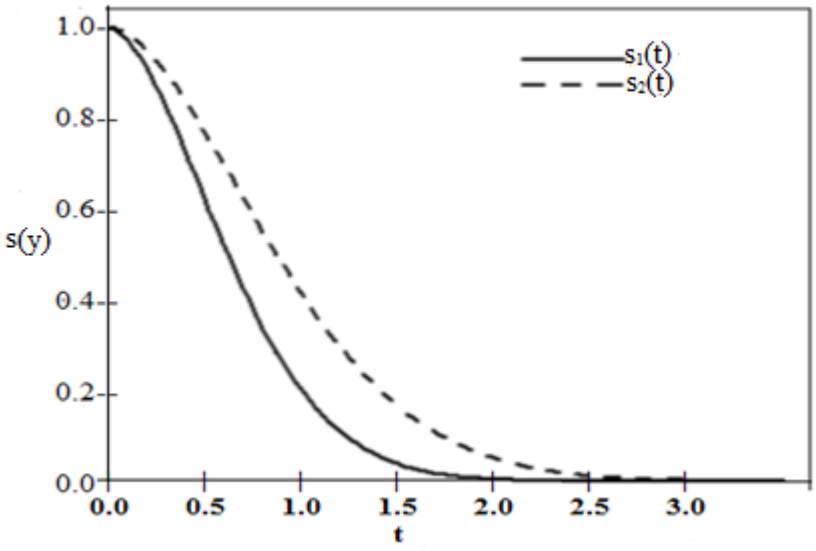


Figura 2.5. Funciones de Supervivencia para Modelos de Regresión Loc-Escala

Muchos modelos en ingeniería cuando la falla es acelerada por la corriente eléctrica, voltaje u otras tensiones son de este tipo y tienen especificaciones lineales $u(\mathbf{x}) = \delta'\mathbf{x}$. Por ejemplo, el modelo con $u(\mathbf{x}) = \delta_0 + \delta_1\mathbf{x}$, y $\mathbf{x} = ln(tension)$ es referenciado como el

modelo de la ley de potencia inversa. Entonces, δ_1 es generalmente negativo, se usa a menudo con tensiones de alto voltaje. Para temperatura como factor de tensión el modelo de Arthenius con $u(\mathbf{x}) = \delta_0 + \delta_1 \mathbf{x}$ y $\mathbf{x} = d^{-1}$ es usado a menudo, donde d es la temperatura en grados Kelvin.

El tiempo de falla acelerado o el Modelo Log-Loc-Escala también son útiles en otros campos de aplicación y dominan en muchas áreas del análisis de regresión.

2.4.2. Modelos de Regresión de Riesgo Proporcional

Hay dos enfoques principales para el tiempo de vida en los modelos de regresión. Uno es usar transformaciones del tiempo, asumiendo que el efecto de covariables es equivalente a alterar el tiempo pasado, solo se hablará de este tipo para el modelo de tiempo de falla acelerado. El segundo enfoque asume las especificaciones en la dirección en que las covariables afectan la función de riesgo para $\boldsymbol{T}$. El modelo más común de este tipo es el modelo de Riesgo Proporcional, para el que la Función de Riesgo para t dada $\mathbf{x}$, es de la forma:

$$h(t|\mathbf{x}) = h_0(t) r(\mathbf{x}). \qquad (2.4.2.1)$$

Donde $r(\mathbf{x})$ y $h_0(t)$ son funciones con valores positivos. La función $h_0(t)$ es usualmente llamada el punto de partida de la Función de Riesgo (o Función de Hazard), ésta es la función de riesgo para un individuo de quien el vector de covariables $\mathbf{x}$ es tal que

$r(\mathbf{x}) = 1$. Una especificación común para $r(\mathbf{x})$ es $e^{(\delta'\mathbf{x})}$, en este caso $h_0(t)$ es la Función de Riesgo cuando $\mathbf{x} = \mathbf{0}$. El nombre riesgo proporcional (RP) proviene de que cualquier individuo tiene Funciones de Riesgo que son múltiplos constantes de otras.

El Modelo RP (completamente paramétrico) especifica $h_0(t, \alpha)$ y $r(\mathbf{x}, \boldsymbol{\delta})$ en (2.4.2.1), de éste se desprende la siguiente relación:

$$s(s|t) = e^{[-H(t|\mathbf{x})]} = e^{\left[-\int_0^t h(u|\mathbf{x})du\right]}.$$

entonces la Función de Supervivencia para t dado $\mathbf{x}$ es:

$$s(t|\mathbf{x}) = s_0(t)^{r(\mathbf{x})} \qquad (2.4.2.2)$$

Donde $s_0 = e^{[-H_0(t)]}$ es la base de la Función de Supervivencia. La Figura 2.6 muestra las Funciones de Riesgo y la Figura 2.7 la correspondiente Función de Supervivencia para dos vectores diferentes de covariables $\mathbf{x}_1$ y $\mathbf{x}_2$. Note que una de las Funciones de Supervivencia debe estar completamente por encima de la otra, por la ecuación (2.4.2.2).

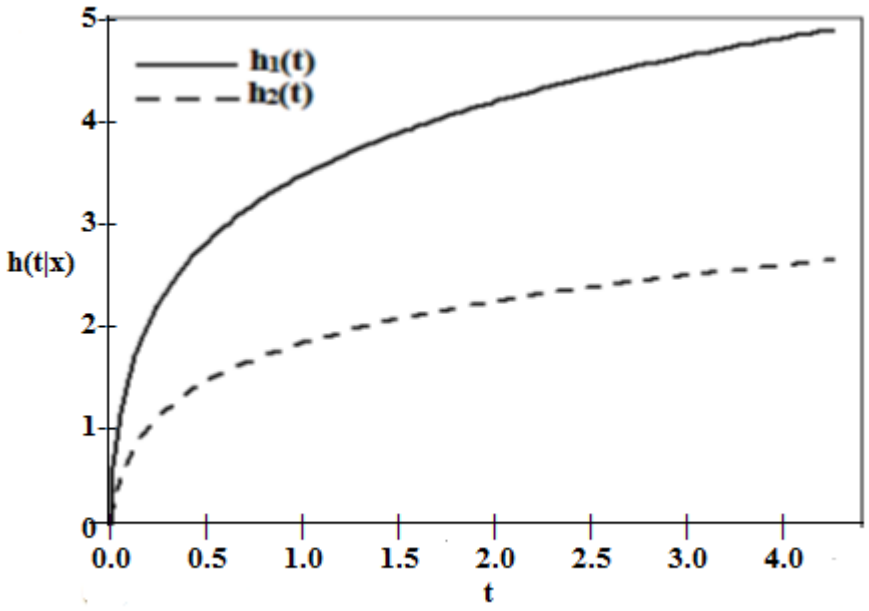


Figura 2.6. Funciones de Riesgo para Modelos de Regresión de Riesgo Proporcional

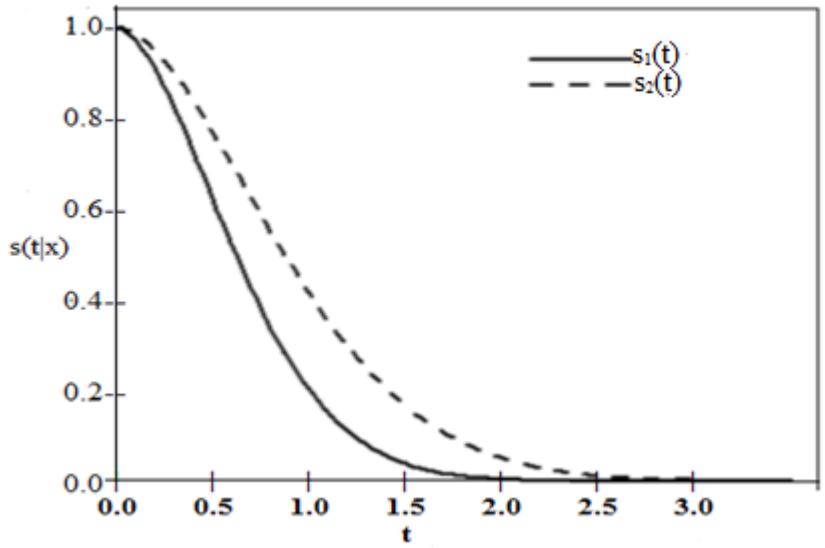


Figura 2.7. Funciones de Supervivencia para Modelos de Regresión de Riesgo Proporcional

Una característica de los modelos RP es que si $s_0(t;\ \alpha)$ está en alguna familia de los modelos paramétricos, entonces $s(t|\mathrm{x})$ no es de la misma familia, aunque si $h_0(t)$ es de la forma $\alpha_1 h_1(t;\ \alpha_2)$ para los parámetros α_1 y α_2. Esto está en contraste con la situación para modelos de tiempo de falla acelerado, y quizá una de las razones del porqué los modelos RP paramétricos se utilizan menos que los semiparamétricos donde $h_0(t)$ en la ecuación (2.4.2.1) es arbitraria.

Un modelo paramétrico usado frecuentemente es el modelo RP de la familia Weibull, se puede verificar que de la función (2.4.2) es un modelo RP con función de riesgo:

$$h(t|\mathrm{x}) = \frac{\beta}{\alpha(\mathrm{x})}\left[\frac{t}{\alpha(\mathrm{x})}\right]^{\beta-1} = \left(\beta t^{\beta-1}\right)\alpha(\mathrm{x})^{-\beta}. \qquad (2.4.2.3)$$

2.5. Modelo de Regresión de Cox

En este contexto, se introduce el método de Verosimilitud Parcial que permite realizar la estimación de los parámetros del Modelo de Regresión de Cox. Se distinguen las tres situaciones posible según el tipo de datos: completos, incompletos y empates.

En los estudios longitudinales que tienen por objetivo el estudio de cambio se recomienda (Tuma y Hannan, 1984, p.22), para la recogida sistemática de la información pertinente, la utilización del plan de observación denominado Event-History Data. La ventaja de este diseño frente al clásico diseño de panel, es el que permite la máxima información de

cada cambio producido, recogida por medio de las secuencias de cambio y del momento temporal en que éste se produce, y permite utilizar asimismo la información de las observaciones en las que no se ha producido ningún cambio (datos incompletos).

En esta situación, el modelo de riesgo proporcional es el modelo más utilizado para representar los efectos de un conjunto de variables explicativas (variables independiente) sobre las variables tiempo de cambio (tiempo de supervivencia), o más bien sobre la probabilidad condicional de cambio, es decir sobre la función de riesgo $h(t)$. Suponemos que para cada sujeto tenemos un vector **x** de variables explicativas. Las componentes de dicho vector pueden representar tratamientos, definidos por medio de variables indicadoras (dummy variables), propiedades intrínsecas de los sujetos, tales como, por ejemplo, la edad, el sexo, características individuales, agrupaciones cualitativas de los sujetos, o bien variables exógenas, como pueden ser las propiedades ambientales del problema.

El Modelo de Regresión de Cox (1972) viene dado por la relación:

$$h(t|\mathrm{x}) = h_0(t)\mathrm{e}^{(x'\beta)} \qquad (2.5.1)$$

Donde la dependencia temporal está incluida en la tasa de riesgo de línea base, $h_0(t)$, y las variables pronostico actúan en forma Log-Lineal, $e^{(x'\beta)}$, donde β es un vector de coeficientes de regresión desconocidos que parametrizan el modelo.

Este modelo puede describirse (Allison, 1984) como semiparamétrico o parcialmente paramétrico. Es paramétrico ya que especifica un modelo de regresión con una forma funcional específica; es no paramétrico en cuanto que no especifica la forma exacta de la distribución de los tiempos de supervivencia.

En este modelo las variables concomitantes actúan sobre la función de riesgo de forma multiplicativa. Las variables explicativas además pueden ser dependientes o independientes del tiempo.

El Modelo de Cox puede utilizarse en los siguientes casos:

- Cuando no se tiene información previa acerca de la dirección temporal de la Función de Riesgo.

- Cuando siendo conocida la dirección, no puede ser determinada por un Modelo Paramétrico.

- Cuando se está únicamente interesado en la magnitud de los efectos de las variables concomitantes, teniendo controlada la dirección temporal.

Debido a las existencias de datos incompletos, los parámetros del modelo de Cox, no pueden ser estimados por el Método Ordinario de Máxima Verosimilitud al ser desconocida la forma específica de la función arbitraria de riesgo, lo cual no se encuentra habitualmente en los textos y pensamos que es un ejercicio muy recomendable ya que permite tener una idea más clara del funcionamiento de dicho proceso.

Cox (1975) propuso un método de estimación denominado Verosimilitud Parcial, siendo las verosimilitudes condicionales y marginales casos particulares del anterior.

El Método de Verosimilitud Parcial se diferencia del Método de Verosimilitud Ordinario en el sentido de que mientras el Método Ordinario se basa en el producto de las verosimilitudes para todos los individuos de la muestra, el método parcial se basa en el producto de las verosimilitudes de todos los cambios ocurridos.

Para estimar los coeficientes β en el Modelo de Cox, en ausencia de conocimiento de $h_0(t)$, éste propuso la siguiente Función de Verosimilitud:

$$L(\beta) = \prod \left[\frac{e^{(x'\beta)}}{\sum e^{(x\beta)}} \right] \qquad (2.5.2)$$

Donde $L(\beta)$ no es una verdadera Función de Verosimilitud ya que no puede derivarse como la probabilidad de algún resultado observado bajo el modelo como indica Cox (1975), tratarse como una Función de Verosimilitud ordinaria a efectos de realizar estimaciones de β.

Dichas estimaciones son consistentes (Cox, 1975; Tsiatis, 1981) y eficientes (Efron, 1977).

Cuando un mismo instante t_i se produce más de un cambio, lo cual puede ocurrir cuando la variable tiempo se mide de forma discreta, la probabilidad de ocurrencia de los d_i cambios observados, condicionados al conjunto de riesgo R_i viene dado (Cox, 1972) por la siguiente ecuación:

$$L(\beta) = \frac{e^{(z'\beta)}}{\sum (Z'\beta)}$$

Donde cada elemento z_i del vector Z es la suma de los valores x_i sobre los d_i individuos que realizan un cambio en el instante t_i y la suma de los denominadores se efectúa sobre R_i sujetos expuestos al riesgo en t_i.

El logaritmo de la Función de Verosimilitud Parcial viene dada, entonces por:

$$ln\big(L(\beta)\big) = \sum_{i=1}^{k}(Z_i\beta) - \sum_{i=1}^{k}\Big[d_i ln(\sum(e^{\beta' X}))\Big] \qquad (2.5.3)$$

Las estimaciones Máximo Verosímiles de β son estimaciones que maximizan la función $ln\big(L(\beta)\big)$.

CAPITULO III. APLICACIONES

INTRODUCCIÓN

En este capítulo se presentan e ilustran algunas aplicaciones, haciendo uso de los métodos estadísticos y el análisis de datos del tiempo en semanas que una persona que depende de las drogas ha estado incluida en el programa "Fe y Esperanza", y el tiempo en horas de vida útil de veinte focos, en la Granja de Pollos de Engorde "Martínez". Con esta información se aplicará la metodología de algunas de las distribuciones de supervivencia vistas en los capítulos anteriores, para estimar sus respectivas funciones de supervivencia de estos datos.

Por otro lado, en los procedimientos de inferencia estadística son usadas las funciones de Verosimilitud basadas en los datos observados. En la primera aplicación se establecerá la forma de Verosimilitud Parcial (*VP*) y en la segunda aplicación las condiciones asociadas con la selección y observación de los individuos y los focos del estudio.

De manera que en las aplicaciones, se hará uso del método de Máxima Verosimilitud para la estimación de parámetros, entre otros cálculos estadísticos.

En la primera aplicación se realizará de forma numérica y exhaustiva, todos los pasos presentes en el proceso de estimación de parámetros, su significación y su interpretación. Para estudiar el proceso de estimación de parámetros en el Modelo de Regresión de Cox tomaremos los datos proporcionados por el Centro de Rehabilitación "Fe y Esperanza",

durante los meses de Enero a Octubre del año 2013, sobre un programa de tratamiento libre de drogas. El tratamiento de drogodependencia es un proceso continuo formado por diferentes fases y programas. Se empieza por la preparación al tratamiento y diagnóstico, seguido por la desintoxicación. Luego se procede a la deshabituación, tras el cual se lleva a cabo el Programa de reinserción socio-laboral y finalmente se realiza un Programa de seguimiento.

En la segunda aplicación, como parte de la investigación, la cual consiste en la observación de la vida útil de 20 focos, en la Granja de Pollos de Engorde "Martínez", durante los meses de Julio a Octubre del año 2013. Se observaron en el periodo establecido el tiempo de vida para determinar la durabilidad de los mismos, el tiempo de falla de estos componentes fue medido en horas, hasta su falla final, por lo cual los datos no presentaron ningún tipo de censura debido a que se les conoció el tiempo exacto de vida útil y no parcial. Por lo tanto hablaremos de datos no-censurados.

En esta aplicación se presenta de forma ilustrativa la estimación de la función de supervivencia, la tasa de fallo y el valor esperado de la Distribución Exponencial, mediante el método de Máxima Verosimilitud, aplicando la teoría antes expuesta en los capítulos anteriores. Así también, el ajuste del Modelo Exponencial, por medio de la estimación de la Función de Distribución Corregida de los datos, y la representación grafica tal que si el modelo elegido es correcto los datos presentan aspecto lineal.

La decisión de usar este Modelo, esta sustentada por las caracteristicas de los datos obtenidos, donde se observó que el tiempo de falla de los focos, no dependía de lo que había

sucedido anteriormente. Cumpliendo así con una de las propiedades de la Distribución Exponencial.

3.1. Primera aplicación

En la primera aplicación la metodología del Modelo de Regresión de Cox es referido en situaciones, en donde se presentan dos o más variables, necesarias para el análisis de interés. Para mayor comodidad, las variables en estudio se presentan en una matriz de datos, y posteriormente en esquemas gráficos, que permiten visualizar el instante de tiempo t_i en que se han producidos los cambios.

En esta aplicación se presenta e ilustra, el modelo estadístico propuesto, Modelo de Regresión de Cox. Tomando los datos proporcionados por el centro de rehabilitación como base, sobre el programa de tratamiento libre de drogas con dicha información, se calcula la función de Verosimilitud Parcial (*VP*) y la función de densidad de probabilidad en un instante t_i, relacionada con la función de riesgo, vista en la ecuación (1.1.2).

Para poder analizar, los tiempos de cambio, la tasa de riesgo, es necesario: el proceso de estimación de parámetros, la estimación del coeficiente β (Método de Newton-Raphson), calculo de valor β^* en cada iteración, cálculos finales para el coeficiente obtenido e interpretación del coeficiente del Modelo de Regresión de Cox.

3.1.1. Método propuesto Modelo de Regresión de Cox

En los estudios longitudinales para el análisis del cambio, el Modelo de Riesgo Proporcional permite estudiar el efecto de un conjunto de variables explicativas sobre la función de riesgo. En este contexto, se introduce el Método de Verosimilitud Parcial (*MVP*) que permite realizar la estimación de los parámetros del Modelo de Regresión de Cox. Se distinguen las tres situaciones posibles según los tipos de datos: completos, incompletos y empates. En esta aplicación se permite seguir, de forma numérica y exhaustiva, todos los pasos presentes en el proceso de estimación de parámetros, su significación y su interpretación.

Para estudiar el proceso de estimación de parámetros en el Modelo de Regresión de Cox tomaremos los datos proporcionados por el Centro de Rehabilitación "Fe y Esperanza", durante los meses de Enero a Octubre del año 2013, sobre el programa de tratamiento libre de drogas, mediante un enfoque bio-psico-social.

El programa de desintoxicación, llevado a cabo desde el marco ambulatorio, puede realizarse con medicación o bien por medio de un soporte psico-social del entorno.

La tabla 3.1 presenta una matriz de datos que incluye las variables necesarias para el análisis.

Las variables son:

1- **TIEMPO**: El Tiempo en semanas de la persona que depende de las drogas ha estado incluido en el programa.

2- **ESTADO:** Define la situación del sujeto en su última observación (0 = Sano, 1 = Recaída), obtenida por la determinación de drogas en la orina.

3- **GRUPO:** Diferencia a los sujetos según el tipo de desintoxicación observada. (0 = Psico: Con soporte psico-social, 1 = Forma: con medicación).

4- **EDAD:** La edad del sujeto esta medida en años.

CASO	TIEMPO	ESTADO	GRUPO	EDAD (años)
1	3	1	1	45
2	5	0	1	20
3	5	0	1	39
4	8	1	1	51
5	8	1	0	30
6	13	0	0	35
7	16	1	1	44
8	19	0	1	28
9	20	1	0	35
10	20	0	0	35
11	21	1	0	38
12	25	0	0	24

Tabla 3.1 Matriz de Datos

Para mayor comodidad, los sujetos se han ordenado en función de la variable tiempo de inclusión en el programa.

La figura 3.1 muestra el esquema gráfico de los datos (0=incompleto, x=cambio) que permite visualizar el momento en que se han producido los cambios, así como saber, en cada instante t_i, el conjunto de riesgo R_i.

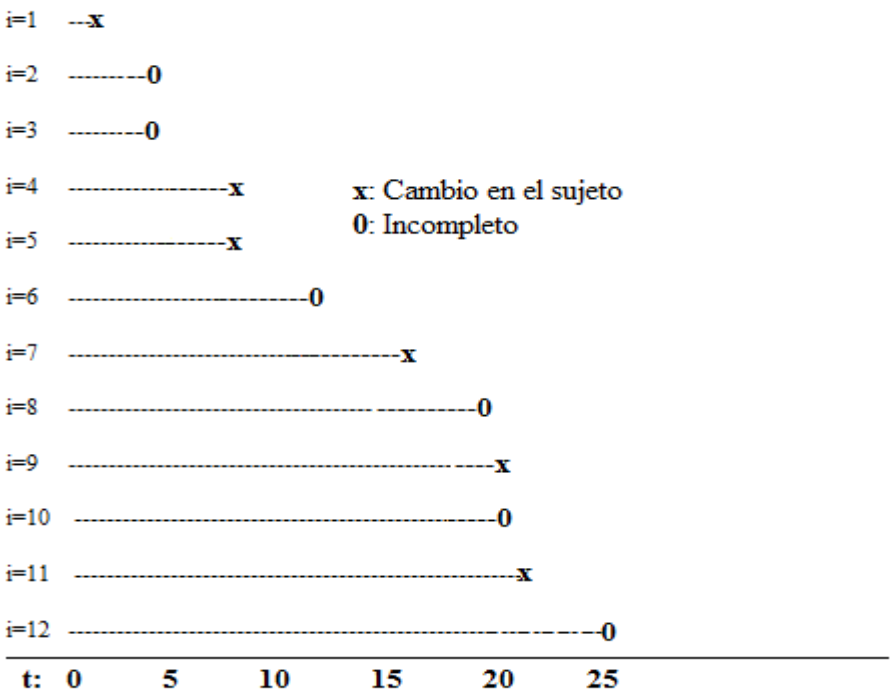


Figura 3.1. Esquema grafico de los datos

Utilizando los datos de la matriz de datos (tabla 3.1) podemos construir la tabla 3.2

t_i	i	k	$R(t_i)$	Sujeto que cambia	Sujetos del conjunto de riesgo
3	(1)	1	12	(1)	(1), (2), (3), (4), (5), (6), (7), (8), (9), (10), (11), (12)
5*	(2) (3)				
8	(4) (5)	3	9	(4) (5)	(4), (5), (6), (7), (8), (9), (10), (11), (12)
13*	(6)				
16	(7)	4	6	(7)	(7), (8), (9), (10), (11), (12)
19*	(8)				
20	(9)	5	4	(9)	(9), (10), (11), (12)
20*	(10)				
21	(11)	6	2	(11)	(11), (12)
25*	(12)				

Tabla 3.2. Sujetos expuestos al riesgo y cambio en los datos.

En la tabla 3.2 se han ordenado los tiempos observados en la muestra en forma ascendente, señalando con un asterisco aquellos tiempos incompletos. Así pues, t_i representa el tiempo observado para el sujeto (i). En la tercera columna los valores de k representan los cambios ocurridos hasta el tiempo t_i inclusive. La cuarta columna $R(t_i)$ representa el número de sujetos expuestos al riesgo de un cambio en cada tiempo t_i. La quinta columna identifica a los sujetos que han realizado un cambio en el instante t_i y la sexta columna enumera a los sujetos expuestos al riesgo en cada instante en el que ocurre un cambio.

Así pues, con la información proporcionada por la tabla 3.2. Se puede calcular la función de Verosimilitud Parcial como un primer paso, que viene dada en función de los parámetros desconocidos y de los datos observados y el segundo paso es hallar el máximo de dicha función, el cual vendrá determinado por un conjunto concreto de valores de los parámetros.

La función de Verosimilitud Parcial (*VP*) viene dada por el producto de las siguientes probabilidades condicionales:

$$VP = P\big(i = k | R(t_i)\big) = \left[\frac{i = k}{R(t_i)}\right] \qquad (3.1.1.1)$$

Aplicando a cada uno de los k cambios observados en la ecuación (3.1.1.1) en la muestra estudiada. Es decir:

$$VP = P\left[\frac{i = 1}{R = 12}\right] P\left[\frac{i = 3}{R = 9}\right] P\left[\frac{i = 4}{R = 6}\right] P\left[\frac{i = 5}{R = 4}\right] P\left[\frac{i = 6}{R = 2}\right]$$

Partiendo de $h(t) = \frac{f(t)}{s(t)}$, se tiene la función de densidad en un instante t_i:

$$f(t) = h(t)s(t) \qquad (3.1.1.2)$$

Donde $f(t)$ representa la probabilidad de que ocurra el cambio en el instante $t\ y\ s(t)$ representa la probabilidad de que el cambio se produzca pasado el tiempo t.

Cada uno de los términos de la expresión anterior que permite obtener la función de Verosimilitud Parcial, *VP*, puede expresarse en términos de riesgo.

Para comprobarlo procedemos a calcular uno de los elementos de la siguiente ecuación $VP = P\big(i = k|R(t_i)\big) = \left[\frac{i=k}{R(t_i)}\right]$

Por ejemplo, el valor de $P\left[\frac{i=5}{R=4}\right]$.

En el instante $t = 20$, correspondiente a $k = 5$, hay cuatro elementos expuestos al riesgo de un cambio: $R(t = 20) = (9), (10), (11), (12)$.

La probabilidad que el cambio le ocurra a $i = (9)$ en lugar de $i = (10)$ ó $i = (11)$ ó $i = (12)$ es:

$$f_9(20)s_{10}(20)s_{11}(20)s_{12}(20) = h_9(20)s_9(20)s_{10}(20)s_{11}(20)s_{12}(20) \qquad (1)$$

Igualmente la probabilidad de que el cambio le ocurriera a $i = (10)$ viene dada por:

$$s_9(20)f_{10}(20)s_{11}(20)s_{12}(20) = s_9(20)h_{10}(20)s_{10}(20)s_{11}(20)s_{12}(20) \quad (2)$$

Y que la ocurriera al sujeto $i = (11)$:

$$s_9(20)s_{10}(20)f_{11}(20)s_{12}(20) = s_9(20)s_{10}(20)h_{12}(20)s_{11}(20)s_{12}(20) \quad (3)$$

y que ocurrirá $i = (12)$

$$s_9(20)s_{10}(20)s_{11}(20)f_{12}(20) = s_9(20)s_{10}(20)s_{11}(20)h_{12}(20)s_{12}(20) \quad (4)$$

Así pues:

$$VP = P\left[\frac{i=5}{R=4}\right] = \frac{f_9 s_{10} s_{11} s_{12}}{f_9 s_{10} s_{11} s_{12} + s_9 f_{10} s_{11} s_{12} + s_9 s_{10} s f_{11} s_{12} + s_9 s_{10} s_{11} f_{12}}$$

Sustituyendo cada término por las igualdades anteriores se simplifica $s_9(20)s_{10}(20)s_{11}(20)s_{12}(20)$ del numerador y del denominador:

$$VP = \frac{h_9(20)s_9(20)s_{10}(20)s_{11}(20)s_{12}(20)}{h_9 s_9 s_{10} s_{11} s_{12} + h_{10} s_9 s_{10} s_{11} s_{12} + h_{11} s_9 s_{10} s_{11} s_{12} + h_{12} s_9 s_{10} s_{11} s_{12}}$$

$$VP = \frac{(s_9 s_{10} s_{11} s_{12}) h_9}{(s_9 s_{10} s_{11} s_{12})(h_9 + h_{10} + h_{11} + h_{12})}$$

Quedando finalmente:

$$VP = \frac{h_9(20)}{h_9(20) + h_{10}(20) + h_{11}(20) + h_{12}(20)} = \frac{e^{(X'\beta)}}{\sum e^{(X'\beta)}}$$

La expresión anterior que depende únicamente de los valores de las variables vistas en la ecuación expresada $h(t|\mathrm{x}) = h_0(t)e^{(\mathrm{x}'\beta)}$ en términos de exponenciales.

Cada uno de los términos de la ecuación (3.1.1.1) de *VP* viene determinado siguiendo este mismo proceso, de manera que la expresión general de la Función de Verosimilitud Parcial será el producto de cada una de las expresiones halladas.

Podemos distinguir, según el Modelo de Regresión Cox, tres situaciones distintas en la obtención de la función de Verosimilitud Parcial:

A continuación se presentan las tres situaciones posibles según los tipos de datos:

Caso I. Datos completos sin empates: son aquellos que provienen de individuos o componentes de los que se les conoce con exactitud su tiempo de supervivencia, sin presentar empate con otros individuo o componentes sobre el suceso estudiado.

Este caso es realista en cuanto la variable tiempo de supervivencia tenga una distribución continua y su valor se registre de forma exacta, desapareciendo la probabilidad de empates.

Sean $t_1 < t_2 < \cdots < t_n$ los n tiempos ordenados correspondientes a los n sujetos del estudio, sea $R(t_i) = [i: t_i \geq t_j]$ el conjunto de riesgo justo antes del tiempo t_j y sea r_j el tamaño de dicho conjunto. La figura 3.2 ilustra el esquema de datos sin empates.

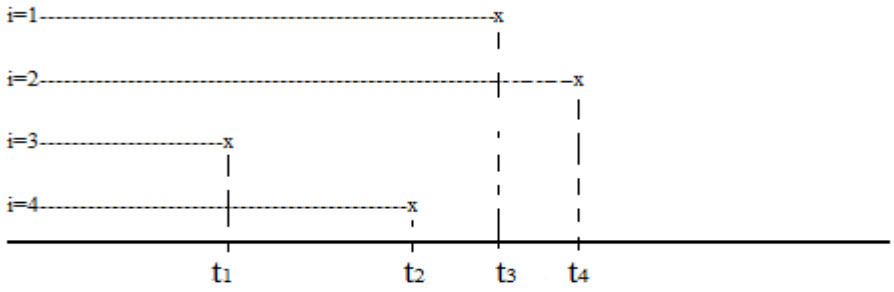


Figura 3.2. Esquema de datos completos sin empates.

En la figura 3.2 podemos ver los tiempos de supervivencia de 4 sujetos: $t_1 < t_2 < t_3 < t_4$ siendo $R(t_1) = [1,2,3,4],\ R(t_2) = [1,2,4],\ R(t_3) = [1,2]$ y $R(t_4) = [2]$. Los conjuntos de riesgo en cada tiempo.

La probabilidad $P(3,\ 4,1,2)$ de obtener la configuración conjunta de la figura 3.2 puede obtenerse por medio de la regla de la cadena para probabilidades condicionales:

$$P(3,4,1,2) = P\left(\frac{3}{1,2,3,4}\right) P\left(\frac{4}{1,2,4}\right) P\left(\frac{1}{1,2}\right) P\left(\frac{2}{2}\right)$$

En la Función de Verosimilitud Parcial cada uno de estos elementos puede expresarse en términos de exponenciales. Así pues:

$$P\left(\frac{3}{1,2,3,4}\right) = \frac{\mathrm{e}^{(3)}}{\mathrm{e}^{(1)} + \mathrm{e}^{(2)} + \mathrm{e}^{(3)} + \mathrm{e}^{(4)}}$$

$$P\left(\frac{4}{1,2,4}\right) = \frac{\mathrm{e}^{(4)}}{\mathrm{e}^{(1)} + \mathrm{e}^{(2)} + \mathrm{e}^{(4)}}$$

$$P\left(\frac{1}{1,2}\right) = \frac{\mathrm{e}^{(1)}}{\mathrm{e}^{(1)} + \mathrm{e}^{(2)}} \qquad y \qquad P\left(\frac{2}{2}\right) = \frac{\mathrm{e}^{(2)}}{\mathrm{e}^{(2)}}$$

De manera que:

$$VP = P\big(i = k | R(t_i)\big) = \left(\frac{i = k}{R(t_i)}\right)$$

$$P(3,4,1,2) = \prod_{i=1}^{n} P\left(\frac{i}{R(t_i)}\right) = \prod_{i=1}^{n} \frac{e^{(i)}}{\sum e^{(R)}}$$

Donde R representa a cada uno de los sujetos expuestos al riesgo en cada tiempo t_i.

Caso II. Datos Incompletos: Datos que contienen información parcial sin empate sobre el suceso estudiado, que se caracteriza por un dato incompleto.

En este caso supongamos que tenemos "**d**" desenlaces (cambios, sucesos) observados en la muestra de tamaño n y ordenados los tiempos $t_1 < t_2 < \cdots < t_d$. La figura 3.3 ilustra los datos de este caso:

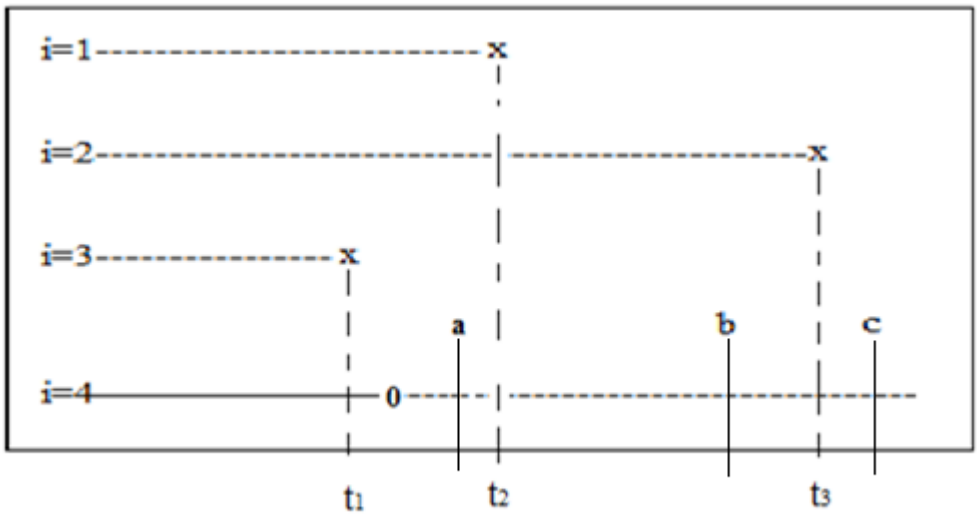


Figura 3.3. Casos de Datos Incompletos

En la figura 3.3 podemos ver los tiempo de 4 sujetos con un dato incompleto y 3 datos completos: $t_1 < t_2 < t_3$ siendo $R(t_1) = [1, 2, 3, 4]$, $R(t_2) = [1, 2\,]$, $R(t_3) = [2]$ el conjunto de riesgo en cada tiempo y a, b, c son las posibles posiciones que puede tomar el tiempo de la observación incompleta.

La probabilidad de obtener la combinación de desenlaces observadas en la figura 3.3, siguiendo el esquema anterior, viene dada por el producto de probabilidades condicionadas al conjunto de riesgo que, como hemos visto anteriormente, puede expresarse en términos de exponenciales:

$$P(3,1,2) = \prod_{i=1}^{d} P\left(\frac{i}{R(t_i)}\right) = \prod_{i=1}^{d} \frac{e^{(i)}}{\sum[e^{(R)}]}$$

De manera que:

$$P(3,1,2) = P\left(\frac{3}{1,2,3,4}\right) P\left(\frac{1}{1,2}\right) P\left(\frac{2}{2}\right)$$

$$P(3,1,2) = \left(\frac{e^{(3)}}{e^{(1)} + e^{(2)} + e^{(3)} + e^{(4)}}\right)\left(\frac{e^{(1)}}{e^{(1)} + e^{(2)}}\right)\left(\frac{e^{(2)}}{e^{(2)}}\right)$$

Esta probabilidad puede obtenerse así mismo como suma de todas las probabilidades condicionales consistente con el patrón de desenlaces y datos incompletos observados. Es decir:

$$P(3,1,2) = P(3,4,1,2) + P(3,1,4,2) + P(3,1,2,4)$$

El cálculo de cada sumando se realiza siguiendo el método descrito en el apartado anterior, a partir del cual se obtiene:

$$P(3,4,1,2) = \left(\frac{e^{(3)}}{e^{(1)} + e^{(2)} + e^{(3)} + e^{(4)}}\right)\left(\frac{e^{(4)}}{e^{(1)} + e^{(2)} + e^{(4)}}\right)\left(\frac{e^{(1)}}{e^{(1)} + e^{(2)}}\right)\left(\frac{e^{(2)}}{e^{(2)}}\right)$$

$$P(3,1,4,2) = \left(\frac{e^{(3)}}{e^{(1)} + e^{(2)} + e^{(3)} + e^{(4)}}\right)\left(\frac{e^{(1)}}{e^{(1)} + e^{(2)} + e^{(4)}}\right)\left(\frac{e^{(4)}}{e^{(2)} + e^{(4)}}\right)\left(\frac{e^{(2)}}{e^{(2)}}\right)$$

$$P(3,1,2,4) = \left(\frac{e^{(3)}}{e^{(1)} + e^{(2)} + e^{(3)} + e^{(4)}}\right)\left(\frac{e^{(1)}}{e^{(1)} + e^{(2)} + e^{(4)}}\right)\left(\frac{e^{(2)}}{e^{(2)} + e^{(4)}}\right)\left(\frac{e^{(4)}}{e^{(4)}}\right)$$

Caso III. Empates: Datos que se caracterizan por tener dos cambios o doble información incompleta o parcial sobre el suceso estudiado.

La Verosimilitud hallada en el apartado anterior, para datos incompletos, no es apropiada para distribuciones discretas del tiempo de supervivencia en las cuales podemos hallar empates en dichos tiempos.

El cálculo de Función de Verosimilitud puede obtenerse sumando todos los términos de la función de verosimilitud marginal que son consistentes con los datos observados.

Así pues si las observaciones 1 y 2 realizan un cambio en el tiempo t siendo el conjunto de riesgo las observaciones 1, 2, 3 y 4, la contribución en la Función de Verosimilitud seria:

$$\frac{d\mathbf{e}^{(1)}\ldots\mathbf{e}^{(d)}}{[\sum(\mathbf{e}^{(R)})]^{d}} = \left(\frac{\mathbf{e}^{(1)}}{\mathbf{e}^{(1)}+\mathbf{e}^{(2)}+\mathbf{e}^{(3)}+\mathbf{e}^{(4)}}\right)\left(\frac{\mathbf{e}^{(2)}}{\mathbf{e}^{(2)}+\mathbf{e}^{(3)}+\mathbf{e}^{(4)}}\right)$$

$$+\left(\frac{\mathbf{e}^{(2)}}{\mathbf{e}^{(1)}+\mathbf{e}^{(2)}+\mathbf{e}^{(3)}+\mathbf{e}^{(4)}}\right)\left(\frac{\mathbf{e}^{(1)}}{\mathbf{e}^{(1)}+\mathbf{e}^{(3)}+\mathbf{e}^{(4)}}\right)$$

Para simplificar los términos se aumentan todas las sumas de los denominadores para todas la R observaciones del conjunto de riesgo, lo cual permite escribir la expresión anterior como:

$$\frac{d!\,\mathbf{e}^{(1)}\ldots\mathbf{e}^{(d)}}{[\sum(\mathbf{e}^{(R)})]^{d}} = \frac{2\mathbf{e}^{(1)}\mathbf{e}^{(2)}}{[\mathbf{e}^{(1)}+\mathbf{e}^{(2)}+\mathbf{e}^{(3)}+\mathbf{e}^{(4)}]^{2}}$$

Esta expresión puede generalizarse para el caso de que existan **d** desenlaces en un instante t:

$$\frac{d!\,e^{(1)} \ldots e^{(d)}}{[\sum(e^{(R)})]^d}$$

3.1.2. Estimación de los parámetros

Para no complicar excesivamente los cálculos y las expresiones numéricas, y para efectos de una mayor simplicidad se realizará el análisis utilizando únicamente una variable explicativa: la edad del sujeto en años.

Así pues, lo que se pretende es averiguar cómo afecta la edad del individuo en el riesgo de que un adicto a las drogas tenga una recaída en su tratamiento de desintoxicación. A continuación vamos a exponer los distintos pasos que se deben realizar para obtener la estimación de los parámetros del modelo.

3.1.2.1. Cálculo del logaritmo de la Función de Verosimilitud Parcial

Sean $t_1 < t_2 < \cdots < t_k$ los k tiempos diferentes en los que ocurre algún cambio entre los *n* tiempo de supervivencia.

Las estimaciones Máximo-Verosímiles se obtienen maximizando la siguiente ecuación:

$$ln\big(L(\beta)\big) = \sum_{i=1}^{k}(Z_i\beta) - \sum_{i=1}^{k}\Big[d_i ln\Big(\sum\big(e^{(X'\beta)}\big)\Big)\Big]$$

Siendo Z_i la suma de los valores de la variable explicativa para todos los d_i individuos que realizan un cambio en el instante t_i.

Desarrollando la expresión general para cada uno de los 5 instantes en los que se produce algún cambio, se obtiene el Logaritmo de la Función de Verosimilitud, en función del parámetro que se desea estimar:

$$ln(L(\beta)) = 45\beta - 1\left(ln\left[e^{(45\beta)} + e^{(20\beta)} + \cdots + e^{(38\beta)} + e^{(24\beta)}\right]\right)$$

$$+81\beta - 2\left(ln\left[e^{(51\beta)} + e^{(30\beta)} + \cdots + e^{(38\beta)} + e^{(24\beta)}\right]\right)$$

$$+44\beta - 1\left(ln\left[e^{(44\beta)} + e^{(28\beta)} + \cdots + e^{(38\beta)} + e^{(24\beta)}\right]\right)$$

$$+35\beta - 1\left(ln\left[e^{(35\beta)} + e^{(35\beta)} + \cdots + e^{(38\beta)} + e^{(24\beta)}\right]\right)$$

$$+38\beta - 1\left(ln\left[e^{(38\beta)} + e^{(24\beta)}\right]\right)$$

3.1.2.2. Cálculo del valor de las primeras derivadas

Las derivadas parciales del logaritmo de la función de Verosimilitud Parcial, respecto a cada uno de los coeficientes, se calcula de acuerdo a las siguientes ecuaciones:

$$U_k(\beta) = \sum_{i=1}^{k}\left[\frac{Z_i - m_i\left(\sum_{R_i}(Xe^{(X'\beta)})\right)}{\sum_{R_i}(e^{(X'\beta)})}\right]$$

La sumatoria general para los *k* instantes en los que se produce algún desenlace. Las otras sumatorias se realizan para los sujetos expuestos al riesgo.

Desarrollando la expresión general para cada uno de los 5 instantes en los que se produce algún cambio, se obtiene el vector de primeras derivadas, en función del parámetro que se desea estimar:

$$U(\beta) = 45 - \frac{45e^{(45\beta)} + \cdots + 24e^{(24\beta)}}{e^{(45\beta)} + e^{(20\beta)} + \cdots + e^{(24\beta)}}$$

$$+81 - 2\left(\frac{51e^{(51\beta)} + \cdots + 24e^{(24\beta)}}{e^{(51\beta)} + e^{(30\beta)} + \cdots + e^{(24\beta)}}\right)$$

$$+44 - \frac{44e^{(44\beta)} + \cdots + 24e^{(24\beta)}}{e^{(44\beta)} + e^{(28\beta)} + \cdots + e^{(24\beta)}}$$

$$+35 - \frac{35e^{(35\beta)} + \cdots + 24e^{(24\beta)}}{e^{(35\beta)} + e^{(35\beta)} + \cdots + e^{(24\beta)}}$$

$$+38 - \frac{38e^{(38\beta)} + \cdots + 24e^{(24\beta)}}{e^{(38\beta)} + e^{(24\beta)}}$$

3.1.2.3. Cálculo de la matriz de segundas derivadas

Las segundas derivadas parciales de la Función de Verosimilitud Parcial, respecto a cada uno de los coeficientes, se calcula de acuerdo a las siguientes ecuaciones:

$$I_{KK'}(\beta) = \sum_{i=1}^{m} \left[m_i \left[\frac{\sum(X_K X_{K'} e^{(X'\beta)})}{\sum_{R_i}(e^{(X'\beta)})} - \frac{\sum(X_K e^{(X'\beta)})\left(\sum_{R_i} X_{K'} e^{(X'\beta)}\right)}{\left(\sum_{R_i} X_{K'} e^{(X'\beta)}\right)^2} \right] \right]$$

Desarrollando la expresión general para cada uno de lo 5 instantes en los que se produce algún cambio, se obtiene la matriz de segundas derivadas, en función del parámetro que se desea estimar:

$$I(\beta) = \left[\frac{(45)(45)e^{(45\beta)} + \cdots + (24)(24)e^{(24\beta)}}{e^{(45\beta)} + \cdots + e^{(24\beta)}} - \frac{\left(45e^{(45\beta)} + \cdots + 24e^{(24\beta)}\right)\left(45e^{(45\beta)} + \cdots + 24e^{(24\beta)}\right)}{(e^{(45\beta)} + \cdots + e^{(24\beta)})^2} \right] + \cdots +$$

$$+ \left[\frac{(38)(38)e^{(38\beta)} + (24)(24)e^{(24\beta)}}{e^{(38\beta)} + e^{(24\beta)}} - \frac{\left(38e^{(38\beta)} + 24e^{(24\beta)}\right)\left(38e^{(38\beta)} + 24e^{(24\beta)}\right)}{(e^{(38\beta)} + e^{(24\beta)})^2} \right]$$

3.2. Cálculo del valor β^* en cada iteración

Pasos para estimar el coeficiente β utilizando el método **(Método de Newton-Raphson).**

1) Se realiza una estimación inicial $\beta = \beta_0$. Generalmente se toma cero como primer valor. Así pues, la primera estimación es $\beta = 0$.

2) Se calculan los valores de U (β_0) y de I (β_0); para ello se sustituye el valor $\beta = 0$ en las expresiones anteriormente halladas del vector de primeras derivadas U (β) y de la matriz de segundas derivadas I (β).

   Para calcular $U(0)$ se sustituyendo el valor inicial $\beta = 0$ en la expresión general del vector de primeras derivadas $U(\beta)$, se obtiene el valor:

$$U(0) = \left(45 - \frac{424}{12}\right) + \left(81 - 2\left(\frac{320}{9}\right)\right) + \left(44 - \frac{204}{6}\right) + \left(35 - \frac{132}{4}\right) + \left(38 - \frac{62}{2}\right)$$

$$U(0) = 38.555$$

Para calcular $I(0)$ se sustituye el valor inicial $\beta = 0$ en la expresión general de la matriz de segundas derivadas así:

$$I(0) = \left(\frac{15862}{12} - \left(\frac{(424)(424)}{12^2}\right)\right)$$

$$+2\left(\frac{11916}{9} - \left(\frac{(320)(320)}{9^2}\right)\right)$$

$$+\left(\frac{7190}{6} - \left(\frac{(204)(204)}{6^2}\right)\right)$$

$$+\left(\frac{4470}{4} - \left(\frac{(132)(132)}{4^2}\right)\right)$$

$$+\left(\frac{2020}{2} - \left(\frac{(62)(62)}{2^2}\right)\right)$$

$$I(0) = 312.827$$

3) Se calcula la siguiente aproximación β^* de β, por medio de la expresión:

$$\beta^* = \beta_0 + I^{-1}(\beta_0)U(\beta_0)$$

La segunda estimación viene dada por:

$$\beta^* = 0 + \left(\frac{38.555}{312.827}\right) = 0.12324881$$

4) Se repiten los pasos 2 y 3 reemplazado β_0 por β^*.

$$U(\beta) = 45 - \frac{68394.50347}{1577.160825} + 81 - 2\left(\frac{51856.71381}{1186.808914}\right) + 44 - \frac{20652.72618}{534.9127812}$$

$$+35 - \frac{9801.841895}{276.8360083} + 38 - \frac{4571.648199}{127.4019026}$$

$$U(\beta) = 1.634414 - 6.388480 + 5.390479 - 0.406673 + 2.116327$$

$$U(\beta) = 2.346067175$$

$$I(\beta) = \left[\frac{3047460.317}{1577.160825} - \frac{(68394.5)(68394.5)}{(1577.16)^2}\right] + 2\left[\frac{2337766.067}{1186.808914} - \frac{(51856.7)^2}{(1186.8)^2}\right]$$

$$+\left[\frac{813621.5166}{534.9127812} - \frac{(20662.7)^2}{(534.9)^2}\right] + \left[\frac{350308.4325}{276.8360083} - \frac{(9801.8)^2}{(276.8)^2}\right]$$

$$+\left[\frac{167251.6532}{127.4019026} - \frac{(4571.6)^2}{(127.4)^2}\right]$$

$$I(\beta) = 51.6704386 + 2(60.6048327) + 30.3407495 + 11.7680724 + 25.1497456$$

$$I(\beta) = 240.1386715$$

La siguiente iteración proporciona la estimación:

$$\beta^* = 0.12324881 + \frac{2.346067175}{240.1386715} = 0.133018445$$

A partir de este valor $\beta^* = 0.133018445$ y siguiendo los pasos anteriores, se llega en la última iteración al valor $\beta^{**} = 0.133293$.

3.3. Cálculos finales para el coeficiente obtenido

Una vez obtenido el valor final $\beta^{**} = 0.133293$. Se puede estimar la Verosimilitud de la distribución muestral del coeficiente, así como proceder al estudio de la significación global del modelo.

Valor del logaritmo de la Función de Verosimilitud Parcial:

$$ln\big(L(\beta)\big) = 5.9985 - ln[402.82 + 14.38 + 181.03 + 158.44 + 24.51]$$

$$+10.7973 - 2ln[896.32 + 54.54 + 106.2 + 158.44 + 24.51]$$

$$+5.8652 - ln[352.55 + 41.78 + 158.44 + 24.51]$$

$$+4.665 - ln[106.2 + 106.2 + 158.44 + 24.51]$$

$$+5.0654 - ln[158.44 + 24.51]$$

$$ln\big(L(\beta)\big) = 5.9985 - ln[2445] + 10.7973 - 2ln[1845.71] + 5.8652 - ln[788.65] +$$

$$+4.6655 - ln[394.32] + 5.0654 - ln[181.92]$$

$$ln\big(L(\beta)\big) = -8.3132$$

Cálculo del error estándar del coeficiente β hallado.

Puesto que el valor hallado del coeficiente β^{**} es una estimación del valor real, vendrá afectado por un determinado error. El valor del error estándar puede ser obtenido por medio de la matriz de información de Fisher. Para ello se obtiene el valor de la matriz de segundas derivadas para el coeficiente $\beta^{**} = 0.133293$ obtenido.

Este valor resulta ser: $I(\beta) = 226.81$

La matriz de varianzas covarianzas de los coeficientes viene dada por la inversa de la matriz de información. En nuestro caso, al ser único coeficiente, la varianza viene dada por la inversa del valor obtenido:

$$I(\beta)^{-1} = Var(\beta) = \frac{1}{226.81} = 0.004409$$

De donde resulta que el error estándar (EE) buscado es:

$$EE(\beta) = 0.0664.$$

Cálculo del test Ji-cuadrado global

A partir del vector de primeras derivadas y de la matriz de segundas derivadas, el test de Rao (score test) permite estudiar la hipótesis nula que el vector de coeficientes es un valor nulo. Bajo la hipótesis nula este test (Miller, 1981, p.125; Lawless, 1982, p.440; Hill et al, 1990, p.83) sigue una Distribución Ji-cuadrado con grados de libertad igual al orden del vector de coeficientes estimado.

En este caso, el orden del vector es 1. Así pues:

$$\chi^2 = U'(0)U^{-1}(0)U(0) = 38.553\left(\frac{1}{312.827}\right)8.55 = 4.75$$

3.4. Interpretación del Coeficiente del Modelo de Regresión de Cox

Los coeficientes del Modelo de Regresión de Cox indican la relación existente entre la variable explicativa correspondiente y la función de riesgo. Un valor positivo del coeficiente supone un aumento en el valor de la función de riesgo para el sujeto, lo cual con lleva una relación negativa con el tiempo de cambio. Es decir, un coeficiente positivo indica un mayor riesgo de que se produzca el cambio.

El siguiente Modelo de Regresión de Cox visto anteriormente queda de la siguiente manera:

$$h(t|\mathbf{x}) = h_0(t)\mathrm{e}^{(0.133293(EDAD))}$$

El valor $\beta^{**} = 0.133293$ significa que el incremento de un año de edad aumenta el logaritmo de la tasa de riesgo en 0.133293 (controlando las otras variables incluidas en la ecuación).

De esta manera concluimos que el Modelo de Regresión de Cox, puede ser utilizado, cuando sea necesario estimar la función de riesgo $h(t|\mathbf{x})$, dentro de un conjunto de variables explicativas (variables independiente) ó mas bien cuando sea necesario la probabilidad condicional del individuo en investigación.

Además, $100(e^{0.133293} - 1) = 100(1.1426 - 1) = 14.26$ proporciona el cambio en porcentaje en la tasa de riesgo que se produce con cada unidad de cambio en la variable independiente. Es decir, que en ocho meses aumenta la tasa de riesgo en un 14.26% respecto a su valor inicial, para los sujetos en estudio del Centro de Rehabilitación "Fe y Esperanza".

3.5. Segunda aplicación

En la segunda aplicación se utilizará la Distribución Exponencial, vista y analizada en el capitulo I la decisión de usar este modelo, esta sustentada por las caracteristicas de los datos obtenidos. Donde se observó que el tiempo de falla de los focos, no dependía de lo que había sucedido anteriormente. Cumpliendo así con una de las propiedades de la Distribución Exponencial.

A continuación se presenta de forma ilustrativa, la estimación de la función de supervivencia, la tasa de fallo y el valor esperado de la Distribución Exponencial, mediante el Método de Máxima Verosimilitud, aplicando así la teoría menciona en los capítulos anteriores.

Distribución Exponencial. Para esta aplicación se comenzará definiendo a ***T*** como una variable aleatoria, que representa el tiempo de vida útil de los componentes electrónicos. Se dice que ***T*** tiene una Distribución Exponencial con parámetros $\lambda > 0$. Optaremos por usar esta Distribución con reparametrización $\theta = \frac{1}{\lambda}$ en este caso haremos uso de la siguiente ecuación:

$$f(t) = \theta^{-1}\mathrm{e}^{\frac{-t}{\theta}}$$

Donde:

$$\theta = \frac{w}{n} = \bar{t}$$

$$w = \sum_{i=t}^{n} t_i$$

Y

$$\lambda = \frac{1}{\theta}$$

Entonces la ecuación $s(t) = \mathrm{e}^{-\lambda t}$ se convierte en:

$$s(t) = \mathrm{e}^{\frac{-t}{\theta}} \qquad (3.5.1)$$

Como ya se menciono en los capítulos anteriores, es el único modelo con tasa de fallo constante.

$$h(t) = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda$$

La probabilidad de fallo condicionada a que el elemento o componente este en uso no varía con el tiempo. Esta propiedad se denomina falta de memoria.

La Distribución Exponencial, es una de las más comunes en los textos estadísticos y en los modelos de supervivencia.

.

Por ejemplo para calcular la probabilidad que un componente electrónico (foco) con vida media de 638.75 horas funcione más de 700 horas. Se usará la ecuación (3.5.1)

Sustituyendo $\frac{1}{\theta} = \frac{1}{638.75}$ en la ecuación (3.5.1) se tiene:

$$s(\boldsymbol{T} > 700) = e^{\frac{-700}{638.75}} = 0.3342$$

3.5.1. Estimación Paramétrica

El proceso de ajuste del modelo estadístico a partir de datos muestrales se obtienen de la siguiente manera:

- Se estudian los datos mediante técnicas de Estadística Descriptiva.
- Se elige un modelo de Distribución de Probabilidad.
- Se estima la función.

En la investigación, y como parte de una de las aplicaciones, se recopiló información en la Granja de Pollos de Engorde "Martínez", donde se tiene un control de la duración de la vida útil de los focos, entre otras observaciones que ahí se utilizan, 20 de estos focos se les conoció el tiempo de falla en horas.

A continuación se presenta una matriz de datos, que contiene los tiempos de vida útil de los 20 focos.

Para mayor comodidad, en la tabla 3.3 el tiempo de vida útil de los focos se han ordenado en función de la variable tiempo de durabilidad de los productos.

CASO	TIEMPO EN HORAS (FOCOS)
1	372
2	388
3	397
4	412
5	427
6	453
7	477
8	498
9	515
10	547
11	591
12	608
13	657
14	702
15	758
16	796
17	867
18	956
19	1076
20	1278

Tabla 3.3 Matriz de Datos recopilados

3.5.2. Estimación de la Función de Supervivencia.

Si conocemos la función de supervivencia, podemos obtener una estimación de la misma a partir de los datos observados mediante el método de la Máxima Verosimilitud.

Para ello, debemos construir la Función de Verosimilitud que dependerá de cómo se recogen los resultados. El caso más sencillo es cuando disponemos del tiempo t_i en el que se presenta el suceso en cada uno de los individuos o componente observados. En este caso, si la función de supervivencia y la correspondiente densidad es la siguiente expresión:

$$s(t,\theta) \rightarrow f(t,\theta) \tag{3.5.2.1}$$

Podemos indicar la Verosimilitud como:

$$L = \prod_{i=1}^{n} f(t,\theta) \tag{3.5.2.2}$$

Si el vector es $\theta = (\theta_1, \dots, \theta_p)$, podemos obtener los estimadores a partir de las ecuaciones:

$$\frac{\partial L}{\partial \theta_j} = 0; \quad j = 1, \dots, p \tag{3.5.2.3}$$

La varianza de los estimadores puede obtenerse calculando la siguiente matriz de datos de Fisher:

$$I(\hat{\theta}) = -\left(\frac{\partial^2 ln(L)}{\partial \theta^2}\right)_{\hat{\theta}}$$

$$V(\hat{\theta}) = I(\hat{\theta})^{-1} \qquad (3.5.2.4)$$

3.5.3. Función de Supervivencia Exponencial sin Censura

Supongamos que $f(t) = \lambda e^{-\lambda t}$. En este caso, si hemos observado el suceso en todos los individuos o componentes y haciendo uso de la ecuación (3.5.2.2) tenemos:

$$L = \prod_{i=1}^{n} \lambda e^{-\lambda t}$$

Aplicando logaritmo natural y sus respectivas derivadas parciales a la ecuación anterior, se obtienen los siguientes estimadores: $\hat{\lambda}$ y la varianza estimada $V(\hat{\lambda})$, respectivamente.

$$ln(L) = \ ln\left(\prod_{i=1}^{n} \lambda e^{-\lambda t}\right)$$

$$ln(L) \ = \sum_{i=1}^{n} ln(\lambda) - \sum_{i=1}^{n} \lambda t_i$$

$$\frac{\partial ln(L)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} t_i \rightarrow \hat{\lambda} = n\left(\sum_{i=1}^{n} t_i\right)^{-1}$$

$$\frac{\partial^2 ln(L)}{\partial \lambda^2} = \frac{n}{\lambda^2} \rightarrow V(\hat{\lambda}) = \frac{\hat{\lambda}^2}{n}$$

Por lo tanto, a partir de las observaciones del momento en que se producen los sucesos, podemos obtener el parámetro requerido para caracterizar la función de supervivencia exponencial.

3.5.4. Función de Supervivencia Exponencial con Censura Tipo I

La situación más común en el análisis de supervivencia se caracteriza por la presencia de censura por la derecha. En este caso, tenemos dos tipos de observaciones. En primer lugar, los focos para los cuales se observa el suceso en un tiempo t_i (no censurados). Por otra, los focos para los cuales no se observa el suceso antes de dicho tiempo (censurados). Por lo tanto, para los focos censurados tenemos el suceso: $\boldsymbol{T} \geq t_i$. En este caso, si disponemos de *n* focos no censurados y *m* focos censurados, la Función de Verosimilitud puede construirse así:

$$L = \prod_{i=1}^{n} f\,(t_i) \prod_{j=1}^{m} s(t_j) \qquad (3.5.4.1)$$

En el caso de la Distribución Exponencial la función de supervivencia correspondiente es de la forma $s(t) = e^{-\lambda t}$, formando la Función de Verosimilitud tenemos:

$$L = \prod_{i=1}^{n} \lambda e^{-\lambda t_i} \prod_{j=1}^{m} e^{-\lambda t_j}$$

$$ln(L) = ln\left(\prod_{i=1}^{n} \lambda e^{-\lambda t_i} \prod_{j=1}^{m} e^{-\lambda t_j}\right)$$

$$ln(L) = nln(\lambda) - \lambda \sum_{i=1}^{n} t_i - \lambda \sum_{j=1}^{m} t_j$$

$$l\,n(L) = n\lambda - \lambda\left(\sum_{i=1}^{n} t_i + \sum_{j=1}^{m} t_j\right)$$

$$ln(L) = n\lambda - \lambda m$$

$$\frac{\partial ln(L)}{\partial \lambda} = \frac{n}{\lambda} - w \rightarrow \hat{\lambda} = \frac{n}{w}$$

$$\frac{\partial^2 ln(L)}{\partial \lambda^2} = -\frac{n}{\lambda^2} \rightarrow V = (\hat{\lambda}) = \frac{\lambda^2}{n}$$

A partir de aquí podemos inferir muchas propiedades de la Distribución Exponencial.

Para la investigación, como los datos observados no presentaron ningún tipo de censura, debido a que se les conoció el tiempo de vida útil, por lo tanto no haremos uso de la ecuación (3.5.4.1), ni de la expresion anterior.

Consideremos las observaciones de la tabla 3.3 de la matriz de datos anterior con $n = 20$

$$W = \sum_{i=1}^{n} t_i = 372 + 388 + 397 + 412 + 427 + 453 + 477 + 498 + 515 + 547 + 591$$

$$+ 608 + 657 + 702 + 758 + 796 + 867 + 956 + 1076 + 1278$$

$$W = 12{,}775$$

Por lo tanto la media muestral correspondientes a estos datos es:

$$\bar{t} = \theta = \frac{W}{n} = \frac{12{,}775}{20} = 638.75$$

Donde:

$$\hat{\lambda} = \frac{n}{W}$$

$$\hat{\lambda} = \frac{20}{12{,}775}$$

$$\hat{\lambda} = \ 0.00156555773$$

Y la varianza correspondiente a estos datos es:

$$V(\hat{\lambda}) = \frac{\hat{\lambda}^2}{n}$$

$$V(\hat{\lambda}) = \frac{(0.00156555773)^2}{20}$$

$$V(\hat{\lambda}) = 0.0000001225485503$$

Por lo tanto la Función de Supervivencia correspondiente a estas observaciones es:

$$s(\boldsymbol{T} > t) = \mathrm{e}^{-(0.00156555773)(t)}$$

Por ejemplo, la probabilidad de que un componente dure más de 800 horas será:

$$s(t) = \mathrm{e}^{-\lambda t}$$

$$s(\boldsymbol{T} > 800) = \mathrm{e}^{-(0.00156555773)(800)} = 0.2858$$

3.5.5. Ajuste del Modelo Exponencial

Al estimar la Función de Distribución Empírica de los datos y representarla en unas escalas tales que si el modelo elegido es correcto los datos presentan aspecto lineal.

Pasos para el Ajuste del Modelo Exponencial

- Ordenamos los datos de menor a mayor

- Estimación de la Función de Distribución corregida mediante:

$$F_i = \frac{(i - 0.3)}{(n + 0.4)} \qquad (3.5.5.1)$$

- Construcción del gráfico adecuado hasta que los datos formen una recta.

Construcción del grafico exponencial. Para comprobar que nuestros datos son Exponenciales, retomamos la Función de Supervivencia $s(t)$.

$$s(t) = e^{\frac{-t}{\theta}}$$

Y aplicando $logaritmo\ natural$ a la ecuación anterior se tiene:

$$ln\big(s(t)\big) = ln\left(e^{-\frac{t}{\theta}}\right)$$

$$ln\big(s(t)\big) = -\frac{t}{\theta}$$

Como:

$$F(t) = 1 - s(t) \Rightarrow s(t) = 1 - F(t)$$

$$s(t) = 1 - F(t)$$

$$e^{\frac{-t}{\theta}} = 1 - F(t)$$

$$ln\left(e^{\frac{-t}{\theta}}\right) = ln\big(1 - F(t)\big)$$

$$ln\big(1 - F(t)\big) = -\frac{t}{\theta}$$

A continuación se realizan los cálculos necesarios, para determinar si los datos recopilados en la investigación son Exponenciales.

Estimación de la Función de Distribución Corregida mediante la siguiente ecuación:

$$F_i = \frac{(i - 0.3)}{(n + 0.4)}.$$

Los n registros de los tiempos de fallo se ordenan de menor a mayor (como están en la tabla 3.4) y se les asigna un número de orden i del 1 a n.

TIEMPOS	ORDEN	$F = \frac{(i - 0.3)}{(20 + 0.4)}$
372	1	0.03
388	2	0.08
397	3	0.13
412	4	0.18
427	5	0.23
453	6	0.28
477	7	0.33
498	8	0.38
515	9	0.43
547	10	0.48
591	11	0.52
608	12	0.57
657	13	0.62
702	14	0.67
758	15	0.72
796	16	0.77
867	17	0.82
956	18	0.87
1076	19	0.92
1278	20	0.97

Tabla 3.4 Matriz de Datos de la Función de Distribución Corregida

TIEMPOS	$ORDEN$	$F = \frac{(i - 0.3)}{(20 + 0.4)}$	$ln(1 - F(t))$
372	1	0.03	0.03
388	2	0.08	0.08
397	3	0.13	0.14
412	4	0.18	0.20
427	5	0.23	0.26
453	6	0.28	0.32
477	7	0.33	0.39
498	8	0.38	0.47
515	9	0.43	0.55
547	10	0.48	0.64
591	11	0.52	0.74
608	12	0.57	0.85
657	13	0.62	0.97
702	14	0.67	1.11
758	15	0.72	1.27
796	16	0.77	1.46
867	17	0.82	1.70
957	18	0.87	2.02
1076	19	0.92	2.44
1278	20	0.97	3.38

Tabla 3.5 Matriz de Datos de la Función de Distribución Corregida con aspecto lineal.

Colocando en el eje y la variable y $= ln(1 - F(t))$ y en eje x la variable t, los datos de la tabla 3.5 se obtiene la siguiente figura.

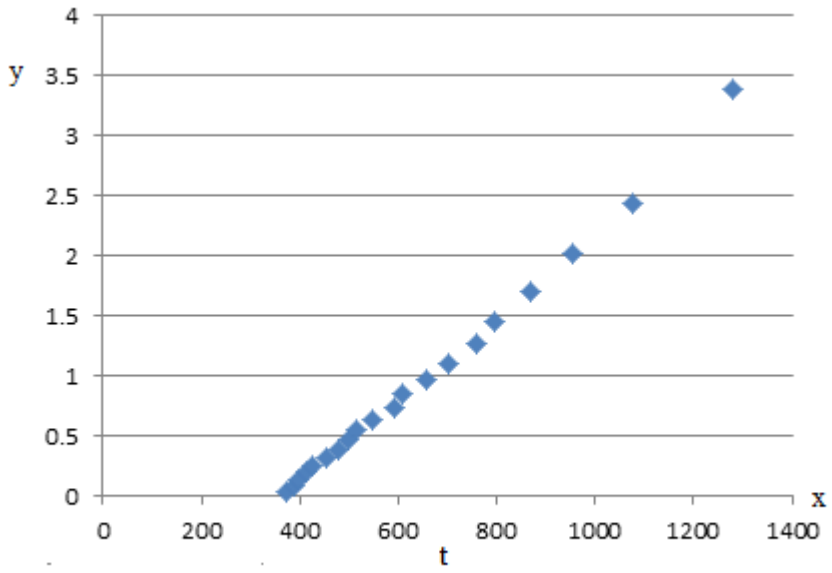


Figura 3.2

Como se puede observar en la figura 3.2, los datos presentan una escala lineal, por lo tanto corresponden a una Distribución Exponencial.

De esta manera concluimos que los datos corresponden a una Distribución Exponencial, además, este Modelo Exponencial, puede ser utilizada cuando se tiene la posibilidad de tener datos que no presentan ningún tipo censura, es decir, que se les conoce el tiempo de falla, y datos censurados a los que no se les conoce el tiempo en que presentaron la falla.

Por lo tanto, la Función de Supervivencia $s(t)$, correspondiente a la Distribución Exponencial, con la información obtenida en la Granja de Pollos de Engorde "Martínez" queda finalmente de la siguiente manera:

$$s(t) = e^{-(0.00156555773)(t)}$$

La cual representa, el valor de la Función de Supervivencia en el tiempo "t" y la probabilidad de que un foco en estudio en la Granja de Pollos de Engorde "Martínez", funcione más allá de t horas.

En muchas situaciones de interés conviene considerar el efecto de un conjunto de variables predictoras (covariables) en la supervivencia. Así, en un estudio acerca de la respuesta a un tratamiento, puede ser interesante estimar la supervivencia en función de la edad, sexo, gravedad, etc. En este caso, debemos especificar el efecto de las variables predictoras.

.

CONCLUSIONES

El uso de la estadística en distintas áreas ha ido adquiriendo cada vez mayor auge, en particular las Ciencias Naturales. El Modelo de Riesgo Proporcional es en ocasiones utilizado para representar los efectos de un conjunto de variables explicativas (variables independientes) sobre las variables tiempo de cambio (tiempo de supervivencia), o más bien sobre la probabilidad condicional de cambio, es decir sobre la Función de Riesgo

Para los estudiosos es de gran interés conocer modelos probabilísticos diferentes a los comunes con los que sea posible realizar predicciones confiables sobre tiempos de vida (muertes/fallas) en individuos o componentes. La Distribución Exponencial es una de las Distribuciones utilizadas para el estudio de tiempos de vida de componentes.

En este trabajo se mostraron las bases teóricas de las principales Distribuciones aplicables en las Funciones de Supervivencia en una forma general para el análisis de tiempos de vida.

Sin embargo, en la naturaleza existen fenómenos aleatorios en los cuales no siempre se puede ajustar satisfactoriamente alguna de las Distribuciones revisadas en el Capítulo I.

Después de recopilar toda la teoría acerca de los Modelos de Supervivencia, se obtuvieron finalmente las aplicaciones para modelar de forma práctica mediante la información facilitada por el Centro de Rehabilitación “Fe y Esperanza”, y por la Granja de Pollos de Engorde “Martínez”, empleando así el Método de Máxima Verosimilitud para la

estimación de los parámetros, del Modelo de Regresión de Cox y de la Distribución Exponencial.

La importancia de este trabajo de investigación se fundamenta en la parte teórica y la parte de aplicaciones. Una vez mencionado lo anterior, estamos seguros de que este trabajo será una buena guía para las personas interesadas en el análisis estadístico de tiempos vida.

REFERENCIAS BIBLIOGRAFICAS

BIBLIOGRAFÍA

- ✓ Elementos de Probabilidad y Estadística (José Hernández Salguero)
- ✓ Control Estadístico de la Calidad (Douglas C.Mongomery)
- ✓ Probabilidad y Estadística para Ingenieros. 4E (Erwin. RMiller / John. E Freund)

PAGINAS WEB

- ✓ Attardi, L., Guida, M., & Pulcini, G. (2005). A mixed-Weibull regression model for the analysis of automotive warranty data. 265-273.
- ✓ Behboodian, J. (1970). On Mixture of Normal Distributions. 7-215.
- ✓ Blischke, W. R. (1962). *Moment estimators for the parameters of a mixture of two binomial*
- ✓ *Distributions* . Annals of Mathematical Statistics.
- ✓ Chambers, R. L., & Skinner, C. J. (2003). *Analysis of Survey Data* (First ed.). Wiley.
- ✓ Charlier., C. V. (1906). *Researches into the theory of probability.*
- ✓ losarios.servidor-alicante.com/terminos-estadistica/datos-censurados.
- ✓ *http://www.formaciononline.us.es/asignaturas/asigedo/apartados/textos/7_2.PDF*