UNIVERSIDAD DE EL SALVADOR

FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICA

ESCUELA DE MATEMÁTICA

DEPARTAMENTO DE ESTADÍSTICA

TRABAJO DE INVESTIGACIÓN TITULADO:

"ANÁLISIS ESTADÍSTICO QUE PREDIGA LAS CAUSAS O FACTORES QUE CONLLEVAN AL INCREMENTO DE PACIENTES CON INSUFICIENCIA RENAL EN EL HOSPITAL MILITAR CENTRAL"

PRESENTADO POR:

CANTOR CAMPOS, VERÓNICA ZULEYMA.

FLORES CHÁVEZ, INGRID XIOMARA.

SANTIAGO, JESSICA MARISOL.

PARA OPTAR AL TITULO:

LICENCIADA EN ESTADÍSTICA.

ASESORA:

MSC. ALBA IDALIA CÓRDOVA CUÉLLAR.

CIUDAD UNIVERSITARIA, SAN SALVADOR EL SALVADOR, NOVIEMBRE 2010.

RECTOR
ING. RUFINO ANTONIO QUEZADA SÁNCHEZ.

SECRETARIO GENERAL
LIC. DOUGLAS VLADIMIR ALFARO CHÁVEZ**.**

FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICA.

DECANO
DR. RAFAEL ANTONIO GÓMEZ ESCOTO.

SECRETARIA
LIC. MARÍA TRINIDAD TRIGUEROS DE CASTRO.

ESCUELA DE MATEMÁTICA.

DIRECTOR
ING. CARLOS MAURICIO CANJURA LINARES.

TRABAJO DE GRADUACIÓN APROBADO POR:

ASESOR
MSC. ALBA IDALIA CÓRDOVA CUÉLLAR.

AGRADECIMIENTOS.

- Damos gracias en primer lugar a Dios por darnos la oportunidad de terminar nuestros estudios de una manera exitosa, y porque a lo largo de nuestra carrera nos ha dado la fortaleza y la perseverancia para alcanzar todas nuestras metas, y superar los obstáculos que se nos han presentado.

- Agradecemos de igual forma a nuestras familias, por el apoyo incondicional que nos han brindado día con día en nuestros estudios y en toda nuestra vida, ya que su apoyo ha sido una de las principales fortalezas que nos ha impulsado a seguir siempre adelante a pesar de todas las dificultades que se nos han presentado.

- Agradecemos a nuestra asesora Msc. Alba Idalia Córdova Cuéllar, ya que ha sido la persona que nos fue guiando durante todo el proceso de nuestro proyecto de Graduación.

ÍNDICE

Contenido Página

INTRODUCIÓN.

El Salvador es un país territorialmente pequeño en donde predomina el desempleo, pobreza y sobrepoblación. Es uno de los países que por diversos factores sociales, educativos, climáticos, hábitos dietéticos y creencias populares, se desarrollan muchas enfermedades y entre ellas se encuentra las infecciones de las vías urinarias que algunas veces dejan como resultado la insuficiencia renal.

En los últimos años la insuficiencia renal ha tenido una gran incidencia en la población salvadoreña; y se encuentra estadísticamente ubicada entre las diez primeras causas de mortalidad institucional (Hospital Militar Central). Entre los factores que afectan a la población están los malos hábitos alimenticios incluyendo la poca ingesta de agua y a la vez el consumo de aguas contaminadas y la automedicación (medicina natural).

Las causas que originan la insuficiencia renal, suelen ser diferentes según áreas geográficas y desarrollo económico. En términos generales la insuficiencia renal puede ser causada por enfermedades que afectan exclusivamente al riñón y por enfermedades sistémicas que terminan comprometiendo la función renal como parte de su evolución natural. En nuestro país, al igual que en otros de la región, la mayor cuantía de pacientes que llegan a una falla renal avanzada, suelen ser por causa de una enfermedad glomerular primaria, es decir; dificultad de la eliminación de los desechos del cuerpo que se acumulan en la sangre. La insuficiencia renal afecta del 2% al 10% de la población salvadoreña; ocasionando así un aumento en el número de muertes por dicha enfermedad. En los pacientes con insuficiencia renal la ingestión de magnesio, es el mayor determinante de su concentración plasmática. Debido al fallo renal se produce una reducción de la filtración neta de este ión, no existiendo vías alternativas para su eliminación. Por otra parte, existen controversias en relación a la absorción intestinal de magnesio en pacientes urémico. Diversos estudios nos muestran diferencias significativas entre individuos sanos e individuos con insuficiencia renal.

En el Hospital Militar Central, se ha registrado un aumento considerable de pacientes con insuficiencia renal. En el año 2006, la insuficiencia renal ocupaba la quinta posición dentro de las 10 primeras causas de morbilidad de adultos (hombres y mujeres) con un total de 180 casos entre las edades de 16 años y más, mientras que el año 2007, esta enfermedad se

situaba en la séptima posición dentro de las 10 primeras causas de morbilidad de adultos (hombres y mujeres) con un total de 201 casos entre las edades de 20 años y más; en el año 2008, se ubicó en la novena posición dentro de las 10 primeras causas de morbilidad de adultos (hombres y mujeres) con un total de 250 casos entre las edades de 16 años y más.[1]

Con base a lo expuesto anteriormente, en esta investigación, se plantea abordar el siguiente problema: Determinar las causas o factores que conllevaron al incremento de casos de pacientes con insuficiencia renal; para ello se cuenta con una base de datos, que se ha construido a partir de los expedientes de los pacientes que padecen de insuficiencia renal y que han sido atendidos en el Hospital Militar Central de San Salvador, dicha base de datos tiene registros de estos pacientes desde el año 2002 hasta el 2008. A las variables estadísticas contenidas en esta base de datos se le realizará un análisis descriptivo y además se le aplicará técnicas estadísticas avanzadas, es decir; métodos especiales de análisis de supervivencia, que pueden ser realizadas, utilizando pruebas como las No paramétricas, entre los que se encuentra la metodología de Kaplan & Meier y la Regresión de Cox, con el objetivo de llegar a conclusiones que ayuden a prevenir esta enfermedad en el país; a partir de la información proporcionada por las autoridades del Hospital Militar Central.

[1] Fuente: Anuarios de los registro de morbilidad de egresos, sección de estadística de los años 2006, 2007 y 2008 del Hospital Militar Central.

OBJETIVOS.

Objetivo General:

- Aplicar las técnicas de análisis de supervivencia para predecir las causas o factores que conllevan a un incremento de pacientes con insuficiencia renal.

Objetivos Específicos:

- Hacer un análisis exploratorio y de supervivencia de las variables contenidas en la base de datos de pacientes con insuficiencia renal del Hospital Militar Central.

- Estimar las funciones de supervivencia y de riesgo para determinar la probabilidad de que a un individuo le ocurra el evento de interés (muerte).

- Ajustar un modelo de regresión que prediga las causas o factores que generaron un incremento de pacientes con insuficiencia renal en el Hospital Militar Central.

- Realizar la aplicación del análisis de supervivencia, utilizando el software estadístico SPSS.

CAPÍTULO I: ANÁLISIS EXPLORATORIO DE DATOS.

PRÓLOGO.

El análisis exploratorio de datos, es un conjunto de estrategias para el análisis de datos; cuya esencia es permitir que los datos muestren resultados, y además se pueda realizar la búsqueda de patrones en dichos datos. En muchos casos el análisis exploratorio de datos puede proceder a una situación de inferencia formal, mientras que otros pueden sugerir preguntas y conclusiones que se podrían confirmar con un estudio adicional.

De acuerdo con lo anterior, el análisis exploratorio de datos puede ser una herramienta muy útil en la generación de hipótesis, conjeturas y preguntas de investigación acerca de los fenómenos de donde los datos fueron obtenidos.

En este capítulo se realiza un enfoque de las técnicas estadísticas que comprende el análisis exploratorio de datos y dentro de este el análisis univariado y el análisis bivariado el primero de estos con el objetivo de observar la procedencia de los datos que se analizaran y ver la forma de cada una de las variables dentro de este se obtendrá las medidas numéricas como son las de posición, tendencia central, dispersión y la construcción de tablas de frecuencias etc. Dentro del análisis bivariado el objetivo es ver la relación que existe entre dos variable y poder observar si entre las variables existe o no dependencia; esta teoría se desarrolla con el fin de ver la forma o la estructura de los datos.

1.1 TEORÍA DEL ANÁLISIS EXPLORATORIO DE DATOS.

En toda investigación, y antes de extraer conclusiones acerca de los objetivos e hipótesis planteados, es necesario llevar a cabo un análisis exploratorio previo de los datos. El cual consiste en el descubrimiento de estructura en los datos por medio de métodos simples como parámetros de estadística descriptiva o técnicas de visualización, cuya finalidad es conseguir un entendimiento básico de los datos y de las relaciones existentes entre las variables analizadas. Para alcanzar este objetivo el análisis exploratorio de datos proporciona métodos o técnicas sistemáticas sencillas para organizar y preparar los datos, es decir; organiza, describe y resume los datos, a través de un análisis numérico y representaciones gráficas.

El análisis exploratorio de datos de divide en:

- ✓ Análisis Unidimensional.
- ✓ Análisis Bidimensional.
- ✓ Análisis Multidimensional.

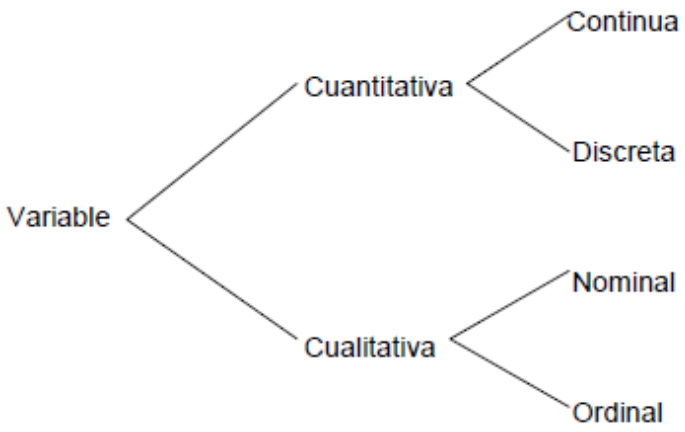
Para cumplir con el objetivo de esta investigación, solo se realizará un análisis unidimensional y bidimensional, los cuales se detallan a continuación:

1.1.1 ANÁLISIS UNIDIMENSIONAL.

El análisis unidimensional de datos es importante porque permite observar a cada variable individualmente y se aprende mucho acerca de la información recopilada. Además; es un buen método para verificar la calidad de los datos. Siempre deben investigarse las inconsistencias o los resultados inesperados, usando los datos originales como punto de referencia.

Los datos son los valores que toma la variable en cada caso. Se llama variable a una característica que se observa en los individuos u elementos de una población o muestra, la cual se desea estudiar para realizar las respectivas conclusiones.

Una variable se puede clasificar según su valor de la siguiente manera:



- **Variables cuantitativas:** Son las variables que pueden medirse, cuantificarse o expresarse numéricamente, es decir, son aquellas que toman valores numéricos. Las variables cuantitativas pueden ser de dos tipos:

  - ✓ **Continuas:** Son aquellas que pueden tomar cualquier valor dentro de un rango numérico determinado.

  - ✓ **Discretas:** Son aquellas que no admiten todos los valores intermedios en un rango. Suelen tomar solamente valores enteros.

- **Variables cualitativas:** Este tipo de variables representan una cualidad o atributo que clasifica a cada caso en una de varias categorías. La situación más sencilla es aquella en la que se clasifica cada caso en uno de dos grupos, son datos dicotómicos o binarios es decir; cuando sólo pueden tomar dos valores posibles. Como resulta obvio, en muchas ocasiones este tipo de clasificación no es suficiente y se requiere de un mayor número de categorías.

En el proceso de medición de estas variables, se pueden utilizar dos escalas:

- ✓ **Escalas nominales**: Ésta es una forma de observar o medir en la que los datos se ajustan por categorías que no mantienen una relación de orden entre sí.

- ✓ **Escalas ordinales**: Son cualidades que representan un orden y jerarquía.

Cuando se realiza un análisis unidimensional, se utilizan técnicas tales como: Las distribuciones de frecuencias y el análisis de las medidas numéricas. A continuación describiremos cada una de ellas.

- ➢ **Distribuciones de Frecuencia.**

La forma de la distribución de los datos de una variable se denomina *distribución de frecuencias*, la cual tiene por objeto la construcción de tablas de frecuencias que podrán utilizarse para una mejor presentación e interpretación de la información contenida en los datos observados en la muestra.
Uno de los primeros pasos que se realiza en cualquier estudio estadístico descriptivo; es la tabulación de resultados; recoger la información de la muestra y resumirla en una tabla, en la que a cada valor de la variable se le asocian determinados números que representan el número de veces que ha aparecido, su proporción con respecto a otros valores de la variable, etc. Estos números se denominan *frecuencias*. En una distribución de frecuencia se pueden encontrar:

- Frecuencia absoluta.
- Frecuencia relativa.
- Porcentaje.
- Frecuencia absoluta acumulada.
- Frecuencia relativa acumulada.
- Porcentaje acumulado.

A continuación se realizará una breve descripción de las frecuencias anteriormente mencionadas:

Frecuencia absoluta:

Se define como el número de veces que aparece en la muestra dicho valor de la variable y se denota por n_i.

Frecuencia relativa:

Se define como el cociente entre la frecuencia absoluta y el tamaño de la muestra, la cual se denota por f_i y se calcula mediante la siguiente fórmula:

$$f_i = \frac{n_i}{n}$$

Donde $n =$ Tamaño de la muestra y $n_i =$ frecuencia absoluta.

Porcentaje:

La frecuencia relativa representa un *tanto por uno*, sin embargo, hoy en día, es bastante común hablar siempre en términos de *tantos por ciento o porcentajes*, por lo que esta medida resulta de multiplicar la frecuencia relativa por 100 la cual se denota por p_i y se calcula mediante la fórmula:

$$p_i = f_i \times 100$$

Frecuencia Absoluta Acumulada:

Se denomina frecuencia absoluta acumulada del valor x_i a la suma de las frecuencias absolutas de los valores inferiores o iguales a él. Se denota por N_i y se calcula mediante la siguiente fórmula:

$$N_i = \sum_{j=1}^{i} n_j$$

Frecuencia Relativa Acumulada:

Al igual que en el caso anterior la frecuencia relativa acumulada, es la frecuencia absoluta acumulada dividido por el tamaño de la muestra, la cual se denota por F_i y se calcula mediante la siguiente fórmula:

$$F_i = \frac{N_i}{n} = \frac{\sum_{j=1}^{i} n_j}{n} \quad \text{ó} \quad F_i = \sum_{j=1}^{i} f_j$$

Porcentaje Acumulado:

Similarmente se define el porcentaje acumulado y se denota por P_i; como la frecuencia relativa acumulada por 100, la cual se calcula mediante la siguiente formula:

$$P_i = F_i \times 100$$

La forma de resumir lo expuesto anteriormente, es mediante la representación estructurada en forma de tabla, en la cual se presenta de manera ordenada los distintos valores de una variable en estudio y sus correspondientes frecuencias. Su forma más común es la siguiente:

Tabla1. Distribución de frecuencia para variables discretas.

Valores de la Variable x_i	Frecuencias Absolutas n_i	Frecuencias Relativas $f_i = \frac{n_i}{n}$	Frecuencias Absolutas Acumuladas $N_i = \sum_{j=1}^{i} n_j$	Frecuencias Relativas Acumuladas $F_i = \sum_{j=1}^{i} f_j$
x_1	n_1	f_1	$N_1 = n_1$	$F_1 = f_1$
x_2	n_2	f_2	$N_2 = N_1 + n_2$	$F_2 = F_1 + f_2$
....	....	....	....	....
x_i	n_i	f_i	$N_i = N_{i-1} + n_i$	$F_i = F_{i-1} + f_i$
....	....	....	....	....
x_k	n_k	f_k	$N_k = n$	$F_k = 1$
	$\sum_i n_i = n$	$\sum_i f_i = 1$		

La elaboración de la distribución de frecuencias de una variable continua plantea algunos problemas que no se dan en el caso de variables discretas. Se trata de decidir el número de intervalos en los que hay que agrupar los valores de la variable, así como la amplitud o recorrido de los mismos debe ser igual. Estas cuestiones no tienen una respuesta determinada de antemano. La solución dependerá de cada caso concreto, por lo que no tiene

sentido entrar en detalle de las distintas situaciones que pudieran darse. Otro problema surge cuando un valor de la variable coincide exactamente con un extremo del intervalo, con lo que hay dudas sobre su inclusión en este intervalo o el siguiente. Como solución a este problema es habitual proceder a definir intervalos abiertos por la izquierda y cerrados por la derecha.

Antes de presentar la tabla de distribución se definen los siguientes conceptos:

a) Si se define el intervalo i-ésimo como $(L_{i-1} - L_i]$ las fronteras del intervalo, se llaman límites inferior y superior de la clase y se denotan por L_{i-1}, L_i respectivamente.

b) Amplitud de la clase: Es la diferencia entre el límite superior e inferior del intervalo. Así para el intervalo i-ésimo, la amplitud vendrá dada por:

$$a_i = L_i - L_{i-1}$$

c) Marca de clase: Es el punto central de cada intervalo y suele representarse por x_i, por lo que para el intervalo i-ésimo será:

$$x_i = \frac{(L_i + L_{i-1})}{2}$$

En general, una distribución de frecuencias para una variable continua será como la que se presentará a continuación:

Tabla 2. Distribución de frecuencia para variables continúas.

Valores intervalos $(L_{i-1} - L]_i$	Amplitud a_i	Marca de Clase x_i	Frecuencias Absolutas n_i	Frecuencias Relativas $f_i = \frac{n_i}{n}$	Frecuencias Absolutas Acumuladas $N_i = \sum_{j=1}^{i} n_j$	Frecuencias Relativas Acumuladas $F_i = \sum_{j=1}^{i} f_j$
$(L_0 - L_1]$	a_1	x_1	n_1	f_1	$N_1 = n_1$	$F_1 = f_1$
$(L_1 - L_2]$	a_2	x_2	n_2	f_2	$N_2 = N_1 + n_2$	$F_2 = F_1 + f_2$
$(L_2 - L_3]$	a_3	x_3	n_3	f_3	$N_3 = N_2 + n_3$	$F_3 = F_2 + f_3$
....	....	....	....	....	....	....
$(L_{i-1} - L_i]$	a_i	x_i	n_i	f_i	$N_i = N_{i-1} + n_i$	$F_i = F_{i-1} + f_i$
....	....	....	....	....	....	....
$(L_{k-1} - L_k]$	a_k	x_k	n_k	f_k	$N_k = n$	$f_k = 1$
			$\sum_i n_i = n$	$\sum_i f_i = 1$		

En resumen, se utilizan las distribuciones de frecuencias para realizar un análisis univariado de la información obtenida de la muestra, cuya finalidad es representar de forma estructurada y en forma de tabla, toda la información que se ha recogido sobre las variables en estudio.

- **Gráficos Estadísticos.**

En estadística se denomina gráficos, a las imágenes que combinando la utilización de sombreado, colores, puntos, líneas, símbolos, números, texto y un sistema de referencia como coordenadas que nos permiten presentar información cuantitativa y cualitativa.
La utilidad de los gráficos es doble, ya que pueden servir como sustituto a las tablas, o bien constituyen por sí mismos una poderosa herramienta para el análisis de los datos, siendo en ocasiones el medio más efectivo no sólo para describir y resumir la información, sino también para analizarla.
Los gráficos son medios popularizados y a menudo los más convenientes para presentar datos, se emplean para tener una representación visual de la totalidad de la información; en forma de dibujo de tal modo que se pueda percibir fácilmente los hechos esenciales y compararlos con otros.
A continuación se describe cada uno de los gráficos que se utiliza según el valor que toma la variable en estudio.

Gráficos para Variables Cualitativas o Atributos.

- **Diagrama de barras:** Recibe el nombre de diagrama de barras el gráfico que asocia a cada valor de la variable una barra, generalmente vertical y proporcional a las frecuencias absolutas o relativas con que se presenta.

- **Diagrama de Sectores:** Se construye tomando un círculo, el cual divide en tantos sectores como clases se tengan, siendo el arco del círculo proporcional a las frecuencias absolutas; también se puede realizar con las frecuencias relativas o porcentajes.

Para determinar el arco circular que corresponde a cada clase se relaciona el total de observaciones con los 360º grados de la circunferencia, cada sector debe tener un área proporcional a su frecuencia que suele venir indicada en tanto por ciento.

El ángulo en grados del sector circular correspondiente al valor i-ésimo observado viene dado por:

$$\alpha_i = 360º \times f_i$$

- **Pictograma:** Quizás es el tipo de gráfico más atractivo a la vista, pues en él aparecen dibujos que hacen alusión al fenómeno estudiado, mediante su tamaño, forma, etc.

  Para realizarlo se representan a diferentes escalas un mismo dibujo teniendo en cuenta que el perímetro del dibujo tiene que ser proporcional a la frecuencia, pero esto puede incurrir en un efecto visual engañoso, ya que a frecuencia doble corresponde un dibujo de área cuádruple, con lo cual tiene un inconveniente debido a la falta de precisión.

  A pesar de este inconveniente este tipo de gráfico, es muy utilizado por los medios de comunicación a la hora de hacer que el público no especializado comprenda temas complejos; sin necesidad de dar una explicación complicada.

Gráficos para Variables Cuantitativas.

Para este tipo de variables, se tiene diferentes gráficos según el tipo de frecuencia que se utilice, y además se debe de tener en cuenta si la variable es discreta o continua.
Según el tipo de frecuencia usada se dividen en:

a) **Diagramas diferenciales:** Representan el número o porcentaje de elementos de una modalidad. Se representan a partir de las frecuencias absolutas o relativas.

b) **Diagramas integrales:** Representan el número de elementos de una modalidad inferior o igual a la dada. Se representan a partir de las frecuencias acumuladas. Este tipo de diagramas no tiene ningún sentido para variables cualitativas.

Gráficos para Variables Cuantitativas Discretas.

- **Diagrama de barras:** Su representación es idéntica a la explicada para variables cualitativas, las barras deben de ser estrechas para mostrar que los valores que toma la variable son discretos. Se usan cuando se pretende hacer un diagrama diferencial utilizando variables discretas.

En el caso de realizar un diagrama integral, es decir, usando frecuencias acumuladas, las barras aparecen formando una escalera.

Gráficos para Variables Cuantitativas Continuas.

- **Histograma**: Es una representación gráfica de una variable en forma de barras, el cual se construye a partir de la tabla de distribución; en el eje de las abscisas se construyen unos rectángulos que tienen por base la amplitud del intervalo y el criterio para calcular la altura de cada rectángulo, es el de mantener la proporcionalidad entre las frecuencias absolutas o relativas de cada intervalo y el área de los mismos.

- **Polígono de frecuencias.** Se construye fácilmente una vez representado el histograma, y consiste en unir los puntos del histograma que corresponden a las marcas de clase de cada intervalo por segmentos de rectas.
  La diferencia esencial entre los histogramas y los polígonos de frecuencias es que estos últimos proporcionan una representación más suavizada de la distribución de frecuencias.

En resumen, dependiendo de la variable estadística que se analice; así se realizará el tipo de gráfico correspondiente, para hacer un análisis visual de la información contenida en la población o muestra, de forma de que ésta información, se pueda observar de una manera más sistemática y resumida.

Por lo tanto en la tabla 3, se resume el tipo de gráfico a realizar, según el tipo de variable estadística que se utilice en cada caso.

Tabla 3. Gráficos según el tipo de variables.

Tipo de Variable	Diagrama o Gráfico
Cualitativa	Barras, Sectores, Pictogramas
Cuantitativa (discreta)	Diferencial (barras) Integral (escalera)
Cuantitativa (continua)	Diferencial (histograma, polígono de frecuencias) Integral (diagramas acumulativos)

- **MEDIDAS NUMÉRICAS.**

Medidas de Posición Central.

Las medidas de posición central tienen como objetivo, el sintetizar los datos en un valor representativo y además indicar con precisión el centro de un conjunto de observaciones. Algunas de las medidas de tendencia central más utilizadas son la media, la mediana y la moda.

- **La Media o Media Aritmética.**

La media de las observaciones $x_1, x_2, ..., x_n$; es la suma de todas las observaciones entre el número total de observaciones y se denota por $\bar{x}$; calculándose de la siguiente manera:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

La media, es una medida apropiada de tendencia central para muchos conjuntos de datos. Sin embargo, dado que todas las observaciones se emplea para su cálculo, el valor de la media; puede afectarse de manera desproporcionada por la existencia de algunos valores extremos.

- **La Mediana.**

La mediana de un conjunto de observaciones; es el valor que divide al conjunto de datos en dos partes iguales; en el cual el 50% es menor que él y el otro 50% es mayor que él.

- ✓ Si el total de valores que toma la variable en estudio es impar, al ordenar los datos de forma creciente o decreciente, el valor que ocupa la posición $\frac{n+1}{2}$ corresponde a la mediana; el cual es el valor central del conjunto de datos.

- ✓ Si el total de valores que toma la variable en estudio es par, al ordenar los datos de forma creciente o decreciente, el promedio que ocupan los valores de las posiciones $\frac{n}{2}$ y $\frac{n}{2}+1$ corresponde a la mediana, el cual es el valor central del conjunto de datos.

- **La Moda.**

La moda de un conjunto de observaciones, es el valor de la observación que ocurre con mayor frecuencia en el conjunto, es decir, el valor que más se repite.

En resumen, el propósito principal de las medidas de posición central, será para estudiar las características de los valores centrales de la distribución y para comparar o interpretar cualquier puntaje en relación con el puntaje central o típico.

Medidas de Posición.

- **Cuantiles.**

Los cuantiles son medidas de posición que determinan las ubicaciones de los valores que dividen un conjunto de observaciones en partes iguales, es decir; en intervalos que comprenden la misma cantidad de valores, los cuales pueden ser cuatro, diez o cien partes iguales.

Los cuantiles más usados son:

- ✓ Cuartiles
- ✓ Deciles
- ✓ Centiles o Percentiles

Para algunos valores *u*, se dan nombres particulares a los cuantiles, Q (u):

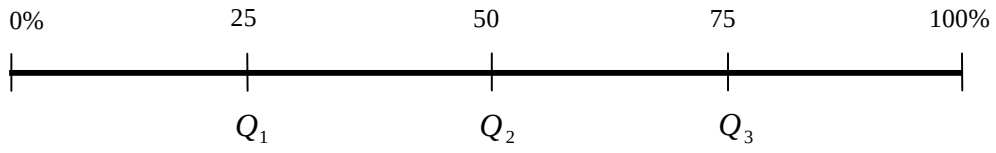
U	***Q(u)***
2	Mediana
1, 2, 3	Cuartiles
1, 2,…., 9	Deciles
1, 2,…,99	Centiles

- **Cuartiles.**

Los cuartiles son los tres valores que dividen al conjunto de datos ordenados en cuatro partes porcentualmente iguales y usualmente son denotados por Q_1, Q_2, Q_3. El segundo cuartil es precisamente la mediana. El primer cuartil, es el valor en el cual o por debajo del cual queda un cuarto de los datos, es decir; un 25% de todos los valores de la sucesión ordenada; el tercer cuartil, es el valor en el cual o por debajo del cual quedan las tres cuartas partes de los datos es decir; un 75% de los datos.

Esquemáticamente se tiene:

Esquema 1. Representación de los Cuartiles.



Q_1 Primer cuartil, Q_2 Segundo cuartil, Q_3 Tercer cuartil

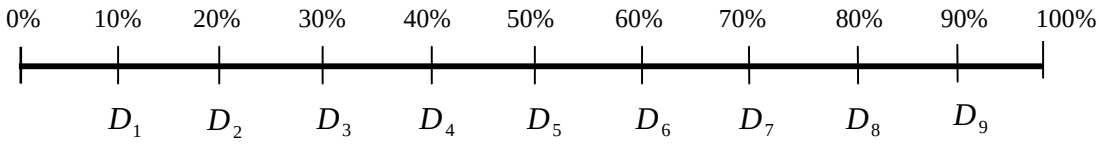
- **Deciles.**

Los deciles son valores que dividen la sucesión de datos ordenados en diez partes porcentualmente iguales, son también un caso particular de los percentiles. Los deciles se denotan por $D_1, D_2, ...D_9$, que se leen primer decil, segundo decil, etc.

El primer decil; Es el valor en el cual o por debajo del cual queda el 10% de los datos y por encima el 90% de ellos; el tercer decil, es el valor en el cual o por debajo del cual queda el 30% de los datos y por encima el 70% de ellos.

Esquemáticamente se tiene:

Esquema 2. Representación de los Deciles.



D_1 Primer decil, D_2 Segundo decil, D_3 Tercer decil,……, D_9 Noveno decil

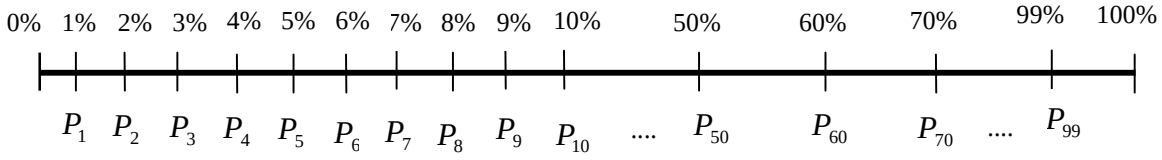
- **Centiles o Percentiles.**

Los percentiles son valores que dividen la sucesión de datos ordenados en cien partes porcentualmente iguales. Los percentiles se denotan por $P_1, P_2 .... P_{99}$, leídos como, primer percentil,..., nonagésimo noveno percentil.

El segundo percentil; es el valor en el cual o por debajo del cual queda el 2% de los datos y por encima el 98% de ellos; el décimo percentil, es el valor en el cual o por debajo del cual queda el 10% de los datos y por encima el 90% de ellos.

Esquemáticamente se tiene:

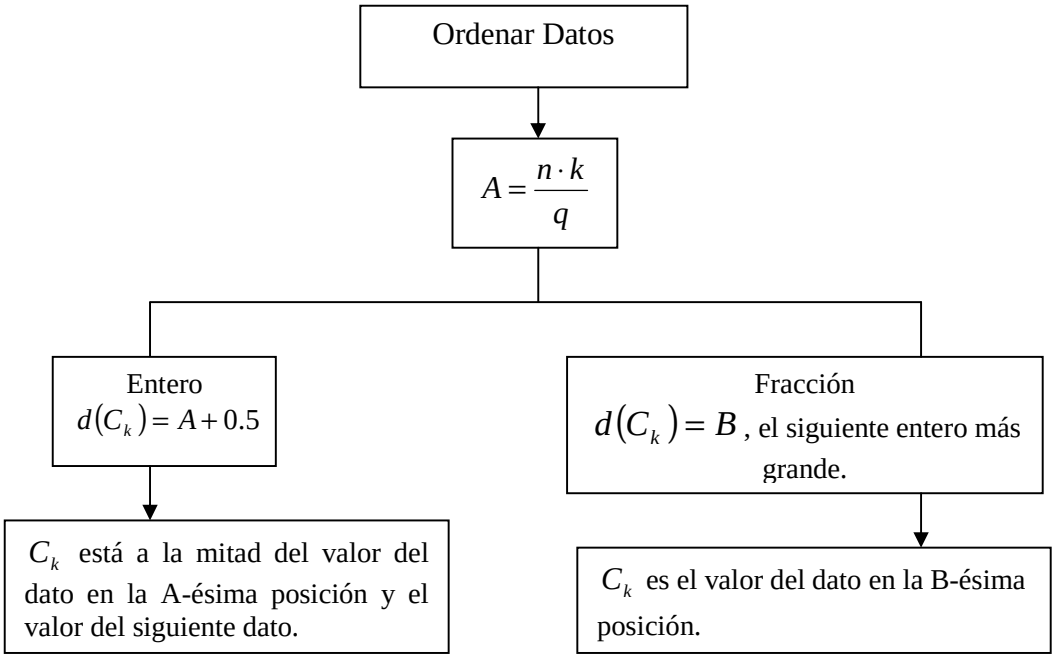
Esquema 3. Representación de los Centiles ó Percentiles.



P_1 Primer percentil, P_2 Segundo percentil, P_3 Tercer percentil,……, P_{99} Nonagésimo noveno percentil.

En forma general se puede calcular los cuantiles mediante el siguiente esquema:

Esquema 4. Algoritmo para calcular los Cuantiles.



Nota: Cuando se desee calcular los cuartiles el valor de k y q será: $k = 1,2,3$ y $q = 4$; para los deciles $k = 1,2,3...,9$ y $q = 10$; para los percentiles $k = 1,2,3...,99$ y $q = 100$.

En resumen, se utilizará las medidas de posición, para proporcionar diferentes medidas cuantitativas de donde está el centro de los datos, y además ayudarán para encontrar medidas que sinteticen las distribuciones de frecuencias. En vez de manejar todos los datos sobre las variables, tarea que puede ser incomoda, se puede caracterizar su distribución de frecuencias mediante algunos valores numéricos, eligiendo como resumen de los datos un valor central alrededor del cual se encuentran distribuidos los valores de la variable. La descripción de un conjunto de datos, incluye como un elemento de importancia la ubicación de éstos dentro de un contexto de valores posibles.

- **Medidas de Dispersión.**

Las medidas de dispersión se utilizan para obtener información complementaria a las medidas de tendencia central y miden la forma de como se distribuyen los datos que integran una población o muestra. Las medidas de dispersión más usadas son: El rango, la varianza y la desviación estándar, las cuales se describen a continuación:

- **El Rango.**

El rango, es la diferencia entre la observación más grande (máximo) y la más pequeña (mínimo) de los datos de una distribución estadística, el cual se calcula mediante la siguiente fórmula:

$$R = N_{máx} - N_{\min}$$

- **La Varianza.**

La varianza de las observaciones $x_1, x_2, ..., x_n$; es en esencia, el promedio del cuadrado de las distancias entre cada observación y la media del conjunto de observaciones, con el fin de eliminar los signos negativos, es decir; sumando todos los cuadrados de las diferencias de cada valor respecto a la media y dividiendo este resultado por el número de observaciones que se tengan. Si la varianza es calculada a una población (Total de componentes de un conjunto), la cual se denota por σ^2, y se calcula mediante la siguiente fórmula:

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

Donde σ^2 representa la varianza poblacional, x_i representa cada uno de los valores, μ representa la media poblacional y N es el número de observaciones o tamaño de la población. En el caso que estemos trabajando con una muestra la ecuación que se debe emplear es:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Donde s^2 representa la varianza muestral, x_i representa cada uno de los valores, $\bar{x}$ representa la media de la muestra y n es el número de observaciones o tamaño de la muestra.

Si se observa la ecuación anteriormente descrita, se muestra que se le resta uno al tamaño de la muestra en el denominador; esto se hace con el objetivo de aplicar una pequeña medida de corrección a la varianza, intentando hacerla más representativa para la población aunque por lo general se utiliza como denominador solo n. Es necesario resaltar que la varianza da como resultado el promedio de la desviación, pero este valor se encuentra elevado al cuadrado.

Por lo tanto, la varianza es una medida razonablemente buena de la variabilidad de los datos; debido a que si muchas de las diferencias son grandes o pequeñas, entonces el valor de la varianza será grande o pequeño. El valor de la varianza puede sufrir un cambio muy desproporcionado; aún, más que la media por la existencia de algunos valores extremos en el conjunto.

- **Desviación Estándar**

Es la raíz cuadrada positiva de la varianza.

La expresión de la desviación estándar poblacional se calcula mediante la siguiente fórmula:

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

Luego la expresión de la desviación estándar muestral se calcula de la siguiente manera:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

La varianza y la desviación estándar, no son medidas de variabilidad distintas, debido a que la última no puede determinarse a menos que se conozca la primera.

En resumen, las medidas de dispersión se utilizan para medir el grado de dispersión que tiene una variable estadística en torno a una medida de posición o tendencia central, y además indica la representatividad de dicha medida.

1.1.2 ANÁLISIS BIVARIADO.

Una distribución se denomina bidimensional si para cada elemento de una población o muestra se considera las medidas referentes a dos variables. Estas dos variables pueden ser del mismo tipo ambas cuantitativas o cualitativas, o bien una de ellas puede ser cualitativa y la otra cuantitativa.

El interés del estudio bidimensional frente a los unidimensionales, reside en la información adicional que proporciona la observación simultánea de ambas variables, como la posible existencia de dependencias estadística y funcional entre las dos variables.

La relación funcional y estadística entre ambas variables se describe a continuación:

- **Relación Funcional**: Cuando es posible predecir con exactitud los valores de una variable a partir de los de la otra, entonces se dice que ambas variables están en relación funcional. Dada las variables (X,Y) existirá una función $f(x)$ tal que $y_i = f(x_i)$. Para cada valor de x_i se puede conocer el valor de y_i.

- **Relación Estadística**: Cuando no es posible expresar mediante una función matemática la relación existente entre ambas variables; es decir; cuando los valores que toma una de las variables están relacionados con los valores que toma la otra, pero no de manera exacta.

DISTRIBUCIONES BIDIMENSIONALES.

Cuando se quiere describir conjuntamente dos variables estadísticas, el primer paso será representar los datos en una tabla de distribución de frecuencias. Ahora, a cada caso le corresponde no un valor, si no dos (uno para cada una de las variables), los pares de valores así formados constituyen la distribución bidimensional.

La distribución de frecuencias es una tabla de doble entrada, en la que se recogen tanto las frecuencias de cada una de las variables por separado, como los pares de valores que cada caso obtiene en ambas variables (Frecuencia conjunta).

Los valores de cada variable pueden aparecer sin agrupar o agrupadas en intervalos, no teniendo por qué ser el número de intervalos de las dos variables iguales entre sí, precisamente como la amplitud de los mismos.

Considérese la variable X que toma los valores $x_1, x_2, ..., x_k$ y la variable Y que toma los valores $y_1, y_2, ..., y_p$. Ahora, la distribución de frecuencias viene determinada por las parejas (x_i, y_j) de valores y sus correspondientes frecuencias absolutas o número de veces que se repiten dichas parejas. Análogamente como el caso unidimensional se pueden definir las frecuencias relativas y acumuladas.

❖ Representación Tabular.

Las representaciones tabulares más usuales en el caso bidimensional son:

1. Cuando el número de observaciones es pequeño, las variables se pueden presentar en forma de *tabla simple* con dos filas o columnas conteniendo las parejas de valores como se representa a continuación:

Tabla 4. Representación de una tabla simple.

Variable X	x_1	x_2	……	x_n
Variable Y	y_1	y_2	……	y_n

2. Cuando el número de observaciones es grande, pero corresponde a pocas parejas distintas, éstas se pueden presentar en forma de *tabla simple* con tres filas o columnas conteniendo las parejas de valores y sus frecuencias correspondientes como se presenta a continuación:

Tabla 5. Representación de una tabla simple.

Variable X	Variable Y	Frecuencia
x_1	y_1	n_1
x_2	y_2	n_2
.	.	.
.	.	.
.	.	.
x_k	y_k	n_k
		n

3. Cuando hay un gran número de observaciones con parejas distintas, los datos se disponen en una tabla de doble entrada, en la que los valores de cruce de cada fila y columna, representan la frecuencia de la correspondiente pareja de valores como se presenta a continuación en la tabla 6:

Tabla 6. Representación de una Tabla de Doble Entrada.

x \ y	y_1	y_2	.....	y_j	.....	y_s	$n_{i\cdot}$	$f_{i\cdot}$
x_1	n_{11}	n_{12}	.....	n_{1j}	.....	n_{1k}	$n_{1\cdot}$	$f_{1\cdot}$
x_2	n_{21}	n_{22}	.....	n_{2j}	.....	n_{2k}	$n_{2\cdot}$	$f_{2\cdot}$
.	.	.	.....	.	.....	.	.	.
.	.	.		.		.	.	.
.	.	.		.		.	.	.
x_i	n_{i1}	n_{i2}	.....	n_{ij}	.....	n_{ik}	$n_{i\cdot}$	$f_{i\cdot}$
.	.	.	.....	.	.....	.	.	.
.	.	.		.		.	.	.
.	.	.		.		.	.	.
x_r	n_{k1}	n_{k2}	.....	n_{kj}	.....	n_{kk}	$n_{k\cdot}$	$f_{k\cdot}$
$n_{\cdot j}$	$n_{\cdot 1}$	$n_{\cdot 2}$	.....	$n_{\cdot j}$	.....	$n_{\cdot k}$	$n_{\cdot\cdot}$	
$f_{\cdot j}$	$f_{\cdot 1}$	$f_{\cdot 2}$	.....	$f_{\cdot j}$	.....	$f_{\cdot k}$		1

TIPOS DE DISTRIBUCIONES.

Cuando se estudian conjuntamente dos variables, surgen tres tipos de distribuciones: Conjuntas, Marginales y Condicionadas.

a) Distribución Conjunta.

Es la sucesión de los distintos valores de una variable bidimensional (X,Y) en la muestra junto con sus respectivas frecuencias absolutas o relativas.

- *La frecuencia absoluta conjunta*; viene determinada por el número de veces que aparece el par ordenado $(x_i,\ y_j)$ en la muestra o la población, y se representa por n_{ij} .

- *La frecuencia relativa conjunta* del par (x_i , y_j); es el cociente entre la frecuencia absoluta conjunta y el número total de observaciones y se denota por f_{ij} la cual se calcula mediante la siguiente fórmula:

$$f_{ij} = \frac{n_{ij}}{n..}$$

Las siguientes propiedades que se cumplen entre las frecuencias de distribución conjunta son:

1) La suma de las frecuencias absolutas conjuntas, extendida a todos los pares es igual al total de observaciones.

$$\sum_{i=1}^{h}\sum_{j=1}^{k} n_{ij} = n..$$

2) La suma de todas las frecuencias relativas conjuntas extendida a todos los pares es igual a la unidad.

$$\sum_{i=1}^{h}\sum_{j=1}^{k} f_{ij} = 1$$

b) Distribuciones Marginales.

Son distribuciones de frecuencia de cada una de las variables de manera individual o independiente.

- *Frecuencia absoluta marginal del valor* x_i *de* X *:* Es el valor $n_{i.}$ que representa el número de veces que aparece el valor x_i de X e Y toma cualquiera de sus valores posibles en la muestra. De forma que:

$$n_{i.} = n_{i1} + n_{i2} + ... + n_{ik}$$

- *Frecuencia absoluta marginal del valor* y_i *de* Y: Es el valor $n_{.j}$ que representa el número de veces que aparece el valor y_i de Y y X toma cualquiera de sus valores posibles en la muestra. De forma que:

$$n_{.j} = n_{1j} + n_{2j} + \ldots\ldots\ldots + n_{hj}.$$

- *Frecuencia relativa marginal de* x_i *de* X*:* Es la proporción de individuos de la muestra para los que X toma el valor x_i e Y toma cualquiera de sus valores posibles en la muestra. Se denota por $f_{i.}$ y viene dada por:

$$f_{i.} = \frac{n_{i.}}{n..}$$

- *Frecuencia relativa marginal de* y_j *de* Y*:* Es la proporción de individuos de la muestra para los que Y toma el valor y_j *y* X toma cualquiera de sus valores posibles en la muestra. Se denota por $f_{.j}$ y viene dada por:

$$f_{.j} = \frac{n_{.j}}{n..}$$

Las siguientes propiedades que se cumplen entre las frecuencias de distribución marginales son:

Se cumplen las siguientes relaciones entre las frecuencias de distribuciones marginales:

1. La suma de frecuencias absolutas marginales de la variable X, es igual al número de observaciones que componen la muestra.

$$\sum n_{i.} = n..$$

2. La suma de las frecuencias relativas marginales de la variable X, es igual a 1.

$$\sum f_{i.} = 1$$

3. Las dos propiedades anteriores se cumplen también para la variable Y.

c) Distribuciones Condicionadas.

Estas distribuciones son de tipo unidimensional y hay que construirlas en términos de una condición previa. En este sentido se tendrá la distribución de los valores de la variable X, condicionada a que la variable Y tome un valor concreto y viceversa.
Si se define la condicionada de la variable X, entonces los valores que puede tomar esta variable son los mismos que los de la marginal, lo único que cambia son sus frecuencias absolutas que se representan por $n_{i/j}$; si se trata de la condicionada de la variable Y, los valores de esta distribución son los mismos que los de la marginal de Y, pero las frecuencias absolutas son distintas y se representan por $n_{j/i}$.
La distribución condicional no es única, al contrario de lo que ocurre con la marginal, habrá tantas como valores tome la variable condicionante.

- *Condicionando la variable X a un determinado valor de Y*: Se considera a los $n_{.j}$ individuos de la población, como la modalidad y_i de la variable Y, obsérvese en la columna j-ésima de la tabla 6, sus $n_{.j}$ elementos que constituyen una población, la cual es un subconjunto de la población total. Sobre este subconjunto se define la distribución de X; condicionada por y_j, que se representa por X/y_j.

- *Frecuencia absoluta condicionada para $X = x_i$ dado que $Y = y_i$*: Es el número de veces que se repite el valor de x_i teniendo en cuenta solo aquellos valores en que $Y = y_i$.

- *Frecuencia relativa condicionada para X dado que* $Y = y_i$; viene dada por:

$$f_{i\,/j} = \frac{n_{ij}}{n_{.j}}$$

- *Condicionando la variable* Y *a un determinado valor de* X**:** Se considera a los $n_{i.}$ individuos de la población, como la modalidad x_i de la variable X, obsérvese en la columna i-ésima de la tabla 6, sus $n_{i.}$ elementos que constituyen una población, la cual es un subconjunto de la población total. Sobre este subconjunto se define la distribución de Y**;** condicionada por x_i, que se representa por Y/x_i.

- *Frecuencia absoluta condicionada para* $Y = y_i$ *dado que* $X = x_i$*:* Es el número de veces que se repite el valor de y_i teniendo en cuenta solo aquellos valores en que $X = x_i$**.**

- *Frecuencia relativa condicionada para* Y *dado que* $X = x_i$*;* viene dada por:

$$f_{j/i} = \frac{n_{ij}}{n_{i.}}$$

Los datos bivariados constan de valores de dos variables diferentes obtenidas de elementos de la misma población. Cada una de las dos variables puede ser de naturaleza *cualitativa* o *cuantitativa*.

TABULACIÓN DE VARIABLES ESTADÍSTICAS BIDIMENSIONALES.

Según el tipo de cada variable se van a considerar tres tipos de combinaciones:

- Ambas variables cualitativas (atributo).
- Una variable cualitativa (atributo) y otra cuantitativa (numérica).
- Ambas variables cuantitativas (ambas numéricas).

A continuación se describirán las tablas y gráficos para representar cada una de estas combinaciones de dos variables.

- **Dos variables Cualitativas**: Cuando los datos bivariados resultan de dos variables cualitativas (de atributo o categóricas), a menudo los datos se disponen en una tabla cruzada o de contingencia; donde, en las filas se ubican los niveles de una de las variables (atributos) y en las columnas las de la otra; en las celdas resultantes del cruce de las filas y las columnas se incluye el número de elementos de la distribución que presentan ambos niveles, la tabla de contingencia tendría la forma de la tabla 6.

  Donde:

  n_{ij}: Es el Número de elementos de la distribución que presentan el nivel i–ésimo de la variable X y el nivel j – ésimo de la variable Y.

  $n_{i.} = \sum_{j=1}^{k} n_{ij}$ es la frecuencia absoluta del valor x_i de la variable X.

  $n_{.j} = \sum_{i=1}^{h} n_{ij}$ es la frecuencia absoluta del valor y_j de la variable Y.

  $f_{i.} = \frac{n_{i.}}{n..}$ es la frecuencia relativa del valor x_i de la variable X.

  $f_{.j} = \frac{n_{.j}}{n..}$ es la frecuencia relativa del valor y_j de la variable Y.

Las tablas de contingencia a menudo presentan porcentajes (frecuencias relativas). Estos porcentajes pueden estar basados en toda la muestra o en las clasificaciones de la submuestra (renglones o columnas) como se presenta a continuación:

- **Porcentajes basados en el gran total (toda la muestra).**

  La tabla de contingencia que se presenta en la tabla 6, puede convertirse fácilmente en porcentajes del total, al dividir cada elemento de la tabla con el gran total y multiplicar por 100 el resultado. La tabla tendría la siguiente forma:

Tabla 7. Tabla cruzada en porcentajes basados en el gran total.

y / x	y_1	y_2	…	y_j	…	y_s	$n_i.$
x_1	$\left(\frac{n_{11}}{n..}\right)\times 100$	$\left(\frac{n_{12}}{n..}\right)\times 100$	…	$\left(\frac{n_{1j}}{n..}\right)\times 100$	…	$\left(\frac{n_{1k}}{n..}\right)\times 100$	$\left(\frac{n_{1.}}{n..}\right)\times 100$
x_2	$\left(\frac{n_{21}}{n..}\right)\times 100$	$\left(\frac{n_{22}}{n..}\right)\times 100$	…	$\left(\frac{n_{2j}}{n..}\right)\times 100$	…	$\left(\frac{n_{2k}}{n..}\right)\times 100$	$\left(\frac{n_{2.}}{n..}\right)\times 100$
. . .	. . .	. . .	…	. . .	…	. . .	. . .
x_i	$\left(\frac{n_{i1}}{n..}\right)\times 100$	$\left(\frac{n_{i2}}{n..}\right)\times 100$	…	$\left(\frac{n_{ij}}{n..}\right)\times 100$	…	$\left(\frac{n_{ik}}{n..}\right)\times 100$	$\left(\frac{n_{i.}}{n..}\right)\times 100$
. . .	. . .	. . .	…	. . .	…	. . .	. . .
x_r	$\left(\frac{n_{h1}}{n..}\right)\times 100$	$\left(\frac{n_{h2}}{n..}\right)\times 100$	…	$\left(\frac{n_{hj}}{n..}\right)\times 100$	…	$\left(\frac{n_{hk}}{n..}\right)\times 100$	$\left(\frac{n_{h.}}{n..}\right)\times 100$
$n_{.j}$	$\left(\frac{n_{.1}}{n..}\right)\times 100$	$\left(\frac{n_{.2}}{n..}\right)\times 100$	…	$\left(\frac{n_{.j}}{n..}\right)\times 100$	…	$\left(\frac{n_{.k}}{n..}\right)\times 100$	100%

- **Porcentajes basados en los totales por renglón**.

  Los elementos de la tabla 6, pueden expresarse como porcentajes de los totales por renglón al dividir cada elemento del renglón con el total de éste y multiplicar por 100 el resultado. A continuación se presenta la estructura de la tabla para este caso:

Tabla 8. Tabla cruzada en porcentajes basados los totales por renglón.

y / x	y_1	y_2	…	y_j	…	y_s	$n_i.$
x_1	$\left(\frac{n_{11}}{n_{1.}}\right)\times 100$	$\left(\frac{n_{12}}{n_{1.}}\right)\times 100$	…	$\left(\frac{n_{1j}}{n_{1.}}\right)\times 100$	…	$\left(\frac{n_{1k}}{n_{1.}}\right)\times 100$	100%
x_2	$\left(\frac{n_{21}}{n_{2.}}\right)\times 100$	$\left(\frac{n_{22}}{n_{2.}}\right)\times 100$	…	$\left(\frac{n_{2j}}{n_{2.}}\right)\times 100$	…	$\left(\frac{n_{2k}}{n_{2.}}\right)\times 100$	100%
. . .	. . .	. . .	…	. . .	…	. . .	. . .
x_i	$\left(\frac{n_{i1}}{n_{i.}}\right)\times 100$	$\left(\frac{n_{i2}}{n_{i.}}\right)\times 100$	…	$\left(\frac{n_{ij}}{n_{i.}}\right)\times 100$	…	$\left(\frac{n_{ik}}{n_{i.}}\right)\times 100$	100%
. . .	. . .	. . .	…	. . .	…	. . .	. . .
x_r	$\left(\frac{n_{h1}}{n_{h.}}\right)\times 100$	$\left(\frac{n_{h2}}{n_{h.}}\right)\times 100$	…	$\left(\frac{n_{hj}}{n_{h.}}\right)\times 100$	…	$\left(\frac{n_{hk}}{n_{h.}}\right)\times 100$	100%
$n_{.j}$	$\left(\frac{n_{.1}}{n..}\right)\times 100$	$\left(\frac{n_{.2}}{n..}\right)\times 100$	…	$\left(\frac{n_{.j}}{n..}\right)\times 100$	…	$\left(\frac{n_{.k}}{n..}\right)\times 100$	100%

- **Porcentajes basados en los totales por columna.**

  Así mismo; los elementos de la tabla 6, también pueden expresarse como porcentajes de los totales por columna, al dividir cada elemento de la columna entre el total de ésta y multiplicar por 100 el resultado. A continuación se presenta la estructura de la tabla para este caso:

Tabla 9. Tabla cruzada en porcentajes basados en los totales por columna.

y / x	y_1	y_2	…	y_j	…	y_s	$n_i.$
x_1	$\left(\frac{n_{11}}{n_{1.}}\right)\times 100$	$\left(\frac{n_{12}}{n_{2.}}\right)\times 100$	…	$\left(\frac{n_{1j}}{n_{.j}}\right)\times 100$	…	$\left(\frac{n_{1k}}{n_{.k}}\right)\times 100$	$\left(\frac{n_{1.}}{n..}\right)\times 100$
x_2	$\left(\frac{n_{21}}{n_{1.}}\right)\times 100$	$\left(\frac{n_{22}}{n_{2.}}\right)\times 100$	…	$\left(\frac{n_{2j}}{n_{.j}}\right)\times 100$	…	$\left(\frac{n_{2k}}{n_{.k}}\right)\times 100$	$\left(\frac{n_{2.}}{n..}\right)\times 100$
. . .	. . .	. . .	…	. . .	…	. . .	. . .
x_i	$\left(\frac{n_{i1}}{n_{1.}}\right)\times 100$	$\left(\frac{n_{i2}}{n_{2.}}\right)\times 100$	…	$\left(\frac{n_{ij}}{n_{.j}}\right)\times 100$	…	$\left(\frac{n_{ik}}{n_{.k}}\right)\times 100$	$\left(\frac{n_{i.}}{n..}\right)\times 100$
. . .	. . .	. . .	…	. . .	…	. . .	. . .
x_r	$\left(\frac{n_{h1}}{n_{1.}}\right)\times 100$	$\left(\frac{n_{h2}}{n_{2.}}\right)\times 100$	…	$\left(\frac{n_{hj}}{n_{.j}}\right)\times 100$	…	$\left(\frac{n_{hk}}{n_{.k}}\right)\times 100$	$\left(\frac{n_{h.}}{n..}\right)\times 100$
$n_{.j}$	100%	100%	…	100%	…	100%	100%

Como a las variables cualitativas no se les puede someter a operaciones básicas por estar expresadas en escalas nominales u ordinales, no tiene sentido entonces de hablar de medias marginales, condicionadas, varianzas, etc.; sí, se puede calcular la moda en el caso de que se empleara una escala nominal y la mediana si se utiliza escalas ordinales.

Para variables cualitativas, la información de estas, puede ser presentada en un gráfico de barras; este gráfico, se usa para representar las frecuencias observadas en clasificaciones dobles, es decir; cuando son dos los criterios de clasificación, para variables cualitativas o cuantitativas discretas. Su forma de construcción es similar a la del gráfico de barras simples, sólo que en este caso se representan dos variables. El hecho de ser doble, triple, cuádruple, etc., parte del número de niveles que tenga la variable, que no es el criterio principal de clasificación. Las barras que integran una barra múltiple se colocan juntas o ligeramente solapadas.

Medidas de relación entre variables ordinales.

- **Coeficiente de correlación de Spearman.**

  Esta prueba estadística permite medir la correlación o asociación de dos variables y es aplicable cuando las mediciones se realizan en una escala ordinal, aprovechando la clasificación por rangos.

  La metodología para calcular el coeficiente de correlación de Spearman, consiste en ordenar todos los casos para cada una de las variables de interés y asignar un rango consecutivo a cada observación de cada una de las variables por separado. Si la asociación lineal entre ambas variables fuera perfecta, se espera que el rango de la variable X; fuera exactamente igual al rango de la variable Y, por lo tanto el coeficiente se calcula en base a las diferencias registradas en los rangos entre ambas variables, esperando que estas diferencias fueran cero.

Conforme mayores son las diferencias observadas en las ordenaciones de ambas variables, más se alejaría la relación de ser perfecta. Para evitar que las diferencias positivas anulen las diferencias negativas y prevenir así las decisiones equivocadas, el estadístico se calcula en función de la suma de las diferencias elevadas al cuadrado, la cual se denota por ρ y se calcula mediante la siguiente fórmula:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)}$$

Donde:

d_i: Es la diferencia entre el rango del caso i en la variable X, y en la variable Y.

n: Tamaño de la muestra expresada en parejas de rangos de las variables.

Propiedades:

- El coeficiente de correlación de Spearman; se encuentra siempre comprendido entre los valores -1 y 1.
- Si $\rho = 1$, hay correlación directa máxima.
- Si $\rho = -1$, hay correlación inversa máxima.

Medidas de relación entre variables nominales.

En muchos casos la relación entre determinadas variables no puede medirse con una escala cuantitativa, al no cuantificarse numéricamente las variables no se puede hablar de una correlación directa o inversa. Por lo tanto; cuando se dice que dos variables nominales X e Y están relacionadas, se quiere decir; que las proporciones de X son diferentes en cada categoría de Y. Si X e Y no están relacionadas, entonces las proporciones de X serán iguales en las distintas categorías de Y.

Las frecuencias que se espera obtener si X e Y estuvieran relacionadas se les denomina *frecuencias observadas* y a las frecuencias que esperaríamos obtener si X e Y no estuvieran relacionadas se les denomina *frecuencias esperadas.*

- **Ji-Cuadrado**.

  La prueba Ji-cuadrado permite determinar si dos variables cualitativas están o no asociadas. Si al final del estudio concluimos que las variables no están relacionadas se puede decir; con un determinado nivel de confianza previamente fijado, que ambas son independientes.

  Para el cálculo de esta prueba, es necesario obtener las frecuencias esperadas (aquellas que deberían haberse observado si la hipótesis de independencia fuese cierta), y compararlas con las frecuencias observadas en la realidad, la cual se denota por χ^2 y se calcula mediante la siguiente fórmula:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

  Donde:

  O_{ij} : Es la frecuencia conjunta observada en la fila i y la columna j.

  E_{ij} : Es la frecuencia conjunta esperada en la fila i y la columna j, suponiendo independencia entre las variables.

  Para calcular E_{ij}; se emplea la siguiente fórmula:

$$E_{ij} = \frac{n_{i.} \times n_{.j}}{n..}$$

  El estadístico χ^2 mide la diferencia entre el valor que debiera resultar si las dos variables fuesen independientes y el que se ha observado en la realidad. Cuanto mayor sea esa diferencia (y, por lo tanto, el valor del estadístico), mayor será la relación entre ambas variables. El hecho de que las diferencias entre los valores observados y esperados estén elevadas al cuadrado en el estadístico χ^2 convierte cualquier diferencia en positiva.

Las hipótesis nula a contrastar, será la independencia entre las variables; siendo la hipótesis alternativa la dependencia entre las variables.

El valor de Ji-Cuadrado calculado se compara con un valor tabulado de una χ^2 de tabla, para un nivel de confianza determinado; y $(n-1)(k-1)$ grados de libertad. Donde n; representa el número de filas de la variable X, y k está definido como el número de columnas de la variable Y.

Si el valor calculado es mayor que el valor de la tabla de una $\chi^2_{(n-1)(k-1)}$, significa que las diferencias en las frecuencias observadas y las frecuencias teóricas o esperadas, son muy elevados y por lo tanto; con un determinado nivel de confianza se concluirá que existe dependencia entre las variables analizadas.

Otra forma para determinar si existe o no dependencia entre las variables analizadas se tiene el p-valor; que usualmente reportan la mayoría de paquetes estadísticos. No es más que la probabilidad de obtener, según esa distribución, un dato más extremo que el que proporciona el test o, equivalentemente, la probabilidad de obtener los datos observados si fuese cierta la hipótesis de independencia. Si el p-valor es muy pequeño (usualmente se considera $p < 0.05$) es poco probable que se cumpla la hipótesis nula y se debería de rechazar.

Interpretación:

- Si $\chi^2 = 0$; las variables son independientes.
- Si $\chi^2 > 0$; las variables están relacionadas entre sí.

❖ **Una variable cualitativa y otra cuantitativa**: Cuando los datos bivariados resultan de una variable cualitativa y otra cuantitativa, el valor cuantitativo es visto como una muestra separada, identificada por los niveles de la variable cualitativa, y estas pueden ser presentadas mediante gráficas de barras apiladas o de caja y bigotes. A continuación se describen cada uno de estos gráficos:

- **Gráfico de barras apiladas**: Muestra todas las series apiladas en una sola barra para cada categoría. El ancho de cada barra es determinado por el total de todos los valores de las variables para cada nivel.

- **Diagramas de cajas o bigotes:** En este tipo de gráfico se representan, los valores máximo y mínimo, los cuartiles inferior y superior (percentil 25 y 75 respectivamente) y la mediana (percentil 50) se representan en una caja rectangular alineada, ya sea horizontal o verticalmente. La caja se extiende del cuartil inferior al superior, y es atravesada de un lado a otro por la mediana. A partir de los extremos de la caja se extienden líneas ("bigotes") hasta los valores máximo y mínimo. En el caso para dos variables se tiene que en el eje de las abscisas, se ubica el nivel o la categoría de la variable cualitativa y se dibuja para cada nivel de esta un gráfico de cajas o bigotes, con los valores correspondientes de la variable cuantitativa para luego comparar los resultados.

Cuando tenemos la información de una variable cualitativa y otra cuantitativa, se dispone de una tabla de contingencia como se mostró anteriormente en la tabla 6.

❖ **Dos variables cuantitativas**: Cuando los datos bivariados resultan de dos variables cuantitativas, se acostumbra expresar los datos matemáticamente emparejados como (X,Y), donde X es la **variable de entrada** (conocida como variable independiente) y la Y es la **variable de salida** (conocida como variable dependiente). Se llaman *emparejados* porque para cada valor X corresponde un valor Y del mismo origen. Estos datos se representan en una tabla de correlaciones, la cual estudia simultáneamente dos variables X e Y; que se representará genéricamente como (x_i; y_j; n_{ij}), donde x_i; y_j, son dos valores cualesquiera y n_{ij} es la frecuencia absoluta conjunta del valor i-ésimo de X con el j-ésimo de Y. La cual tiene la estructura de la tabla 6.

Donde:

n_{11} indica, el número de veces que aparece x_1 conjuntamente con y_1; mientras que n_{12}, muestra la frecuencia conjunta de x_1 con y_2, etc.

Las representaciones gráficas más importantes para las distribuciones bidimensionales de variables cuantitativas son: Diagrama de dispersión, diagrama de frecuencias y el estereograma. A continuación se describen cada uno de ellos.

- **Diagrama de dispersión**: Consiste en la representación de los distintos pares de valores sobre un eje cartesiano. De esta forma, cada par viene representado por un punto del plano xy que forman una nube de puntos. La frecuencia de cada par de puntos puede representarse utilizando distintos tamaños de puntos.

- **Diagrama de Frecuencias**: Este tipo de representación está indicado para variables discreta, el cual consiste en una representación en tres dimensiones, dos para las variables y una tercera para las frecuencias. El resultado son una serie de puntos o barras ubicadas en el punto del plano xy correspondiente al par de valores y cuya altura representa la frecuencia absoluta o relativa.

- **Estereograma**: Se utiliza para representar variables continuas distribuidas en intervalos. Se realiza análogamente al diagrama de frecuencias utilizando paralelepípedos, en vez de barras o puntos, cuyo volumen representa la frecuencia absoluta o relativa correspondiente.

Medidas de relación entre variables cuantitativas.

- **Covarianza**: Se llama covarianza de X e Y a la media aritmética de los productos de las desviaciones de cada variable respecto de la media. También se denomina varianza conjunta o sincronizada de las variables X e Y.

La covarianza, mide la forma en que varían conjuntamente dos variables X e Y, la cual se denota por S_{xy} **y** se expresa mediante la siguiente fórmula:

$$S_{xy} = \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{(x_i - \bar{x})(y_j - \bar{y})n_{ij}}{n..}$$

Donde:

$\bar{x}$: Es la media de la variable X.

$\bar{y}$: Es la media de la variable Y.

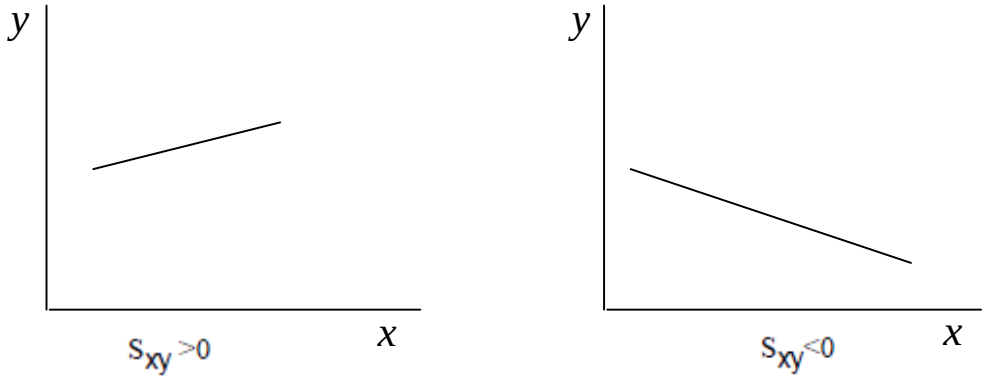
$n..$: Es el número total de pares observados.

n_{ij}: Es la frecuencia absoluta conjunta del valor i-ésimo de X con el j-ésimo de Y

Propiedades:

- Si $S_{xy} > 0$ hay dependencia directa (positiva), es decir; las variaciones de las variables tienen el mismo sentido.
- Si $S_{xy} = 0$ las variables están incorreladas, es decir; no hay relación lineal, pero podría existir otro tipo de relación.
- Si $S_{xy} < 0$ hay dependencia inversa o negativa, es decir; las variaciones de las variables tienen sentido opuesto.

Cuando el valor de $S_{xy} > 0$ ó $S_{xy} < 0$, los puntos se ajustan a una recta, gráficamente tendrá la siguiente forma:



Otra forma de calcular la Covarianza es: $S_{xy} = \dfrac{\sum_{i=1}^{h}\sum_{j=1}^{k} x_i y_j n_{ij}}{n..} - \overline{xy}$

La covarianza no es un parámetro acotado, y puede tomar cualquier valor real, por lo que su magnitud no es importante; lo significativo es el signo que adopte esta.

- **Coeficiente de correlación de Pearson**.

  Es un índice estadístico, que mide la relación lineal entre dos variables X e Y. A diferencia de la covarianza, la correlación de Pearson es independiente de la escala de medida de las variables, este parámetro se mide en términos de covarianza de X e Y y se denota por r_{xy} el cual se expresa mediante la siguiente fórmula:

$$r_{xy} = \frac{S_{XY}}{S_X S_y}$$

  Donde:

  S_{XY}: Es la covarianza de (X,Y).

  S_X y S_y: Son las desviaciones típicas de las distribuciones marginales.

  <u>Propiedades:</u>

  - Es un coeficiente adimensional, es decir; que es independiente de las unidades en que están expresadas las variables. Por ello se utiliza como un valor de comparación, aunque las variables vengan expresadas en unidades diferentes.

  - Si $r_{xy} = 1$, existe una correlación lineal positiva perfecta y directa entre X e Y.

  - Si $r_{xy} = -1$, existe una correlación lineal negativa perfecta entre X e Y.

- Si $r_{xy} = 0$, no existe correlación lineal, por lo que puede existir otro tipo de relación.

- Si $-1 < r_{xy} < 0$, existe correlación lineal negativa y dependencia inversa, mayor cuanto más se aproxime a -1.

- Si $0 < r_{xy} < 1$, existe correlación positiva, y dependencia directa, mayor cuanto más se aproxime a 1.

Interpretación gráfica.

- Si $r_{xy} = 1$, los puntos (X, Y) forman una línea ascendente.
- Si $r_{xy} = -1$, los puntos (X, Y) forman una línea descendente.
- Si $r_{xy} = 0$, la nube de puntos sigue una distribución totalmente aleatoria (circular).
- Si $r_{xy} < 0$, los puntos (X, Y) forman una nube de puntos descendente más cerca a una recta, cuando más cercano sea este valor a -1.
- Si $r_{xy} > 0$, los puntos (X, Y) forman una nube de puntos ascendente más cerca a una recta, cuando más cercano sea este valor a 1.

CAPÍTULO II: ANÁLISIS DE SUPERVIVENCIA.

PRÓLOGO.

El origen del análisis de supervivencia se remonta al siglo XVII; con la construcción de las Tablas de Vida, y se debe primordialmente a que en muchas aplicaciones, el suceso de interés era entonces la muerte. No obstante el Análisis de Supervivencia que actualmente se estudia, se desarrolló inicialmente mediante aproximaciones semiparamétricas (Cox, 1972; Prentice y Gloeckler, 1978).

Se conoce como análisis de supervivencia, al análisis de datos en el que se recoge el período de tiempo que transcurre desde un punto de partida fijado hasta el momento en que acontece un suceso particular. Además las técnicas de Análisis de Supervivencia se fundamentan en dos distribuciones de probabilidad específica, las cuales son las funciones de Supervivencia y de Riesgo.

El análisis de supervivencia centra el interés, en uno o varios grupos de individuos para los cuales se define un evento, a menudo llamado fracaso o falla, que ocurre después de un intervalo de tiempo, llamado tiempo de fracaso o falla. Para determinar el tiempo de fracaso se hace necesario definir los siguientes tres requerimientos:

- ✓ Un tiempo inicial, que debe estar inequívocamente definido.
- ✓ Una escala para medir el transcurso del tiempo.
- ✓ Tener bien definido el evento o fracaso falla.

Usualmente, existe una definición clara del final del período de observación, pero el comienzo es menos evidente. Por ejemplo, rara vez se conoce el momento exacto de inicio de la enfermedad de un individuo, por lo tanto, la fecha de diagnóstico es a menudo, una alternativa para resolver este problema.

En muchos estudios de supervivencia, cuando se llega al final del período de observación fijado por el investigador previamente (o en el transcurso del estudio), hay individuos a los cuales no les ha ocurrido el evento de interés, por lo tanto; no se conoce el tiempo real de supervivencia para esos individuos, sólo se conoce el tiempo de supervivencia hasta el final del estudio; a tales tiempos de supervivencia se les llama tiempos censurados o observaciones incompletas. También ocurre, en algunos casos, que los individuos abandonan el estudio antes de concluir el período de análisis por motivos ajenos a la investigación, por ejemplo, los cambios de domicilio, las muertes por otras causas como

son los accidentes, etc.; estos tiempos también se consideran como censurados. Los tiempos censurados indican que el período de observación, es más corto que el tiempo de supervivencia real. Los datos censurados contribuyen con información valiosa y no deben ser omitidos en el análisis.

Debido a la existencia de datos censurados, en las investigaciones clínicas, los datos relevantes para un análisis de supervivencia son: El estado del sujeto en la última observación (respondió o fue censurado) y el intervalo de tiempo de seguimiento del sujeto.

Los períodos de seguimiento en este tipo de análisis son casi siempre diferentes, ya que los pacientes se van incorporando al estudio durante todo el período de la observación, por lo que los últimos en hacerlo, serán observados durante un período de tiempo menor que los que entraron al principio. El tiempo de fracaso de cada individuo es medido a partir de su fecha de entrada al estudio. Para cada paciente se dispone de un tiempo real, que corresponde a la diferencia entre la fecha en la que se incorpora al estudio y su última observación, y de un tiempo *t*; que representa el tiempo de seguimiento del paciente (en años, meses, días, etc.). Un requerimiento universal de los tiempos reales es que son no negativos.

En el análisis de supervivencia, los datos pueden ser analizados utilizando técnicas paramétricas y no paramétricas.

En este capítulo se realizará una introducción a la teoría del Análisis de Supervivencia, presentando las definiciones y el comportamiento gráfico de la función de Riesgo y de Supervivencia, seguidamente de las técnicas paramétricas (Distribuciones: Exponencial, Gamama, weibull y Lognormal) y no paramétricas (Tabla de Vida, Kaplan & Meir y Regresión de Cox).

2.1 TERMINOLOGÍA DEL ANÁLISIS DE SUPERVIVENCIA.

- **Fecha inicial:** Fecha de diagnóstico de inicio del tratamiento o de remisión completa.

- **Fecha de última noticia:** Fecha correspondiente a la última información que se tiene del caso.

- **Seguimiento:** Es la observación de los individuos de un grupo a partir de la fecha inicial, para conocer su estado vital (vivo, fallecido o desconocido).

- **Período de seguimiento:** El tiempo transcurrido entre la fecha de inicio y la fecha de corte del estudio; que determina la duración del estudio.

- **Fecha de finalización del estudio:** Fecha fijada por el investigador para el término del seguimiento de los individuos.

- **Tiempo de supervivencia:** Es el intervalo de tiempo transcurrido entre las fechas de inicio y de última noticia.

- **Censurada:** Ocurre cuando existe pérdida del seguimiento, muerte por otras causas, exclusión del estudio sin haber ocurrido el evento, o no ocurre el evento de interés durante el período de observación.

- **Evento terminal:** Generalmente, se llama evento terminal a la definición o muerte, la fecha de recaída, la fecha de alta, la remisión de la enfermedad, fallo o cualquier otro incidente que pueda tener dos estados bien definidos **“vivo”** y **“fallecido”**.

- **Sujetos "retirados vivos“:** Cuando se termina un estudio, hay sujetos que se han seguido regularmente y que en el momento del cierre del estudio no han presentado el evento terminal.

- **Sujetos "perdidos“:** Son aquellos sujetos, que bien han cambiado de domicilio o porque han fallecido por otras razones no relacionadas con el estudio, también producen tiempos incompletos.

2.2 TIPOS DE DATOS CENSURADOS.

Los datos correspondientes a estudios de Análisis de Supervivencia presentan una particularidad**,** que dificulta su análisis estadístico con los métodos clásicos de análisis exploratorio o estadística inferencial. Esta particularidad es la presencia de datos censurados; ya que solo se conoce el tiempo de fallo para una fracción, que puede ser pequeña, de los individuos de la muestra, mientras que del resto se dispone solo de información parcial, habitualmente que el "tiempo de vida" es mayor que un valor dado.
Los datos censurados pueden clasificarse como se describe a continuación:

- Un dato se dice que esta **censurado a la derecha de *L*,** si se desconoce el valor exacto de la observación y solo se sabe que **éste** es mayor que *L*.

- Un dato se dice que esta **censurado a la izquierda de *L*,** si solo se sabe que la observación es menor que el valor *L*.
  La censura a la derecha es mucho más frecuente que la censura a la izquierda.

- Un dato se dice que está **censurado arbitrariamente o por intervalos** si durante el estudio se observa un individuo en el instante t_i y se registra vivo, pero al volver a observarlo en un instante t_s ya ha fallecido. Suele hablarse de censura arbitraria o por intervalos para referirse a datos que han sido censurados tanto a la derecha como a la izquierda; es decir, de ellos sólo sabemos que entran en el estudio, cuando ésta ya estaba inicializada y salen de la misma, sin haber fallecido, antes de que finalice.

2.3 DATOS CENSURADOS.

Una de las características del análisis de supervivencia es que algunos individuos en el estudio no han experimentado el suceso de interés al final del estudio. Los tiempos exactos de supervivencia de estos individuos son desconocidos, solo se sabe que exceden un cierto valor y son llamados observaciones censuradas o tiempos censurados.

Los tipos de censura se pueden clasificar en tres formas, los cuales se describen a continuación:

- **Censura de tipo I**. El estudio o periodo de observación está fijado, de forma que al final del mismo no todos los individuos han experimentado el suceso de interés. Para ello los tiempos de supervivencia no son conocidos pero se sabe que al menos son iguales a la longitud del periodo de estudio.

- **Censura de tipo II.** El estudio o periodo de observación no se encuentra fijado y la otra opción es seguir el estudio hasta que se registran un número determinado de acontecimientos. En este caso, las observaciones censuradas son iguales a las observaciones no censuradas más grandes.

- **Censura de tipo III.** En muchos estudios el periodo de observación está fijado y los individuos entran en el estudio en diferentes instantes durante dicho periodo. De ellos algunos pueden estar vivos al final del estudio y otros pueden abandonar antes de que finalice el mismo; en ambos casos los tiempos de supervivencia son desconocidos, siendo los tiempos de censura distintos.

La necesidad de que el mecanismo de censura sea independiente de la observación del fenómeno, es un requisito imprescindible para la validez de las conclusiones.
Hay que tener en cuenta que la variable en estudio es el tiempo hasta que ocurre un evento, y está definida por la duración del intervalo temporal entre los instantes en que empieza la observación y ocurre el evento.
En estudios médicos o científicos se desea conocer el tiempo de ocurrencia de un evento específico de interés, ya sea éste beneficioso (curación, alta hospitalaria), perjudicial (muerte, rechazo del trasplante) o incluso indiferente (cambio de tratamiento).
Aparece entonces el concepto de *seguimiento,* ya sea de una enfermedad determinada en un paciente definido o en una serie de estudios de investigación, donde lo que se valora es el tiempo transcurrido desde un momento inicial como el diagnóstico, el inicio de un tratamiento o la aleatorización en un ensayo clínico hasta un tiempo final en el que acaba la recolección de los datos donde se puede demostrar o no la aparición de un suceso. A este tiempo habitualmente se lo conoce como *"tiempo de supervivencia:* la ocurrencia no

significa sólo el concepto de defunción, porque pueden ser también la duración de permeabilidad de un puente.

La característica más importante de este tipo de datos (tiempo transcurrido hasta la aparición del suceso) es que muy probablemente, al final del período de observación, no todos los pacientes habrán presentado el suceso objeto de estudio. Algunos tal vez se hayan perdido del seguimiento por otras causas, sin poder determinar su estado o si se produjo la muerte, pudo deberse a causas distintas a las que se analizan. Además es habitual, que se vayan incorporando nuevos pacientes durante todo el período de observación, por lo cual estos últimos tendrán un período de observación menor que los iníciales y, por lo tanto, la probabilidad de que les ocurra el suceso es menor. También hay que considerar que al finalizar el estudio, habrán pacientes que no presentan el suceso y por lo tanto el tiempo hasta su ocurrencia es desconocido.

Si se define que el tiempo de seguimiento termina antes de producirse la muerte o antes de completar el período de observación, se dice que el paciente está censurado (ver figura 1). Existen tres motivos por los que pueden aparecer los datos censurados:

a) El paciente decide no participar más en el estudio y lo abandona.

b) El paciente se pierde y no hay información.

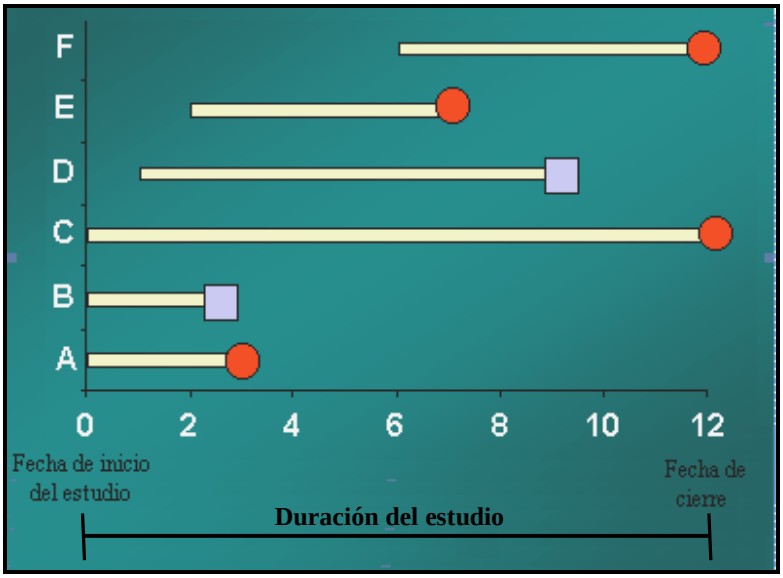
c) El estudio termina antes de aparecer el evento.

Si los tiempos de supervivencia no se conocen con exactitud, los datos se consideran censurados. Las fechas de inicio y cierre son diferentes para cada individuo, pues los pacientes o personas incluidas en el estudio se incorporan en momentos temporales diferentes. En las observaciones censuradas (o incompletas) el evento de interés no se produjo.

Por consiguiente, es necesario definir apropiadamente el origen o inicio del seguimiento, la escala de tiempo a usar y el evento de interés.

Para comprender mejor lo expuesto anteriormente se presenta la siguiente figura 1 donde, la observación no comienza en el mismo instante para todos los individuos.

Figura 1. Esquema de los tiempos de fallo.



La figura 1; corresponde a un estudio de supervivencia en una intervención quirúrgica durante 12 meses. Con el círculo se representan las pérdidas y con el cuadrado las muertes (ocurrencia del evento).

De la figura se puede observar lo siguiente:

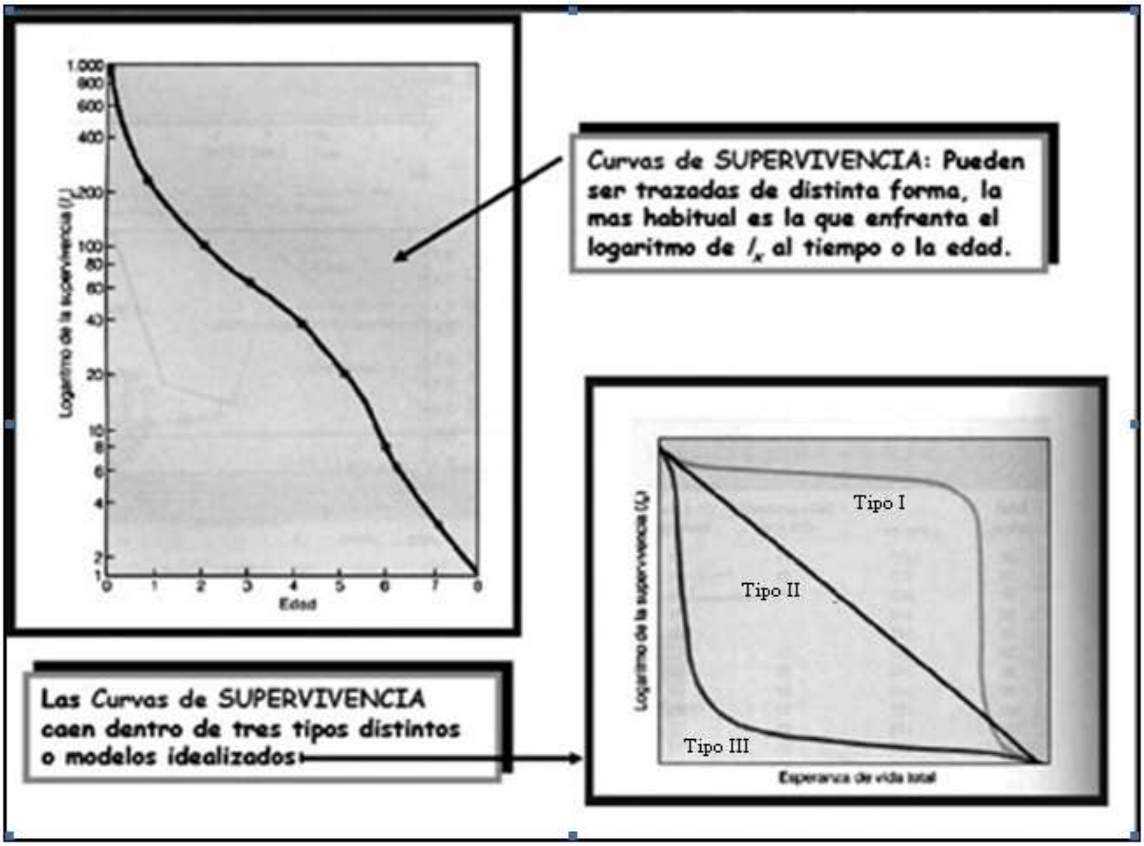
- Individuo **A** desaparece del estudio 3 meses después de la intervención (sería una pérdida en sentido estricto).
- Individuo **B** fallece a los dos meses y medio.
- Individuo **C** sigue vivo al acabar el estudio (sería una pérdida a los 12 meses por fin del estudio).
- Individuo **D**, al que se le interviene en el mes 1, fallece en el noveno mes, el tiempo de supervivencia sería 8 meses (hay 1 mes de pérdida por la izquierda).
- Individuo **E**, al que se le interviene en el 2° mes, se pierde en el 7° mes (sería una pérdida a los 5 meses, ya que hay pérdida en sentido estricto y pérdida por la izquierda).

- Individuo **F**, al que se le interviene en el 6° mes, sigue vivo al acabar el estudio, sería una pérdida a los 6 meses (existe pérdida por fin del estudio y pérdida por la izquierda).

Los datos del estudio pueden entonces estar sesgados por las censuras o las discontinuidades (también llamados truncamientos), estos últimos motivados en que se entró al estudio después del hecho que define el origen en todos los individuos analizados. Se tendría que haber empezado con anterioridad ya que la enfermedad habría comenzado antes.

Existen distintos tipos de curvas de supervivencia las cuales se representan en la figura 2.

Figura 2. Modelos de curvas de supervivencia.



- ✓ La curva Tipo I describe la situación en la que la mortalidad se halla concentrada al final del tiempo de análisis de supervivencia la cual representa una baja mortalidad y es la curva ideal de cualquier tratamiento.

- ✓ La curva Tipo II hay un número constante de fallecidos desde el inicio hasta el final del seguimiento.

- ✓ La curva Tipo III se indica una mortalidad temprana intensa, mientras los individuos que sobreviven tienen una elevada tasa de supervivencia.

Estas curvas son modelos representativos de una curva de supervivencia, aunque en la práctica se presentan combinaciones de ellas.

2.4 FUNCIONES IMPORTANTES DEL ANÁLISIS DE SUPERVIVENCIA.

2.4.1 MODELOS CONTINUOS.

Los modelos continuos se caracterizan por representar la evolución de las variables de interés de forma continua. En general suelen utilizarse ecuaciones diferenciales ordinarias, si se considera simplemente la evolución de una propiedad respecto al tiempo, o bien ecuaciones en derivadas parciales si se considera también la evolución respecto al espacio.
Diremos que una variable aleatoria es continua; si su espacio de resultados es un subconjunto continuo de R. En las definiciones siguientes se considera T como el tiempo transcurrido desde la entrada del individuo al ensayo hasta su salida del mismo y el espacio de resultados es $\xi = [0;\infty)$. La probabilidad de que una variable aleatoria continúa T tome un valor específico t es cero, donde t son los valores específicos que toma la variable aleatoria T.

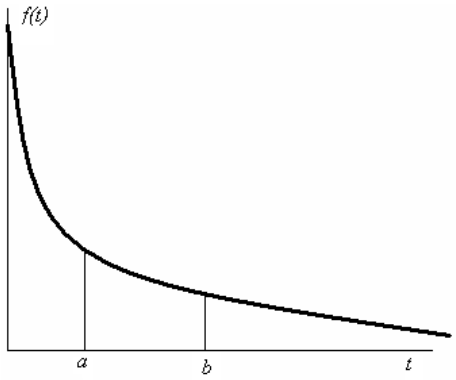
- **Función de Densidad.**

La función $f(t)$, cuya gráfica es la curva límite que se obtiene para un número muy grande de observaciones y para una amplitud de intervalo muy pequeño, es la función de densidad de probabilidad para una variable aleatoria continua T, ya que la escala vertical se elige de tal manera que el área total bajo la curva es igual a uno. La función de densidad de probabilidad de una variable aleatoria continua T se define formalmente de la siguiente manera:

Si f (t) es la función de densidad de probabilidad de la variable aleatoria continua ***T,*** entonces para cualquier $a,\ b \in R$ se tiene:

1. $f(t) \geq 0,\ -\infty < t < \infty,$
2. $\int_{-\infty}^{\infty} f(t)dt = 1$
3. $P(a \leq T \leq b) = \int_{a}^{b} f(t)dt$

La gráfica de la Función de Densidad se representa de la siguiente manera:



Donde los valores más elevados de f (t), se sitúan en la zona en la que falla una mayor proporción de los individuos en la población original, además se observa en la gráfica que f es monótona decreciente.

- **Función de Distribución.**

Sea T una variable aleatoria de tipo continuo que toma un número infinito de valores sobre la recta real y cuya función de densidad es $f(t)$. Se define la función de distribución acumulativa de la variable aleatoria T, que se denota por $F(t)$, como la probabilidad de que la variable aleatoria continua T tome valores menores o iguales a t, es decir:

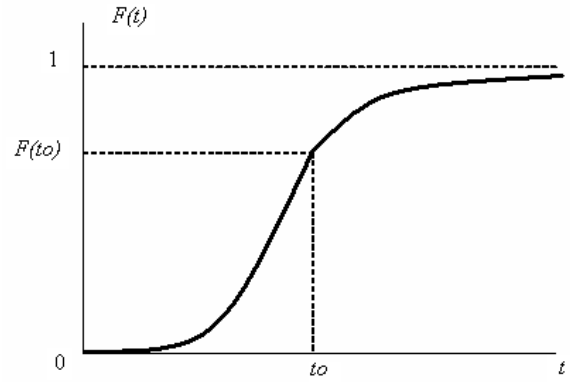
$$F(t) = P(T \leq t) = \int_{-\infty}^{t} f(t)dt$$

Por lo tanto; la función de distribución acumulativa $F(t)$; es el área acotada por la función de densidad $f(t)$ que se encuentra a la izquierda de la recta $T = t$. La función $F(t)$; proporciona la probabilidad de que ocurra el evento de interés antes o en un momento *t*, del período de estudio que se ha fijado.

La distribución acumulativa $F(t)$; es una función no decreciente de los valores de la variable aleatoria con las siguientes propiedades.

1. $F(-\infty) = 0;$
2. $F(\infty) = 1;$
3. $P(a \leq T \leq b) = \int_a^b f(t)\,dt = F(b) - F(a);$
4. $\frac{dF(t)}{dt} = f(t).$

Las propiedades anteriormente descritas son fáciles de verificar, siendo la última expresión una consecuencia del teorema fundamental del cálculo integral.
La grafica de la Función de Distribución viene dada de la siguiente manera:



Donde el $F(t_0)$ es la proporción de individuos en la población cuya duración hasta el fallo es inferior a t_0, además $F(t)$ es creciente y cumple que $P(a < T < b) = F(b) - F(a).$
Se observa en la grafica que la función $F(t)$ es monótona no decreciente y verifica que $\lim_{t \to \infty} F(t) = 1$. En este caso, por ser *T* no negativa, $F(t) = 0$ para $t < 0$.

2.5 FUNCIÓN DE SUPERVIVENCIA.

Sea T una variable aleatoria no negativa, que representa el tiempo de vida de los individuos de alguna población. Usualmente, se asume que T es continua, en el intervalo $[0,\infty)$.

Sea $f(t)$ la función de densidad de probabilidad de T; donde la función de distribución acumulada se define de la siguiente manera:

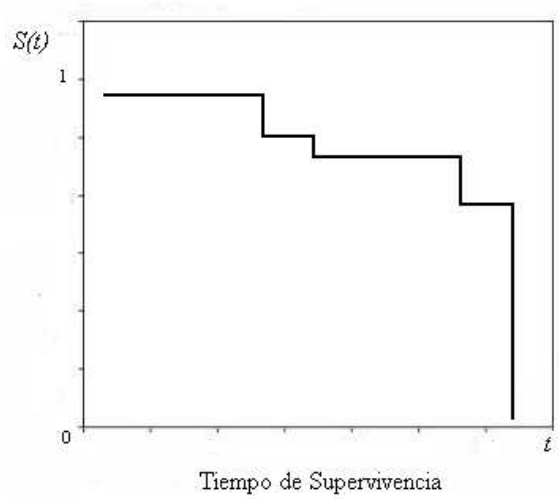
$$F(t) = \Pr(T \le t) = \int_0^t f(x)dx$$

Además la función de supervivencia $S(t)$ es la complementaria de la función de distribución y se define como la probabilidad de que una persona sobreviva (no le ocurra el evento de interés) al menos hasta el tiempo t; es decir, que indica la probabilidad de que un individuo supere cierto tiempo de vida.

Entonces se tiene que la función de supervivencia $S(t)$ es:

$$S(t) = 1 - F(t) = P[T > t]$$

La grafica de la Función de Supervivencia viene dada de la siguiente manera:



Tiempo de Supervivencia

Donde $S(t)$ es la función de supervivencia e indica la probabilidad de que un individuo siga vivo al cabo de un periodo de estudio de tiempo, además esta función es monótona no creciente y verifica que $S(0)=1$ *y* $S(\infty)=\lim_{t\to\infty} S(t)=0$.

2.5.1 ERROR ESTÁNDAR E INTERVALOS DE CONFIANZA PARA LA SUPERVIVENCIA.

Si se desea calcular un intervalo de confianza para la estimación de la supervivencia a un determinado tiempo se puede realizar a partir del error estándar de cada estimación de la supervivencia acumulada. Este error estándar para cada tiempo EE_{st} es el producto de la supervivencia estimada para ese tiempo por la raíz de la suma de los cocientes entre el número de fallecidos en cada momento y el producto de supervivientes y pacientes a riesgo en ese tiempo. Es decir,

$$EE_{S_t} = S_t\sqrt{\left(\sum \frac{n_i - S_i}{n_i S_i}\right)}$$

Una aproximación poco fina pero conservadora para estimar los intervalos de confianza al 95% será aplicar la siguiente expresión:

$$IC\ 95\%\ \ para\ S_t = \sup erv_t \pm 1.96 EE$$

(1.96 es el valor Z de la normal para un error alfa bilateral del 5%)

Pero, el método simplista de sumar y restar 1.96 veces el error estándar a la supervivencia estimada no es aconsejable porque proporciona intervalos de confianza que son negativos y otros que exceden de 1.0, lo cual es absurdo. Se puede usar otra expresión más adecuada, calculando un error estándar transformado (EE_t).

2.5.2 MÉTODO RECOMENDABLE PARA ESTIMAR LOS INTERVALOS DE CONFIANZA DE LA SUPERVIVENCIA.

$$EE_t = \sqrt{\frac{1}{\left(\ln[S_t]\right)^2} * \sum \frac{n_i - S_i}{n_i S_i}} \qquad IC\ 95\%\ para\ S_t = S_t^{e^{(\pm 1.96 EE_t)}}$$

Donde ln significa logaritmo natural.

2.6 FUNCIÓN DE RAZÓN DE RIESGOS.

La función de razón de riesgos o tasa instantánea de fallas $h(t)$, representa la evolución de la probabilidad de fallo en relación con la edad de los individuos e indica la probabilidad por unidad de tiempo, de representar el lapso subsiguiente condicionado a que el evento no se haya presentado antes y se define como el cociente entre la función de densidad y la función de supervivencia:

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

Se interpreta como la probabilidad de que a un individuo le ocurra el evento de interés en la siguiente unidad de tiempo Δt, dado que ha sobrevivido hasta el tiempo t.

Dicha función proviene de la tasa media de fallas, dada la probabilidad condicional de fallas en el período $(t, t + \Delta t)$, dado que la persona sobrevive en el período (0; t), la tasa media de fallas (**TMF**) se define como:

$$\textbf{TMF} = \frac{F(t+\Delta t) - F(t)}{\Delta t}\left(\frac{1}{S(t)}\right)$$

Tomando límites para $\Delta t \to 0$, queda:

$$h(t) = \lim_{\Delta t \to 0} TMF = \frac{F'(t)}{S(t)} = \frac{f(t)}{S(t)}$$

La función de riesgo acumulada $H(t)$, se define como:

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

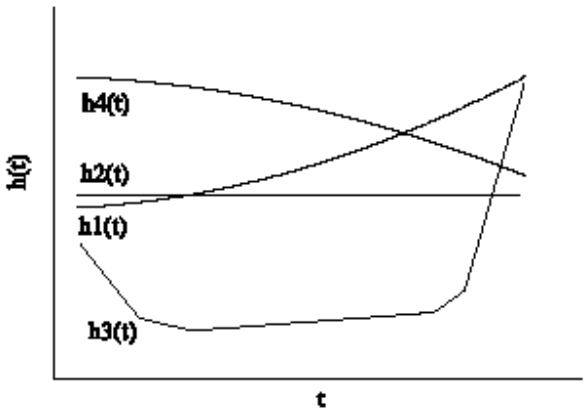
Y esta relacionada con la función de supervivencia mediante la siguiente expresión:

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

Los datos de supervivencia suelen presentarse en la forma (t_i, δ_i), donde t_i es el tiempo de observación y, δ_i =0, si la observación es censurada y δ_i =1 cuando se observa la ocurrencia del evento de interés.

Para una interpretación más amplia de la función de riesgo, se explicarán cuatro ejemplos de poblaciones cuyo perfil de riesgo corresponde a cada uno de los tipos de curvas que se muestran en la figura 3:

Figura 3. Distintas Funciones de Riesgo.



- *h1(t):* El conjunto de personas mayor de 65 años. Esta población presenta una función de riesgo creciente que indica que la tasa de fallo tiende a aumentar con el transcurso del tiempo. Por ejemplo, la probabilidad de que un individuo con 70 años

viva más de 71, es mayor que la probabilidad de que un individuo con 80 años viva más de 81años.

- *h2(t):* Una población de individuos sanos entre los 20 y 40 años de edad, para los que el único riesgo de muerte, en la práctica, viene dado por distintos tipos de accidentes (laborales, deportivos, de tráfico, etc.). En esta población, la función de riesgo es prácticamente constante.
- *h3(t):* Esta población presenta una función de riesgo con forma de $\cup$, llamada también riesgo "bañera", típica de las tablas de vida poblacionales. Inicialmente se tiene un período con tasa de fallo alta, correspondiente a la etapa neonatal e infantil, que va decreciendo hasta estabilizarse. El riesgo permanece bajo y aproximadamente constante, hasta una cierta edad, en torno a los 40 años, a partir de la cual comienza a aumentar con el tiempo.

- *h4(t):* Una población de personas jóvenes que padece cierto defecto congénito y que es sometida a un proceso quirúrgico complicado para corregirlo, analizada mientras dura el periodo de recuperación. Esta población presentará una tasa de riesgo decreciente ya que en estos casos, el principal riesgo de muerte aparece como consecuencia de la intervención o de sus complicaciones inmediatas.

2.7 RELACIÓN ENTRE FUNCIONES.

Las siguientes expresiones muestran la relación que existe entre cada una de las funciones anteriormente descritas.

$$F(t) = \int_0^t f(x)dx$$

$$f(t) = \frac{dF(t)}{dt}$$

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

$$f(t)=\frac{dF(t)}{dt}=\frac{d\left(1-S(t)\right)}{dt}$$
$$f(t)=-\frac{d}{dt}S(t)=-S'(t)$$

$$h(t)=\frac{f(t)}{S(t)}=-\frac{S'(t)}{S(t)} \qquad (1)$$
$$h(t)=-\frac{d}{dt}\ln S(t)$$

Integrando la expresión (1), de 0 a t y usando el hecho que $S(0)=1$ se tiene que:

$$-\int_0^t h(t)dt=\ln S(t)$$

ó

$$S(t)=e^{-\int_0^t h(x)dx}$$

Estas ecuaciones muestran que las funciones contienen la misma información sobre la experiencia de la mortalidad de la generación. Si una de ellas es conocida, las otras dos pueden ser obtenidas.

2.8 MÉTODOS PARA LA ESTIMACIÓN DE LA FUNCIÓN DE SUPERVIVENCIA.

El análisis de datos para estudios de supervivencia requiere métodos de análisis específicos por dos razones fundamentales:

- Los investigadores muy frecuentemente analizan los datos antes de que todos los pacientes hayan muerto, ya que si no habría que esperar muchos años para realizar dichos estudios.

- La segunda razón por la que se necesitan métodos especiales de análisis; es porque típicamente los pacientes no inician el tratamiento o entran al estudio al mismo tiempo.

En el análisis de supervivencia, el análisis de los datos puede ser realizado utilizando métodos paramétricos y no paramétricos como las siguientes:

• **Paramétricos:**

- Distribución Exponencial.
- Distribución de Weibull.
- Distribución Gamma.
- Distribución Lognormal.

• **No paramétricos:**

- Tabla de Vida.
- Kaplan & Meier.
- Regresión de Cox.

A continuación se describirán cada uno de estos métodos paramétricos y no paramétricos:

2.9 MÉTODOS PARAMÉTRICOS.

Los métodos paramétricos son basados en el muestreo de una población con parámetros específicos, como la media (μ), la desviación estándar (σ) o la proporción (*p*). Estos métodos paramétricos usualmente tienen que ajustarse a algunas condiciones completamente estrictas, así como el requisito de que los datos de la muestra provengan de una población normalmente distribuida; es decir que los métodos paramétricos requieren supuestos acerca de la naturaleza o forma de las poblaciones involucradas. Dentro de éstos se pueden diferenciar según su función de riesgo en los modelos Exponenciales, Gamma, Weibull y Lognormal.

Los problemas que plantean los modelos paramétricos tanto por la variación de la tasa de riesgo; como por la variación de las variables explicativas a lo largo del tiempo es solucionado por el modelo de COX, a menudo denominado modelo semiparamétrico o parcialmente paramétrico.

A continuación se describe brevemente cada uno de los métodos no paramétricos:

Distribución Exponencial.

La distribución de probabilidad más sencilla es la que presenta una función de riesgo constante, $h(t)=\lambda$ para $0 \le t < \infty$, con λ una constante positiva. Las restantes funciones que caracterizan este modelo son:

$$H(t)=\lambda t$$
$$S(t)=e^{-\lambda t}$$
$$f(t)=\lambda e^{-\lambda t} \qquad para \quad 0 \le t < \infty.$$
$$F(t)=1-e^{-\lambda t}$$

Esta distribución se denomina Exponencial de parámetro λ. Su media es $1/\lambda$, su varianza $1/\lambda^2$, y su coeficiente de variación la unidad. Una propiedad importante, característica de esta distribución, es la ausencia de memoria; en cualquier instante t, la variable tiempo de vida restante, $R_t = T - t / T \ge t$, sigue también una distribución $Exp(\lambda)$.

Distribución Weibull

En la mayor parte de los fenómenos de interés la hipótesis de que la función de riesgo sea constante, resulta demasiado restrictiva. La distribución de Weibull define un modelo más general cuya función de riesgo es:

$$h(t)=\lambda\gamma(\lambda t)^{\gamma-1} \qquad para \quad 0 \le t < \infty$$

Donde los parámetros λ y γ, se denominan parámetros de escala y forma respectivamente y toman valores positivos. Esta función es siempre monótona; es creciente si γ>1 y

decreciente si γ<1. Si γ=1, la función de riesgo es constante y corresponde al modelo $Exp(\lambda)$.

Las restantes funciones características de la distribución son:

$$S(t) = e^{-(\lambda t)^{\gamma}}$$

$$f(t) = \lambda\gamma(\lambda t)^{\gamma-1} e^{-(\lambda t)^{\gamma}} \qquad para \quad 0 \le t < \infty.$$

$$F(t) = 1 - e^{-(t/\gamma)^{\lambda}}$$

$$H(t) = (\lambda t)^{\gamma}$$

El nombre de esta distribución proviene del físico sueco que la introdujo por primera vez en 1939, en relación con experimentos de resistencia de materiales. La media de la distribución es:

$$E(T) = \frac{\Gamma\left(1+\gamma^{-1}\right)}{\lambda}$$

Donde $\Gamma(x)$ es la función gamma, definida para todo $x > 0$, por la integral,

$$\Gamma(x) = \int_0^{\infty} u^{x-1} e^{-u} du$$

La gran variedad de formas que puede tomar esta distribución depende del valor de γ, y la relativa sencillez de sus funciones, hacen que sea una de las distribuciones más utilizadas en el análisis paramétrico de tiempos de fallo.

Distribución Gamma

La distribución Gamma tiene dos parámetros positivos γ y λ; y su función de densidad es: $f(t) = \frac{\lambda}{\Gamma(\gamma)}(\lambda t)^{\gamma-1} e^{-\lambda t} \quad para \quad t > 0.$

Su media y varianza son $\frac{\gamma}{\lambda}$ y $\frac{\gamma}{\lambda^2}$ respectivamente y su coeficiente de variación es $\frac{1}{\sqrt{\gamma}}$, independiente del valor de λ.

El modelo Gamma es creciente si γ>1 y decreciente si γ<1; en ambos casos $h(t)$ tiende a λ al crecer t. Si γ=1 se obtiene la distribución exponencial. En el caso particular en que γ toma valores enteros, esta distribución suele denominarse de Erlang y aparece con frecuencia en los modelos de teoría de colas. El caso particular $\gamma = n/2$ y $\lambda = 1/2$, corresponde a la distribución χ^2 con n grados de libertad.

Aunque la distribución Gamma es una de las distribuciones continuas que toman valores positivos más importantes, en Fiabilidad no es muy utilizada, ya que las expresiones de las correspondientes funciones de riesgo y supervivencia son complicadas. La distribución de Weibull proporciona en muchos casos resultados similares a los que se obtienen con la distribución Gamma y la inferencia con ella es más sencilla.

Distribución Log-Normal.

La distribución Log-Normal está estrechamente relacionada con la distribución normal de tal manera que la función de densidad de la distribución Log-Normal coincide con la función de densidad de una variable aleatoria cuyo logaritmo sigue una distribución normal.

Otra forma de formular un modelo para el tiempo de supervivencia, T, consiste en especificar una distribución para la variable $Y = \ln(T)$, que toma valores en toda la recta real. Una posibilidad es considerar que $\ln(T)$ tiene una distribución normal de media μ y varianza σ^2; en este caso se dice que T sigue una distribución Log-Normal de parámetros μ y σ. Desafortunadamente, esta distribución presenta el mismo inconveniente que la distribución Gamma, ya que su función de supervivencia no tiene una expresión explícita.

2.10 MÉTODOS NO PARAMÉTRICOS.

Los métodos de estimación no paramétricos más utilizados son el estimador de Kaplan & Meier (1958), y la estimación de las tablas de vida conocida también como estimación actuarial. Mediante estos dos procedimientos se estiman las funciones del análisis de duración, sin realizar ningún supuesto sobre la distribución de la duración. Además, estos métodos pueden ser aplicados a una amplia variedad de situaciones; ya que no tienen los requisitos rígidos de los métodos paramétricos correspondientes. En particular, los métodos no paramétricos no requieren poblaciones normalmente distribuidas y pueden frecuentemente ser aplicados a datos no numéricos, tal como el género de los que contestan una encuesta.

El Análisis de Supervivencia, es la estimación no paramétrica de la función de supervivencia $S(t)$. Esta función es la base para estimar la mayor parte de las funciones y parámetros de interés en el análisis del tiempo de vida.

Si la muestra no contiene observaciones censuradas, la función de supervivencia se estima mediante la función de supervivencia empírica, dicha función se calcula de la siguiente forma:

$$\hat{S}(t) = \frac{N^{o}\ de\ individuos\ en\ la\ muestra\ que\ sobreviven\ al\ instante\ t}{N^{o}\ total\ de\ individuos\ en\ la\ la\ muestra}$$

Este estimador es una función no creciente, toma el valor de uno en todo instante anterior al tiempo de fallo más pequeño y cero a partir del máximo tiempo de fallo observado; la función permanece constante entre dos instantes de fallo consecutivos y presenta un salto descendente en cada tiempo de fallo observado. Si no hay igualdad en la muestra, todos los saltos de la función son de altura $\frac{1}{n}$; mientras que si se observan d tiempos de vida iguales a t_i, el salto de $\hat{S}(t)$ en ese instante será de altura $\frac{d}{n}$.

Cuando en la muestra existen observaciones censuradas la función de supervivencia empírica no es un estimador adecuado, porque tiende a subestimar la función de supervivencia.

Los métodos no paramétricos pueden ser aplicados a una amplia variedad de situaciones porque ellos no tienen los requisitos rígidos de los métodos paramétricos correspondientes. En particular, los métodos no paramétricos no requieren poblaciones normalmente distribuidas y pueden frecuentemente ser aplicados a datos no numéricos, tal como el género de los que contestan una encuesta.

En el análisis de supervivencia, el análisis de los datos puede ser realizado utilizando métodos no paramétricos como las siguientes:

- Tabla de Vida.
- Kaplan & Meier.
- Regresión de Cox.

A continuación se describen cada uno de estos métodos.

2.10.1 TABLA DE VIDA.

Las tablas de vida son un procedimiento clásico para describir la mortalidad que experimenta una población. Este método, cuyo origen se atribuye a Halley (1963), sigue siendo una herramienta muy utilizada en campos como la demografía o los seguros de vida. El objetivo de una tabla de vida es expresar el patrón de mortalidad que experimenta un colectivo de individuos en unas condiciones dadas.

Las tablas de vida en el análisis de supervivencia tienen una estructura análoga a las tablas poblacionales y sirven para estimar la supervivencia de una población a partir de una muestra.

Para crear una tabla de vida se procede como se explica a continuación:

- Se subdivide el intervalo temporal de observación desde el punto inicial en intervalos menores, por ejemplo en años.

- Se contarán las personas que han sobrevivido al menos hasta algún punto de ese intervalo para calcular determinadas probabilidades relacionadas con el momento terminal.

- Las probabilidades se utilizan para calcular o estimar la probabilidad genérica de que una persona viva en un momento determinado.

Lo expuesto anteriormente, presenta dos tipos de dificultades a la hora de efectuar el análisis de los datos los cuales son:

- El origen de tiempo no es el mismo para los diferentes individuos objeto de estudio.

- La ausencia de información, en relación al tiempo de supervivencia, de algunos de los individuos objeto de estudio.

Estas dificultades serán corregidas por el conjunto de métodos y técnicas propios del análisis de supervivencia.

Para construir una tabla de vida han de cumplirse las siguientes hipótesis:

1. Las condiciones experimentales de supervivencia no cambian a lo largo del estudio.

2. Un individuo o máquina que se comienza a estudiar en el momento ha de responder de la misma forma que si se hubiera introducido en el estudio por ejemplo hace cinco años.

3. Las observaciones censuradas no difieren de las que no lo son. De no ser así, esto significaría que las muertes o fallos se producen de manera no aleatoria influenciada por una variable que no se ha tomado en cuenta.

2.10.1.1 REPRESENTACIÓN SIMBÓLICA DE LOS DATOS.

La matriz de datos del diseño se puede representar como la Tabla 1, donde una de las variables indicará el tiempo de supervivencia (t), y la otra el estado de entidad (δ), en el momento del cierre del estudio, codificada conforme a los siguientes criterios:

- $1 \rightarrow$ muerte/fallo
- $0 \rightarrow$ censura.

Cada uno de los n individuos o unidades estadísticas que son objeto de estudio corresponden a un caso y viene representado en una fila (ver Tabla 1) donde los datos contienen $n(k+2)$ observaciones, n por cada una de las $k+2$ variables o atributos que describen el contexto de la investigación y que caracterizan a los individuos objeto de estudio.

Tabla 1. Formato típico de las variables y de los casos.

Casos	Variables						
	t	δ	1	2	3	...	k
1	t_1	δ_1	x_{11}	x_{12}	x_{13}	⋯	x_{1k}
2	t_2	δ_2	x_{21}	x_{22}	x_{23}	⋯	x_{2k}
⋮	⋮	⋮	⋮	⋮	⋮	...	⋮
n	t_n	δ_n	x_{n1}	x_{n2}	x_{n3}	⋯	x_{nk}

Si se realiza la tabla de vida con el programa estadístico SPSS, será necesario proporcionar la variable tiempo de supervivencia, así como la variable estado, especificando el valor correspondiente a la ocurrencia del suceso muerte/fallo. La variable tiempo puede estar medida en cualesquier unidad y ha de tener siempre valores positivos. Dichos valores corresponden al tiempo de supervivencia de los no censurados y al tiempo de seguimiento para los censurados.

En el programa SPSS la ocurrencia del suceso (muerte/fallo) viene dada por un valor/valores de la(s) variable(s) estado, de modo que por exclusión todos los demás casos se consideran censurados. Si algún tiempo está ausente (missing), ese caso debería considerarse de modo especial en el estudio.

Para hacer la tabla de vida es necesario definir la longitud de los intervalos y el momento final de estudio. Puesto que a partir de la tabla pueden construirse las distintas funciones de supervivencia y riesgo, su construcción suele ser una opción gráfica de los programas estadísticos, que en particular el SPSS proporciona. Los resultados del análisis pueden sintetizarse de forma general como se presenta en la tabla 2:

Tabla 2. Representación general de una tabla de vida.

t_{i-1}	n_i'	c_i	n_i	d_i	q_i	p_i	S_i	f_i	H_i	ES	Ef	EH
$[t_{i-1}, t_i)$												
$[t_i, t_{i+1})$												
$\vdots$												
$[t_{k-1}, t_k)$												

Ahora se detallará lo que representa cada una de las columnas de la tabla de vida dada anteriormente:

- **Punto inicial del intervalo** t_{i-1}**:** Extremo izquierdo de cada intervalo.

- **Entradas en el intervalo** n_i' : Supervivientes hasta el inicio de ese intervalo. Son los que llegaron al intervalo anterior, menos los que se perdieron o murieron en ese intervalo.

- **Censurados en el intervalo** c_i: Número de observaciones censuradas en el intervalo. Podríamos distinguir dos casos posibles, aquellos que al final del estudio no han muerto (w_i) y aquellos que en un momento dado del estudio se han perdido (l_i). A estos se les suele llamar abandonos, si el abandono es debido a causas ajenas a la enfermedad, dichos datos se tratarán como censurados, de no ser así, como ya se ha comentado más arriba, deberían tratarse en cada análisis determinado como datos perdidos. Nosotros supondremos siempre que los abandonos no están motivados por el estudio y por lo tanto se trataran como censurados, de modo que $c_i = l_i + w_i$.

- **Expuestos a riesgo** n_i: Número de casos que entran en el intervalo menos la mitad de los censurados en ese intervalo $\left(n_i = \frac{n_i' - c_i}{2} \right)$. Es una estimación del número de los que tienen riesgo de morir en algún momento del periodo observado. Para hacer esta estimación a los censurados se les ha asignado una probabilidad de ½ de morir si se les hubiera observado durante todo el periodo de observación, se está suponiendo que el tiempo de permanencia en el intervalo se distribuye uniformemente.

- **Sucesos terminales** d_i: Número de sucesos en ese intervalo, es decir, de observaciones terminales en el intervalo.
  Por tanto el $n_i' = n_{i-1}' - c_{i-1} - d_{i-1}$.

- **Proporción de sucesos terminales** $q_i = \frac{d_i}{n_i}$: Es una estimación de la probabilidad de que un paciente que entra en el intervalo muera dentro de él.

- **Proporción de supervivientes** $p_i = 1 - q_i$: Estimación de la probabilidad de supervivencia de un paciente que entra en un intervalo.

- **Proporción acumulada de supervivientes** S_i: Estimación de la probabilidad de supervivencia al final del intervalo ***i***-ésimo. Es un estimador de la función de supervivencia en el instante t_i y nos referimos a menudo a ella como la razón de supervivencia acumulada y se define como $S_0 = 1$ y $S_i = p_i S_{i-1,} \quad i = 1,2,...,k$.

- **Densidad de probabilidad** f_i : Estimación de la probabilidad del suceso (muerte/fallo) por unidad de tiempo, es decir, de la función de densidad del tiempo de ocurrencia del suceso, esta estimación se hace de la manera siguiente: $f_i = \frac{(S_{i-1} - S_i)}{b_i}$

- **Razón de riesgo** h_i: Estimación de la probabilidad del suceso (morir) por unidad de tiempo supuesto que el paciente ha sobrevivido hasta el inicio del intervalo. Es una estimación de la función de riesgo en el instante t_i. Esta función ofrece en cada momento una idea del riesgo que un paciente tiene de morir "en breve". Esta función es el cociente entre la densidad y la supervivencia. Un estimador del riesgo en cada intervalo viene dado por la relación: $h_i = \frac{f_i}{0.5(S_{i-1} + S_i)}$. Teniendo en cuenta que S_i es la supervivencia al final del intervalo, para calcular el riesgo se toma la supervivencia media del intervalo.

- **Error típico de la Supervivencia Acumulada (ES):** Estimación del error típico de la estimación S_i , que vendrá expresado de la siguiente manera:

$$ES(S_i) \approx S_i \sqrt{\sum \left(\frac{d_i}{n_i (n_i - d_i)} \right)}$$

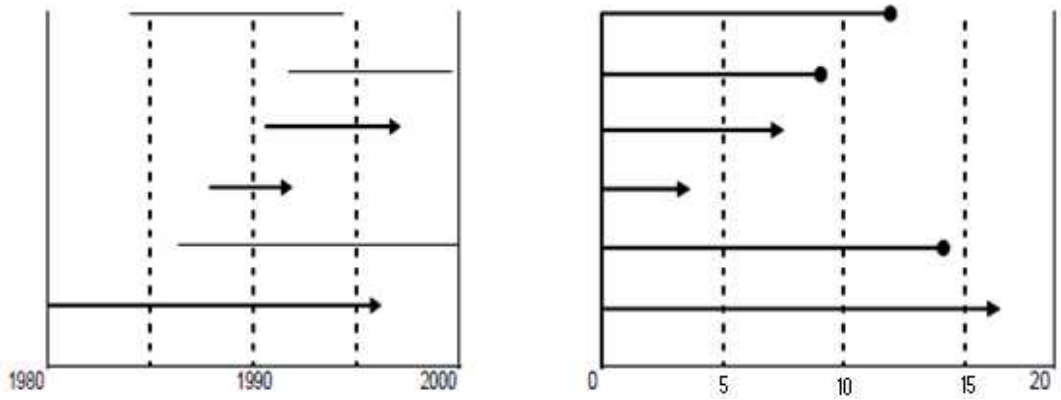
- **Error típico de la densidad de probabilidad (EF):** Estimación del error típico de la estimación f_i.

- **Error típico de la razón de riesgo (EH):** Estimación del error típico de la estimación H_i.

Para comprender mejor lo antes expuesto se presenta el siguiente ejemplo, donde la representación puede hacerse en tiempo real como se muestra en la figura 4, en el gráfico de la izquierda a lo largo de los años 1980 a 2000, con momentos iniciales y terminales reales.

En el gráfico de la derecha la representación se hace poniendo a cero todos los momentos iniciales. En este caso una **"►"** indicará que el suceso (muerte) ha ocurrido, mientras que **"●"** indicará que el dato está censurado.

Figura 4. Representación de tiempos de vida.



Del grafico de la derecha se puede observar que el periodo de observación se subdivide en intervalos de cinco años, y a partir de el se obtiene la información organizada en la siguiente tabla 3.

Tabla 3. Tiempos de vida correspondientes al gráfico 6.

Tiempo	Estado
16	1
13	0
4	1
6	1
9	0
12	0

En este ejemplo, la tabla de vida es realizada con los datos de la tabla 3 y la figura 4, ya que si se parte en cuatro intervalos el gráfico de la derecha de la figura 4; y se toma el límite inferior se obtiene la primera columna de la tabla de vida, además al saber cuantos casos entran en cada intervalo se obtiene la segunda columna y si se sabe cuantos han sido censurados en cada intervalo se obtiene la tercera columna, la cuarta columna son los expuestos a riesgo, mientras que la quinta columna representa los sucesos terminales, el resto de columnas representan cálculos estadísticos cuya explicación se describió anteriormente.

A partir de ahora se supone que los datos vienen representados como el gráfico de la derecha de la figura 4, donde se presentan los tiempos que medían entre el inicio de la observación y el momento en que se produce la muerte o la censura, habría que precisar que los intervalos en cada columna se entienden cerrados por la izquierda y abiertos por la derecha, en otras palabras, en cada intervalo se considera el punto inicial, pero no el punto final.

El límite superior del último intervalo corresponde aproximadamente con el momento en que se interrumpe el estudio, aunque para algunos individuos no se haya producido el suceso. Entonces, se supone que el periodo de observación se divide en k intervalos de tiempo, de t_0 a t_1 el primero, de t_1 a t_2 el segundo,… y de t_{k-1} a t_k el último. Así el intervalo correspondiente a la población i se representa como $I = [t_{i-1}, t_i)$, donde está incluido t_{i-1}, pero no t_i.

Se supone que la longitud de cada uno de los intervalos se mantiene fija a lo largo del estudio, aunque podría considerarse de otra forma, el punto medio del intervalo ***i***-ésimo será t_{mi} . Su conocimiento es necesario, pues las funciones de supervivencia y de riesgo serán representadas en estos puntos.

La amplitud o longitud del intervalo se llamará $b_i = t_i - t_{i-1}$, y se utilizará para la estimación de las funciones de densidad y de riesgo, en cada intervalo pueden darse cuatro casos posibles para las observaciones que entran en él. Para ilustrar los casos se tomara como ejemplo las observaciones en el segundo intervalo, del gráfico de la derecha de la figura 6, comenzando la numeración de abajo hacia arriba.

a) Se produce el suceso dentro del intervalo (observación 4): **Observaciones terminales en el intervalo.**

b) No se produce el suceso en el intervalo, pero sí en algún intervalo posterior (observación 1): **Observaciones terminales en un momento posterior al intervalo.**

c) No se produce el suceso en el intervalo ni en ningún intervalo posterior del estudio (observación 2 y 6): **Observaciones censuradas en un momento posterior al intervalo.**

d) Se produce la censura de la observación dentro del intervalo (observación 5): **Observaciones censuradas en el intervalo.**

Recuerde que las censuras en los dos últimos casos podrían producirse por ser supervivientes al final del estudio o bien por tratarse de abandonos. A continuación se presenta la tabla de vida realizada con los datos de la tabla 3:

Tabla 4. Tabla de vida.

Intervalos	n_i'	c_i	n_i	d_i	q_i	p_i	S_i	f_i	H_i	ES	Ef	EH
[0,5)	6	0	6.0	1	0.1667	0.8333	0.8333	0.0333	0.0364	0.1521	0.0304	0.0362
[5,10)	5	1	4.5	1	0.2222	0.7778	0.6481	0.0370	0.0500	0.2017	0.0334	0.0496
[10,15)	3	2	2.0	0	0.0000	1.0000	0.6481	0.0000	0.0000	0.2017	0.0000	0.0000
[15,20)	1	0	1.0	1	1.0000	0.0000	0.0000	0.1296	0.4000	0.0000	0.0403	0.0000

A continuación se realizará una breve explicación de algunos resultados de la tabla 4 y además la forma de cómo fueron obtenidos algunos de ellos.

Se tiene que al inicio del tercer intervalo hay 3 supervivientes y aparecen 2 casos censurados. Se observa además que en el segundo intervalo la estimación de los expuestos a riesgo será $n_2 = \frac{5-1}{2} = 4.5$.

También se tiene que para el tercer intervalo no se produce ninguna muerte o fallo, mientras que en los otros tres se produce una muerte en cada uno. Luego para la proporción de sucesos terminales para el tercer intervalo, dicha probabilidad se hace cero, puesto que en él no hay ninguna disminución, esto no es lo habitual cuando se dispone de una mayor cantidad de datos.

La supervivencia acumulada en el segundo intervalo será $0.8333 * 0.7778 = 0.6481$, que corresponde a la probabilidad de que sobreviva hasta el final del primer intervalo y hasta el final del último. Puesto que hasta el final del cuarto intervalo no llega ningún paciente, la probabilidad de supervivencia al final de este intervalo es nula, es decir, la probabilidad de sobrevivir más de 20 meses es cero. Con estas estimaciones se representará después la función de supervivencia.

La densidad de probabilidad en el ejemplo correspondería a la probabilidad de morir en un periodo de un año en el intervalo correspondiente. Quiere expresarse así una cierta probabilidad instantánea propia de la densidad de probabilidad, y que se estima mediante una distribución discreta. En el ejemplo solamente cuatro valores, correspondientes a los cuatro intervalos nos dan una estimación de dicha densidad de probabilidad. Así, si se centra en el intervalo de tiempo de 5 a 10 años la probabilidad de que una persona muera en un año cualquiera de ese intervalo será de un 3.7%.

El programa SPSS proporciona además de la tabla de vida la mediana del tiempo de supervivencia, que en el ejemplo es 16.14, esto significa que el 50% de los pacientes considerados en la muestra sobrevive 16.14 años o más. Mientras que casi un 85% de los pacientes sobrevive los primero 5 años, aproximadamente un 65% llega a los 15 años. Por estar algunos datos censurados esta no es la simple mediana de los números que aparecen en la variable tiempo, que en el ejemplo sería 10.5. Se han de tener en cuenta los datos que están censurados. De no hacerlo así estaríamos actuando como si los datos censurados correspondieran a muertes y por tanto se produciría un sesgo hacia la izquierda. Un estimador de la mediana correcta será el punto temporal para el cual la función de supervivencia vale 0.5.

Como es habitual en el cálculo de la mediana, se utiliza interpolación lineal en el intervalo en el que ha de estar contenida. Podría ocurrir que la función de supervivencia al final del último intervalo no llegara a 0.5. Esto significa que la mediana no puede calcularse, es decir, hay demasiados datos censurados. En estos casos los programas informáticos suelen identificar el comienzo del último intervalo para señalar que la mediana excede dicho valor. El programa SPSS proporciona en este caso el punto inicial del último intervalo con un signo + para indicar que la mediana está más adelante.

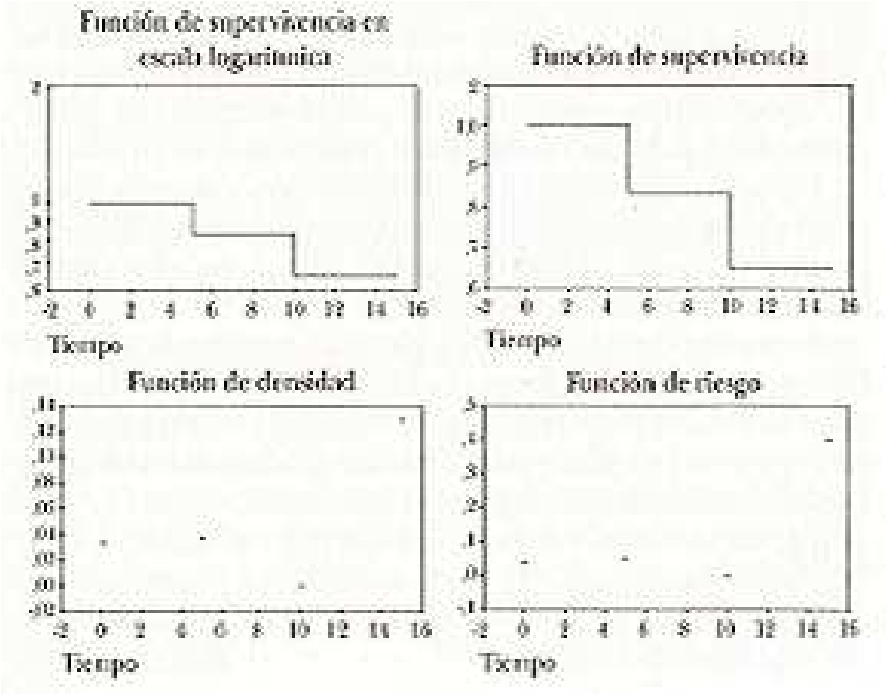
Gráficos de las funciones de supervivencia y riesgo.

La tabla de vida presentada anteriormente ha proporcionado estimaciones de las funciones de supervivencia, densidad y riesgo. Los valores estimados servirán ahora para dibujar una primera aproximación de dichas funciones, lo que permitirá estudiarlas de una manera más adecuada. Los valores de la columna correspondiente a la supervivencia acumulada proporcionan la función de supervivencia. Hay que tener en cuenta que la columna

proporciona valores al final del intervalo. Por este motivo, como se puede observar en la figura 7, en el primer intervalo el valor de la función es siempre uno, y solo a partir del final del primero y durante todo el segundo intervalo tomará el valor 0.8333. En el tercer intervalo la probabilidad correspondiente es de 0.6481 que se mantiene a lo largo del último intervalo esa es la razón por la que en este caso en el gráfico no aparece el intervalo. La representación gráfica de esta función en escala logarítmica es más representativa de manera visual, como se presenta en la figura 5.

La representación de la función de densidad da una idea de las zonas de mayor riesgo y viene dada por los valores de la tabla en el inicio de cada intervalo. En el ejemplo considerado (figura 5) se observa un crecimiento grande del riesgo de morir en la fase final.

Figura 5. Gráficos de las funciones de supervivencia, densidad y riesgo.



La función de riesgo viene dada también por los valores de la tabla en el inicio de cada intervalo, y como ya se comentó en el primer apartado, proporciona una idea del riesgo de muerte en cada momento para aquellos que han llegado hasta ahí.

2.10.2 PRODUCTO LÍMITE DE KAPLAN & MEIER.

El estimador de Kaplan & Meier propone el método del producto límite, con el objetivo de resolver los problemas planteados por la ausencia de información en los problemas de análisis de datos de supervivencia.

El estimador de la función de supervivencia del método de Kaplan & Meier; viene dado de la siguiente manera**:**

$$\hat{S}_{KM}(t) = \prod_{t_i \le t} \frac{n(t_i) - d(t_i)}{n(t_i)}$$

Donde $n(t_i)$ y $d(t_i)$ son**:** El número de individuos en riesgo y el número de muertes (o de ocurrencia del evento de interés en el momento t_i), respectivamente.

El estimador de Kaplan & Meier, suele emplearse también para estimar la probabilidad de que un individuo incluido en el estudio en el tiempo 0; no haya alcanzado el suceso de interés en el tiempo ***t***; además da proporciones exactas de supervivencia; debido a que utiliza tiempos de supervivencia precisos.

El estimador de la curva de supervivencia propuesto por Kaplan & Meier se basa en el mismo principio que el actuarial que consiste en calcular la supervivencia como producto de probabilidades condicionadas, pero llevando la partición del tiempo de estudio en intervalos al caso extremo de considerar que cada intervalo contenga sólo la observación correspondiente a un individuo, sea ésta muerte o censura.

En el estimador de Kaplan & Meier los datos observados corresponden a los tiempos $0 = t_0 < t_1 < ... < t_n$, y se considera la partición determinada por los intervalos $(t_{i-1}, t_i]$, al que pertenece el instante de tiempo t_i pero no el t_{i-1}, y además el interior de estos intervalos está siempre libre de censuras, que sólo ocurrirán, en su caso, en un extremo.

Si llegan $n_i^{'}$ individuos con vida al intervalo, $(t_{i-1}, t_i]$, el estimador de la probabilidad de fallo en ese intervalo, condicionada a que ha sobrevivido hasta entonces, será:

$$q_i = \begin{cases} \dfrac{1}{n_i^{'}} & \textit{si en } t_i \textit{ se produce una muerte.} \\ 0 & \textit{si en } t_i \textit{ se produce una censura.} \end{cases}$$

Los intervalos que no contienen muertes no contribuyen a la construcción de $S(t)$, ya que para ellos la estimación de la probabilidad condicionada de supervivencia en el intervalo es igual a 1. La existencia de censuras sí influye en el número de individuos expuestos al riesgo de morir al comienzo del intervalo siguiente, que se ve disminuido en una unidad.

Si un determinado día "i" no hubo ningún fallecimiento: La probabilidad de sobrevivir ese día habiendo sobrevivido hasta el "$i-1$" será: $S_{i/i-1} = 1$

Si al día "i" llegan $n_i^{'}$ individuos pero fallece uno de ellos, la probabilidad de sobrevivir ese día habiendo sobrevivido hasta el "$i-1$" será: $S_{i/i-1} = \dfrac{n_i^{'} - 1}{n_i^{'}}$

Este proceso se repite para todos los días y se multiplican todas estas estimaciones: $S_i = S_1 \cdot S_{2/1} \cdot S_{3/2} \cdot S_{4/3} \cdot \cdots \cdot S_{i/i-1}$ (De este producto se podrán suprimir los días con probabilidad de supervivencia igual a 1).

Este método tiene el inconveniente de considerar los datos incompletos como sometidos al riesgo de fallecer durante el intervalo en que finaliza su participación. Esto obliga a utilizar intervalos de tiempo pequeños en relación con el lapso de tiempo en el que van ocurriendo los fallecimientos. Hay que evitar la posibilidad de igualdad entre individuos con diferente causa de finalización.

La proporción de individuos que han sobrevivido al instante t se denota por p_i y se calcula mediante la siguiente fórmula:

$$p_i = \frac{n_i^{'} - d_i}{n_i^{'}} = 1 - \frac{d_i}{n_i^{'}}$$

Donde:

$n_i^{'}$: Número de individuos que llegan al comienzo del intervalo.

d_i: Número de fallecidos en el intervalo $(t, t+1]$.

Por lo tanto p_i es la probabilidad condicional de sobrevivir el i-ésimo tiempo, habiendo sobrevivido hasta el $(i-1)-ésimo$, antes denotado por $S_{i/i-1}$.

Por lo tanto; la probabilidad de supervivencia después del instante t_i viene dada por:

$$S(t_i) = p_1 \cdot p_2 \cdots p_i = \prod_{j=1}^{i} \left(1 - \frac{d_j}{n_j^{'}}\right) = S(t_{i-1}) \cdot \left(1 - \frac{d_i}{n_i^{'}}\right) = S(t_{i-1}) \cdot p_i$$

Cuando $t = 0$, $S(0) = 1$; es decir, todos los individuos comienzan vivos el estudio, las estimaciones conseguidas con el método de las tablas de vida coinciden con las estimaciones de Kaplan & Meier al aumentar el número de intervalos hacia el infinito y reducir la longitud de los intervalos. Esta sería otra posibilidad para obtener el estimador de Kaplan & Meier (Collet (1994) Cap.2), por eso se dice que el estimador de Kaplan & Meier es el caso límite del estimador por tablas de vida.

Debido a que el estimador de Kaplan & Meier es el que se utilizará en esta investigación, se desarrollará el cálculo de su varianza. El cual consiste en calcular dicha varianza de la aproximación de primer orden de la función a partir de su desarrollo en serie de Taylor.

Los valores de $\hat{q}_i$, $\hat{p}_i$ y $\hat{p}(t_i)$, son estimaciones sujetas a la variabilidad inherente al proceso de muestreo, por lo que deben completarse con información relativa a su precisión.

Bajo determinadas hipótesis sobre los mecanismos de censura es posible, aunque complicado, deducir estimaciones de sus varianzas. Por esta razón, aunque la metodología de las tablas de vida clínicas es antigua, el estudio teórico de las propiedades estadísticas de sus estimadores es reciente y está aún por completar.

Fórmula de Greenwood: La estimación más empleada de la varianza de $\hat{p}(t_j)$ es la propuesta por Greenwood en 1926, se tiene que:

$$Var\left[\hat{P}(t_j)\right] \approx \left[\hat{P}(t_j)\right]^2 \sum_{i=1}^{j} \frac{\hat{q}_i}{\hat{p}_i n_i} = \left[\hat{P}(t_j)\right]^2 \sum_{i=1}^{j} \frac{d_i}{n_i (d_i - n_i)}$$

Esta estimación, es resultado de una aproximación asintótica que se comentará más adelante, es razonable utilizar esta aproximación cuando el valor esperado de n_j no es demasiado pequeño. La fórmula de Greenwood tiende a subestimar la varianza de $\hat{P}(t_j)$, especialmente en los intervalos de la cola derecha de la distribución donde el valor esperado de n_j suele ser pequeño. No obstante, en esos casos su cálculo no es adecuado ya que la distribución de $\hat{P}(t_j)$ suele ser muy sesgada y, en consecuencia, la varianza no es una buena medida de precisión de la estimación.
A continuación se realizará la demostración de la Fórmula propuesta por Greenwood, para ello se utilizará algunos supuestos sobre la distribución binomial y los polinomios de Taylor.

Justificación: La estimación actuarial de la función de supervivencia es, $\hat{p}(t_j) = \hat{p}_j \hat{p}_{j-1} ..... \hat{p}_1$. Para obtener su varianza aproximada se aplica un procedimiento, denominado *método delta*, que consiste en calcular la varianza de la aproximación de primer orden de la función obtenida a partir de su desarrollo en serie de Taylor.

De forma general para el caso de una función de una sola variable, se tiene que:

$$g(X) \approx g(\theta) + g'(\theta)(X - \theta)$$

Por lo que la varianza aproximada viene dada de la siguiente manera:

$$Var[g(X)] \approx [g'(\theta)]^2 Var(X)$$

Donde θ es un parámetro tal que, asintóticamente, $E(X - \theta) = 0$.

En el caso de una función de varias variables, aplicando un procedimiento análogo, se tiene,

$$Var[g(X)] \approx \sum_{i=1}^{j} \sum_{k=1}^{j} \frac{\partial g}{\partial \theta_i} \frac{\partial g}{\partial \theta_k} Cov(X_i, x_k) \qquad (2)$$

Donde $\frac{\partial g}{\partial \theta_i}$ denota $\left.\frac{\partial g(x)}{\partial x_i}\right|_{x=\theta}$ y θ es un vector de parámetros tal que, asintóticamente, $E[X - \theta] = 0$. Generalmente, el vector θ no es conocido, por lo que las derivadas se evalúan en su estimador $\hat{\theta}$.

Aplicando este método descrito anteriormente a la expresión de $\hat{p}(t_j)$, se obtiene que la varianza aproximada esta dada por:

$$Var[p(t_j)] \approx \sum_{i=1}^{j} \sum_{k=1}^{j} \frac{\partial \hat{p}(t_j)}{\partial p_i} \frac{\partial \hat{p}(t_j)}{\partial p_k} Cov(\hat{p}_i, \hat{p}_k)$$

Para calcular esta expresión se hará resolviendo cada parte de sus términos de forma individual de la siguiente manera:

1. Estimadores de p_j y q_j. Los más utilizados son:

$$\hat{p}_j = \frac{(n_j - d_j)}{n_j}$$

$$\hat{q}_j = \frac{d_j}{n_j}$$

En el caso sin censura, éstos son los estimadores máximo verosímiles; este resultado se obtiene utilizando el hecho que el vector $(d_1, d_2, ..., d_{s+1})$sigue una distribución Multinomial de parámetros n_1, el número de individuos que inician el estudio, y $\pi_1, \pi_2, ..., \pi_s$ la diferencia de proporciones de supervivencia en el tiempo t, donde:

$$n_1 = \sum_{i=1}^{s+1} d_i$$

$$\pi_j = p(t_{j-1}) - p(t_j) = p_1 .... p_j \qquad j = 1,2,....., s+1$$

2. Cálculo de las derivadas parciales de la función de supervivencia respecto a cada componente del vector $p = (p_1, p_2, ..., p_s)$, es decir:

$$\frac{\partial \hat{p}(t_j)}{\partial p_i} = \frac{\hat{P}(t_j)}{\hat{P}_i}$$

Demostración:

Sea $\hat{P}(t_i) = (\hat{P}_j, \hat{P}_{j-1} .... \hat{P}_1)$ el vector estimador de la función de supervivencia, si se le aplica derivadas parciales, será:

$$\frac{\partial \hat{P}(t_j)}{\partial \hat{P}_i} = \frac{\partial (\hat{P}_j \hat{P}_{j-1} .... \hat{P}_1)}{\partial \hat{P}_i} = (\hat{P}_j \hat{P}_{j-1} .... \hat{P}_{i-1} \hat{P}_{i+1} ... \hat{P}_1) \frac{\partial \hat{P}_i}{\partial \hat{P}_i} = \frac{\hat{P}(t_j)}{\hat{P}_i}$$

3. Cálculo de las varianzas y covarianzas de los estimadores $\hat{p}_i$. Presentamos los resultados en el caso de muestras sin observaciones censuradas ya que su justificación en el caso general resulta complicada.

- Proposición: Bajo la hipótesis $n_j > 0$, d_j sigue una distribución Binomial de parámetros n_j y q_j; se tiene que:

  a) $E[\hat{q}_j] = q_j$

  b) $E[\hat{p}_j] = p_j$

  c) $Var[\hat{q}_j] = Var[\hat{p}_j] = \frac{p_j q_j}{n_j}$

  d) $Cov(\hat{q}_i, \hat{q}_j) = Cov(\hat{p}_i, \hat{p}_j) = 0 \quad con\ i < j$

Demostración:

a) Como $d_j \sim B(n_j, q_i)$ al aplicar valor esperado a la primera expresión se tiene que:

$E[\hat{q}_j] = E\left(\frac{d_j}{n_j}\right)$, donde $E\left(\frac{d_j}{n_j}\right) = \frac{1}{n_j} E(d_j) = \frac{1}{n_j} n_j q_j$ entonces:

$E[\hat{q}_j] = q_j$

b) Utilizando siempre el supuesto que sigue una distribución binomial.

$$E[\hat{p}_j] = p_j$$

Sustituyendo al valor de $\hat{p}$ y aplicando valor esperado se tiene:

$$E[\hat{p}_j] = E\left[1 - \frac{d_j}{n_j}\right] = E[1] - E\left[\frac{d_j}{n_j}\right]$$

Y utilizando el resultado de la primera ecuación tendremos:

$$E\left[\hat{p}_j\right]=1-q_j=p_j$$

En muestras con observaciones censuradas, los resultados anteriores no son ciertos aunque, asintóticamente, se verifica que $Cov\left(\hat{p}_i,\hat{p}_j\right)=0$ y la estimación de $Var\left(\hat{p}_j\right)$ puede aproximarse mediante $\frac{\hat{p}_j\hat{q}_j}{n_j}$.

Sustituyendo los resultado obtenidos anteriormente en la expresión (2) de $Var\left(\hat{p}_j\right)$ se obtiene la fórmula de Greenwood.

$$Var\left[\hat{P}\left(t_j\right)\right]\approx\left[\hat{P}\left(t_j\right)\right]^2\sum_{i=1}^{j}\frac{\hat{q}_i}{\hat{p}_i n_i}=\left[\hat{P}\left(t_j\right)\right]^2\sum_{i=1}^{j}\frac{d_i}{n_i\left(d_i-n_i\right)}$$

Una vez estimada la función de supervivencia de la variable duración, ésta puede ser utilizada, mediante procedimientos gráficos, para tratar de observar la posible distribución a la que se ajustan los datos. Muchas veces resulta interesante determinar si dos o más muestras tienen funciones de supervivencia similares.

El intervalo de confianza del estimador de Kaplan & Meier calculado por defecto por los programas estadísticos es el de identidad o de escala plana, dado, para un nivel de confianza del 90%, por:

$$\hat{S}_{KM}^{\,2}(t)\pm 1.645 ee(\hat{S}_{KM}(t))$$

Donde $ee(\hat{S}_{KM}(t))$; es el error estándar de estimación del estimador de Kaplan & Meier

Sobrevida media y mediana.

La sobrevida media o media de la supervivencia puede ser estimada mediante la siguiente expresión:

$$\hat{\mu}=\int_0^T \hat{S}_{KM}(t)dt$$

Donde T es tiempo máximo de seguimiento observado durante el estudio.

La sobrevida mediana o mediana de la supervivencia se define como el primer tiempo t que satisface la siguiente condición:

$$\hat{S}_{KM} \leq 0.5$$

La validez de este método descansa en dos suposiciones:

1) Las personas que se retiran del estudio tienen un destino parecido a las que quedan.
2) El período de tiempo durante el cual una persona entra en el estudio no tiene efecto independiente en la respuesta.

Dentro de esta investigación se utilizará solo una técnica de estimación no paramétrica, la cual es el estimador de Kaplan & Meier, ya que este método calcula la supervivencia de forma individual cada vez que un paciente muere, mientras que el análisis actuarial divide el tiempo en intervalo y calcula la supervivencia en cada intervalo. El procedimiento de Kaplan & Meier da proporciones exactas de supervivencia debido; a que utiliza tiempos de supervivencias precisos; el análisis actuarial da aproximaciones, debido a que agrupa los tiempos de supervivencia en intervalos, anteriormente el método actuarial era más fácil de usar para un número muy grande de observaciones, mientras que el método de Kaplan & Meier se utiliza cuando la muestra es menor de 30 o bien para muestras mayores de 30 y se conocen los tiempos individuales de los censurados y no censurados, de manera que se calcula la supervivencia cada vez que un paciente muere o alcanza el evento.

El método actuarial implica dos premisas en los datos: la primera de ellas es que todos los abandonos durante un intervalo dado ocurre aleatoriamente durante dicho intervalo. Esta premisa es de escasa importancia cuando se analizan intervalos de tiempos cortos, sin embargo, puede haber un sesgo importante cuando los intervalos son grandes, si hay numerosos abandonos o si los abandonos no ocurren a mitad del intervalo; el método de Kaplan & Meier supera estos problemas. La segunda premisa es que aunque la supervivencia en un tiempo dado depende de la supervivencia en todos los periodos previos, la probabilidad de la misma en un período de tiempo es independiente de la probabilidad de supervivencia en los demás períodos.

La aplicación del método de Kaplan & Meier consiste en la elaboración de una tabla la cual tiene la siguiente forma general:

Tabla 5. Tabla general del método de supervivencia de Kaplan & Meier.

Columna 1	Columna 2	Columna 3	Columna 4	Columna 5
Tiempo de supervivencia en meses	Nº de orden	Orden de las observaciones no censuradas	$\frac{n-r}{n-r+1}$	$S(t)$

Donde:

Columna 1: Se hace una lista con todos los tiempos de supervivencia, censurada o no censurada, en orden de menor a mayor. Se coloca un signo positivo al lado de cada observación censurada. Para observaciones censuradas y no censuradas que tienen el mismo tiempo de supervivencia, se debe colocar la observación no censurada primero.

Columna 2: Una vez ordenados de menor a mayor los datos, en esta columna se numeran las observaciones.

Columna 3: Colocar el número de orden (rango) de las observaciones no censuradas.

Columna 4: Calcular la proporción de pacientes que sobrevive a cada intervalo.

$$\frac{n-r}{n-r+1}$$

Donde n es el tamaño de la muestra y r el rango no censurado.

En esta columna se calcula la probabilidad de supervivencia para cada tiempo.

Columna 5: Calcular el estimador de la proporción acumulativa que sobrevive $S(t)$. Se realiza multiplicando los valores de la columna anterior (columna 4).

De este modo, la probabilidad de vivir un cierto período de tiempo (hasta el instante t) desde el principio del estudio, es el producto de la probabilidad acumulada de sobrevivir

hasta el período del tiempo anterior a t, $(t-1)$, multiplicado por la probabilidad de sobrevivir durante el intervalo $(t-1;t)$.

Finalmente, siempre es recomendable representarlo gráficamente. La representación gráfica de la función de supervivencia en el método de Kaplan & Meier se tiene que en el eje X se sitúan los tiempos observados y en el eje Y el valor de la supervivencia acumulada.

Se debe empezar con una supervivencia de 1, que se mantiene hasta que se produce el primer fallecimiento. En ese momento la gráfica da un salto correspondiente al descenso de la supervivencia a partir de ese momento y así sucesivamente, es decir; que esta se supone que es constante hasta el tiempo en que vuelve a suceder un nuevo acontecimiento de interés. En el momento "0", el 100% de los sujetos están vivos.

Cuando alguien sigue vivo al final del periodo de observación, se deja una línea horizontal y si alguien ha fallecido se traza una línea vertical y si se da el caso en el que todos han fallecido se traza una línea vertical hasta el punto cero de supervivencia.

Para poner en práctico lo del método de Kaplan & Meier se presentan los siguientes ejemplos.

Ejemplo1.

El ejemplo se basa en datos publicados por Pratt, et al[2]. Se recogieron los intervalos libres de enfermedad (tiempos de remisión) de 20 pacientes con osteosarcoma, a los que se trataba con tres meses de quimioterapia después de amputación obteniéndose:

- ✓ 11 pacientes que recayeron a los 6, 8, 10, 11, 12, 13, 13, 22, 32, 34, 36 meses.
- ✓ 8 pacientes se retiraron vivos al final del estudio contribuyendo 3, 7, 7, 11, 14, 16, 20, 20 meses de observación, sin haber sufrido recaídas.
- ✓ Un paciente rehusó continuar la terapia a los 11 meses y se retiró del estudio libre de enfermedad.

[2] Pratt C, Shanks E, Hustu O, Rivera G, Smith J, Kumar AP. Adjuvant multiple drug chemotherapy for osteosarcoma of the extremity. Cáncer 1977; 39(1):51-57.

Utilizando el procedimiento descrito anteriormente del método de Kaplan & Meier se obtiene la siguiente tabla 6.

Tabla 6. Método para calcular la cuerva de supervivencia de Kaplan & Meier.

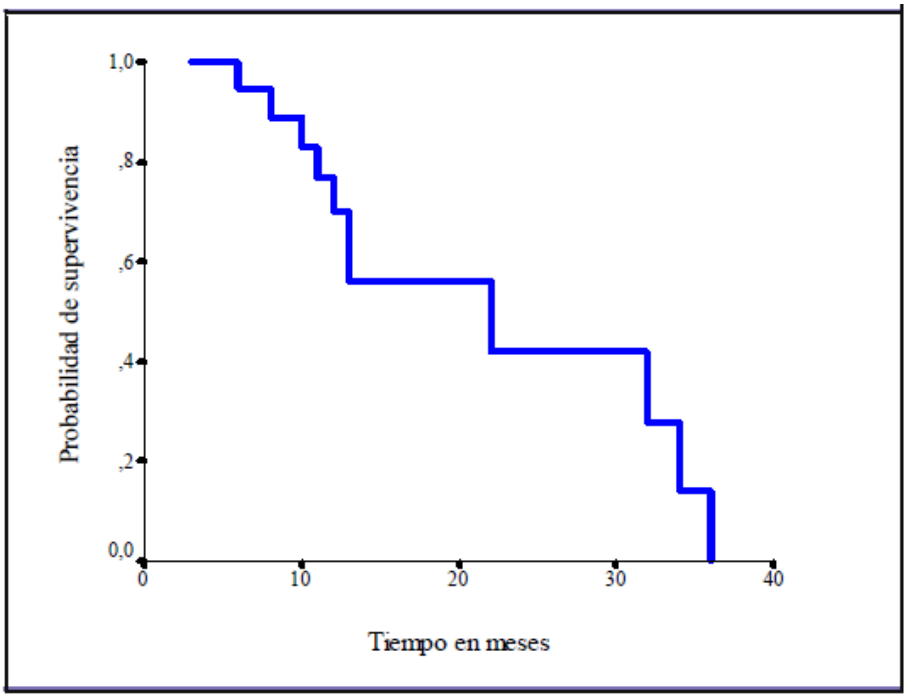
Columna 1	Columna 2	Columna 3	Columna 4	Columna 5
Tiempo de supervivencia en meses	Nº de orden	Orden de las observaciones no censuradas	$\frac{n-r}{n-r+1}$	$S(t)$
3+	1	--	--	--
6	2	2	$\frac{18}{19}=0.95$	0.95
7+	3	--	--	--
7+	4	--	--	--
8	5	5	$\frac{15}{16}=0.94$	0.89
10	6	6	$\frac{14}{15}=0.93$	0.83
11	7	7	$\frac{13}{14}=0.93$	0.77
11+	8	--	--	--
11+	9	--	--	--
12	10	10	$\frac{10}{11}=0.91$	0.70
13	11	11	$\frac{9}{10}=0.90$	0.63
13	12	12	$\frac{8}{9}=0.89$	0.56[3]
14+	13	--	--	--
16+	14	--	--	--
20+	15	--	--	--
20+	16	--	--	--
22	17	17	$\frac{3}{4}=0.75$	0.42
32	18	18	$\frac{2}{3}=0.67$	0.28
34	19	19	$\frac{1}{2}=0.50$	0.14
36	20	20	0	0.0

[2] Cuando hay un tiempo de supervivencia (13 meses) con valores de supervivencia diferente se utilizará como estimador el valor más bajo (0.56).

En la tabla 6, se tiene que 0.95 estima la probabilidad de sobrevivir 6 o más meses. La última columna corresponde al estimador de Kaplan & Meier y va multiplicando los cocientes de la columna 4 de cada tiempo por el producto previo. Así, se puede decir que la supervivencia acumulada a los 6 meses era del 95%, a los 8, del 89% y a los 36 meses del 0%.

La probabilidad de supervivencia puede representarse gráficamente como se muestra en la figura 6.

Figura 6. Curva de Kaplan & Meier.



En la figura 6, se observa la curva de Kaplan & Meier representando la supervivencia acumulada durante el seguimiento de 20 pacientes. Puede observarse que, como es lógico, sólo hay cambios en la supervivencia cuando muere algún paciente. Se han observado 10 muertes los otros 10 pacientes están censurados.

Ejemplo 2.

Supongamos ahora que disponemos de los datos de supervivencia de 10 pacientes que han sido aleatoriamente asignados a los tratamientos A y B (datos hipotéticos).

Tratamiento:

A. 3, 5, 7, 9+, 18

B. 12, 19, 20, 20+, 33+

"9+" indica dato censurado y, por tanto, no ha presentado el evento (en este caso morir de cáncer), como tampoco lo han presentado las observaciones 20+ y 33+.

Con estos datos se construye la tabla 7 para calcular la proporción acumulativa que sobreviven hasta el tiempo t, o tasa de supervivencia acumulativa, de la misma forma que se indicó en el ejemplo 1.

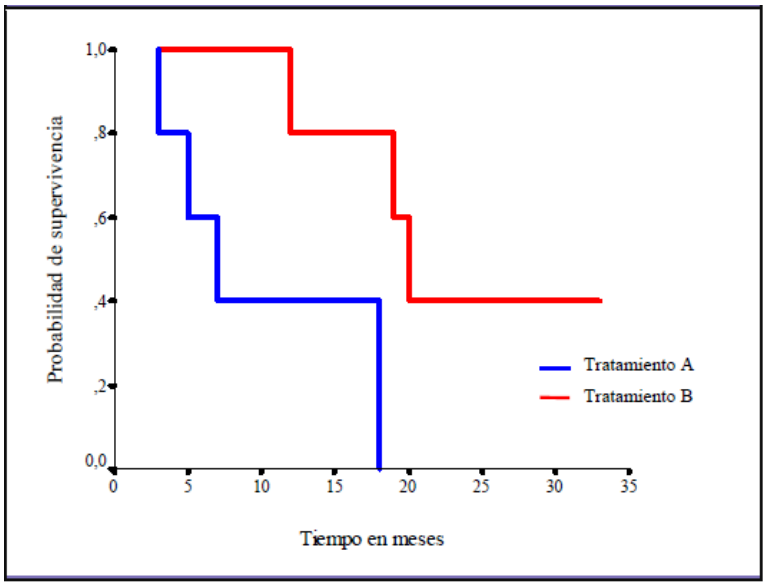
Tabla 7. Método para calcular la curva de supervivencia de Kaplan & Meier.

Columna 1	Columna 2	Columna 3	Columna 4	Columna 5
Tiempo de supervivencia en meses	Nº de orden	Orden de las observaciones no censuradas (r)	$\frac{n-r}{n-r+1}$	$S(t)$
Tratamiento A				
3	1	1	$4/5 = 0.80$	0.8
5	2	3	$3/4 = 0.75$	0.6
7	3	3	$2/3 = 0.67$	0.4
9+	4	--	--	--
18	5	5	0	0.0
Tratamiento B				
12	1	1	$4/5 = 0.80$	0.80
19	2	2	$3/4 = 0.75$	0.60
20	3	3	$2/3 = 0.67$	0.40
20+	4	--	--	--
30+	5	--	--	--

En la tabla 7, se tiene que 0.80 estima la probabilidad de sobrevivir 3 o más meses con el tratamiento A, mientras que con el tratamiento B se tiene que 0.80 estima la probabilidad de sobrevivir 12 o más meses, es decir que es mucho mayor la probabilidad de sobrevivir con el tratamiento B que con el A.

Una vez calculada la probabilidad de supervivencia, ésta puede representarse gráficamente como se presenta en la figura 7.

Figura 7. Curvas de Kaplan-Meier



En la figura 7, se observa que en el caso del tratamiento B el estimador no llega a cero ya que la última observación es censurada caso contrario en el tratamiento A que los peldaños de la grafica descienden. Además el tratamiento B proporciona mayor probabilidad de vida, por estar ubicada su gráfica sobre la del tratamiento A.

2.10.2.1 COMPARACIÓN DE CURVAS DE SUPERVIVENCIA.

En los estudios de supervivencia, frecuentemente se comparan las experiencias de supervivencia de dos o más grupos de pacientes. Estos grupos posiblemente diferirán con respecto a cierto factor. El efecto de este factor en la supervivencia es el evento de interés al comparar los grupos. Para detectar la diferencia entre los grupos se utiliza la prueba "Log-rank". Esta prueba está diseñada para detectar si existen o no diferencias entre las curvas de supervivencia de los grupos. Estas diferencias ocurren cuando la tasa de mortalidad en un grupo es consistentemente mayor que la tasa correspondiente en un segundo grupo, y la razón de estas dos tasas es constante a través del tiempo.

La prueba "Log-rank" es un método estadístico no paramétrico, en el cual se compara la experiencia de supervivencia de dos o más grupos. Además esta prueba es bastante robusta contra desviaciones de azar proporcional o lo que es lo mismo, desviaciones de los logaritmos de las curvas de supervivencia, pero debe tenerse cuidado. Si las curvas de supervivencia de Kaplan-Meier se cruzan, entonces esto presenta una divergencia del azar proporcional, por lo que la prueba Log-rank no puede utilizarse; en este caso se utilizan otras pruebas tales como la Breslow, también llamada prueba de Gehan o de Wilcoxon generalizado. Por lo tanto; se tiene que estos tests presentan las siguientes características comunes:

- Hipótesis nula (H_0): La supervivencia de los grupos que se comparan (2 ó más) es la misma.

- Hipótesis alternativa (H_1): Al menos uno de los grupos tiene una supervivencia diferente.

El estadístico utilizado para estas hipótesis es la prueba del chicuadrado para analizar las pérdidas observadas y esperadas, es decir; que esta prueba compara en esencia el número de eventos (muertes, fracasos) en cada grupo con el número de fracasos que podría esperarse de las pérdidas en los grupos combinados.

Si no hubiese observaciones censuradas la prueba no paramétrica de suma de rangos de Wilcoxon podría ser apropiada para comparar dos muestras independientes. Como la mayoría de las veces hay datos censurados debemos utilizar otras técnicas.

La prueba de la t de Student para datos independientes comparando la supervivencia en uno y otro grupo tampoco es apropiada, pues los tiempos de supervivencia no presentan una distribución normal.

Para el cálculo de la prueba Log-rank se disponen los datos de tal forma que se objetive en cada grupo los pacientes en riesgo y los eventos presentados.

Esta prueba consiste en comparar k grupos y para cada grupo se calcula los totales para pérdidas observadas y esperadas, y el procedimiento para probar la hipótesis nula de que las distribuciones de supervivencia son iguales en los dos grupos es el siguiente:

Primero se obtienen los valores de las muertes observadas y esperadas, luego se totalizan de la siguiente forma:

$$O_j = \sum_{i=1}^{t} d_{ij} \quad y \quad E_j = \sum_{i=1}^{t} e_{ij}$$

Donde:

O_j = El número total de muertes observadas durante la duración del estudio para el j-ésimo grupo.

E_j = El número total de muertes esperadas durante la duración del estudio para el j-ésimo grupo.

i = Cada tiempo en el que ocurre una muerte.

t = El número total de tiempos cuando ocurren las muertes.

d_{ij} = El número de muertes observadas en el i-ésimo tiempo para el j-ésimo grupo.

e_{ij} = El número de muertes esperadas en el i-ésimo tiempo para el j-ésimo grupo.

Luego de obtener el número total de muertes, tanto observadas como esperadas, se calcula la estadística de la prueba "Log-rank".

$$X^2 = \sum_{j=1}^{g} \frac{(O_j - E_j)^2}{E_j},$$

Donde g = el número total de grupos.

El test χ^2 sigue una distribución chi cuadrado con un grado de libertad.

Este estadístico se compara con una distribución ji-cuadrado (χ^2) con k-1 grados de libertad y un nivel de significancia (α); donde k es el número total de grupos, y si $\chi^2 > \chi^2_{k-1,\alpha}$ se rechaza la hipótesis (H_0) que las distribuciones de supervivencia son iguales en los grupos y se concluye que la distribución de supervivencia son diferentes.

Estimación de riesgo (OR).

La prueba "Log-rank" es ampliamente utilizada para comparar la experiencia de supervivencia de dos o más grupos de individuos. Sin embargo, la misma es simplemente una prueba de hipótesis y no provee información directa de cómo los grupos difieren. Para medir la supervivencia relativa entre dos grupos, se comparan los eventos observados con los esperados.

Al calcular la razón entre el número total de muertes observadas y el número total de muertes esperadas en un mismo grupo, se obtiene la tasa observada de mortalidad en ese grupo como una proporción de la tasa esperada de mortalidad si la hipótesis nula de la prueba "Log-rank" es cierta. Utilizando esta razón entre las tasas de mortalidad, se puede calcular la experiencia relativa de supervivencia de dos grupos, conocida también como la razón de riesgo de la siguiente manera:

$$OR = \frac{O_1/E_1}{O_2/E_2}$$

Para poner en práctica lo expuesto anteriormente, se utilizará los datos del ejemplo 2 construyendo la tabla de la siguiente forma:

- ✓ En la 1ª columna se ponen los meses en los que se presentaron los eventos (muertes). Se trata por lo tanto de tiempos no censurados.

- ✓ En la 2ª y 3ª columna debe colocarse el nº de pacientes en cada grupo que estuvieron en riesgo hasta la presencia del evento.

- ✓ En la columna 4ª se ubica el número total de pacientes que estuvieron en riesgo de ambos grupos.

- ✓ En las columnas 5ª y 7ª se ubican los pacientes que tuvieron el evento en ese tiempo y el total.

- ✓ Se calculan los totales para pérdidas observadas y esperadas; y el test siguiente puede utilizarse para probar la hipótesis nula de que las distribuciones de supervivencia son iguales en los dos grupos.

El número esperado de pérdidas para un grupo se calcula multiplicando el número total de pérdidas en un período dado por la proporción de pacientes en ese grupo. Así por ejemplo, en el mes quinto hay una pérdida; de modo que $\frac{1*4}{9} = 0.44$ es el número de pérdidas que se espera para el grupo A y $\frac{1*5}{9} = 0.56$ es el número de pérdidas que se espera para el grupo B.

Tabla 8. Test de long-rank para comparar la probabilidad de supervivencia entre grupos.

Mes del evento	Pacientes en riesgo			Pérdidas observadas			Pérdidas esperadas		
	Tratamiento			Tratamiento			Tratamiento		
	A	B	Total	A	B	Total	A	B	Total
3	5	5	10	1	0	1	0.50	0.50	1
5	4	5	9	1	0	1	0.44	0.56	1
7	3	5	8	1	0	1	0.38	0.62	1
12	1	5	6	0	1	1	0.16	0.83	1
18	1	4	5	1	0	1	0.20	0.8	1
19	0	4	4	0	1	1	0.0	1.0	1
20	0	3	3	0	1	1	0.0	1.0	1
				4	**3**	**7**	**1.68**	**5.31**	**7**

Las hipótesis a probar son las siguientes:

H_0 : La supervivencia del grupo que se le aplicó el tratamiento A es igual al grupo que se le aplicó el tratamiento B.

H_1 : La supervivencia del grupo que se le aplicó el tratamiento A es diferente al grupo que se le aplicó el tratamiento B.

Luego calculando el test χ^2 se tiene:

$$\frac{(1-1.68)^2}{1.68}+\frac{(3-5.31)^2}{5.31}=3.20+1.005=4.21$$

Consultando las tablas de una distribución χ^2 con un grado de libertad y un nivel de significancia del 5% ($\alpha=0.05$), se tiene que $\chi^2_{2-1,0.05}=3.84$ por lo tanto se concluye que la diferencia es significativa; ya que el valor obtenido mediante la prueba Long-rank es mayor que el valor encontrado en tablas, es decir que existe diferencia en la supervivencia de la aplicación de ambos tratamientos.

Los datos generados permiten a su vez realizar una estimación del riesgo (OR) y favorece la determinación de qué tratamiento es el más adecuado, se tiene que:

$$OR=\frac{O_1/E_1}{O_2/E_2}=\frac{4/1.68}{3/5.31}=4.21$$

Al grupo que se le aplicó el tratamiento B sobreviven 4.21 veces más que los del grupo que se les aplicó el tratamiento A.

2.11 EL MODELO DE REGRESIÓN DE COX.

Una de las técnicas estadísticas más utilizadas para evaluar la relación entre un conjunto de variables explicativas y el tiempo de supervivencia, es el modelo de regresión de riesgos proporcionales; conocido también como el Modelo de Cox. Este modelo es el más utilizado para representar los efectos de un conjunto de variables explicativas sobre la variable tiempo de cambio (tiempo de supervivencia), o más bien sobre la probabilidad condicional de cambio, es decir; sobre la función de riesgo $\lambda_0(t)$. Suponemos que para cada sujeto tenemos un vector X de variables explicativas, concomitantes o pronosticas. Las componentes de dicho vector pueden presentar tratamientos, definidos por medio de variables indicadoras, 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.

La aplicación de modelos de regresión de riesgos proporcionales tiene dos vertientes, la primera como herramienta de investigación no experimental, para medir un efecto de forma precisa mediante el control de las variables de confusión de la modelización de las interacciones, y la segunda como procedimiento para seleccionar variables predictoras y construir un modelo que permita describir, explicar o predecir la respuesta (Y) de los sujetos, así como también evaluar la contribución de cada una de las variables predictoras.

Además el modelo de riesgo proporcional, permite analizar variables predictoras dependientes del tiempo, es decir, variables pronosticas que pueden tomar diferentes valores durante el seguimiento del sujeto, como por ejemplo; la presencia de una complicación o el cambio del estado clínico de un paciente durante la evolución de la enfermedad. Estas variables pronosticas pueden o no ser dependientes del tiempo; las no dependientes son aquellas variables iniciales que no cambian a lo largo de la evolución del sujeto, como por ejemplo; el sexo, la edad de inicio al consumo de drogas, etc. Un ejemplo de variable pronostica dependientes del tiempo podría ser, el trabajo estable, ya que un paciente en un momento determinado del seguimiento podría abandonar o incorporarse a un trabajo.

La regresión de Cox, consiste en obtener una función lineal de las variables independientes que permita estimar, en función del tiempo, la probabilidad de que ocurra dicho suceso. Se supone que existe un conjunto de variables independientes $x_1.....x_p$; cuyos valores influyen en el tiempo que transcurre hasta que ocurre el suceso final.

En este modelo el riesgo para el i-ésimo individuo se define mediante la siguiente expresión:

$$\lambda(t; Z_i(t)) = \lambda_0(t) r_i(t)$$

El primer producto de la última expresión depende exclusivamente del tiempo, mientras que el segundo depende sólo de las variables explicativas, es decir $Z_i(t)$; el cual es el vector de covariables para el $i-$ésimo individuo en el tiempo t, que actúa multiplicativamente sobre la función de riesgo básica. Se dice que es un modelo semiparamétrico debido a que incluye una parte paramétrica y otra no paramétrica:

La parte paramétrica es $r_i(t)$, donde $r_i(t) = e^{\beta' Z_i(t)}$, llamada puntaje de riesgo y β es el vector de parámetros de la regresión desconocidos que parametrizan el modelo. Es una función exponencial, cuyo exponente es la combinación lineal, sin término constante de las $Z_i(t)$.

La parte no paramétrica es $\lambda_0(t)$, llamada función de riesgo base. Por tanto se puede apreciar que el efecto de las variables regresoras consiste en multiplicar a la función de riesgo por un factor de escala y es una función arbitraria no especificada, que solo depende del tiempo (t), definida así porque representa las tasas instantáneas de riesgo de un sujeto hipotético con valor 0 en todas las variables predictivas (ya que el termino exponencial es $e^0 = 1$).

El problema consistirá entonces en estimar los parámetros desconocidos $\beta_{1,.....,}\beta_P$.

Obsérvese que, si las estimaciones de todos los parámetros fueran nulas, significaría que las variables independientes no influyen en el tiempo transcurrido hasta que ocurre el suceso final. En dicho caso, la función $r_i(t)$ sería igual a 1 y en consecuencia $\lambda(t; Z_i(t)) = \lambda_0(t) r_i(t) = \lambda_0(t)$.

El modelo de regresión de Cox, es también llamado modelo de riesgos proporcionales; debido a que el cociente entre el riesgo para dos sujetos con el mismo vector de covariables es constante sobre el tiempo, es decir:

$$\frac{\lambda(t; Z_i(t))}{\lambda(t; Z_j(t))} = \frac{\lambda_0(t) e^{\beta' Z_i(t)}}{\lambda_0(t) e^{\beta' Z_j(t)}} = \frac{e^{\beta' Z_i(t)}}{e^{\beta' Z_j(t)}}$$

Si ha ocurrido una muerte en el tiempo t^*, entonces la verosimilitud de que la muerte le ocurra al i-ésimo individuo y no a otro es:

$$L_i(\beta) = \frac{\lambda_0(t^*) r_i(t^*)}{\sum_j Y_j(t^*) \lambda_0(t^*) r_j(t^*)} = \frac{r_i(t^*)}{\sum_j Y_j(t^*) r_j(t^*)}$$

El producto $L(\beta) = \prod L_i(\beta)$ se llama verosimilitud parcial y fue introducida por Cox (1975), ya que los parámetros del modelo de Cox no pueden ser estimados por el método de máxima verosimilitud al ser desconocida la forma específica de la función arbitraria de riesgo. La maximización de $\log(L(\beta))$ proporciona una estimación para β, sin necesidad de estimar los parámetros de ruido $\lambda_0(t)$.

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.

2.11.1 INTERPRETACIÓN DE LOS PARÁMETROS DEL MODELO.

Una vez presentado el Modelo de Cox, se tiene de manera general que interpretar los coeficientes β. Suponiendo el modelo más sencillo, con una sola variable pronostica binaria $Z_i(0/1)$.

Si se calcula el cociente entre la tasa instantánea de riesgo para $Z=1$ y para $Z=0$, es decir; para un incremento de Z igual a 1, se observa que la función $\lambda_0(t)$ se simplifica, porque no depende de Z, y se obtiene:

$$\frac{\lambda(t;Z=1)}{\lambda(t;Z=0)}=\frac{\lambda_0(t)e^{\beta}}{\lambda_0(t)e^{0}}=e^{\beta}\rightarrow \quad \lambda(t;Z=1)=e^{\beta}\lambda(t;Z=0)$$

Es decir; e^{β} es el factor por el que se multiplica la tasa instantánea de riesgo cuando Z se incrementa en una unidad.

A esta misma conclusión se llega planteando, un modelo con p variables explicativas y calculando el cambio de la tasa de riesgo instantáneo; de un sujeto cuando la variable z se incrementa en una unidad y las restantes variables permanecen constantes:

$$\frac{\lambda\left(t;z_1,z_2,....,z+1,.....,z_p\right)}{\lambda\left(t;z_1,z_2,....,z,.....,z_p\right)}=\frac{\lambda_0(t)\ e^{\beta(z+1)}}{\lambda_0(t)\ e^{\beta z}}=\frac{\lambda_0(t)\ e^{\beta z}e^{\beta}}{\lambda_0(t)\ e^{\beta z}}=e^{\beta}$$

En resumen, un parámetro β con signo positivo; indica un aumento de la tasa instantánea de riesgo, cuando se incrementa el valor de la variable z. Un parámetro β con signo negativo, indica un descenso de la tasa instantánea de riesgo; cuando se incrementa el valor de la variable z.

2.11.2 CONDICIONES DE APLICACIÓN DEL MODELO.

Una vez presentado el Modelo de Cox e interpretado sus parámetros, se examinaran con detalle los supuestos básicos que deben cumplir los datos para poder aplicar dicho modelo.

La parte no paramétrica del modelo no impone ningún supuesto sobre la forma de distribución de los tiempos de supervivencia. Sin embargo, la parte paramétrica del modelo implica un supuesto muy fuerte y es la contribución de las diferentes variables explicativas en la predicción de la supervivencia, o más precisamente, que de la tasa instantánea de riesgo, es la misma en cualquier momento de tiempo del seguimiento.

La necesidad de este supuesto se debe a la propia estructura del modelo de cox, formada por el producto de dos términos; el cual depende exclusivamente del tiempo, mientras que el otro depende sólo de las variables explicativas z .

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

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.

2.11.3 TESTS DEL MODELO DE REGRESIÓN DE COX.

A continuación se van a exponer algunos tests para comprobar la significación del modelo y de los parámetros. Los más desarrollados e implementados en la metodología son: La *prueba de razón de verosimilitudes,* el *test de Wald* y la *prueba de Rao* ("score test"). Estos test son asintóticamente equivalentes, pero esto no siempre suceda en la práctica, ya que se tienen distintos tamaños de muestras de individuos.

2.11.3.1 TEST DE RAZÓN DE VEROSIMILITUD.

Si se desea contrastar una variable o un grupo de variables, a fin de ver si son significativas, se puede construir un contraste de la razón de verosimilitudes comparando los máximos de la función de verosimilitud para los modelos con y sin estas variables.
Este contraste se define de la siguiente manera:

$$H_0 : \beta_p = 0$$
$$H_1 : \beta_p \neq 0$$

El contraste de razón de verosimilitudes, representa una mayor confiabilidad y se realiza de la siguiente forma: Se calcular $\lambda = 2L(H_1) - 2L(H_0)$, donde $L(H_1)$ es el máximo del soporte cuando se estima los parámetros bajo H_1 y $L(H_0)$ es el máximo cuando estimamos los parámetros bajo H_0, si H_0 es cierta entonces λ tiene una distribución χ_s^2 con s grados de libertad que corresponden al núumero de dimensión de β_p.

Una manera equivalente de definir el contraste es llamar $D(H_0) = -2L(\hat{\beta}_0)$ a la desviación cuando el modelo se estima bajo H_0, es decir; suponiendo que $\beta_p = 0$ y $D(H_1) = -2L(\hat{\beta})$ a la desviación bajo H_1. La desviación será menor con el modelo con más parámetros y si H_0 es cierta, la diferencia de desviaciones es: $\chi_s^2 = D(H_0) - D(H_1)$ se distribuye como una χ_s^2 con s grados de libertad. En particular este test puede aplicarse para comprobar si un parámetro es significativo en el modelo.

Los criterios formales son los siguientes:

Si $\lambda \leq \chi_{\alpha,s}$ se concluye que el modelo ajustado es el adecuado.

Si $\lambda > \chi_{\alpha,s}$ se concluye que el modelo ajustado no es el adecuado.

Otra forma de contrastar la significancia de los parámetros es utilizando el p-valor asociado al estadístico de Verosimilitud, si este es menor que α se rechazará la hipótesis nula al nivel de significación de α.
Bajo este punto de vista, en cada etapa del proceso de selección de variables, la candidata a ser eliminada será la que presente el máximo p-valor asociado al estadístico de Verosimilitud, será eliminada si dicho máximo es mayor que un determinado valor crítico prefijado.

2.11.3.2 TEST DE WALD.

El Test Wald es una prueba estadística, normalmente utilizada para probar si existe un efecto o no.
Este contraste se define de la siguiente manera:

$$H_0 : \beta_p = 0$$
$$H_1 : \beta_p \neq 0$$

El estadístico de contraste se define mediante:

$$\left(\hat{\beta} - \beta_0\right)' \hat{\sum}_{\hat{\beta}}^{-1} \left(\hat{\beta} - \beta_0\right)$$

donde $\hat{\sum}_{\hat{\beta}}$ es la matriz de varianzas y covarianzas estimada.

El estadístico de Wald, para las variables incluidas en la ecuación del modelo de regresión de Cox, juega exactamente el mismo papel que en la regresión logística. Es decir, para cualquier variable independiente X_i seleccionada, si β es el parámetro asociado en la ecuación de regresión, el estadístico de Wald permite contrastar la hipótesis nula de que β =0.

La interpretación de dicha hipótesis es que la información que se perdería al eliminar la variable X_i no es significativa. Si el p-valor asociado al estadístico de Wald es menor que α se rechazará la hipótesis nula al nivel de significación α.

Bajo este punto de vista, en cada etapa del proceso de selección de variables, la candidata a ser eliminada será la que presente el máximo p-valor asociado al estadístico de Wald, o si dicho máximo es mayor que un determinado valor crítico prefijado.

2.11.3.3 TEST DE LOS PUNTAJES (SCORE TEST).

El tercer contraste es el conocido como test de los puntajes (Score Test), éste se efectúa a partir del cálculo previo de la siguiente expresión definida como $U^{'}IU$, donde U es el vector de derivadas del $\log\left(L\left(\beta\right)\right)$ dado por:

$$X_S^2 = U^T(0)I^{-1}(0)U(0)$$

En la práctica, las conclusiones sobre la significación de los coeficientes deberían ser las mismas según el test seleccionado. Sin embargo, existen situaciones en las cuales esto no ocurre.

En estos casos, el test de la razón de verosimilitudes se ha mostrado como el más robusto.

En cuanto a la selección del modelo final, es recomendable usar procesos como el procedimiento paso a paso, en el que las variables se van añadiendo o eliminando una a una, mediante el cálculo en cada uno de los pasos del estadístico -2(log(L)) con el que se decide si la variable en cuestión es introducida o eliminada del modelo.

2.11.4 INTERPRETACIÓN DEL MODELO DE COX.

La interpretación del modelo de Cox; no se hace directamente a través de su coeficiente estimado sino de su exponencial, $\exp\left(\hat{\beta}\right)$.

2.11.4.1 INTERPRETACIÓN PARA COVARIABLES DICOTÓMICAS.

Para cada covariable dicotómica, $\exp\left(\hat{\beta}\right)$ es un estimador de la razón de riesgos y se interpreta como la cantidad de riesgo que se tiene con la presencia de la covariable en relación a la ausencia de la covariable. Los intervalos de confianza del 90% para $\exp\left(\hat{\beta}\right)$ se obtienen mediante:

$$\exp\left(\hat{\beta} \pm 1.645 ee\left(\hat{\beta}\right)\right)$$

Donde $ee\left(\hat{\beta}\right)$ es el error estándar de $\hat{\beta}$.

2.11.4.2 INTERPRETACIÓN PARA COVARIABLES CONTÍNUAS.

Para el caso de covariables continuas, $\exp\left(\hat{\beta}\right)$ representa la razón de riesgos al incrementar en una unidad la covariable. En el caso de las covariables continuas suele resultar más interesante estimar la razón de riesgos al incrementar la covariable en ***c*** unidades y esto se hace mediante $\exp\left(c\hat{\beta}\right)$, siendo su intervalo de confianza del 90% de la forma:

$$\exp\left(c\hat{\beta} \pm 1.645|c| ee\left(\hat{\beta}\right)\right)$$

2.11.5 CASO PARTICULAR DEL MODELO DE COX: COMPARACIÓN DE Z TRATAMIENTOS.

Para este caso particular del modelo de Cox, se considera una única variable Z con z valores posibles; es decir:

$$Z_i = \begin{cases} 1 \quad Si \quad i \; ha \quad recibido \quad el \quad tratamiento \quad 1 \; (el \; estándar) \\ 0 \quad Si \quad i \; ha \quad recibido \quad el \quad tratamiento \quad 2 \; (uno \; nuevo \; en \; investigación) \end{cases}$$

Bajo el modelo de Cox se tiene que:

- La función de Azar con el tratamiento 1 es: $\lambda(t;1) = \lambda_0(t)\ e^{\beta' 1} = \lambda_0(t)\ e^{\beta'}$ y denominando e^{β^T} se tiene $\lambda(t;1) = \lambda_0(t)\ e^{\beta^{T1}} = \lambda_0(t)\ e^{\beta^T} = \lambda_0(t)\ r$

- La función de Azar con el tratamiento 2 es: $\lambda(t;0) = \lambda_0(t)\ e^{\beta' 0} = \lambda_0(t)$

Nótese que la función de azar con el tratamiento 2 no es más que la función de riesgo base.

2.12 ESTUDIO DE RESIDUOS EN EL ANÁLISIS DE SUPERVIVENCIA.

Una de las ventajas que han surgido del enfoque del análisis de supervivencia es la posibilidad de efectuar análisis de residuos. Recuérdese que el residuo es una cantidad que se calcula para cada individuo y proporciona información en cuanto a la diferencia entre el valor de supervivencia observado para ese individuo y el valor estimado por la ecuación de regresión, cuanto mayor es esa diferencia mayor será el valor del residuo, con su signo correspondiente.

Los residuos se pueden utilizar para:

- Descubrir la forma funcional correcta de un predictor continuo.
- Identificar los sujetos que están pobremente pronosticados por el modelo.
- Identificar los puntos o individuos de influencia.
- Verificar el supuesto de riesgo proporcional.

A continuación, se realizará una descripción de los cuatro residuos que comprende el modelo se Cox, los cuales se denominan residuos de Cox-Snell, de Martingala, Deviance, y Schoenfeld.

2.12.1 RESIDUOS DE COX-SNELL.

El residuo de Cox-Snell, es el más utilizado en análisis de supervivencia. Su nombre proviene de un caso particular de la definición general de residuos, dada por Cox y Snell (1968). El residuo de Cox-Snell viene definido para la *i-ésima* observación $i = 1,\ 2,\ ....,\ n$ como:

$$r_{C_i} = e^{\hat{\beta}' x_i} \hat{H}_0(t_i) = \hat{H}_i(t_i) = -Ln(S_i(t_i))$$

Donde $\hat{H}_0(t_i)$ es el estimador de la función de riesgo acumulada de referencia en el tiempo t_i, representando el tiempo de supervivencia observado para el elemento muestral *i-ésimo*. De esta forma, la expresión anterior no es más que la estimación del riesgo acumulado para un individuo con valores en las covariables dados por x_i. Estos residuos se derivan de un resultado general obtenido de las estadísticas matemáticas sobre la distribución de una función de una variable aleatoria. De acuerdo con este resultado, si T es la variable aleatoria que recoge el tiempo de supervivencia de los elementos muestrales, y S

(t) es la correspondiente función de supervivencia, la variable aleatoria $Y = -\ln S(t)$ tiene una distribución exponencial con media uno, sin tener en cuenta la forma de S (t).

De esta forma, la interpretación está basada en la propiedad de que la variable -ln S(t) tiene una distribución exponencial de parámetro 1, para cualquier variable aleatoria t, independientemente de la forma de t y de S(t). Por ello, si el ajuste del modelo es satisfactorio, los residuos de Cox-Snell definirán aproximadamente una distribución exponencial de parámetro 1. En caso contrario, cuando el modelo ajustado no es bueno, las estimaciones de S (t) posiblemente no tendrán la forma correcta, es decir, no serán monótonamente decrecientes, por ejemplo, y los residuos de Cox-Snell no se ajustarán a una distribución exponencial. En este último caso, suele ser relativamente frecuente la presencia de outliers. Basándose en las propiedades de la distribución exponencial, se pueden construir los gráficos de residuos de Cox-Snell, estos gráficos, denominados también "index plots", se obtienen de la siguiente forma:

1. Se obtiene la función de riesgo acumulada a partir de la estimación de Kaplan & Meier de la función de supervivencia.

2. En el caso en el que el modelo sea correcto, representando gráficamente en el eje de abscisas; el logaritmo neperiano de los residuos y en el eje de ordenadas; el logaritmo neperiano de menos el logaritmo de la estimación de la tasa de riesgo acumulada de los residuos, el cual debería mostrar una línea recta con pendiente uno. Esto se debe a que los residuos deberían seguir una distribución exponencial de parámetro la unidad.

Baltazar-Aban y Peña (1995) han mostrado que este procedimiento de análisis, no proporciona resultados adecuados; cuando los tiempos de duración se ajustan a una distribución de Weibull, o también, cuando se utilizan estimaciones no paramétricas del riesgo acumulado.

A modo de conclusión, los residuos de Cox-Snell tienen las siguientes propiedades: No están simétricamente distribuidos en torno al cero, no pueden tomar valores negativos, tienen una fuerte asimetría debido a la suposición de que tienen una distribución

exponencial con media y varianza uno, y cuando el tiempo de duración más grande es un tiempo completo, el valor estimado de la función de supervivencia para dicho tiempo es cero; por lo que el residuo de Cox-Snell no está definido.

2.12.2 RESIDUOS DE MARTINGALA.

Probablemente en una salida de ordenador de un análisis de supervivencia nos encontraremos con un tipo de residuos denominados **residuos martingala**. Estos se construyen basándose a su vez en los denominados **residuos de Cox-Snell**:

$$rm_i = d_i - rc_i;\ \ donde\ \ rc_i = \text{Re}\, siduo\ \ Cox\ \ Snell$$

$$d_i = \begin{cases} 1, & suceso\ \ (muerto). \\ 0, & observación\ \ censurada\ \ o\ \ incompleta. \end{cases}$$

Los residuos de Cox-Snell sirven para chequear el ajuste del modelo global. Son muy útiles para verificar la adecuación de la función elegida en el ajuste de modelos paramétricos pero no muy informativo para los modelos de Cox que se estiman por verisimilitud parcial.

No se entrará en el cálculo matemático de estos residuos, pero sí se comentará las propiedades de los **residuos martingala**. Evidentemente por definición el residuo martingala para un individuo incompleto (censurado) será negativo.

Para los individuos fallecidos (observaciones completas) el valor de los residuos puede ir desde $-\infty$ hasta 1. Si la muestra es grande la suma de estos residuos es cero, no están correlacionados y el valor esperado es cero. Sin embargo no se distribuyen de forma simétrica en torno a cero, aunque el modelo sea correcto, lo que complica la interpretación de los gráficos.

Estos residuos se representan gráficamente frente al valor de la variable x_i para las observaciones. La curva estimada alisada proporciona un indicador de la función. En el caso en que el gráfico obtenido sea lineal no se necesita realizar ninguna transformación a

la variable, mientras que si se observa la existencia de umbrales, entonces sería recomendable utilizar una versión discreta de la variable.
Los residuos de martingala; son muy asimétricos y con una cola muy larga hacia la derecha, particularmente para datos de supervivencia para un solo evento. Se usan para estudiar la forma funcional de una covariable continua.

2.12.3 RESIDUOS DE DESVÍOS.

Los residuos de desvíos se obtienen mediante una transformación de normalización de los de martingala y son similares en forma a los residuos de desvíos (deviances) en la regresión de Poisson.
Recuérdese que la desviación (deviance) de un modelo de regresión, es el estadístico que se utiliza para cuantificar hasta qué punto el modelo actual que se ha estimado se aleja (desvía) de un modelo teórico, que se ajuste perfectamente a los datos; denominado *modelo completo* o *modelo saturado*. Cuanto menos se aleje el modelo de ese otro modelo *"ideal"* mejor será el ajuste. La desviación de un modelo se calcula como:

$$D = -2\left(\ln L_A - \ln L_S\right)$$

Donde:

$L_{A:}$ Corresponde a las funciones de verosimilitud para el modelo actual.

L_S: Corresponde a las funciones de verosimilitud para el modelo saturado.

La suma de los cuadrados de los residuos de desviación corresponde al valor de la desviación del modelo, por lo tanto; $D = \sum rd_i^2$.

Este tipo de residuos de desviación se construyen transformando los residuos martingala de tal manera que produzcan valores simétricos en torno de 0, y ahora el rango de valores va desde $-\infty$ hasta $+\infty$. Sin embargo aunque los residuos de desviación se distribuyen simétricamente en torno de cero si el modelo es adecuado, pero no necesariamente tienen

por qué sumar cero. Los residuos de desvíos se utilizan para la detección de valores atípicos (outliers).

Un residuo con un valor negativo grande corresponderá a individuos que tienen un tiempo de supervivencia grande, y para los que sin embargo el valor estimado por el modelo a partir de los factores pronóstico indica una supervivencia mucho menor. Por el contrario, un residuo con un valor negativo pequeño corresponde a individuos con un tiempo de supervivencia menor, contrariamente a lo que nos sugiere el modelo.

Si en los datos se calculan los residuos para cada individuo y se ordenan, puede ser interesante revisar los individuos que tienen valores extremos tanto positivos como negativos ya que, aunque pueden ser correctos, en ocasiones permiten detectar errores en la introducción de datos.

Conviene representar los residuos de desviación en el eje Y frente al *índice de individuo* en el eje X (denominamos *índice del individuo* simplemente al número de orden en el que se ha ido registrando cada observación en el estudio o en la base de datos).

2.12.4 RESIDUOS DE SCHOENFELD.

Otro tipo de residuos que se emplean para verificar el modelo de regresión de Cox son los denominados **residuos de Schoenfeld**, siendo éstos los más efectivos en cuanto a detectar anomalías para cada una de las variables que intervienen en el modelo. El método de Schoenfeld obtiene residuos para cada variable y para cada individuo, es decir; que si se tiene un modelo de Cox con tres factores pronóstico, se calcularán tres residuos de Schoenfeld por individuo. Estos residuos valen cero para las observaciones incompletas, por lo que para facilitar su interpretación se suelen presentar en las salidas de ordenador sólo para los individuos fallecidos. Es posible modificar estos residuos con el fin de que no valgan cero para las observaciones incompletas, obteniéndose entonces los denominados **residuos Schoenfeld corregidos o escalados**.

Además, estos residuos se caracterizan por no depender del tiempo de supervivencia observado, no requiriendo una estimación de la función de riesgo acumulada. Su cálculo viene recogido mediante la siguiente expresión, donde se muestra la *i-ésima* puntuación residual para la *p-ésima* variable explicativa

$$r_{S_{pi}} = \delta\left[x_{p_i} - \frac{\sum_{r\in R(t_i)} x_{p_r} e^{\beta' x_r}}{\sum_{r\in R(t_i)} e^{\beta' x_r}} \right]$$

Donde x_{p_i} es el valor de la *p-ésima* variable explicativa para la *i-ésima* observación y R_{t_j} es el conjunto de observaciones en riesgo en el tiempo j.

La suma de los residuos de Schoenfeld es cero y en muestras grandes el valor esperado del residuo es cero, estando incorrelados entre si.

Los residuos de puntajes se utilizan para verificar la influencia individual y para la estimación robusta de la varianza.

2.13 VERIFICACIÓN DE LA HIPÓTESIS DE RIESGOS PROPORCIONALES.

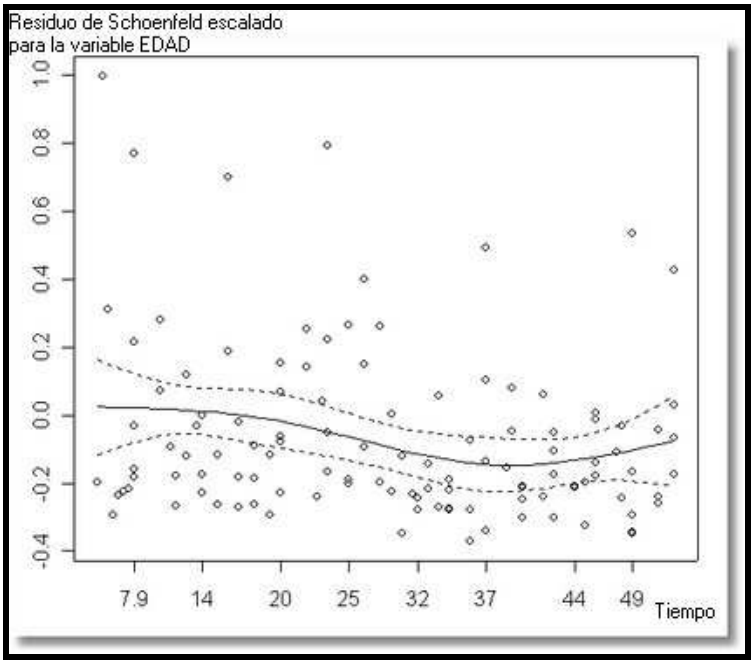
Una de las principales hipótesis del modelo de Cox; es precisamente que la función de riesgo es proporcional dados dos perfiles de factores pronóstico distintos, y por tanto se debe mantener a lo largo del tiempo. Esto es algo que se puede verificar también en las gráficas de residuos.

Para facilitar la interpretación de estos gráficos se suele superponer una curva de ajuste, utilizando alguna función de ajuste local, de "*alisado*", que suelen estar disponibles en la mayor parte de los programas estadísticos, del tipo *ajuste por splines* o también como gráficas tipo *LOWESS* o *LOESS*.

En la figura 8, se presenta un ejemplo de esta representación gráfica; en este caso se trata de los residuos de Schoenfeld para uno de los factores pronóstico del modelo, que es la EDAD en función del tiempo de supervivencia, y se ha ajustado una curva por el sistema de alisado por splines, junto con dos líneas adicionales a ±2 error estándar. Si se cumple la hipótesis de riesgos proporcionales, los residuos debieran agruparse de forma aleatoria a ambos lados del valor 0 del eje Y, y la curva ajustada debería ser próxima a una línea recta.

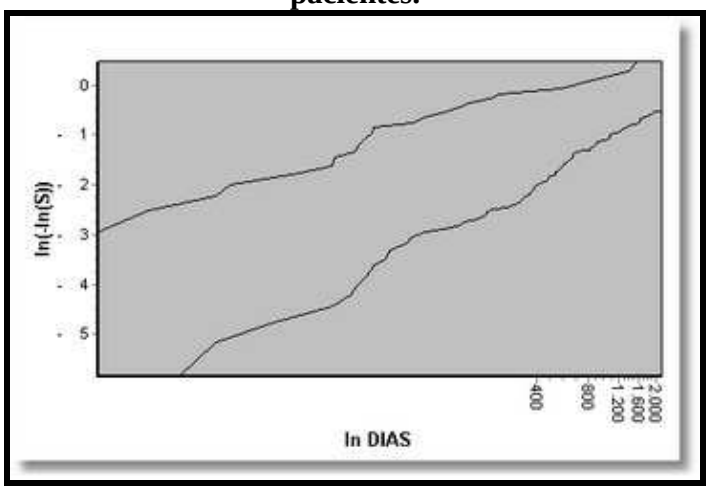
Aquí, se observa que el efecto de la edad; disminuye con el tiempo, lo que está en contradicción con la hipótesis de riesgo constante a lo largo del tiempo de un modelo de Cox correcto.

Fig. 8 Gráfico de los residuos de Schoenfeld escalados de una de las variables que intervienen en el modelo, en función del tiempo.



Otra manera de verificar de forma gráfica la hipótesis de riesgos proporcionales en el modelo de Cox cuando la variable es cualitativa, consiste en representar ln(-lnS(t)) en función del ln(Tiempo) para cada uno de las categorías. Si se cumple la hipótesis de riesgos proporcionales éstas curvas tienen que ser aproximadamente paralelas. Si se analiza una variable numérica habrá que estratificarla previamente.

Fig. 9 Gráfica para verificar la hipótesis de riesgos proporcionales en dos grupos de pacientes.



Lo dicho anteriormente se deduce a partir de que la función de riesgo acumulada viene dada por la siguiente expresión:

$$H_i(t) = H_0(t)e^{(b'x_i)} \quad (2)$$

Por lo tanto; tomando logaritmos y teniendo en cuenta la relación entre la supervivencia y la función de riesgo acumulada se tiene:

$$H(t) = -\ln S(t) \quad (3)$$

$$\ln H_i(t) = \ln H_0(t) + b'x_i \quad (4)$$

$$\ln\left(-\ln S_i(t)\right) = \ln\left(-\ln S_o(t)\right) + b'x_i \quad (5)$$

Luego según la fórmula (5) las curvas en cada grupo seguirán la forma de la supervivencia base S_0 y se mantendrán paralelas de forma aproximada, separadas por la distancia marcada por el coeficiente b.

Si no se puede asumir que es correcta la hipótesis de riesgos proporcionales, una alternativa consiste en incluir en el modelo un elemento de interacción entre esa variable y el tiempo. Si para simplificar sólo tenemos la variable x_1, la función de riesgo quedaría formulada de la siguiente manera:

$$h_i(t) = h_0(t) e^{(\beta_1 x_{1i} + \beta_2 x_{1i} t)} \qquad (6)$$

Donde interviene el producto $x_{1i}t$ que es por tanto una covariante dependiente del tiempo.

2.14 APLICACIÓN DEL MODELO DE COX.

Los datos que se analizan en este ejemplo corresponden a 246 pacientes (n=246) en tratamiento de Diálisis Peritoneal, que acudían al Servicio del Hospital Clínico Universitario de Caracas entre 1980 y 1997. Se hizo un seguimiento a los pacientes desde el comienzo de sus sesiones de diálisis, hasta alcanzar la muerte como evento de interés, o hasta la terminación del estudio. En el análisis inicial se incluyeron 100 covariables dicotómicas y 16 continuas. Además se ajustaron varios modelos de Cox para obtener las covariables significativas, eliminando las variables no significativas mediante el procedimiento paso a paso hacia atrás. Se aplicó también un análisis de residuos a los modelos definitivos para verificar los supuestos del modelo.

Para los cálculos de este ejemplo, muestran algunas tablas que son los resultados aplicado a los datos, de los 246 pacientes, utilizando el software S-PLUS 2000.

2.14.1 MODELO PARA DIÁLISIS PERITONEAL CON MUERTE COMO EVENTO DE INTERÉS.

Estimación de la función de supervivencia a través del estimador de Kaplan & Meier.

La estimación de la función de supervivencia, se obtiene para los 246 individuos, con una mediana de la supervivencia de 61 meses, es decir; que al menos la mitad de los individuos que estaban recibiendo diálisis peritoneal lograron sobrevivir hasta el mes 61 de seguimiento (ver Tabla 9).

Tabla 9. Valores resumen en la estimación de la función de supervivencia para diálisis peritoneal según meses.

n	Eventos	Media	ee(media)	Mediana	LCI (95%)	LCS (95%)
246	64	67.2	4.46	61	55	NA

Donde:

n: Número de individuos.

Eventos: Número total de muertes.

Media: Es la sobrevida media.

ee(media): Error estándar de la media.

Mediana: Es la sobrevida mediana.

LCI (0.95): Límite de Confianza Inferior de 95% para la sobrevida mediana.

LCS (0.95): Límite de Confianza Superior de 95% para la sobrevida mediana.

La Tabla 10, muestra la función de supervivencia estimada, mediante ésta se puede observar, que la fracción de individuos que lograron sobrevivir hasta el primer año fue de 91.6%, hasta el segundo año logran sobrevivir el 78.8% y hasta el quinto año lo logran el 50.4% y a partir de los 110 meses lo logran el 17.7%.

Tabla 10. Función de supervivencia estimada mediante el estimador de Kaplan & Meier para diálisis peritoneal según meses.

Tiempo	n.riesgo	n.eventos	Supervivencia	err.est.	LCI (95%)	LCS (95%)
0	246	2	0.992	0.00573	0.9807	1
1	240	1	0.988	0.00704	0.974	1
3	228	4	0.97	0.01102	0.949	0.992
4	221	1	0.966	0.01182	0.9431	989
5	215	1	0.962	0.01259	0.9372	0.987
6	209	1	0.957	0.01334	0.9311	0.983
7	202	1	0.952	0.01409	0.925	0.98
8	197	1	0.947	0.01483	0.9187	0.977
9	193	1	0.942	0.01554	0.9125	0.973
10	188	1	0.937	0.01625	0.9061	0.97
11	180	3	0.922	0.01831	0.8866	0.958
12	171	1	0.916	0.01898	0.88	0.954
13	161	1	0.911	0.0197	0.8729	0.95
14	151	4	0.887	0.02257	0.8435	0.932
15	144	1	0.88	0.02324	0.8361	0.927
17	135	3	0.861	0.02532	0.8127	0.912
19	124	1	0.854	0.02605	0.8044	0.907
20	119	2	0.84	0.02752	0.7873	0.895
21	115	3	0.818	0.02956	0.7617	0.878
22	110	3	0.795	0.03143	0.7361	0.859
23	104	1	0.788	0.03205	0.7274	0.853
25	94	1	0.779	0.03278	0.7177	0.846
26	90	1	0.771	0.03354	0.7077	0.839
28	81	1	0.761	0.03445	0.6966	0.832
30	78	2	0.742	0.03623	0.6739	0.816
31	75	4	0.702	0.03933	0.6291	0.784
33	63	1	0.691	0.04025	0.6164	0.775
34	59	1	0.679	0.04124	0.6031	0.765
37	53	2	0.654	0.04348	0.5737	0.745
39	50	1	0.641	0.04453	0.5589	0.734
42	43	1	0.626	0.04592	0.5418	0.722
47	38	1	0.609	0.04757	0.5227	0.71
52	33	1	0.591	0.04958	0.5011	0.696
55	31	1	0.572	0.05152	0.4791	0.682
59	26	2	0.528	0.05616	0.4283	0.65
60	22	1	0.504	0.0585	0.4012	0.632
61	21	1	0.48	0.06044	0.3748	0.614
65	18	1	0.453	0.06268	0.3455	0.594
77	11	1	0.412	0.092	0.2963	0.573
96	7	1	0.353	0.08054	0.2258	0.552
110	4	2	0.177	0.09701	0.0601	0.518

Donde:

Tiempo: Es el último mes de seguimiento.

n.riesgo: Número de individuos en riesgo antes del Tiempo.

n.evento: Número de muertes entre el Tiempo y el mes siguiente.

Supervivencia: Probabilidad de que un individuo sobreviva por un número de meses mayor al Tiempo.

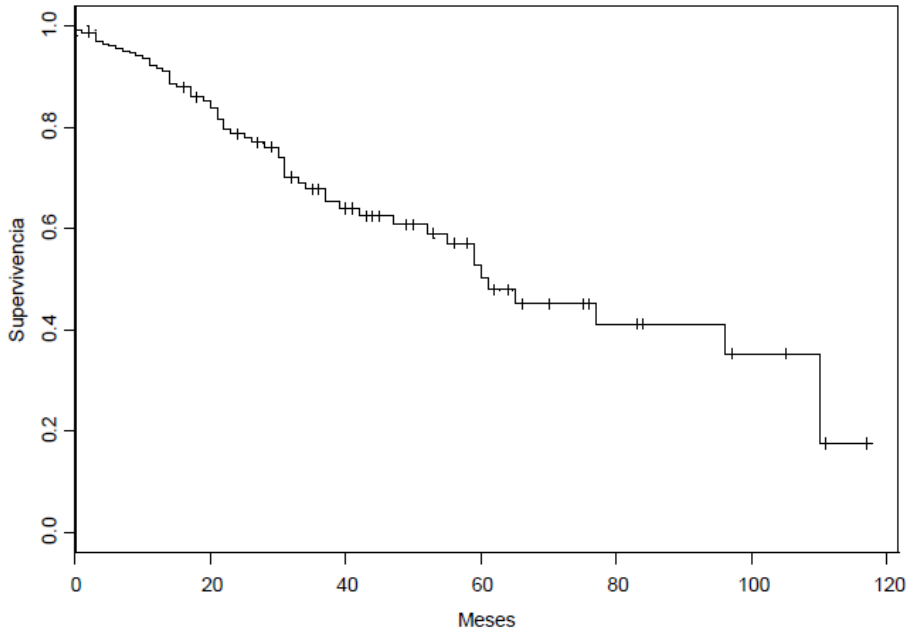
err.est: Error estándar de la Supervivencia.

LCI(95%): Límite de Confianza Inferior del 95% para la Supervivencia.

LCS(95%): Límite de Confianza Superior del 95% para la Supervivencia.

El Gráfico 1; muestra un patrón decreciente casi lineal de la función de supervivencia, lo cual pareciera estar indicando que las muertes por diálisis peritoneal tienen un comportamiento uniforme en el tiempo.

Gráfico No. 1. Función de supervivencia (KM) para Diálisis Peritoneal según meses.



2.14.2 MODELO DEFINITIVO DE REGRESIÓN DE RIESGO PROPORCIONAL (MODELO DE COX) PARA DIÁLISIS PERITONEAL SEGÚN MESES.

Para obtener el modelo de regresión de Cox, se construyó primero un modelo, en donde se incluyeron todas las variables dicotómicas que tenían una frecuencia mayor o igual que cinco y las covariables continuas. De este modelo, se fueron excluyendo las variables que no resultaban significativas al 10%, y se continuó el proceso de exclusión de covariables en otros modelos sucesivos, posteriormente se probaron modelos incluyendo todas las variables excluidas tomando a cada una por separado, obteniéndose al final un modelo que sólo incluye las covariables que resultaron significativas al 10%.
Adicionalmente a la significación de cada covariable, también fue tomada en cuenta la significación global del modelo; fijando en este caso un nivel de significación del 5%.

Covariables incluidas en el modelo: Diabetes, Edad y QUETELLET.

Puede afirmarse que las variables en estudio en este ejemplo son: Edad, QETELLET (Índice de masa corporal) y Diabetes son significativas al 10%, debido a que los p-valores obtenidos son todos menores que 0.10. Además se puede observar el coeficiente (coef) de las variables Diabetes y Edad que son positivos; estos representan que el riesgo de muerte por causas asociadas a la diálisis peritoneal de un individuo, es mayor cuando aumenta su edad y la enfermedad de la diabetes.
De forma contraria, el significado del coeficiente negativo que corresponde a la variable QUETELLET (Índice de masa corporal) con un valor de -0.0969; indica que el riesgo disminuye conforme aumenta el índice de masa corporal, lo cual se observa en la siguiente tabla 11.

Tabla 11. Estimación de los coeficientes para el modelo definitivo de Cox para diálisis peritoneal según meses.

Covariables	coef	exp(coef)	ee(coef)	z	p
Diabetes	0.5492	1.732	0.3208	1.71	0.087
Edad	0.0315	1.032	0.0097	3.25	0.0011
QUETELLET	-0.0969	0.908	0.0389	-2.49	0.013

Donde:

coef: Coeficiente estimado mediante el modelo.

exp(coef): Exponencial del coef y se interpreta como el riesgo.

ee(coef): Error estándar del coeficiente.

z: Estadístico de contraste para la significación del coeficiente.

p: El p-valor o valor de probabilidad de la significación del coeficiente.

Este modelo resulta significativo por cualquiera de los tres criterios (test) para un 10% de significancia, debido que los p-valores son todos menores que 0.10; como se muestra en la siguiente tabla 12.

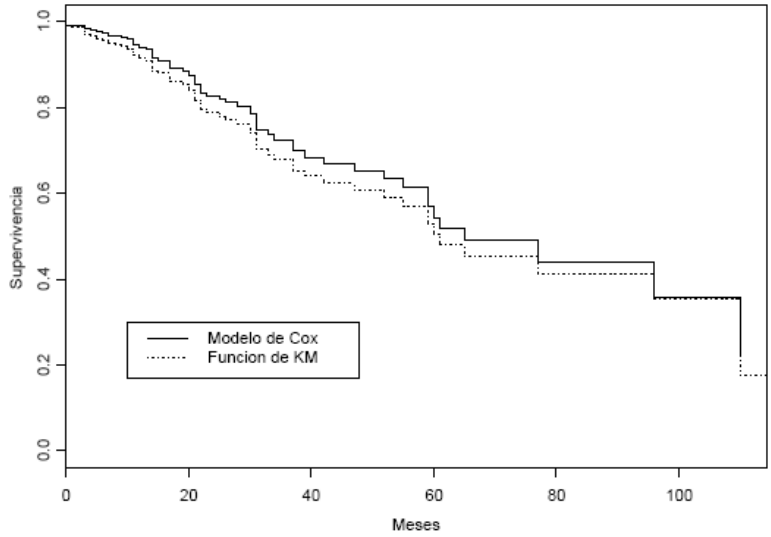
Tabla 12. Significación del modelo definitivo de Cox para diálisis peritoneal según meses.

Test	Estadístico	G. de libertad	p-valor
Test de razón de verosimilitud	18.8	3	0.000308
Test de wald	19.4	3	0.000229
Test de puntajes	19.8	3	0.000184

Para este ejemplo; se observa el test de razón de verosimilitud, con un p-valor de 0.000308, para el test de Wald fue de 0.000229 y para el test de puntajes fue de 0.000184 todos menores que 0.10, por lo que el modelo resulta significativo.

La figura 10; muestra el estimador de Kaplan & Meier de la función de supervivencia y el ajuste, mediante el modelo de Cox, para realizar las comparaciones entre ellas. En esta figura puede observarse que el ajuste obtenido mediante el modelo de Cox se ubica sistemáticamente por encima de la estimación de Kaplan & Meier.

Figura 10: Comparación del ajuste del modelo de Cox y el estimador de KM para Los pacientes de Diales peritoneal según meses.



Interpretación de los Coeficientes Estimados.

De la tabla 13, puede extraerse la información para analizar los riesgos y sus intervalos de confianza. La interpretación se hace de manera diferente para covariables dicotómicas y para covariables continuas. Los exponenciales de los coeficientes estimados pueden interpretarse de la manera siguiente:

Tabla 13. Exponencial de los coeficientes para el modelo definitivo de Cox para diálisis peritoneal según meses.

Covariables	exp(coef)	exp(-coef)	LCI(95%)	LCS(95%)
Diabetes	1.732	0.577	0.924	3.248
Edad	1.032	0.969	1.013	1.052
QUETELLET	0.908	1.102	0.841	0.979

Tipo de Covariables en estudio:

- ✓ Diabetes:

  Esta es una variable dicotómica entonces el estimador de riesgo se calcula como $e^{(0.5492)} = 1.732$; en donde $\hat{\beta} = 0.5492$. Este resultado está indicando que la presencia de Diabetes, aumenta el riesgo de muerte por causas asociadas a diálisis peritoneal en 1.732 veces, es decir; que un individuo con Diabetes tiene 1.732 veces más riesgo de morir por causas asociadas a dicha enfermedad, que un individuo que no tenga Diabetes.

  El intervalo de confianza del 95% es: $e^{(0.5492 \pm 1.96(0.3208))}$ es decir; que se ubica el riesgo entre 0.924 y 3.248.

- ✓ Edad:

  Para el caso de la Edad, por cada año que aumenta la edad del paciente el riesgo de morir por causas asociadas a la diálisis peritoneal es $e^{(0.0315)} = 1.032$ veces que los de un año inmediatamente anterior. Como esta es una variable continua, la interpretación puede hacerse para un período de distinto tamaño (c), pudiera decirse que al aumentar la edad de un paciente en 5 años el riesgo de morir por causas asociadas a diálisis peritoneal es $e^{(5*0.0315)} = 1.171$, es decir; $c = 5$ *años*.

  Mientras que para un aumento de 10 años es de $e^{(10*0.0315)} = 1.370$.

  El intervalo de confianza del 95%; para el riesgo de un año viene dado mediante la siguiente expresión: $e^{(1*0.0315 \pm 1.96*\|1\|(0.0097))}$; dicho valor se ubica entre 1.013 y 1.052.

- ✓ QUETELLET:

  La interpretación del índice de QUETELLET, es análoga a la de la covariable edad, es decir; que se trata de una covariable continua, por lo tanto; al aumentar dicha covariable en una unidad, $(c = 1)$ el riesgo es $e^{(1*(-0.0969))} = 0.908$ veces, en comparación con la unidad menor.

Este resultado está indicando que el índice de QUETELLET, pareciera ser un factor de protección, en lugar de un factor de riesgo, es decir; que mientras mayor es el índice, menor es la mortalidad por causas asociadas a diálisis peritoneal. El intervalo de confianza del 95% para el riesgo es $e^{(-0.0969 \pm 1.96(0.0389))}$, el cual ubica entre los valores 0.841 y 0.979.

Por lo tanto; el modelo final para este ejemplo es el siguiente:

$$\lambda(t/z) = \lambda_0 e^{(0.5492\,Diabetes + 0.0315\,Edad - 0.0969\,Quetellet)}$$

Debido a que el modelo resulta ser significativo, así como las covariables que intervienen en el modelo, es necesario llevar a cabo un análisis de residuos.

2.14.3 ANÁLISIS DE RESIDUOS.

Lo Primero que se analiza es si las covariables y el modelo satisfacen los supuestos de riesgos proporcionales. Esto puede hacerse utilizando la información de la tabla 14.

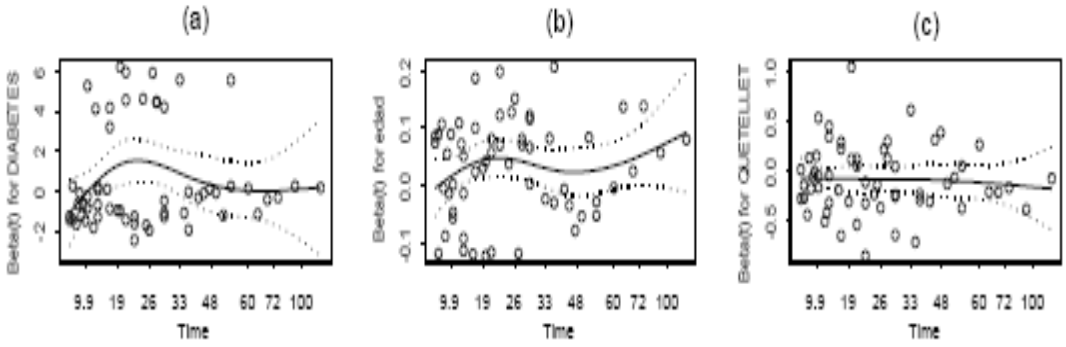
Tabla 14. Test de riesgos proporcionales para el modelo definitivo de Cox para diálisis peritoneal según meses.

Covariables	rho	chi2	p-valor
Diabetes	0.0357	0.0808	0.776
Edad	0.1165	1.0519	0.305
QUETELLET	-0.0540	0.2278	0.633
Global	No disponible	1.3791	0.71

Como puede observarse en la tabla anterior, al comparar los p-valores con un nivel de significación del 5%; observamos que cada una de las covariables se acepta, es decir; la hipótesis nula de los supuestos de riesgos proporcionales.

Al analizar el p-valor global, se concluye que no se viola el supuesto de riesgos proporcionales del modelo. A continuación se presenta el análisis de residuos mediante gráficos incluidos en la figura 11.

Figura 11. Verificación de los supuestos del modelo de Cox.



Donde:

(a) Es el gráfico para el test de riesgos proporcionales para diabetes.

(b) Es el gráfico para el test de riesgos proporcionales para edad.

(c) Es el gráfico para el test de riesgos proporcionales para Índice de Masa Corporal.

Supuestos de Riesgos Proporcionales.

Una de las principales hipótesis del modelo de Cox, es que la función de riesgo sea proporcional; para verificar este supuesto se utilizan los gráficos de los residuos de Schoenfeld versus el tiempo.

La verificación de los supuestos de riesgos proporcionales puede verse mediante los gráficos (a), (b) y (c) de la figura 11. En estos gráficos no se observa una violación del supuesto en cada una de las covariables ya que los residuos se agrupan de forma aleatoria a ambos lados del valor 0 del eje Y, sin presentar una tendencia con cambios bruscos.

La verificación del supuesto de riesgos proporcionales puede efectuarse a través de un contraste de hipótesis; donde la hipótesis nula esta asociada al cumplimiento del supuesto

de riesgos proporcionales. Los resultados de este contraste indican, que no se viola el supuesto de riesgos proporcionales para ninguna de las tres covariables.

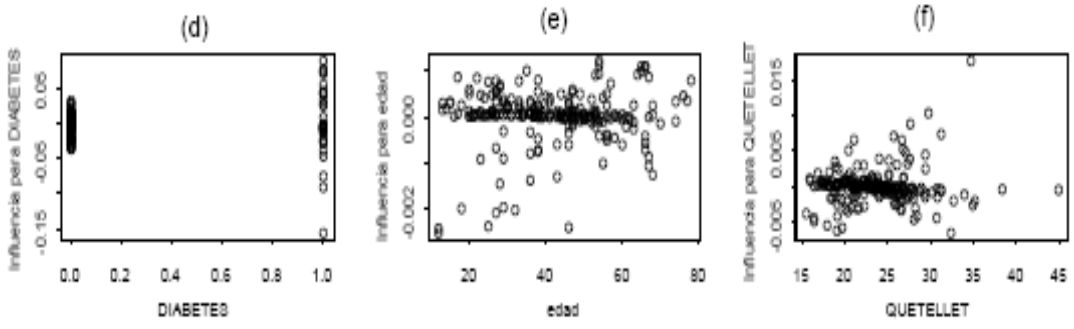
El resultado de los p-valores que se presenta en la tabla 14, realizados en el S-PLUS, asociados a este contraste para diabetes, edad e índice de masa corporal son 0.776, 0.305 y 0.633, respectivamente, observándose que todos son mayores al nivel de significancia que se ha tomado, que es del 10% es decir; que no se estaría rechazando la hipótesis de riesgos proporcionales para ninguna de las covariables.

Este contraste permite verificar la violación global del supuesto de riesgos proporcionales de todas las covariables.

Influencia de Individuos en la Estimación de los Coeficientes.

El supuesto de no influencia de los individuos sobre la estimación de cada coeficiente puede estudiarse graficando los residuos tipo score versus el correspondiente valor de cada covariable. En este ejemplo esta influencia puede verse a través de los gráficos (d), (e) y (f) de la figura 12.

Figura 12. Verificación gráfica de los residuos de scores.



Donde:

(d) Es el gráfico de influencias para diabetes.

(e) Es el gráfico de influencias para edad.

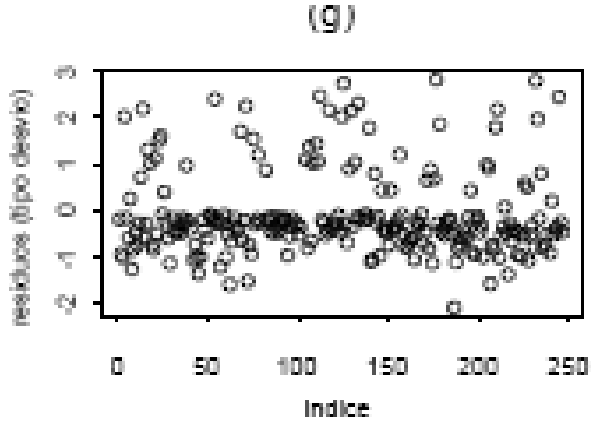
(f) Es el gráfico de influencias para Índice de Masa Corporal.

En estos gráficos se puede observar que para la diabetes y edad, no existen individuos que estén influyendo en la estimación de sus respectivos coeficientes, ya que no se observan valores extremos respecto al eje Y, lo cual indica que no existe alguna influencia de los individuos en la estimación de cada coeficiente del modelo. Para el Índice de Masa Corporal (Gráfico (f)), se observa que existe una observación que podría ser un dato u observación atípica, ya que éste, se aleja de los demás individuos, el cual se encuentra en la parte superior del gráfico, el cual está influyendo en la estimación de su coeficiente.

Influencia de Individuos en la Estimación del Modelo.

La verificación del supuesto de que no existen valores influyentes sobre la estimación del modelo, se hace graficando los residuos tipo deviance versus el índice (individuos).
En el gráfico (g) de la figura 13, no se observa ningún individuo que esté influyendo en la estimación del modelo, ya que el patrón es de una nube de puntos y además no se observan valores atípicos, puede verificarse el supuesto de que los individuos no afectan de modo negativo la estimación del modelo.

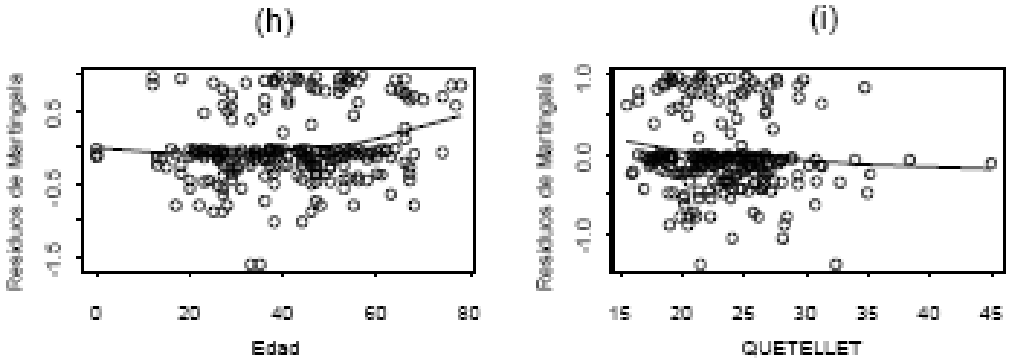
Figura 13. Verificación gráfica de los residuos deviance (desvío).



Forma Funcional de las Covariables Continuas.

Para verificar el supuesto de la forma funcional de cada covariable continua que interviene en el modelo, se utiliza el gráfico de los residuos de martingala versus el valor correspondiente a cada una de las covariables, acompañado de la curva suavizada.

Figura 14. Verificación gráfica de los residuos de martingala.



Donde:

(h) Es el gráfico para la verificación de la adecuacidad de la forma funcional de la edad.

(i) Es el gráfico para la verificación de la adecuacidad de la forma funcional del Índice de Masa Corporal.

En los gráficos (h) e (i) de la figura 14, se observa que la forma funcional es correcta en el modelo; ya que la línea que se traza en cada gráfico de estas variables (Edad y Quetellet) tiende al ajuste casi de una línea recta.

Conclusión.

El análisis de supervivencia clásico, es adecuado para la estimación de funciones de supervivencia y el ajuste de modelos de regresión para la obtención de covariables significativas, lo cual queda evidenciado en este ejemplo.

La incorporación del enfoque de procesos de conteo al análisis de supervivencia ha permitido el desarrollo de nuevas herramientas. En este caso se ha considerado sólo el análisis de los residuos para la verificación de los supuestos del modelo de Cox. Los aportes de este enfoque al análisis de supervivencia son muchos.

Puede concluirse que para el caso de los pacientes que acudían al Servicio de diálisis peritoneal del Hospital Clínico Universitario de Caracas, Venezuela entre los años 1980 y 1997, las covariables significativas en el modelo de Cox, fueron: Diabetes, Edad y el Índice de Masa Corporal. Estas covariables son las que estarían modificando el riesgo de muerte en los pacientes en diálisis peritoneal.

Se concluye además que el modelo de riesgos proporcionales presentado es adecuado ya que todos los supuestos se verifican.

CAPÍTULO III:

ANÁLISIS DE RESULTADOS

- **ANÁLISIS EXPLORATORIO.**
- **MÉTODO DE KAPLAN & MEIER.**
- **MODELO DE REGRESIÓN DE COX.**

PRÓLOGO.

En el capítulo anterior se realizó el enfoque teórico del análisis exploratorio de datos y el análisis de supervivencia; el cual será utilizado para realizar el análisis de la información que contiene la base de datos de los pacientes con Insuficiencia Renal del Hospital Militar Central.

En el presente capítulo se realizará el análisis exploratorio de datos y análisis de supervivencia utilizando el Método de Kaplan & Meier para determinar la probabilidad de vida de los pacientes; y así también se hará una comparación de la supervivencia de los pacientes con o sin Diabetes Mellitus esto utilizando el método adecuado de comparación entre dos grupos.

Por último se presentará un Modelo de Cox que estará compuesto por aquellas variables o factores más significativos que influyen en el incremento de la Insuficiencia Renal, Todo este análisis se realiza con el software estadístico SPSS, con el objetivo de realizar los cálculos de una forma más sencilla y rápido.

3.1 OBTENCIÓN DE LOS DATOS.

La información que se analiza en esta investigación, se ha recopilado de los expedientes de todos los pacientes del Hospital Militar Central, que padecen insuficiencia renal y han sido atendidos entre los años de 2002 a 2008.

De los pacientes antes mencionados se toma una muestra que incluye a aquellos pacientes que por lo menos en los meses de Julio a Septiembre de 2009, tuvieron una cita de control de su enfermedad con el médico que lo atiende; esta fue la única restricción para ser tomado en cuenta dentro de este estudio.

3.1.1 DESCRIPCIÓN DE LA BASE DE DATOS.

Los pacientes con Insuficiencia Renal del Hospital Militar Central; fueron registrados entre los años 2002 a 2008. Se considera una muestra de 180 pacientes de los cuales se registraron información que es organizada en un conjunto de variables, unas contenidas en la base de datos y otras obtenidas mediante la encuesta (Ver anexo 1); las cuales para efecto de análisis o estudio se agrupan en tres categorías que son:

- Datos Generales.
- Diagnóstico y Tratamiento.
- Información Complementaria.

A continuación se hará una descripción de las variables de cada categoría que se utilizará para todo el análisis.

Nombre de las Variables en la Base de Datos.

Datos Generales:

Variables	Descripción.	Tipo de variable
Expediente	Número correlativo que se le asigna al paciente en el hospital.	Cuantitativa
F_ Diagnostico	Fecha de diagnostico de la enfermedad.	Fecha
Sexo	Género del paciente.	Cualitativa
Edad	Edad actual del paciente.	Cuantitativa
Área	Área de la que procede el paciente.	Cualitativa
Zona	Zona geográfica a la cual pertenece el paciente.	Cualitativa
Nivel_Educ	Nivel educativo.	Cualitativa
Ocupación	Ocupación del paciente.	Cualitativa

Diagnostico y Tratamiento:

Variables	Descripción.	Tipo de variable
Tipo_Enfermedades	Enfermedades que el paciente con IR padece.	Cualitativa
Tipo_Medicamentos	Clases de medicamentos que ha usado el paciente.	Cualitativa
Tiempo_Enfermedad	Periodo de años con la enfermedad de IR.	Cuantitativa
Tipo_Tratamiento	Clase de tratamiento que el paciente se realiza.	Cualitativa
Tiempo_Tratamiento	Cantidad en años que el paciente tiene de administrarse el tratamiento.	Cuantitativa

Información Complementaria:

Variables	Descripción.	Tipo de variable
Ingreso_Economico	Ingreso económico del paciente	Cuantitativa
Utililiza_ProdAgri	Si el paciente ha utilizado productos para la agricultura.	Cualitativa
Tipo_Product	Clase de productos que ha utilizado para la agricultura.	Cualitativa
Periodo_Utililización	Periodo de años en que ha utilizado los productos para la agricultura.	Cuantitativa
Abastecimiento	Forma en que el paciente se abastece de agua para el consumo diario.	Cualitativa
Consumo_Agua	Cantidad de agua en vasos que consumía el paciente antes de padecer de IR.	Cuantitativa
Bebidas_Enlatadas	Consumo de bebidas enlatadas.	Cualitativa
Conocimiento_Enfermedad	Si el paciente tenía conocimiento de la enfermedad de IR.	Cualitativa

3.2 ANÁLISIS DESCRIPTIVO DE LAS VARIABLES.

A continuación se presenta el análisis univariado de las variables que contiene la base de datos, de los pacientes con Insuficiencia Renal del Hospital Militar Central; la cual fue fortalecida a través de una encuesta realizada a dichos pacientes. Estos resultados se presentan mediante tablas de frecuencia con sus respectivos gráficos, así como las medidas numéricas de algunos de los indicadores.

- **Variable sexo.**

Tabla 1. Sexo del Paciente.

Sexo	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Femenino	74	41,1	41,1	41,1
Masculino	106	58,9	58,9	100,0
Total	180	100,0	100,0	

Como se puede observar en la tabla 1, del total de la muestra; existen 74 pacientes con Insuficiencia Renal los cuales pertenecen al sexo femenino y 106 al sexo masculino, lo cual se representa como el 41.11% y 58.89% respectivamente. Por lo que se observa que la enfermedad de insuficiencia renal se presenta en mayor porcentaje, en pacientes del sexo masculino que del sexo femenino, es decir; que ser hombre es un factor de riesgo para padecer de la enfermedad, probablemente por el mayor contacto con otros factores exógenos que la mujer.

- **Variable edad.**

Tabla 2. Edad del Paciente con Insuficiencia Renal.

Edad	Frecuencia	Percentaje	Percentaje válido	Porcentaje acumulado
16-26 años	2	1,1	1,1	1,1
26-36 años	8	4,4	4,4	5,6
36-46 años	18	10,0	10,0	15,6
46-56 años	21	11,7	11,7	27,2
56-66 años	27	15,0	15,0	42,2
66-76 años	43	23,9	23,9	66,1
76-86 años	47	26,1	26,1	92,2
86-96 años	14	7,8	7,8	100,0
Total	180	100,0	100,0	

Gráfico 1. Edad del Paciente con Insuficiencia Renal.

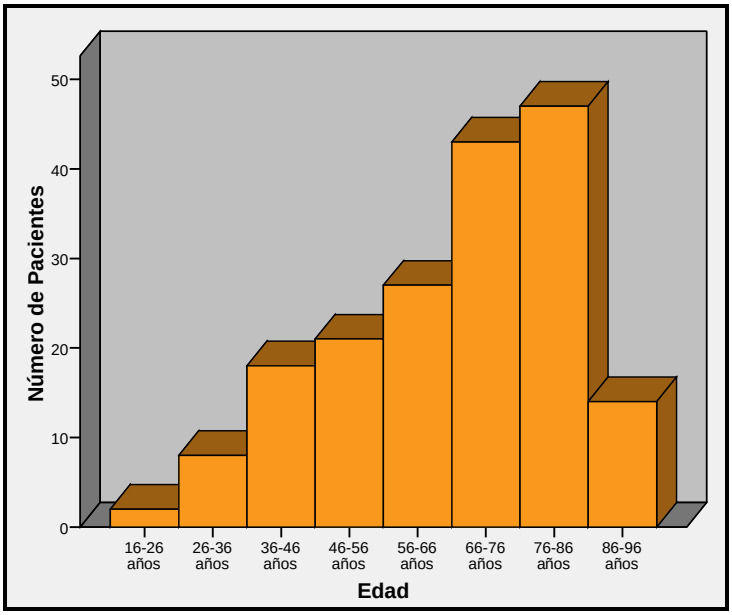


Tabla 3. Medidas numéricas.

Media		64,93
Mediana		68,50
Moda		43[a]
Deviación Estándar		16,576
Varianza		274,749
Rango		78
Percentiles	10	42,10
	20	48,20
	25	51,25
	30	56,30
	40	63,40
	50	68,50
	60	73,00
	70	76,00
	75	78,00
	80	80,00
	90	83,90

a. Multiple modes exist. The smallest value is shown

Nótese que en tabla 2, del total de la muestra; se tiene que el mayor número de pacientes con IR se encuentra en el rango de 76 a 86 años con un total de 47 de estos pacientes, esto expresa que a mayor edad mayor incidencia de la enfermedad, lo que pudiera significar que durante los años precedentes algunos factores estuvieren presentes a nivel renal; en el histograma (gráfico 1) se nota claramente que la mayor parte de las observaciones caen en la séptima clase, es decir en al séptima barra. El principal objetivo de la representación gráfica de las frecuencias relativas, es mostrar el perfil de la distribución de los datos y en este caso es notorio que la distribución de las edades de los pacientes con Insuficiencia Renal no es uniforme a través de todo el intervalo de valores.

En la tabla 3 se presentan las medidas numéricas de dicha variable, en la cual se refleja que la edad de los pacientes con IR; es aproximadamente en promedio 65 años. Además se tiene que la mitad de los pacientes tiene aproximadamente una edad mayor de 69 años mientras que el resto tienen una edad menor que dicho valor. También se tiene que la edad que más se observa en los pacientes es de 43 años, es decir; que la mayoría de pacientes tiene esta edad.

En el caso de las medidas de dispersión se tiene que las edades de los pacientes con IR tienen una distancia promedio de aproximadamente de 17 años en relación a la edad media que es de aproximadamente 65 años. Además se observa que 78 es la diferencia de años entre el paciente con mayor y menor edad. Sin embargo se tiene que el 25 % de pacientes se encuentran en edades por debajo de los 51 años.

- **Variable Área**

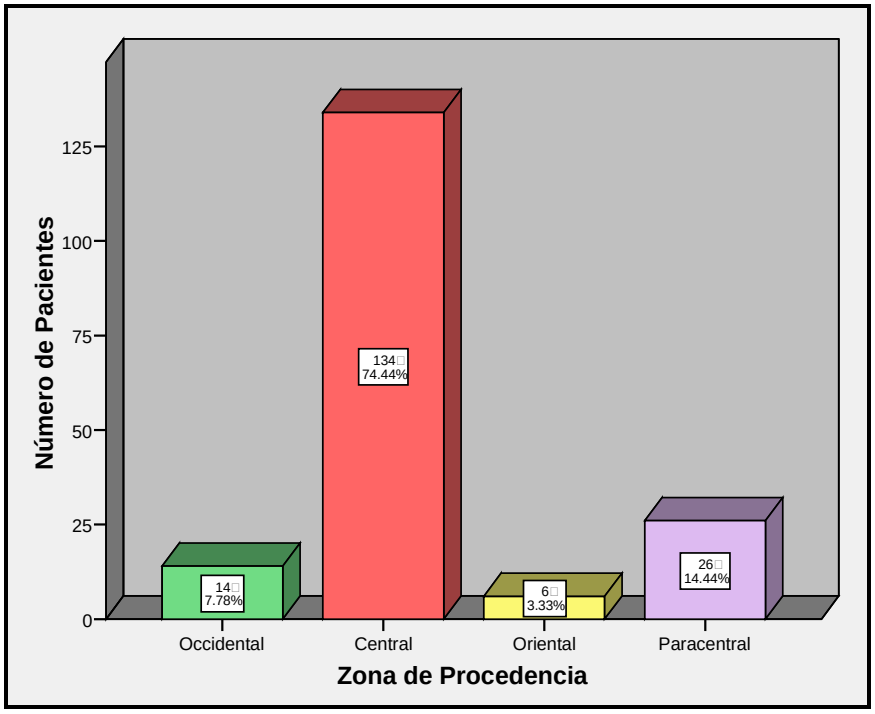
Tabla 4. Área de Procedencia del Paciente.

Área	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Rural	154	85,6	85,6	85,6
Urbana	26	14,4	14,4	100,0
Total	180	100,0	100,0	

En la tabla 4; se muestra el área de procedencia del paciente, en la cual se observa que 154 de estos, provienen del área rural y 26 del área urbana, esto corresponde a un 85.56% y 14.44% respectivamente del total de la muestra, es decir; que la mayor parte de pacientes proviene del área rural de El Salvador; lo que puede llegar a indicar que el lugar de procedencia de los pacientes es un factor de riesgo.

- **Variable Zona**

Gráfico 2. Zona de Procedencia del Paciente.



Nótese que en el gráfico 2; las zonas de mayor procedencia de los pacientes con IR; son la Zona Central y Paracentral del país; de las cuales la zona central sobresale con 134

pacientes, lo que es representado en el gráfico como la barra de mayor altura, el cual corresponde a un 74.44% de la muestra en estudio.

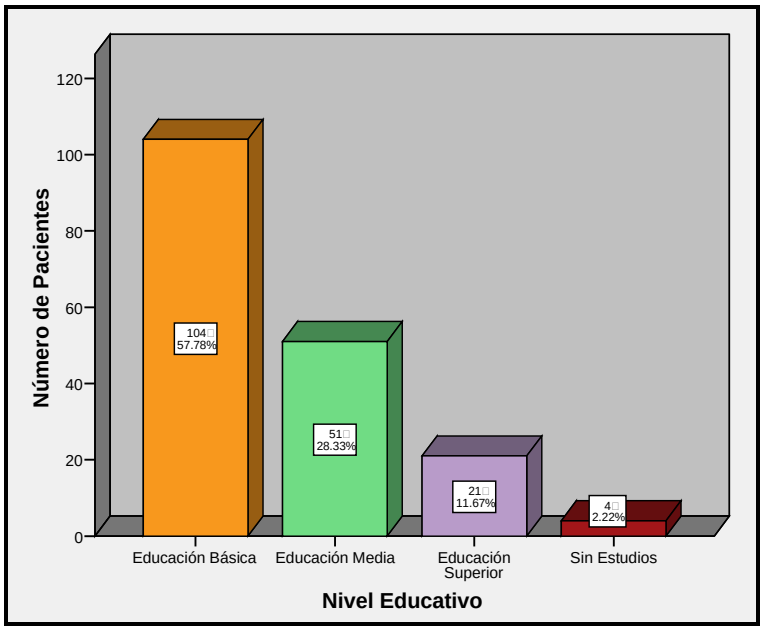
Se considera que la zona central esta compuesta por los departamentos de La Libertad, San Salvador, Chalatenango y la zona paracentral por los departamentos de Cabañas, La Paz, Cuscatlán, San Vicente y Usulután.

Por lo tanto se puede suponer que los pacientes que padecen IR, que provienen de la zona central están más expuestos a la contaminación ambiental, industrial, etc; que pueden ser considerados factores de riesgo.

Con respecto a la zona paracentral cabe mencionar que se han realizados estudios que muestran que en el departamento de La Paz; se usan cantidades industriales de plaguicidas para el algodón y la caña de azúcar, que son cultivos tradicionales de esa región; y por tanto a mayor contacto con tóxicos, esto se convierte en un posible factor de riesgo para los pacientes que residen en esa zona, sobre todo cuando se ha observado con anterioridad que la mayoría provienen de áreas rurales (ver tabla 4).

- **Variable Nivel_Educ**

Gráfico 3. Nivel Educativo de los Pacientes con Insuficiencia Renal.



En el gráfico 3; se observa que la mayoría de pacientes con IR su nivel educativo corresponde a una educación básica (1° a 9°), con un total de 104 pacientes, tal y como se muestra en el gráfico que corresponde a la barra más alta, lo que representa en términos de porcentaje un 57.8% del total de la muestra. También se tiene que 4 pacientes no han cursado ningún nivel educativo esto corresponde al 2.2% de la muestra, como se puede observar es la barra más pequeña del gráfico, es decir; que no tienen estudios académicos, lo que para este grupo podría representar un factor de riesgo, ya que a menor preparación, se tendrá un menor conocimiento de las enfermedades y de su prevención así, como también un menor interés por la salud, por lo que existirá una mayor incidencias de enfermedades.

- **Variable Ocupación.**

Tabla 5. Ocupación de los Pacientes con Insuficiencia Renal.

Ocupación	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Agricultor	13	7,2	7,2	7,2
Secretaria	12	6,7	6,7	13,9
Ama de Casa	23	12,8	12,8	26,7
Cocinera	9	5,0	5,0	31,7
Costurera	11	6,1	6,1	37,8
Motorista	5	2,8	2,8	40,6
Domestica	4	2,2	2,2	42,8
Enfermera	4	2,2	2,2	45,0
Comerciante	3	1,7	1,7	46,7
Electricista	2	1,1	1,1	47,8
Contador	2	1,1	1,1	48,9
Ingeniero Civil	1	,6	,6	49,4
Abogado	1	,6	,6	50,0
Ganadero	1	,6	,6	50,6
Albañil	1	,6	,6	51,1
Mecanico	1	,6	,6	51,7
Carpintero	6	3,3	3,3	55,0
Profesor	4	2,2	2,2	57,2
Telefonista	1	,6	,6	57,8
Estudiante	1	,6	,6	58,3
Jornalero	1	,6	,6	58,9
Grado Militar	74	41,1	41,1	100,0
Total	180	100,0	100,0	

En la tabla 5, se tiene que la ocupación más realizada por los pacientes encuestados, corresponde a Grado Militar; en la cual están agrupadas todas aquellas ocupaciones militares (soldados, sargentos, capitanes, coroneles, etc), que suman 72 casos, que corresponde al 41.1% de los pacientes en estudio, esto sucede porque el Hospital Militar

Central es parte de la Fuerza Armada de El Salvador, y la mayoría de pacientes laboran o laboraban en ramas afines a la Fuerza Armada. Se observa además que el 12.8% son Ama de Casa con 23 casos; luego sigue la ocupación de agricultor con un 7.2%; esta ocupación tiene una influencia importante en la incidencia de Insuficiencia Renal, ya que estudios realizados presentan que al tener contacto con plaguicidas, es un factor de riesgo importante en el desarrollo de dicha enfermedad.

- **Variable Ingreso_Económico.**

Tabla 6. Ingreso Económico de los Pacientes con Insuficiencia Renal.

Ingreso económico	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
< $100 al mes	2	1,1	1,1	1,1
$100-$149 al mes	25	13,9	13,9	15,0
$150-$300 al mes	101	56,1	56,1	71,1
$300 al mes	52	28,9	28,9	100,0
Total	180	100,0	100,0	

Gráfico 4. Ingreso Económico de los Pacientes con Insuficiencia Renal.

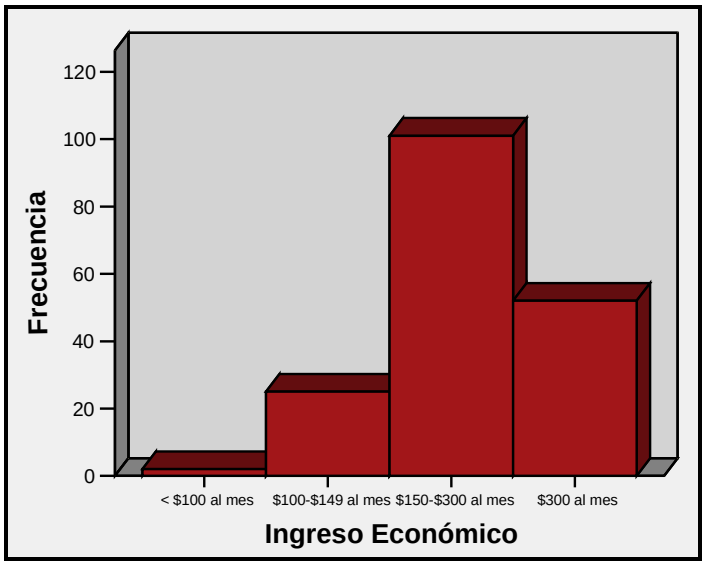


Tabla 7. Medidas numéricas.

Media		273,66
Mediana		284,77
Moda		225
Deviación Estándar		118,050
Varianza		13936,800
Rango		259
Percentiles	10	110,10
	20	150,20
	25	184,25
	30	200,30
	40	232,40
	50	252,50
	60	270,00
	70	285,00
	75	290,00
	80	352,00
	90	435,00

Obsérvese que en la tabla 6; la mayoría de pacientes, su ingreso económico se encuentra en el rango de $150 a $300; el cual representa el 56.1%, esto se observa en el gráfico 4 en el cual la mayor parte de las observaciones caen en la tercera clase, esto se debe a que la mayor parte de pacientes, su ingreso económico depende de la pensión que estos reciben, además la distribución de esta variable, no tiene un comportamiento uniforme a través de todo el intervalo de valores.

En la tabla 7 se presentan las medidas numéricas de dicha variable, en la cual se refleja que el ingreso de los pacientes con IR; es en promedio $273.66. Además se tiene que la mitad de los pacientes tiene un ingreso económico mayor de $284.77 mientras que el resto tienen un ingreso menor que dicho valor. También se tiene que $225 es el ingreso que más se observa, es decir; que la mayoría de pacientes tiene este ingreso.

En el caso de las medidas de dispersión se tiene el ingreso económico de los pacientes con IR tienen una distancia promedio de $118.05 en relación al ingreso promedio que es de $273.66. Además se observa que $259 es la diferencia de entre el paciente con mayor y menor ingreso económico. Sin embargo se tiene que el 25 % de los ingresos de los pacientes se encuentran por debajo de los $184.25.

- **Variable Utiliza_ProdAgri.**

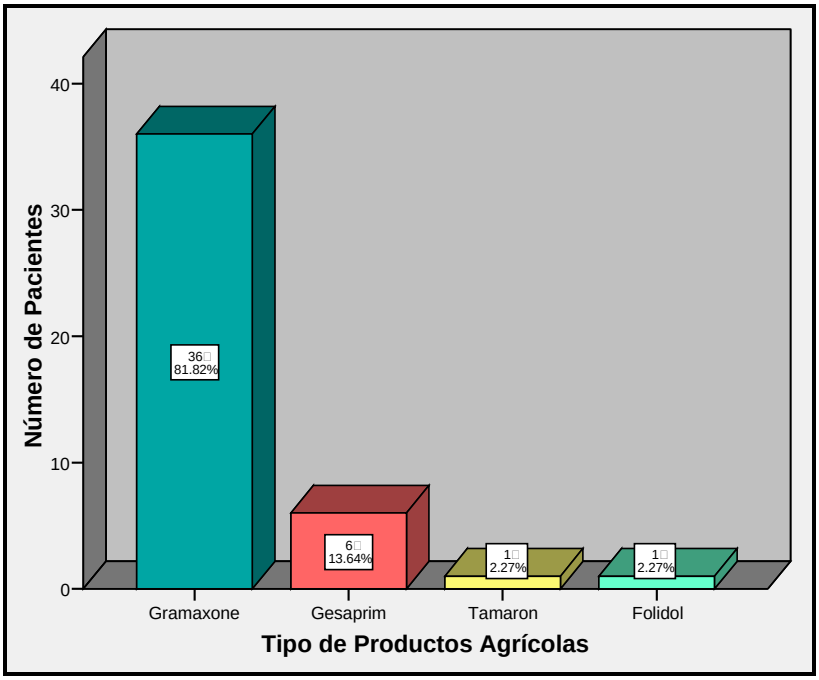
Tabla 8. Utiliza productos agrícolas.

Utilización de Productos Agrícolas	Frecuencia	Porcentaje	Porcentaje Acumulado
Si	44	24,4	24,4
No	136	75,6	100,0
Total	180	100,0	

De los pacientes con Insuficiencia Renal el 24.44%, manifiestan en algún momento o tiempo de su vida haber utilizado productos agrícolas, estos pacientes tienen un mayor factor de riesgo de desarrollar la enfermedad; debido a que han estado en contacto directo con dichos productos, luego se tiene que el 75.56% expresaron que no utilizaron. Por lo tanto; se muestra que una mayor proporción de estos pacientes no han tenido contacto con productos agrícolas.

- **Variable Tipo_porduct.**

Gráfico 5. Clase de productos agrícolas que fue utilizado por los pacientes con insuficiencia renal.



Del total de la muestra; 44 pacientes expresaron haber utilizado algún producto agrícola de los cuales el 81.82 % hizo uso del producto Gramoxone; el cual se muestra en el gráfico 5 como la barra más alta. Además se tiene que los productos menos utilizados fueron el Tamaron y Folidol ambos con un 2.3%.

Por lo tanto; se tiene que el uso de este tipo de productos podría convertirse en un factor de riesgo, ya que estudios revelan que el uso de los abonos y pesticidas tienen efectos negativos en la salud humana.

- **Variable Período_utilización.**

Tabla 9. Período de años en que el paciente con IR ha utilizado los productos para la agricultura.

Período de años en que el paciente con IR ha utilizado los productos para la agricultura	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
5-10 años	13	29.5	29.5	29.5
10-15 años	6	13.6	13.6	43.2
15-20 años	10	22.7	22.7	65.9
20-25 años	11	25.0	25.0	90.9
más de 25 años	4	9.1	9.1	100.0
Total	44	100.0	100.0	

Tabla 10. Medidas Numéricas

Media		14.80
Mediana		15.00
Moda		5
Deviación Estándar		8.601
Varianza		73.980
Rango		35
Percentiles	10	5.00
	20	5.00
	25	8.00
	30	9.00
	40	10.00
	50	15.00
	60	15.00
	70	20.00
	75	20.00
	80	20.00
	90	27.50

En la Tabla 9, se observa que del total de la muestra solamente 44 pacientes con IR han utilizado los productos agrícolas durante un periodo de más de 25 años. Entonces se tiene

que durante el período comprendido de 5 a 10 años, un 29.5% del total de los pacientes encuestados manifestaron haber utilizado productos agrícolas. Mientras que en el periodo de 20 a 25 años el porcentaje de pacientes fue de 25%.

En la tabla 10, se refleja que el periodo de tiempo en que utilizaron los productos agrícolas los pacientes con IR es aproximadamente en promedio un período de 15 años.

Además se tiene del total de pacientes que se tomaron para el estudio; 44 de ellos utilizaron productos agrícolas durante su vida de los cuales aproximadamente 22 de ellos lo han hecho por más de 15 años y el resto por menos de 15 años.

Nótese también que el periodo de tiempo que más se observa en que los pacientes utilizaron los productos agrícolas es de 5 años, es decir; que la mayoría de pacientes con IR utilizó los productos durante ese periodo de tiempo.

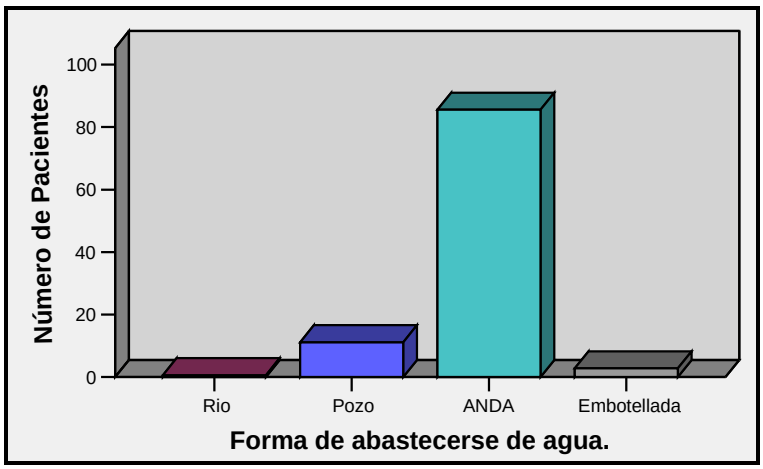
Se tiene que el periodo de tiempo en que fueron utilizados los productos agrícolas por los pacientes con IR, tienen una distancia promedio de aproximadamente de 9 años en relación al periodo de tiempo medio que fue de 15 años. Luego se observa que 35 es la diferencia de años entre el paciente con mayor y menor periodo de tiempo de haber utilizado dichos productos. Sin embargo se tiene que el 25% de pacientes se encuentran en periodos de tiempo de utilización menores que los 8 años.

- **Variable Abastecimiento.**

Tabla 11. Forma en que el paciente con IR se abastece de agua para el consumo diario.

Forma en que se abastece de agua.	Frecuencia	Porcentaje	Porcentaje Acumulado
Rio	1	,6	,6
Pozo	20	11,1	11,7
ANDA	154	85,6	97,2
Embotellada	5	2,8	100,0
Total	180	100,0	

Gráfico 6. Forma en que el paciente con IR se abastece de agua para el consumo diario.



En la tabla 11; se observa que del total de pacientes encuestados, 154 expresaron que la forma en que ellos se abastecían de agua para su consumo fue por medio del servicio de agua por cañería, es decir; por medio del abastecimiento público ANDA, siendo este el 85.6%, el cual corresponde a la barra más alta en el gráfico 6. Mientras que un 11.1% se abastece de agua por medio de pozo. Además se tiene que las formas de abastecimiento de agua que menos utilizan los pacientes fueron rio y agua embotellada ambas suman un valor de 3.4%. Por lo tanto; puede suponerse que la forma en que los pacientes se abastecen de agua por medio de ANDA puede ser un factor de riesgo, ya que esta no posee los estándares de calidad en el tratamiento del agua para el consumo humano y podría estar contaminada a la hora de ser ingerida provocando diferentes enfermedades en el organismo.

- **Variable Tipo_ Enfermedades.**

  - **Diabetes Mellitus.**

Tabla 12. Padece de la enfermedad de Diabetes Mellitus.

Padece de la enfermedad de Diabetes Mellitus	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Si	114	63.3	63.3	63.3
No	66	36.7	36.7	100.0
Total	180	100.0	100.0	

En la tabla 12, se tiene que más de la mitad de pacientes encuestados padecían de Diabetes Mellitus antes de que se les diagnosticara la enfermedad de IR, lo cual representa un total de 114, siendo este un 63.3%; mientras que del total de los pacientes encuestados un 36.67% manifiestan que no padecen dicha enfermedad. Como se observa la mayoría de pacientes padecían de diabetes; esta es silenciosa y sus efectos son muy peligrosos si no se toma conciencia de la realidad de dicha enfermedad. Estudios realizados muestran, que el exceso de azúcar en la sangre puede producir daños en ojos, riñones y en el sistema circulatorio**,** entre otros. Por lo tanto; la diabetes podría ser un factor de riesgo importante para el desarrollo de una insuficiencia renal.

- **Tiempo de padecer Diabetes Mellitus.**

Tabla 13. Cantidad de años que el paciente con IR tiene de padecer de la enfermedad de Diabetes Mellitus.

Cantidad de años que el paciente con IR tiene de padecer de la enfermedad de Diabetes Mellitus.	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
1-5 años	2	1.8	1.8	1.8
5-10 años	30	26.3	26.3	28.1
10-15 años	30	26.3	26.3	54.4
15-20 años	21	18.4	18.4	72.8
20-25 años	14	12.3	12.3	85.1
25-30 años	17	14.9	14.9	100.0
Total	114	100.0	100.0	

Tabla 14. Medidas Numéricas

Media		13.98
Mediana		13.00
Moda		10[a]
Deviación Estándar		7.269
Varianza		52.832
Rango		28
Percentiles	10	5.00
	20	8.00
	25	8.00
	30	10.00
	40	10.00
	50	13.00
	60	15.00
	70	16.50
	75	20.00
	80	20.00
	90	25.00

a. Multiple modes exist. The smallest value is shown

La tabla 13, muestra la cantidad de años que el paciente con IR tiene de padecer de la enfermedad de Diabetes Mellitus; se tiene que en el rango de 5 a 10 y de 10 a 15 años cada uno con un total de 30 pacientes, es decir; un 26.3% respectivamente. Luego le sigue un 18.4% que corresponde al tiempo de 15 a 20 años, en estos rangos de tiempo es donde se concentra la mayoría de pacientes.

En la tabla 14, se refleja que el periodo de tiempo en que tienen con la enfermedad de diabetes mellitus es aproximadamente en promedio un período de 14 años.

Además se tiene del total de pacientes que se tomaron para el estudio; 114 se les diagnostico diabetes antes de la enfermedad de IR de los cuales aproximadamente 57 de ellos tienen la enfermedad por más de 13 años y el resto por menos de 13 años.

Nótese también que el periodo de tiempo que más se observa que los pacientes tienen con la enfermedad de diabetes es de 10 años, es decir; que la mayoría de pacientes tienen esa enfermedad en ese periodo de tiempo.

Se tiene que el tiempo que tienen de padecer de diabetes tienen una distancia promedio de 7 años en relación al periodo de tiempo medio que fue de aproximadamente 14 años. Luego se observa que 28 es la diferencia de años entre el paciente con mayor y menor periodo de

tiempo con la enfermedad de diabetes. Sin embargo se tiene que el 75% de pacientes se encuentran en periodos de tiempo con la enfermedad menores que los 20 años.

Por lo tanto, la Diabetes Mellitus es un factor de riesgo de gran importancia para el deterioro de la función renal ya que como se sabe, una persona enferma de diabetes que no recibe los cuidados y el tratamiento adecuado esto puede reducir su calidad de vida en forma importante, llegando incluso a padecer problemas en su organismo. Además se tiene que la IR se puede evitar, ya que numerosos estudios han demostrado que las personas con diabetes pasan hasta 15 años sin saber que el daño renal en su organismo está progresando y que hay forma de detenerlo por completo.

❖ Hipertensión Arterial (HA).

Tabla 15. Padece de la enfermedad de Hipertensión Arterial.

Padece de la enfermedad de Hipertensión Arterial	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Si	138	76.7	76.7	76.7
No	42	23.3	23.3	100.0
Total	180	100.0	100.0	

Nótese que la tabla 15, se presenta que 138 pacientes del total de la muestra padecen de hipertensión arterial, el cual representa un 76.7%, mientras que un 23.3% declaro que no han padecido de dicha enfermedad.

Por lo tanto; la hipertensión arterial es un factor de riesgo para llegar a padecer de IR ya que estudios realizados presentan que el problema más grande de esta enfermedad es que, en la mayoría de los casos, el enfermo no sabe y al no recibir un tratamiento adecuado, la hipertensión puede provocarle daños en órganos vitales, como los riñones, los ojos, el corazón, entre otros.

- **Tiempo de padecer Hipertensión Arterial.**

Tabla 16. Cantidad de años que el paciente con IR tiene de padecer de Hipertensión Arterial.

Cantidad de años que el paciente con IR tiene de padecer de Hipertensión Arterial.	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
1-5 años	7	5.1	5.1	5.1
5-10 años	21	15.2	15.2	20.3
10-15 años	41	29.7	29.7	50.0
15-20 años	34	24.6	24.6	74.6
20-25 años	12	8.7	8.7	83.3
25-30 años	23	16.7	16.7	100.0
Total	138	100.0	100.0	

Tabla 17. Medidas Numéricas.

Media		3.67
Mediana		3.50
Moda		3
Deviación Estándar		1.421
Varianza		2.019
Rango		5
Percentiles	10	2.00
	20	2.00
	25	3.00
	30	3.00
	40	3.00
	50	3.50
	60	4.00
	70	4.00
	75	5.00
	80	5.00
	90	6.00

La tabla 16, representa el tiempo de padecer de la enfermedad de Hipertensión Arterial en pacientes con Insuficiencia Renal, aquí se observa que la mayoría de pacientes están concentrados en los periodos de tiempo de 10 a 15 y de 15 a 20 años, es decir; un 54.3% del total de la población encuestada, estos tiempos corresponden a la tercera y cuarta clase respectivamente.

En la tabla 17, se refleja que el periodo de tiempo en que tienen con la enfermedad de Hipertensión Arterial es aproximadamente en promedio un período de 4 años.

Además se tiene que del total de pacientes que se tomaron para el estudio; 138 se les diagnostico hipertensión Arterial antes de la enfermedad de IR de los cuales 69 de ellos tienen la enfermedad aproximadamente por más de 4 años y el resto por menos de 4 años.

Nótese también que el periodo de tiempo que más se observa que los pacientes tienen con la enfermedad de HA es de 3 años, es decir; que la mayoría de pacientes tienen esa enfermedad en ese periodo de tiempo.

Se tiene que el tiempo que tienen de padecer de HA tienen una distancia promedio de aproximadamente 1.5 años en relación al periodo de tiempo medio que fue aproximadamente de 4 años. Luego se observa que 5 es la diferencia de años entre el paciente con mayor y menor periodo de tiempo con la enfermedad de HA. Sin embargo se tiene que el 75% de pacientes se encuentran en periodos de tiempo con la enfermedad menores que los 5 años y el resto mayor.

❖ Infección de Vías Urinarias (IVU).

Tabla 18. Padece de la enfermedad de Infección de vías Urinarias.

Padece de la enfermedad de Infección de vías Urinarias	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Si	65	36.1	36.1	36.1
No	115	63.9	63.9	100.0
Total	180	100.0	100.0	

Se observa que en la tabla 18, del total de la muestra 115 expresaron no padecer de infección en las vías urinarias el cual representa un 63.89%, mientras que solamente un 36.11% manifestaron que sí han padecido de dicha enfermedad. Para este grupo de pacientes representa un factor de riesgo para el desarrollo de la enfermedad de IR, ya que estudios médicos señalan que las infecciones repetidas no tratadas, llegan a producir cicatrices y lesiones en los riñones que pueden dañarlos y afectar su función.

- **Tiempo de padecer infección de vías urinarias.**

Gráfico 7. Cantidad de años que el paciente con IR tiene de padecer de Infección de Vías Urinarias.

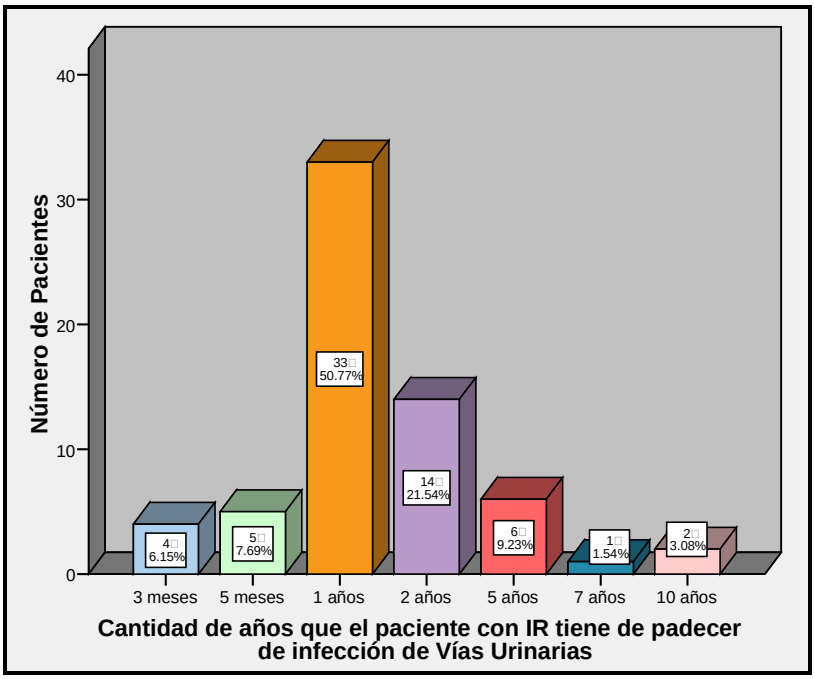


Tabla 19. Medidas Numéricas

Media		22.65
Mediana		12.00
Moda		12
Deviación Estándar		24.254
Varianza		588.263
Rango		117
Percentiles	10	5.00
	20	12.00
	25	12.00
	30	12.00
	40	12.00
	50	12.00
	60	12.00
	70	24.00
	75	24.00
	80	24.00
	90	60.00

Nótese que en el gráfico 7, del total de la muestra 33 pacientes tienen 12 meses de estar con Infección de Vías Urinarias, es decir; un 50.8%, mientras que sólo existe un paciente que ha

venido desarrollando dicha enfermedad en un periodo de 84 meses. Así mismo se observa que estos periodos de tiempo corresponden a la tercera y sexta barra.

Por lo tanto; éste tiene una mayor probabilidad de que se le desarrolle una insuficiencia renal por padecer mucho tiempo con infección de vías urinarias.

En la tabla 19, se refleja que el período de tiempo en que tienen con la enfermedad de IVU es aproximadamente en promedio un período de 23 meses.

Además, se tiene que del total de pacientes que se tomaron para el estudio; 65 se les diagnosticó IVU antes de la enfermedad de IR de los cuales aproximadamente 33 de ellos tienen la enfermedad por más de 12 meses y el resto por menos de 12 meses.

Nótese también que el período de tiempo que más se observa que los pacientes tienen con la enfermedad de IVU, es de 12 meses. Es decir; que la mayoría de pacientes han tenido esa enfermedad en ese período de tiempo.

Se observa que el tiempo de padecer de IVU, tiene una distancia promedio de 24 meses en relación al período de tiempo medio que fue aproximadamente de 23 meses. Luego se muestra que 117 meses, es la diferencia entre el paciente con mayor y menor período de tiempo con la enfermedad de IVU. Sin embargo se tiene que el 75% de pacientes se encuentran en periodos de tiempo con la enfermedad menores que los 24 meses.

- **Variable Tipo_Medicamento.**

  - **Medicamento Ibuprofeno.**

Tabla 20. Consumo del medicamento Ibuprofeno.

Ha consumido el medicamento Ibuprofeno	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Si	115	63.9	63.9	63.9
No	65	36.1	36.1	100.0
Total	180	100.0	100.0	

En la tabla 20, se muestra que 115 pacientes, se han suministrado el medicamento Ibuprofeno, mientras que 65 de ellos manifestaron no haber consumido dicho medicamento.

Por lo tanto; según investigaciones medicas, las personas que toman medicamentos antiinflamatorios no esteroides (AINE) como el ibuprofeno, pueden tener efectos adversos en su organismo.

- **Tiempo de consumir el medicamento Ibuprofeno.**

Tabla 21. Cantidad de meses que el paciente con IR tiene de consumir el medicamento Ibuprofeno.

Cantidad de meses que el paciente con IR tiene de consumir el medicamento Ibuprofeno	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
1 mes	15	13.0	13.0	13.0
2 meses	18	15.7	15.7	28.7
5 meses	5	4.3	4.3	33.0
6 meses	2	1.7	1.7	34.8
8 meses	1	.9	.9	35.7
12 meses	13	11.3	11.3	47.0
24 meses	13	11.3	11.3	58.3
36 meses	11	9.6	9.6	67.8
60 meses	30	26.1	26.1	93.9
72 meses	1	.9	.9	94.8
120 meses	5	4.3	4.3	99.1
360 meses	1	.9	.9	100.0
Total	115	100.0	100.0	

Tabla 22. Medidas Numéricas.

Media		32.56
Mediana		24.00
Moda		60
Desviación Estándar		43.489
Varianza		1891.267
Rango		359
Percentiles	10	1.00
	20	2.00
	25	2.00
	30	3.00
	40	12.00
	50	24.00
	60	36.00
	70	60.00
	75	60.00
	80	60.00
	90	60.00

Nótese que en la tabla 21, del total de la muestra 30 pacientes tienen 60 meses de estar consumiendo el medicamento Ibuprofeno, es decir; un 26.1%, mientras que sólo existe un paciente que ha venido consumiendo dicho medicamento en los períodos de 8, 72 y 360 meses.
En la tabla 22, se refleja que el período de tiempo que tienen los pacientes de consumir el medicamento ibuprofeno es aproximadamente en promedio un período de 33 meses.

Además, se tiene que del total de pacientes que se tomaron para el estudio; 115 han consumido ibuprofeno de los cuales aproximadamente 58 de ellos tienen de haber consumido dicho medicamento por mas de 24 meses y el resto por menos de 24 meses.
Nótese también que el período de tiempo que más se observa que los pacientes tienen de haber consumido el medicamento ibuprofeno, es de 60 meses. Es decir; que la mayoría de pacientes tienen ese periodo de tiempo de estar consumiendo este medicamento.
Luego se observa que el tiempo de consumir dicho medicamento, tiene una distancia promedio de 44 meses en relación al período de tiempo medio que fue aproximadamente de 33 meses. Se muestra además que 359, es la diferencia de meses entre el paciente con mayor y menor período de tiempo de haber consumido el medicamento ibuprofeno. Sin embargo se tiene que el 75% de pacientes se encuentran en periodos de tiempo de consumo menores que los 60 meses.

- **Medicamento Acetaminofen.**

Tabla 23. Consumo del medicamento Acetaminofén.

Ha consumido el medicamento Acetaminofén	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Si	141	78.3	78.3	78.3
No	39	21.7	21.7	100.0
Total	180	100.0	100.0	

En la tabla 23, se observa que 141 pacientes se han suministrado el medicamento Acetaminofén, es decir un 78.3% del total de la muestra, mientras que solo un 21.67% manifestaron no haber consumido dicho medicamento.

Por lo tanto; el consumo de este medicamento podría ser un factor de riesgo en el desarrollo del IR, ya que este medicamento es suministrado como un antiinflamatorio y estudios médicos reflejan, que el uso frecuente de este medicamento puede provocar daños hepáticos porque la ingesta del componente, a largo plazo, afecta las encimas de los órganos del cuerpo humano como el hígado y riñón, entre otros.

- **Tiempo de consumir el medicamento Acetaminofén.**

Tabla 24. Cantidad de meses que el paciente con IR tiene de consumir el medicamento Acetaminofén.

Cantidad de meses que el paciente con IR tiene de consumir el medicamento Acetaminofen.	Frecuencia	Percentaje	Porcentaje válido	Porcentaje acumulado
2 meses	24	17.0	17.0	17.0
3 meses	11	7.8	7.8	24.8
6 meses	3	2.1	2.1	27.0
12 meses	23	16.3	16.3	43.3
24 meses	28	19.9	19.9	63.1
36 meses	13	9.2	9.2	72.3
60 meses	30	21.3	21.3	93.6
72 meses	1	.7	.7	94.3
120 meses	7	5.0	5.0	99.3
360 meses	1	.7	.7	100.0
Total	141	100.0	100.0	

Tabla 25. Medidas Numéricas

Media		32.53
Mediana		24.00
Moda		60
Desviación Estándar		40.616
Varianza		1649.651
Rango		358
Percentiles	10	2.00
	20	3.00
	25	4.50
	30	12.00
	40	12.00
	50	24.00
	60	24.00
	70	36.00
	75	60.00
	80	60.00
	90	60.00

Nótese que en la tabla 24, del total de la muestra 30 pacientes tienen 60 meses de estar consumiendo el medicamento Acetaminofén, es decir; un 21.3%, mientras que sólo existe

un paciente que ha venido consumiendo dicho medicamento en los períodos de 72 y 360 meses.

Por lo tanto; éste tiene una mayor probabilidad de que se le desarrolle una IR por consumir por mucho tiempo este medicamento.

En la tabla 25, se refleja que el período de tiempo que tienen los pacientes de consumir el medicamento Acetaminofén es aproximadamente en promedio un período de 33 meses. Además, se tiene que del total de pacientes que se tomaron para el estudio; 141 han consumido este medicamento, de los cuales aproximadamente 60 de ellos tienen de haberlo consumido por más de 24 meses y el resto por menos de 24 meses.

Nótese también que el período de tiempo que más se observa que los pacientes tienen de haber consumido el medicamento Acetaminofén, es de 60 meses. Es decir; que la mayoría de pacientes han utilizado dicho medicamento en ese período de tiempo.

Se observa que el tiempo de consumir el medicamento Acetaminofen, tiene una distancia promedio de 41 meses en relación al período de tiempo medio que fue aproximadamente de 33 meses. Luego se muestra que 60, es la diferencia de meses entre el paciente con mayor y menor cantidad de meses de estar consumiendo dicho medicamento.

Sin embargo se tiene que el 75% de pacientes se encuentran en periodos de tiempo de consumo menores que los 60 meses.

- **Medicamento Diclofenac.**

Tabla 26. Consumo del medicamento Diclofenac.

Ha consumido el medicamento Diclofenac	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Si	42	23.3	23.3	23.3
No	138	76.7	76.7	100.0
Total	180	100.0	100.0	

En la tabla 26, se observa que 42 pacientes se han suministrado el medicamento Diclofenac, es decir un 23.33% del total de la muestra, mientras que solo un 76.67% manifestaron no haber consumido dicho medicamento.

Por lo tanto; el consumo de este medicamento podría ser un factor de riesgo en el desarrollo del IR, ya que estudios médicos muestran que este medicamento, al ser suministrado por largo tiempo puede provocar daños en la función renal y hepática.

- **Tiempo de consumir el medicamento Diclofenac.**

Gráfico 8. Cantidad de meses que el paciente con IR tiene de consumir el medicamento Diclofenac

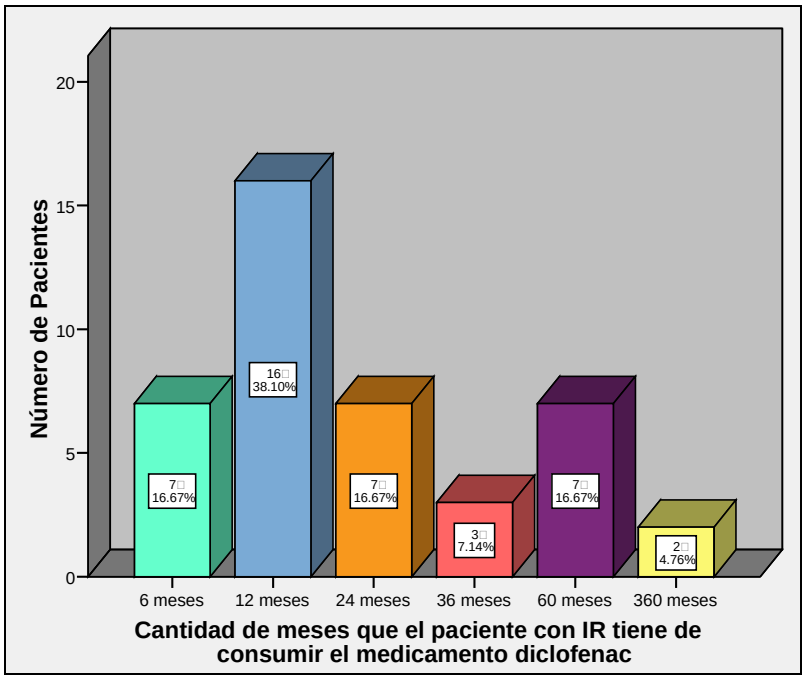


Tabla 27. Medidas Numéricas.

Media		3.05
Mediana		2.00
Moda		2
Deviació Estándar		1.860
Varianza		3.461
Rango		6
Percentiles	10	1.00
	20	2.00
	25	2.00
	30	2.00
	40	2.00
	50	2.00
	60	3.00
	70	3.10
	75	4.00
	80	6.00
	90	6.00

Nótese que en el gráfico 8, del total de la muestra 16 pacientes tienen 12 meses de estar consumiendo el medicamento Diclofenac, es decir; un 38.10%, mientras que sólo existen

siete paciente que han venido consumiendo dicho medicamento en los períodos de 6, 24 y 60 meses. Así mismo se observa en el gráfico 23, que estos periodos de tiempo corresponden a las barras medianas.

Por lo tanto; éste tiene una mayor probabilidad de que se le desarrolle una IR por consumir por mucho tiempo este medicamento.

En la tabla 27, se refleja que el período de tiempo que tienen los pacientes de consumir el medicamento Acetaminofén es aproximadamente en promedio un período de 3 meses. Además, se tiene que del total de pacientes que se tomaron para el estudio; 42 han consumido este medicamento, de los cuales aproximadamente 21 de ellos tienen de haberlo consumido por más de 2 meses y el resto por menos de 2 meses.

Nótese también que el período de tiempo que más se observa que los pacientes tienen de haber consumido el medicamento Diclofenac, es de 2 meses. Es decir; que la mayoría de pacientes han utilizado dicho medicamento en ese período de tiempo.

Se observa que el tiempo de consumir el medicamento Diclofenac, tiene una distancia promedio de 4 meses en relación al período de tiempo medio que fue aproximadamente de 3 meses. Luego se muestra que 6, es la diferencia de meses entre el paciente con mayor y menor cantidad de meses de estar consumiendo dicho medicamento.

Sin embargo se tiene que el 75% de pacientes se encuentran en periodos de tiempo de consumo menores que los 4 meses.

❖ Medicamento Enalapril

Tabla 28. Consumo del medicamento Enalapril.

Ha consumido el medicamento Enalapril	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Si	88	48.9	48.9	48.9
No	92	51.1	51.1	100.0
Total	180	100.0	100.0	

En la tabla 28, se observa que 88 pacientes se han suministrado el medicamento Enalapril, es decir un 48.89% del total de la muestra, mientras que sólo un 51.11% manifestaron no haber consumido dicho medicamento.

Por lo tanto; el consumo del medicamento Enalapril; se utiliza para tratar la presión arterial y el fallo congestivo del corazón. El uso de este medicamento podría ser un factor de riesgo en el desarrollo del IR, ya que estudios médicos muestran que al ser suministrado por largo tiempo puede provocar daños en la función renal.

Tabla 29. Cantidad de meses que el paciente con IR tiene de consumir el medicamento Enalapril.

Cantidad de meses que el paciente con IR tiene de consumir el medicamento Enalapril	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
2 meses	1	1,1	1,1	1,1
6 meses	1	1,1	1,1	2,3
12 meses	10	11,4	11,4	13,6
24 meses	11	12,5	12,5	26,1
36 meses	11	12,5	12,5	38,6
48 meses	1	1,1	1,1	39,8
60 meses	13	14,8	14,8	54,5
72 meses	11	12,5	12,5	67,0
120 meses	29	33,0	33,0	100,0
Total	88	100,0	100,0	

Tabla 30. Medidas Numéricas.

Media		6,57
Mediana		7,00
Moda		9
Desviación Estándar		2,338
Varianza		5,467
Rango		8
Percentiles	10	3,00
	20	4,00
	25	4,00
	30	5,00
	40	6,60
	50	7,00
	60	8,00
	70	9,00
	75	9,00
	80	9,00
	90	9,00

Nótese que en la tabla 29, del total de la muestra 29 pacientes tienen 120 meses de estar consumiendo el medicamento Enalapril, es decir; un 33.0%, mientras que en los periodos

comprendidos de 2, 6 y 48 meses sólo existen un paciente que han venido consumiendo dicho medicamento en esos períodos.

En la tabla 30, se refleja que el período de tiempo que tienen los pacientes de consumir el medicamento Enalapril es aproximadamente en promedio un período de 7 meses. Además, se tiene que del total de pacientes que se tomaron para el estudio; 88 han consumido este medicamento, de los cuales aproximadamente 44 de ellos tienen de haberlo consumido por más de 7 meses y el resto por menos de 7 meses.

Nótese también que el período de tiempo que más se observa que los pacientes tienen de haber consumido el medicamento Enalapril, es de 9 meses. Es decir; que la mayoría de pacientes han utilizado dicho medicamento en ese período de tiempo.

Se observa que el tiempo de consumir el medicamento Enalapril, tiene una distancia promedio de 2 meses en relación al período de tiempo medio que fue aproximadamente de 7 meses. Luego se muestra que 8, es la diferencia de meses entre el paciente con mayor y menor cantidad de meses de estar consumiendo dicho medicamento.

Sin embargo se tiene que el 75% de pacientes se encuentran en periodos de tiempo de consumo menores que los 9 meses.

- **Variable periodo de tiempo con la enfermedad de IR.**

Tabla 31. Período de tiempo con la enfermedad de IR.

Periodo de tiempo con la enfermedad de IR.	Frecuencia	Porcentaje	Porcentaje válidos	Porcentaje acumulado
1 año	42	23,3	23,3	23,3
2 años	55	30,6	30,6	53,9
3 años	37	20,6	20,6	74,4
4 años	13	7,2	7,2	81,7
5 años	18	10,0	10,0	91,7
6 años	4	2,2	2,2	93,9
7 años	4	2,2	2,2	96,1
8 años	2	1,1	1,1	97,2
9 años	2	1,1	1,1	98,3
10 años	1	,6	,6	98,9
11 años	1	,6	,6	99,4
12 años	1	,6	,6	100,0
Total	180	100,0	100,0	

Tabla 32. Medidas Numéricas.

Media		2,91
Mediana		2,00
Moda		2
Deviación Estándar		2,009
Varianza		4,037
Rango		11
Percentiles	10	1,00
	20	1,00
	25	2,00
	30	2,00
	40	2,00
	50	2,00
	60	3,00
	70	3,00
	75	4,00
	80	4,00
	90	5,00

En la tabla 31, se observa que el tiempo donde se concentra la mayoría de pacientes con Insuficiencia Renal es en 2 años; lo que representa un 30.6% del total de la muestra., es decir; que a medida va transcurriendo el tiempo; el número de pacientes con la enfermedad, va disminuyendo. Esto puede ser debido a que los pacientes tienen los cuidados y el control respectivo, en el tratamiento de la enfermedad para una mejor calidad de vida.

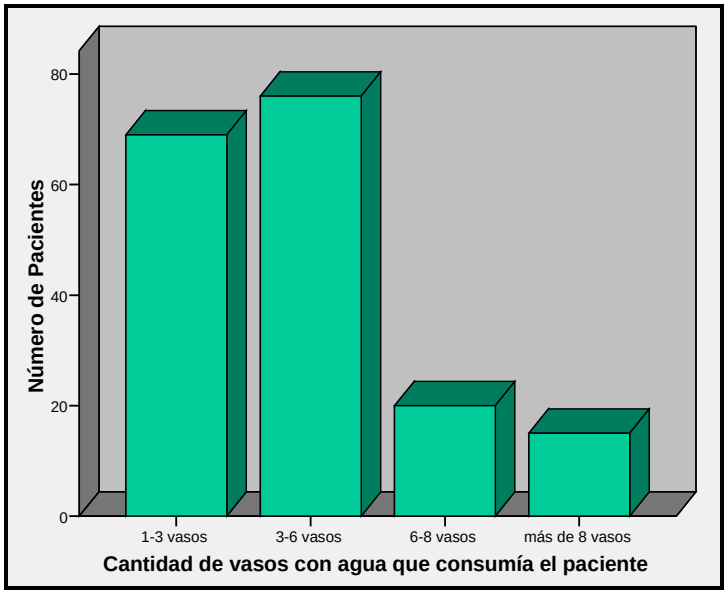
En la tabla 32, se presentan las medidas numéricas de dicha variable, en la cual se refleja que el tiempo de padecer de IR; es aproximadamente en promedio 3 años. Además se tiene que la mitad de los pacientes tiene un período de tiempo con esta enfermedad de 2 años. Nótese también que la mayor cantidad de tiempo que se observó de que los pacientes tienen de padecer de IR, es de 2 años. Luego se muestra que el tiempo que tienen los pacientes de tener la enfermedad es de 2 años; la cual representa la distancia promedio con relación al valor medio que fue aproximadamente de 3 años.. Se observa que 11 es la diferencia de años entre el paciente con mayor y menor tiempo de padecer dicha enfermedad. Así mismo, el 25% de pacientes con IR, tienen de padecer de dicha enfermedad menos de 2 años; mientras que el 75% de pacientes restante es de 4 años respectivamente.

- **Variable cantidad de vasos con agua que consumía el paciente antes de padecer de IR.**

Tabla 33. Cantidad de vasos con agua que consumía el paciente antes de padecer de IR.

cantidad de vasos con agua que consumía el paciente antes de padecer de IR	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
1-3 vasos	69	38,3	38,3	38,3
3-6 vasos	76	42,2	42,2	80,6
6-8 vasos	20	11,1	11,1	91,7
mas de 8 vasos	15	8,3	8,3	100,0
Total	180	100,0	100,0	

Gráfico 9. Cantidad de vasos con agua que consumía el paciente antes de padecer de IR.



La tabla 33, muestra que 76 pacientes, consumían entre 3 a 6 vasos con agua diariamente antes de que se le diagnosticara la enfermedad de Insuficiencia Renal, es decir; un 42.2% del total de la muestra; además tenemos que sólo 69 de los pacientes dijo haber ingerido de 1 a 3 vasos con agua, siendo este un 38.3%, mientras que un 8.3% de ellos manifestó haber ingerido más de 8 vasos con agua al día, lo cual refleja un número más pequeño de pacientes del total de la muestra encuestada. Así mismo, se observa que la mayoría de pacientes están concentrados hacia la izquierda del gráfico 9 las cuales representan la primera y segunda clase.

Por lo tanto; por estudios médicos se tiene, que el no beber la cantidad suficiente de agua provoca en nuestro organismo una serie de malestares, indicando con ello que el agua que les estamos suministrando no es suficiente. Diariamente nuestro cuerpo realiza un sin número de procesos en los que se pierde agua, por lo que la piel tiende a resecarse, además disminuye la humedad de nuestro organismo, necesaria para funcionar bien.

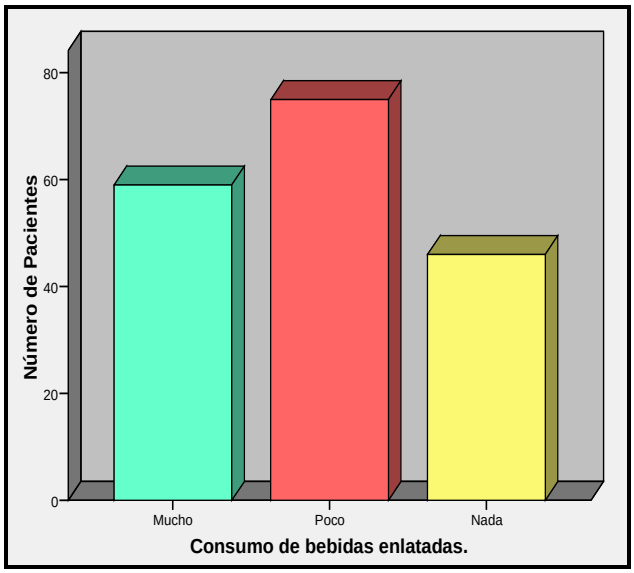
En condiciones normales y con un adecuado funcionamiento del riñón, una persona pierde alrededor de 1450 mililitros de agua al día, es decir; 1.5 litros de agua aproximadamente. Para ello se recomienda beber las cantidades recomendables de agua (por lo menos 8 vasos con agua al día); para que el hígado, los riñones y el sistema digestivo e inmunológico cumplan muy bien con sus funciones.

- **Variable bebidas enlatadas.**

Tabla 34. Consumo de bebidas enlatadas.

Consumo de productos bebibles enlatados.	Frecuencia	Porcentaje	Porcentaje Válido	Porcentaje Acumulado
Mucho	59	32,8	32,8	32,8
Poco	75	41,7	41,7	74,4
Nada	46	25,6	25,6	100,0
Total	180	100,0	100,0	

Gráfico 10. Consumo de Bebidas Enlatadas.



En esta tabla 34; se observa que 46 pacientes manifestaron no haber ingerido ningún tipo de bebidas enlatadas, es decir; un 25.6% del total de la muestra, mientras que 75 pacientes dijeron haber ingerido pocas bebidas enlatadas, el cual se presenta en el gráfico 10 como la barra más alta.

- **Variable Conocimiento_Enfermedad.**

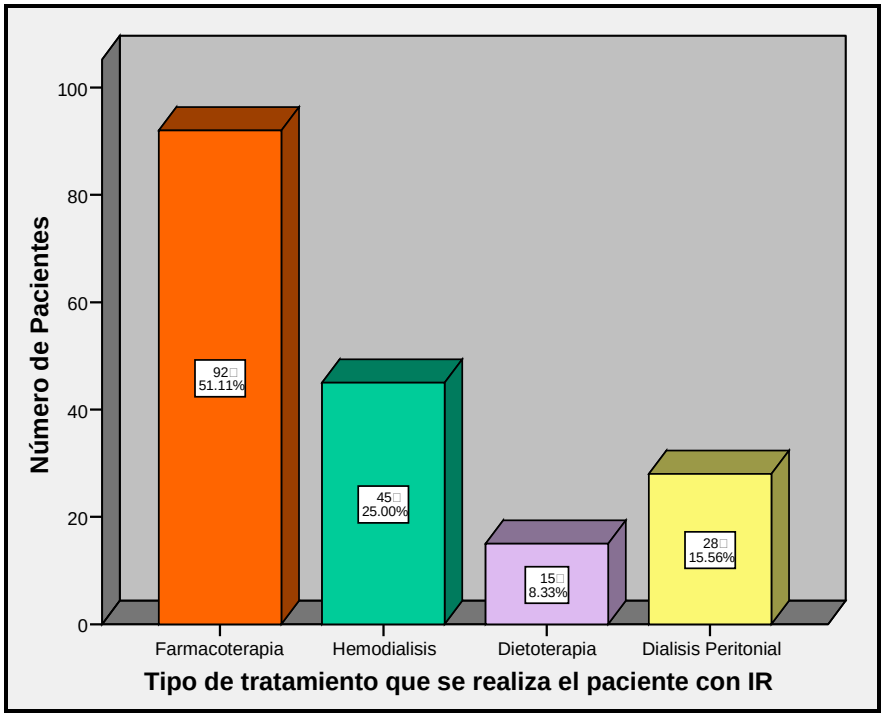
Tabla 35. Conocimiento de la enfermedad de Insuficiencia Renal.

Conocimiento de la enfermedad de Insuficiencia Renal	Frecuencia	Porcentaje	Porcentaje Válido	Porcentaje Acumulado
Poco	16	8,9	8,9	8,9
Nada	164	91,1	91,1	100,0
Total	180	100,0	100,0	

La tabla 35; muestra que 164 pacientes desconocían de qué se trataba la enfermedad de Insuficiencia Renal, el cual representa un 91.1% del total de la muestra, mientras que sólo un 8.9% conocía un poco de dicha enfermedad. Por lo tanto; se sabe que si las personas conocieran sobre los riesgos de padecer de IR, habría una disminución en la incidencia de este tipo de enfermedades crónicas; porque el no conocer sobre la enfermedad, se convierte en un factor de riesgo, ya que a menor conocimiento de dicha enfermedad, mayor es el efecto en la salud de las personas.

- **Variable Tipo_tratamiento.**

Gráfico 11. Tipo de tratamiento que se realiza el paciente con Insuficiencia Renal.



Obsérvese que el gráfico 11; la mayoría de pacientes se encuentra bajo los tratamientos de farmacoterapia y dietoterapia los cuales suman 107 pacientes del total de la muestra.

Luego los tratamientos en los cuales se encuentra un menor número de pacientes son Hemodiálisis y Diálisis peritoneal ambos suman 73, estos tratamientos tienen un mayor efecto en el estado físico del paciente.

- **Variable Tiempo que tiene con el tratamiento el paciente con IR.**

Tabla 36. Tiempo que tiene con el tratamiento el paciente con IR.

Tiempo que tiene con el tratamiento el paciente con IR.	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
1/2 mes	3	1,7	1,7	1,7
2 meses	3	1,7	1,7	3,3
4 meses	2	1,1	1,1	4,4
12 meses	65	36,1	36,1	40,6
24 meses	55	30,6	30,6	71,1
36 meses	27	15,0	15,0	86,1
48 meses	4	2,2	2,2	88,3
60 meses	11	6,1	6,1	94,4
72 meses	6	3,3	3,3	97,8
84 meses	3	1,7	1,7	99,4
132 meses	1	,6	,6	100,0
Total	180	100,0	100,0	

Tabla 37. Medidas Numéricas.

Media		26,62
Mediana		24
Moda		12
Desviación Estándar		20,27
Varianza		410,74
Rango		131,5
Percentiles	10	12
	20	12
	25	12
	30	12
	40	12
	50	24
	60	24
	70	24
	75	36
	80	36
	90	60

En la tabla 36, se observa que el tiempo donde se concentra la mayoría de pacientes con el tratamiento para la Insuficiencia Renal es de 2 años; es decir; el 36.1% del total de la muestra. Además se muestra; que a medida va transcurriendo el tiempo; el número de pacientes que tiene tiempo con los tratamiento para dicha enfermedad va disminuyendo. Esto puede ser debido a que los pacientes tienen los cuidados y el control respectivo, en el tratamiento de la enfermedad para una mejor calidad de vida.

En la tabla 37, se presentan las medidas numéricas de dicha variable, en la cual se refleja que el tiempo de estar con un tratamiento; es aproximadamente en promedio 27 meses. Además se tiene que la mitad de los pacientes tiene un período de tiempo con el tratamiento de 24 meses. Nótese también que la mayor cantidad de tiempo que se observó de que los pacientes tienen con un tratamiento, es de 12 meses. Luego se muestra que el tiempo que tienen los pacientes con los tratamientos es de 20 meses; la cual representa la distancia promedio con relación al valor medio que fue aproximadamente de 27 meses. Se observa que aproximadamente 132 es la diferencia de meses entre el paciente con mayor y menor tiempo de estar con un tratamiento. Así mismo, el 25% de pacientes con IR, tienen de estar con el tratamiento menos de 12 meses mientras que el 75% de pacientes restante es de 36 respectivamente.

3.3. ANÁLISIS BIVARIADO.

En este apartado se mostraran una serie de tablas de correlación, con el objetivo de observar algunas relaciones entre las variables, como pueden ser sus datos personales, hábitos alimenticios y el desarrollo de Insuficiencia Renal, además de observar el comportamiento de algunos factores que pudieran influir en esta enfermedad, ya que es necesario determinar si existe alguna relación entre dos rasgos diferentes en los que la muestra ha sido clasificada. Es importante mencionar que la forma de vida que los pacientes han tenido antes de padecer de Insuficiencia Renal, es de mucha importancia, ya que se puede dar una pauta de algunos factores más influyentes en el desarrollo de esta enfermedad.

Para obtener un mejor parámetro de lo antes expuesto, se iniciará con la descripción del comportamiento que tienen las variables en estudio, las cuales serán obtenidas a través del programa estadístico SPSS, que se muestran en las siguientes tablas de correlación.

- **Variable Sexo y Nivel-Educ.**

Tabla 38. Sexo y Nivel Educativo.

Sexo	Nivel Educativo				
	Educación Básica	Educación Media	Educación Superior	Sin Estudios	Total
Femenino	47	18	5	4	74
	45,2%	35,3%	23,8%	100,0%	41,1%
Masculino	57	33	16	0	106
	54,8%	64,7%	76,2%	,0%	58,9%
Total	104	51	21	4	180
	100,0%	100,0%	100,0%	100,0%	100,0%

En la tabla 38, se observa que, de los 106 pacientes de sexo masculino existen 57 que tienen una educación básica, es decir; un 54.8% de los 104 que tienen este nivel educativo. Además se tiene que 74 pacientes del sexo femenino que corresponde al 41.1% del total de la muestra, 4 de estas no han cursado ningún nivel educativo, representando el 100% dentro de la categoría que no posee ningún nivel educativo.

- **Variable Sexo y Tipo_Enfer (Diabetes Mellitus).**

Tabla 39. Sexo y Diabetes Mellitus.

Sexo	Diabetes Mellitus		
	Si	No	Total
Femenino	55	19	74
	48,2%	28,8%	41,1%
Masculino	59	47	106
	51,8%	71,2%	58,9%
Total	114	66	180
	100,0%	100,0%	100,0%

En la tabla 39, se muestra que de los 114 pacientes que padecen de Diabetes Mellitus 55 pertenecen al sexo femenino, el cual representa un 48.2%; mientras que 59 corresponden al sexo masculino siendo este un 51.8%; lo cual se concluye que la diferencia de padecer esta enfermedad es mínima entre ambos sexos, siendo esta aproximadamente de un 4%. Así mismo se tiene que de los 180 pacientes encuestados 66 no padecen de Diabetes Mellitus de los cuales el 28.8% pertenecen al sexo femenino y el 71.2% al sexo masculino.

- **Variable Sexo y Tipo_Enfer (Hipertensión).**

Tabla 40. Sexo e Hipertensión Arterial.

Sexo	Hipertención Arterial		
	Si	No	Total
Femenino	54	20	74
	39,1%	47,6%	41,1%
Masculino	84	22	106
	60,9%	52,4%	58,9%
Total	138	42	180
	100,0%	100,0%	100,0%

En la tabla 40, se observa que 138 pacientes padecen de hipertensión arterial de los cuales 54 pertenecen al sexo femenino; el cual representa un 39.1% mientras que 84 corresponden al sexo masculino siendo este un 60.9%. Obteniéndose así una diferencia porcentual aproximadamente del 30% de padecer esta enfermedad en ambos sexos.

- **Variable Sexo y Tipo_Enfer (Vias urinarias).**

Tabla 41. Sexo e Infección de Vías Urinarias.

Sexo	Infección de Vias Urinarias		
	Si	No	Total
Femenino	25	49	74
	38,5%	42,6%	41,1%
Masculino	40	66	106
	61,5%	57,4%	58,9%
Total	65	115	180
	100,0%	100,0%	100,0%

En la tabla 41, muestra que la mayor parte de pacientes que desarrollaron una infección de vías urinarias antes de que se les diagnosticara la insuficiencia renal son 65 del total de la muestra (180), de los cuales 25 son del sexo femenino y 40 del sexo masculino, lo que representa en términos de porcentajes un 38.5% y 61.5% respectivamente, siendo el sexo masculino el que muestra mayor frecuencia, ya que su diferencia porcentual con las del sexo femenino es de aproximadamente 23%.

- **Variable Sexo y Bebidas Enlatadas.**

Tabla 42. Variable Sexo y Consumo de Bebidas Enlatadas.

Sexo	Consumo de Bebidas Enlatadas			
	Mucho	Poco	Nada	Total
Femenino	28	37	9	74
	47,5%	49,3%	19,6%	41,1%
Masculino	31	38	37	106
	52,5%	50,7%	80,4%	58,9%
Total	59	75	46	180
	100,0%	100,0%	100,0%	100,0%

La tabla 42, muestra que del total de la muestra (180), solamente 46 de ellos no consumen ningún tipo de bebidas enlatadas; de estos el 19.6% son del sexo femenino y el 80.45% son del sexo masculino siendo este el que presenta una mayor frecuencia. Obsérvese también que 75 pacientes con IR consumen pocas bebidas enlatadas de los cuales el 49.3% son femeninos y 50.75% son masculinos. Además 59 pacientes del total de la muestra son los que representan un mayor consumo de bebidas enlatadas, de este ultimo el 47.5% son del sexo femenino y 52.55 pertenecen al sexo masculino, siendo su diferencia porcentual de aproximadamente 5%

- **Variable Zona y Consumo_agua.**

Tabla 43. Variable Zona y Formas de Abastecimiento de Agua para su Consumo.

Zona	Forma de Abastecimiento de Agua para Consumo				
	Rio	Pozo	ANDA	Embotellada	Total
Occidental	0	2	12	0	14
	,0%	10,0%	7,8%	,0%	7,8%
Central	1	10	118	5	134
	100,0%	50,0%	76,6%	100,0%	74,4%
Oriental	0	3	3	0	6
	,0%	15,0%	1,9%	,0%	3,3%
Paracentral	0	5	21	0	26
	,0%	25,0%	13,6%	,0%	14,4%
Total	1	20	154	5	180
	100,0%	100,0%	100,0%	100,0%	100,0%

En la tabla 43, muestra que del total de la muestra (180); 154 pacientes con IR se abastece de agua para su consumo por medio de ANDA; de los cuales el 7.8% pertenece a la zona Occidental, 76.6% a la Central, 1.9% a la Oriental y 13.6% a la zona Paracentral del El Salvador, nótese que el que presenta mayor porcentaje es la zona Central del país. Además el abastecimiento de agua por medio de río es el que presenta menor frecuencia, siendo este solo de un paciente de la zona Central.

- **Variable Área y Zona**

Tabla 44. Variable Área y Zona de procedencia.

Área	Zona de Procedencia				
	Occidental	Central	Oriental	Paracentral	Total
Rural	9	123	3	19	154
	64,3%	91,8%	50,0%	73,1%	85,6%
Urbana	5	11	3	7	26
	35,7%	8,2%	50,0%	26,9%	14,4%
Total	14	134	6	26	180
	100,0%	100,0%	100,0%	100,0%	100,0%

En la tabla 44, se observa que del total de la muestra 134 pertenece a la zona central de los cuales el 91.8% (123) viven en áreas rurales y el 8.2% (11) en áreas urbanas, siendo la mayoría de áreas rurales, mientras que de los 6 pacientes que viven en la zona Oriental, el

50% viven en áreas urbanas y el otro 50% en áreas rurales. En general tenemos que el 85.6% de la muestra (180) viven en áreas rurales y el 14.4% en áreas siendo su diferencia porcentual del 60%.

- **Variable Tipo_product y Período_utilizacion.**

Tabla 45. Variable Tipo de producto agrícola y Tiempo de haber utilizado los productos agrícolas.

Tipo de Producto Agrícola		Tiempo de Utilizacion de Productos para la Agricultura						Total
		5 años	6 años	8 años	10 años	15 años	20 años	
	Gramaxone	11	5	7	10	3	0	36
		84,6%	83,3%	63,6%	100,0%	75,0%	,0%	20,0%
	Gesaprim	2	0	3	0	1	0	6
		15,4%	,0%	27,3%	,0%	25,0%	,0%	3,3%
	Tamaron	0	0	1	0	0	0	1
		,0%	,0%	9,1%	,0%	,0%	,0%	,6%
	Folidol	0	1	0	0	0	0	1
		,0%	16,7%	,0%	,0%	,0%	,0%	,6%
	No utiliza	0	0	0	0	0	136	136
		,0%	,0%	,0%	,0%	,0%	100,0%	75,6%
Total		13	6	11	10	4	136	180
		100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

En la tabla 45, se tiene que 13 pacientes del total de la muestra ha utilizado productos para la agricultura durante cinco años, de los cuales el 84.6% utiliza el producto llamado Gramaxone y el 15.4% el producto Gesaprim. Además 11 pacientes han utilizado estos productos durante 8 años; de estos el 63.6% ha utilizado el producto Gramaxone, el 27.3% el Gesaprim y el 9.1% utilizo Tamaron.

3.4 MEDIDAS DE RELACIÓN ENTRE VARIABLES NOMINALES.

A continuación se estudia la relación entre diferentes variables, para ello se empleará la prueba Ji-cuadrado; que permite determinar si estas dos variables cualitativas están o no asociadas.

- **variables Sexo y Nivel educ.**

Las hipótesis a contrastar se muestran a continuación:

H_0 : *La variable Sexo y Nivel Educativo son independientes.*

H_1 : *La variable Sexo y Nivel Educativo son dependientes.*

De esta manera el valor de la estadística χ^2 es 9.755 de acuerdo al programa estadístico SPSS mostrado en la siguiente tabla:

Estadístico	Valor	g.l	Sig. Asintótica (bilateral)
Chi-cuadrado de Pearson	9,755	3	,021
N de casos válidos	180		

Este valor de χ^2 =9.755, se compara con un valor de χ^2 con 3 grados de libertad (g.l) para un nivel de confianza determinado (Ver tabla del anexo 2), en este caso se determina un nivel de confianza de 0.1 el cual es $\chi^2_{0.1,3} = 6.25$. De esta forma, el valor que se observa en la estadística de prueba se encuentra dentro de la región crítica, y la hipótesis nula debe rechazarse.

De acuerdo con lo anterior, existe una razón para creer que la variable Sexo con el Nivel Educativo de los pacientes con Insuficiencia Renal no son independientes ya que el valor ji-cuadrado calculado es mayor que el ji-cuadrado de tablas, es decir que existe una dependencia entre esas variables lo que significa que las diferencias en las frecuencias observadas y las frecuencias teóricas o esperadas, son muy elevados y por lo tanto; existe dependencia entre las variables analizadas, es decir; que el sexo es la variable dependiente.

- **Sexo y Tipo_Enfermedades (Diabetes Mellitus).**

Las hipótesis a contrastar se muestran a continuación:

H_0: *La variable Sexo y la enfermedad de Diabetes Mellitus son independientes.*

H_1: *La variable Sexo y la enfermedad de Diabetes Mellitus son dependientes.*

De esta manera el valor de la estadística χ^2 se presenta en la siguiente tabla:

Estadístico	Valor	g.l	Sig. Asintótica (bilateral)
Chi-cuadrado de Pearson	6,537	1	,011
N de casos válidos	180		

Nótese que el valor de χ^2 es de 6.537 y determinando un nivel de confianza de 0.05 el valor de ji-cuadrado de tabla es de 3.84 de esta forma, el valor que se observa en la estadística de prueba se encuentra dentro de la región crítica, y la hipótesis nula debe rechazarse. Por lo tanto, existe una razón para creer que la variable Sexo con Diabetes Mellitus de los pacientes con Insuficiencia Renal no son independientes, es decir que existe una dependencia entre esas variables.

- **Sexo y Tipo_Enfermedades (Hipertensión Arterial**.)

Las hipótesis a contrastar se muestran a continuación:

H_0: *La variable Sexo y la enfermedad de Hipertensión Arterial son independientes.*

H_1: *La variable Sexo y la enfermedad de Hipertensión Arterial son dependientes.*

De esta manera el valor de la estadística χ^2 se muestra en la siguiente tabla:

Estadístico	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	0.952	1	,021
N de casos válidos	180		

Se tiene que el valor de χ^2 es 0.952, con 1 grado de libertad, determinando un nivel de confianza de 0.05 ($\alpha = 0.05$); el valor crítico de tablas es de 3.84. De esta forma, el valor que se observa en la estadística de prueba se encuentra fuera de la región crítica, y la

hipótesis nula debe aceptarse, por lo tanto la variable Sexo con el Hipertensión Arterial de los pacientes con Insuficiencia Renal son independientes.

- **Sexo y Tipo_Enfermedades (Vías Urinarias).**

Las hipótesis a contrastar se muestran a continuación:

H_0: *La variable Sexo y la enfermedad de Infección de Vías Urinarias son independientes.*

H_1: *La variable Sexo y la enfermedad de Infección de Vías Urinarias son dependientes.*

El valor de la estadística χ^2 se muestra en la siguiente tabla:

Estadístico	Valor	g.l	Sig. Asintótica (bilateral)
Chi-cuadrado de Pearson	,295	1	,587
N de casos válidos	180		

Se tiene que el valor de χ^2 es 0.295, con 1 grado de libertad; determinando un nivel de confianza de 0.05 ($\alpha = 0.05$); el valor crítico de tabla es de 3.84. De esta forma, el valor que se observa en la estadística de prueba se encuentra fuera de la región crítica, y la hipótesis nula debe aceptarse, de acuerdo con lo anterior, existe una razón para creer que la variable Sexo con la enfermedad de Infección de Vías Urinarias de los pacientes con Insuficiencia Renal son independientes

3.5 APLICACIÓN DEL MÉTODO DE KAPLAN & MEIER.

Para realizar la aplicación del Método de Kaplan & Meier se obtuvo un total de 180 pacientes de insuficiencia renal que consultan en el hospital militar central. De estos pacientes se obtiene la condición de ingreso y egreso del hospital desde el año 2002 hasta el año 2009, es decir si el paciente salió vivo o muerto. De lo que se puede conocer el número de meses que el paciente ha estado en control en el hospital. Con este método se pretende determinar la probabilidad de vida de los pacientes con IR, obteniendo así proporciones exactas de supervivencia.

A continuación se presenta el resumen del procesamiento de los casos, el cual se puede observar en la tala 46, dicha tabla fue obtenida mediante el método de Kaplan & Meier.

Tabla 46. Resumen del procesamiento de los casos.

Nº Total	Nº de Eventos	Censurados	
		Nº	Porcentaje
180	65	115	63.9%

En la tabla 46, se presenta el total de la muestra; el cual corresponde a 180 pacientes con Insuficiencia Renal, con la que se realizó el Método de Kaplan & Meier para estimar la función de supervivencia de los pacientes enfermos con dicha enfermedad, además en esta tabla se identificaron 65 muertes y 115 datos censurados (vivos), que corresponde a un 63.9% del total de la muestra.

En la tabla 47 se muestra la supervivencia de los pacientes con IR; ésta contiene el tiempo de seguimiento de los pacientes, el cual es la diferencia entre la fecha de ingreso y la de finalización del estudio; esta variable está expresada en meses, luego se tiene el estado de los pacientes al final del estudio; esta columna indica si se ha producido el evento de interés, en este caso etiquetados como “muerte” en caso contrario aparece el valor “vivo”, que corresponde a los datos censurados.

Además se presenta la proporción de pacientes acumulada, que sobrevive hasta el momento, la cual esta formada por dos columnas; la primera muestra la supervivencia acumulada para los que ha tenido lugar el evento de interés en cada tiempo determinado; la segunda contiene el error típico correspondiente a la estimación puntual de Kaplan & Meier en cada tiempo. Nótese que para valores de tiempos iguales, en los que se ha dado el evento de interés los resultados del análisis de estas columnas son iguales para los anteriores

eventos, es por ello que no aparecen en la tabla. Posteriormente se observan los desenlaces acumulados, esto es, el valor de la suma de acumulada del suceso de interés, es decir; el número de pacientes que han muerto hasta ese tiempo. Por último se encuentra la columna del número de pacientes que quedan en cada momento, sin que haya ocurrido en ellos el evento de interés, y representan además los que están en riesgo en el siguiente periodo.

Tabla 47. Supervivencia.

Número	Tiempo	Estado	Proporción acumulada que sobrevive hasta el momento		Nº de eventos acumulados	Nº de casos restantes
			Estimación	Error típico		
1	8	Muerte	.	.	1	179
2	8	Muerte	.	.	2	178
3	8	Muerte	.	.	3	177
4	8	Muerte	.978	.011	4	176
5	8	Censurado	.	.	4	175
6	8	Censurado	.	.	4	174
7	9	Muerte	.972	.012	5	173
8	9	Censurado	.	.	5	172
9	9	Censurado	.	.	5	171
10	10	Muerte	.	.	6	170
11	10	Muerte	.	.	7	169
12	10	Muerte	.	.	8	168
13	10	Muerte	.	.	9	167
14	10	Muerte	.944	.017	10	166
15	10	Censurado	.	.	10	165
16	10	Censurado	.	.	10	164
17	10	Censurado	.	.	10	163
18	10	Censurado	.	.	10	162
19	11	Muerte	.	.	11	161
20	11	Muerte	.932	.019	12	160
21	11	Censurado	.	.	12	159
22	11	Censurado	.	.	12	158
23	11	Censurado	.	.	12	157
24	11	Censurado	.	.	12	156
25	11	Censurado	.	.	12	155
26	11	Censurado	.	.	12	154
27	11	Censurado	.	.	12	153
28	11	Censurado	.	.	12	152
29	11	Censurado	.	.	12	151
30	11	Censurado	.	.	12	150
31	11	Censurado	.	.	12	149
32	12	Muerte	.	.	13	148
33	12	Muerte	.	.	14	147
34	12	Muerte	.913	.021	15	146
35	12	Censurado	.	.	15	145
36	12	Censurado	.	.	15	144
37	12	Censurado	.	.	15	143
38	12	Censurado	.	.	15	142
39	13	Muerte	.	.	16	141
40	13	Muerte	.	.	17	140
41	13	Muerte	.	.	18	139
42	13	Muerte	.888	.024	19	138
43	13	Censurado	.	.	19	137
44	13	Censurado	.	.	19	136
45	13	Censurado	.	.	19	135
46	13	Censurado	.	.	19	134
47	13	Censurado	.	.	19	133
48	13	Censurado	.	.	19	132
49	14	Muerte	.	.	20	131
50	14	Muerte	.874	.026	21	130

Número	Tiempo	Estado	Proporción acumulada que sobrevive hasta el momento		Nº de eventos acumulados	Nº de casos restantes
			Estimación	Error típico		
51	14	Censurado	.	.	21	129
52	14	Censurado	.	.	21	128
53	14	Censurado	.	.	21	127
54	14	Censurado	.	.	21	126
55	14	Censurado	.	.	21	125
56	14	Censurado	.	.	21	124
57	14	Censurado	.	.	21	123
58	14	Censurado	.	.	21	122
59	14	Censurado	.	.	21	121
60	14	Censurado	.	.	21	120
61	14	Censurado	.	.	21	119
62	15	Muerte	.	.	22	118
63	15	Muerte	.	.	23	117
64	15	Muerte	.	.	24	116
65	15	Muerte	.	.	25	115
66	15	Muerte	.837	.029	26	114
67	15	Censurado	.	.	26	113
68	15	Censurado	.	.	26	112
69	15	Censurado	.	.	26	111
70	15	Censurado	.	.	26	110
71	15	Censurado	.	.	26	109
72	15	Censurado	.	.	26	108
73	15	Censurado	.	.	26	107
74	16	Muerte	.830	.030	27	106
75	16	Censurado	.	.	27	105
76	16	Censurado	.	.	27	104
77	16	Censurado	.	.	27	103
78	17	Censurado	.	.	27	102
79	17	Censurado	.	.	27	101
80	17	Censurado	.	.	27	100
81	18	Muerte	.	.	28	99
82	18	Muerte	.813	.032	29	98
83	19	Muerte	.	.	30	97
84	19	Muerte	.	.	31	96
85	19	Muerte	.	.	32	95
86	19	Muerte	.	.	33	94
87	19	Muerte	.772	.035	34	93
00	19	Censurado	.	.	34	92
89	19	Censurado	.	.	34	91
90	19	Censurado	.	.	34	90
91	20	Muerte	.763	.036	35	89
92	20	Censurado	.	.	35	88
93	20	Censurado	.	.	35	87
94	20	Censurado	.	.	35	86
95	20	Censurado	.	.	35	85
96	20	Censurado	.	.	35	84
97	21	Muerte	.754	.037	36	83
98	21	Censurado	.	.	36	82
99	21	Censurado	.	.	36	81
100	21	Censurado	.	.	36	80

Número	Tiempo	Estado	Proporción acumulada que sobrevive hasta el momento		Nº de eventos acumulados	Nº de casos restantes
			Estimación	Error típico		
101	22	Censurado	.	.	36	79
102	22	Censurado	.	.	36	78
103	22	Censurado	.	.	36	77
104	23	Muerte	.744	.037	37	76
105	23	Censurado	.	.	37	75
106	23	Censurado	.	.	37	74
107	23	Censurado	.	.	37	73
108	23	Censurado	.	.	37	72
109	24	Censurado	.	.	37	71
110	24	Censurado	.	.	37	70
111	25	Muerte	.733	.038	38	69
112	25	Censurado	.	.	38	68
113	25	Censurado	.	.	38	67
114	25	Censurado	.	.	38	66
115	25	Censurado	.	.	38	65
116	25	Censurado	.	.	38	64
117	25	Censurado	.	.	38	63
118	26	Muerte	.	.	39	62
119	26	Muerte	.	.	40	61
120	26	Muerte	.699	.041	41	60
121	26	Censurado	.	.	41	59
122	26	Censurado	.	.	41	58
123	26	Censurado	.	.	41	57
124	27	Censurado	.	.	41	56
125	30	Muerte	.686	.043	42	55
126	30	Censurado	.	.	42	54
127	31	Muerte	.	.	43	53
128	31	Muerte	.661	.045	44	52
129	31	Censurado	.	.	44	51
130	32	Muerte	.648	.046	45	50
131	32	Censurado	.	.	45	49
132	32	Censurado	.	.	45	48
133	32	Censurado	.	.	45	47
134	32	Censurado	.	.	45	46
135	33	Muerte	.	.	46	45
136	33	Muerte	.620	.048	47	44
137	34	Muerte	.605	.049	48	43
138	37	Muerte	.	.	49	42
139	37	Muerte	.	.	50	41
140	37	Muerte	.563	.051	51	40
141	43	Censurado	.	.	51	39
142	44	Muerte	.549	.052	52	38
143	44	Censurado	.	.	52	37
144	45	Muerte	.534	.052	53	36
145	49	Muerte	.519	.053	54	35
146	50	Censurado	.	.	54	34
147	50	Censurado	.	.	54	33
148	51	Censurado	.	.	54	32
149	52	Censurado	.	.	54	31
150	52	Censurado	.	.	54	30

Número	Tiempo	Estado	Proporción acumulada que sobrevive hasta el momento		N° de eventos acumulados	N° de casos restantes
			Estimación	Error típico		
151	54	Censurado	.	.	54	29
152	55	Muerte	.501	.054	55	28
153	56	Censurado	.	.	55	27
154	57	Censurado	.	.	55	26
155	58	Censurado	.	.	55	25
156	58	Censurado	.	.	55	24
157	59	Censurado	.	.	55	23
158	59	Censurado	.	.	55	22
159	59	Censurado	.	.	55	21
160	60	Muerte	.477	.057	56	20
161	62	Muerte	.	.	57	19
162	62	Muerte	.430	.060	58	18
163	74	Muerte	.406	.061	59	17
164	76	Muerte	.382	.062	60	16
165	78	Censurado	.	.	60	15
166	78	Censurado	.	.	60	14
167	78	Censurado	.	.	60	13
168	79	Muerte	.	.	61	12
169	79	Muerte	.323	.065	62	11
170	79	Censurado	.	.	62	10
171	79	Censurado	.	.	62	9
172	79	Censurado	.	.	62	8
173	80	Censurado	.	.	62	7
174	80	Censurado	.	.	62	6
175	81	Muerte	.	.	63	5
176	81	Muerte	.215	.076	64	4
177	81	Censurado	.	.	64	3
178	82	Muerte	.144	.077	65	2
179	82	Censurado	.	.	65	1
180	82	Censurado	.	.	65	0

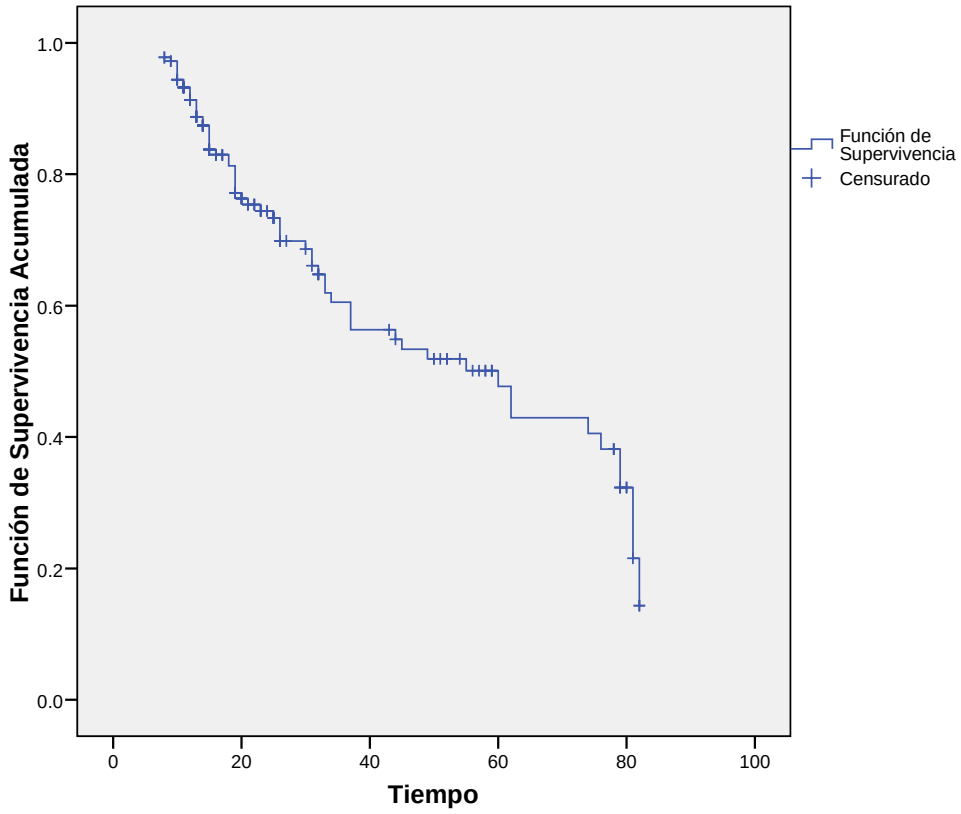
En la tabla 47, se observa que la probabilidad de vida de los pacientes con IR a los 8 meses de haber iniciado el estudio es del 97.8%; esto significa que en media aproximadamente 98 de cada 100 pacientes con IR están vivos hasta este tiempo.

La probabilidad de supervivencia de los pacientes es de 73.3% luego de dos años de haber iniciado el estudio; esto quiere decir que 73 de cada 100 pacientes están vivos hasta este tiempo.

Aproximadamente a los cuatro años se tiene una probabilidad de vida del 51.9% esto significa que en media, 52 de cada 100 pacientes se encuentran vivos hasta ese periodo de tiempo y al final del estudio se observa una probabilidad de vida de 14.4% es decir; que 14 de cada 100 pacientes se encuentran vivos hasta ese periodo de tiempo.

Por lo tanto; se tiene que la probabilidad de vida para los pacientes con IR va disminuyendo a medida avanza el tiempo.

Gráfico 12. Función de Supervivencia.



El gráfico 12; muestra la función de supervivencia, donde se observa que a medida el tiempo transcurre, la probabilidad de que el paciente sobreviva a la enfermedad de insuficiencia renal es menor. Además se puede observar que el último valor de la función de supervivencia aproximadamente en el mes de estudio 82, se encuentra un dato censurado, es decir; que este paciente ha sobrevivido hasta ese periodo de tiempo; esto podría deberse a que el paciente sigue adecuadamente los cuidados y la aplicación de su tratamiento.

3.5.1 COMPARACIÓN DE LAS CURVAS DE SUPERVIVENCIA.

Con la muestra de los 180 pacientes de Insuficiencia Renal, se desea comparar las curvas de supervivencia y de la función de riesgo tomando dos grupos de paciente, el primer grupo esta constituido por los pacientes que tienen Diabetes Mellitus y el otro grupo los que no poseen esta enfermedad; con el objetivo de poder asumir como hipótesis que los pacientes proceden de la misma población.

Para realizar la comparación de estos dos grupos, se plantean las siguientes hipótesis:

Para la función de supervivencia

$$H_0 : S_{posee\ Diabetes\ Mellitus} = S_{no\ posee\ Diabetes\ Mellitus}$$

$$H_1 : S_{posee\ Diabetes\ Mellitus} \neq S_{no\ posee\ Diabetes\ Mellitus}$$

Donde:

H_0: Las funciones de supervivencia para ambos grupos coinciden en el intervalo de tiempo observado.

H_1: Las funciones de supervivencia para ambos grupos no coinciden en el intervalo de tiempo observado.

Para la función de riesgo

$$H_0 : h_{posee\ Diabetes\ Mellitus} = h_{no\ posee\ Diabetes\ Mellitus}$$

$$H_1 : h_{posee\ Diabetes\ Mellitus} \neq h_{no\ posee\ Diabetes\ Mellitus}$$

Donde:

H_0: Las funciones de riesgo para ambos grupos coinciden en el intervalo de tiempo observado.

H_1: Las funciones de riesgo para ambos grupos no coinciden en el intervalo de tiempo observado.

Para realizar la comparación de los dos grupos, primero se obtiene el resumen del procesamiento de los casos; los cuales se presentan en la tabla 48.

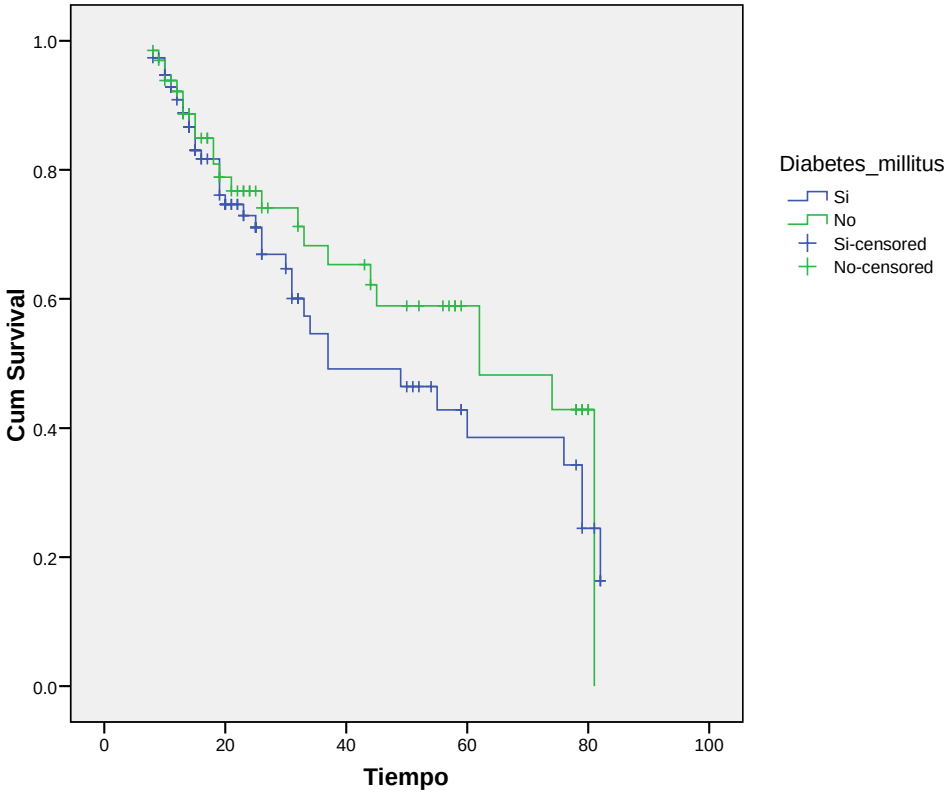
Tabla 48. Resumen del procesamiento de los casos.

Padece de Diabetes Mellitus	Nº Total	Nº de Eventos	Censurado	
			Nº	Porcentaje
Si	114	41	73	64.0%
No	66	24	42	63.6%
Overall	180	65	115	63.9%

En esta tabla se observa que existe un 64% de datos censurados para el grupo de pacientes que padecen de Diabetes Mellitus, mientras que un 63.6% corresponde al grupo que no padecen de dicha enfermedad. Por lo tanto; existe una diferencia porcentual mínima de un 4%; esto refleja que existe un mayor número de pacientes que padecen de Diabetes Mellitus los cuales sobreviven a la enfermedad de IR que del grupo que no padece de Diabetes Mellitus. Nótese que del total de la muestra, existe un 63.9% de pacientes de ambos grupos, que sobreviven a dicha enfermedad.

Para efectos de probar las hipótesis anteriormente planteadas y verificar el método más adecuado de aplicación para realizar las conclusiones entre ambos grupos; se partirá del gráfico de supervivencia que presentan dichos métodos de comparación, el cual se muestra a continuación.

Gráfico 13. Función de Supervivencia.



En el gráfico 13, se observa que las curvas de la función de supervivencia estimadas para ambos grupos; existen periodos de tiempo, en las cuales hay mejor supervivencia en un grupo y luego en el otro, lo cual se nota en el gráfico que las curvas no son aproximadamente paralelas (se cruzan) por lo tanto; el método de Long Rank no es el más indicado de aplicar para detectar diferencias, ya que las curvas se cruzan; en su lugar es adecuado utilizar la prueba Breslow, ya que este método si detecta las diferencias existentes en ambos grupos.

3.5.1.1 MÉTODO BRESLOW.

En la siguiente tabla 49, se presenta el resumen de la Breslow; la cual se utilizará para realizar el contraste de las hipótesis anteriormente planteadas.

Tabla 49. Comparaciones globales.

	Chi-Square	df	Sig.
Breslow (Generalized Wilcoxon)	.555	1	.456

Test of equality of survival distributions for the different levels of Diabetes_millitus.

Utilizando la tabla de la ji-cuadrado con un nivel de significancia del 5% y un grado de libertad, se obtiene un valor de $\chi^2_{0.05,1} = 3.84$ (ver tabla en anexo 2); y utilizando la prueba Breslow, se observa un estadístico asociado de la ji-cuadrado igual a 0.555. Al realizar la comparación de ambos estadísticos se muestra que el valor calculado es menor al valor de tabla de la ji-cuadrado; lo que significa que se acepta H_0, es decir:

- Las curvas de supervivencia para el grupo de pacientes con Diabetes Mellitus y sin Diabetes Mellitus tienen el mismo comportamiento y por lo tanto coinciden en el mismo intervalo de tiempo observado.
- Las curvas de riesgo para ambos el grupo de pacientes con Diabetes Mellitus y sin Diabetes Mellitus tienen el mismo comportamiento y por lo tanto coinciden en el mismo intervalo de tiempo observado.

3.6 MODELO DE COX.

Los datos que se analizan en esta investigación corresponden a 180 pacientes (n=180) con Insuficiencia Renal, que consultaban en el Hospital Militar Central en los años 2002 al 2009. Se realiza un seguimiento a los pacientes desde el comienzo de sus consultas de la enfermedad, hasta alcanzar la muerte como evento de interés, o hasta la terminación del estudio.

A continuación se presenta la estimación de la función de supervivencia que se realiza a través del Modelo de Cox.

3.6.1 ESTIMACIÓN DE LA FUNCIÓN DE SUPERVIVENCIA POR EL MÉTODO DE KAPLAN & MEIER.

La estimación de la función de supervivencia, se obtiene para los 180 pacientes, con una mediana de la supervivencia de 60 meses, es decir; que el 50% de los pacientes con Insuficiencia Renal sobreviven más de 60 meses y el resto sobrevive menos de 60 meses. A demás se observa que el tiempo promedio de supervivencia de los pacientes con IR corresponde a 51 meses aproximadamente (ver Tabla 50).

Tabla 50. Valores resumen en la estimación de la función de supervivencia para diálisis peritoneal según meses.

N	Eventos	Media	ee(media)	Mediana	LCI (95%)	LCS (95%)
180	65	51.474	2.85	60	39.255	80.745

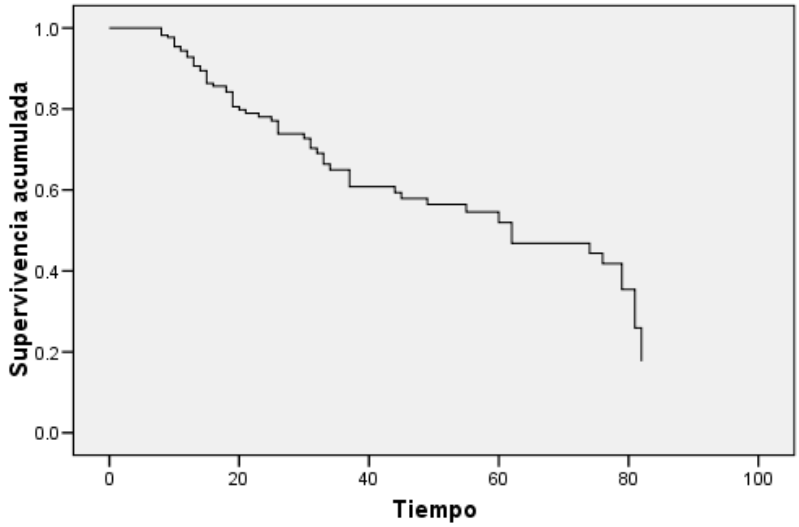
La Tabla 51, muestra la función de supervivencia estimada donde se observa que luego de haber transcurrido 8 meses del inicio del estudio se presentaron cuatro pacientes que fallecieron quedando en riesgo de fallecer 176 con una probabilidad de vida de 0.978. También se muestra que 55 pacientes se encuentran en riesgo después de haber transcurrido aproximadamente dos años y medio de haber iniciado el estudio, con una probabilidad de vida de 0.686 teniendo hasta este momento 42 pacientes con IR fallecidos.

Tabla 51. Función de supervivencia estimada mediante el estimador de Kaplan & Meier

Tiempo	n.riesgo	n.eventos	Supervivencia	err.est.
8	176	4	0.978	0.011
9	173	5	0.972	0.012
10	166	10	0.944	0.017
11	166	12	0.932	0.019
12	146	15	0.913	0.021
13	138	19	0.888	0.024
14	130	21	0.874	0.026
15	114	26	0.837	0.029
16	106	27	0.830	0.030
18	98	29	0.813	0.032
19	93	34	0.772	0.035
20	89	35	0.763	0.036
21	83	36	0.754	0.037
23	76	37	0.744	0.037
25	69	38	0.733	0.038
26	60	41	0.699	0.041
30	55	42	0.686	0.043
31	52	44	0.661	0.045
32	50	45	0.648	0.046
33	44	47	0.620	0.048
34	43	48	0.605	0.049
37	40	51	0.563	0.051
44	38	52	0.549	0.052
45	36	53	0.534	0.052
49	35	54	0.519	0.053
55	28	55	0.501	0.054
60	20	56	0.477	0.057
62	18	58	0.430	0.060
74	17	59	0.406	0.061
76	16	60	0.382	0.062
79	11	62	0.323	0.065
81	4	64	0.215	0.076
82	2	65	0.144	0.077

El Gráfico 14; muestra un patrón decreciente casi lineal de la función de supervivencia, lo cual pareciera estar indicando que las muertes por IR tienen un comportamiento uniforme en el tiempo.

Gráfico 14. Función de supervivencia (KM) para pacientes con IR según meses.



3.6.2 ESTIMACIÓN DEL MODELO DE COX.

Para el proceso de construcción del modelo de Cox, inicialmente se incluyeron 25 covariables (dicotómicas y continuas), con las cuales se ajustaron varios modelos para así obtener las covariables significativas a un nivel de significancia del 10%; eliminando aquellas covariables que no resultaban ser significativas mediante el procedimiento paso a paso hacia atrás (Ver Anexo 3); en las tablas del anexo se presentan los 24 pasos realizados por el SPSS utilizando el procedimiento antes mencionado. Este consiste en observar el p-valor (Sig); si este valor resulta ser mayor al 10% (nivel de significancia) esta covariable, se excluye del modelo y en el siguiente paso quedando solo aquellas que fueran menores a un 10% y este proceso continua de esta manera hasta llegar al modelo definitivo. Para comprender mejor lo antes expuesto se hará la descripción de los últimos tres pasos del proceso (ver tabla 52).

Tabla 52. Últimos pasos del proceso de selección del modelo definitivo de Cox.

Covariable		B	SE	Wald	df	Sig.	Exp(B)
Paso 21	sexo	,465	,285	2,673	1	,102	1,593
	EDAD1	,040	,009	17,804	1	,000	1,041
	cuanto_tiempo	,125	,091	1,899	1	,168	1,133
	Diabetes_millitus	-,703	,353	3,977	1	,046	,495
	Vias_urinarias	-,394	,267	2,184	1	,139	,674
Paso 22	sexo	,289	,263	1,212	1	,271	1,335
	EDAD1	,038	,009	17,881	1	,000	1,039
	Diabetes_millitus	-,675	,354	3,641	1	,056	,509
	Vias_urinarias	-,353	,264	1,781	1	,182	,703
Paso 23	EDAD1	,038	,009	17,158	1	,000	1,038
	Diabetes_millitus	-,634	,352	3,240	1	,072	,531
	Vias_urinarias	-,368	,264	1,937	1	,164	,692
Paso 24	EDAD1	,037	,009	17,004	1	,000	1,038
	Diabetes_millitus	-,569	,349	2,665	1	,086	,566

En esta tabla, se observa que en el paso 21 se encuentran cinco covariables las cuales podrían ser seleccionadas para formar parte del modelo de Cox; estas covariables son sexo, EDAD1, cuanto tiempo, Diabetes_millitus y vías_urinarias; aquí se observa los p-valores (Sig) de cada una de ellas. Luego la variable a ser eliminada será aquella que presente mayor significancía es decir; que el Sig sea mayor que 0.10, entonces la covariable eliminada en este paso será cuanto_tiempo; ya en el paso 22 no se observa esta covariable quedando solamente las otras 4 restantes; las cuales cumplen que su Sig es menor que 0.10, de estas la que presenta mayor significancía es la covariable sexo, es decir; que en el siguiente paso esta variable ya no será parte de las posibles covariables que formaran el modelo, de la misma forma se ha realizado el proceso de eliminación en el paso 23, siendo eliminada la covariable Vias_urinarias y por último en el paso 24 del proceso de selección, las dos covariables que resultan significativas al 10% son la Edad y Diabetes Mellitus, por lo tanto; serán las que formaran parte del modelo definitivo de Cox.

A continuación se presenta en la tabla 53 el análisis de las covariables incluidas en el modelo definitivo, para verificar los supuestos del modelo de riesgos proporcionales.

Tabla 53. Estimación de los coeficientes para el modelo definitivo de Cox.

Covariable	B	SE	Wald	df	Sig.	Exp(B)
EDAD1	,037	,009	17,004	1	,000	1,038
Diabetes_millitus	-,569	,349	2,665	1	,086	,566

En la tabla anterior se observa que para cada una de las covariables incluidas en el modelo definitivo; estas resultan ser significativas, ya que el p-valor (Sig) de cada una de ellas es menor para un $\alpha = 0.1$.

Además se puede observar de la tabla 53; el coeficiente (B) de la covariable Edad (0.037) es positivo mientras que el de la covariable Diabetes Mellitus (-0.569) es negativo, esto indica que el riesgo de muerte por causas asociadas a la insuficiencia renal de los pacientes, es mayor cuando aumenta su edad y no la enfermedad de la diabetes.

Por lo tanto; el modelo final para los datos de los pacientes con IR incluye las covariables Edad y Diabetes Mellitus con coeficientes 0.037 y -0.569 respectivamente, por lo que el modelo de cox se expresa de la siguiente manera:

$$\lambda(t/z) = \lambda_0 e^{(0.037\,Edad - 0.569\,Diabetes\ Mellitus)}$$

Interpretación de los Coeficientes Estimados.

La tabla 54, presenta la información para analizar los riesgos y sus intervalos de confianza. La interpretación de las covariables en el modelo se realizará de forma diferente ya que Edad es un covariable continua mientras que la covariable Diabetes Mellitus es dicotómica los exponenciales de los coeficientes estimados pueden interpretarse de la manera siguiente:

Tabla 54. Exponencial de los coeficientes para el modelo definitivo de Cox para pacientes con IR según meses.

Covariables	Exp(B)	Exp(-B)	95.0% IC para Exp(B)	
			Inferior	Superior
Edad	1.038	.964	1.020	1.056
Diabetes Mellitus	.566	1.767	.286	1.121

Tipo de Covariables en estudio:

- ✓ Edad:

  Para el caso de la covariable Edad, como ésta es de tipo continuo, el riesgo de morir por causas asociadas a la Insuficiencia Renal se obtiene como $e^{(0.037)} = 1.038$ veces, donde $\hat{\beta} = 0.037$; lo que significa que por cada año que aumenta la edad del paciente, el riesgo de morir es 1.038 veces que los de un año inmediatamente anterior. También la interpretación puede hacerse para un periodo de distinto tamaño (c), pudiera decirse que al aumentar la edad de un paciente en 5 años el riesgo de morir por causas asociadas a la insuficiencia renal $e^{(5*0.037)} = 1.20$, es decir; $c = 5$ *años*, para un paciente particular cinco años antes de su edad actual el riesgo de morir era 1.20 veces menor que el riesgo de morir que tiene actualmente. Mientras que para un aumento de 10 años el riesgo de morir es de $e^{(10*0.037)} = 1.45$ veces, por lo tanto; puede observarse que a medida que aumenta la edad de los pacientes el riesgo de morir se incrementa.

El intervalo de confianza del 95%; para el riesgo de un pacientes de morir dentro de un año viene dado mediante la siguiente expresión: $e^{(1*0.037\pm1.96*\|1\|(0.009))}$; por lo tanto el riego de morir estaría ubicado entre 1.019 y 1.056.

- ✓ Diabetes:

  Es variable dicotómica por tanto el estimador de riesgo se calcula como $e^{(-0.569)} = 0.566$; en donde $\hat{\beta} = -0.569$. Este resultado está indicando que la presencia de Diabetes Mellitus en un individuo con IR, aumenta el riesgo de muerte en 0.566 veces, que uno que no tenga Diabetes Mellitus.

  El intervalo de confianza del 95% viene dado por: $e^{(-0.569\pm1.96(0.349))}$, es decir; que el riesgo de morir por causas asociadas a la Diabetes Mellitus se ubica entre el intervalo de 0.286 y 1.122.

A continuación se realizará el análisis de residuos para verificar la validez del modelo definitivo de Cox.

3.6.3 ANÁLISIS DE RESIDUOS.

Lo primero que se analiza es, si las covariables y el modelo satisfacen los supuestos de riesgos proporcionales. Esto puede hacerse utilizando la información de la tabla 55.

Tabla 55. Test de riesgos proporcionales para el modelo definitivo de Cox para según meses.

Covariables	Chi-Cuadrado	gl	Sig.
Edad	20.281	1	.212
Diabetes Mellitus	3.002	1	.172

Como puede observarse en la tabla anterior, al comparar los p-valores (Sig) con un nivel de significación del 5%; se observa que cada una de las covariables incluidas en el modelo se acepta, ya que los p-valores correspondientes a las covariables son mayores que 0.05.

A continuación se presenta el análisis de residuos, mediante los siguientes gráficos incluidos en la figura 1. En la gráfico (a) y (b) se relaciona el tiempo con los residuos parciales de la covariable Edad y Diabetes Mellitus respectivamente.

Figura 1. Verificación de los supuestos del modelo de Cox.

Gráfico (a)

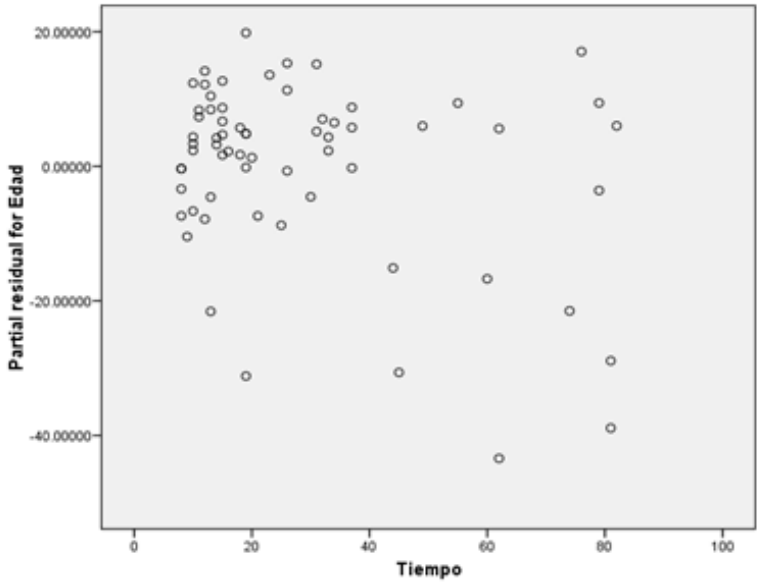
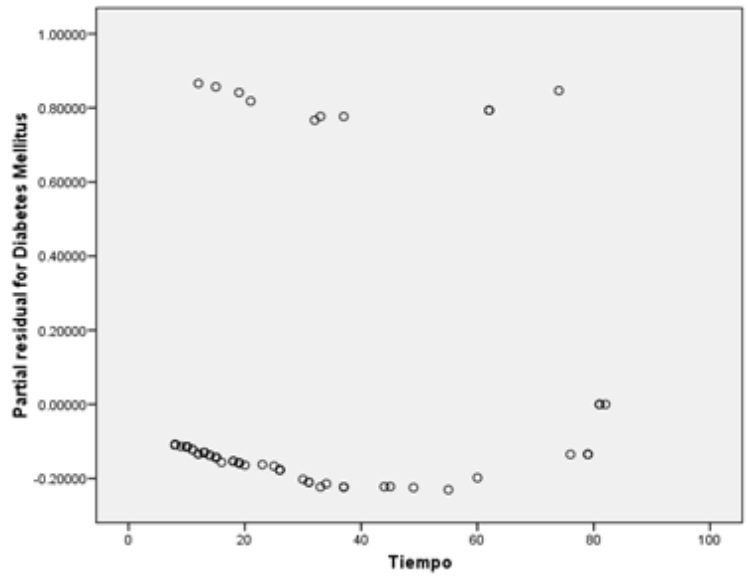


Gráfico (b)



Supuestos de Riesgos Proporcionales.

La verificación de los supuestos de riesgos proporcionales puede observarse mediante los gráficos (a), (b) de la figura 1. En estos gráficos no se presenta una violación del supuesto de riesgos proporcionales; el cual consiste en que los residuos deberán de agruparse en forma aleatoria a ambos lados del valor cero del eje Y, por lo tanto; se observa que las covariables Edad y Diabetes Mellitus, presentan este comportamiento y además estas gráficas muestra un buen comportamiento alrededor de cero del eje Y, sin ninguna tendencia clara predeterminada.

La verificación del supuesto de riesgos proporcionales puede efectuarse a través de un contraste de hipótesis; donde la hipótesis nula está asociada al cumplimiento del supuesto de riesgos proporcionales, el cual consiste en que la función de riesgo es proporcional dados dos perfiles de factores pronóstico distintos, y por tanto se debe mantener a lo largo del tiempo. Los resultados de este contraste indican, que no se viola el supuesto de riesgos proporcionales para ninguna de las dos covariables ya que el resultado de los p-valores que se presenta en la tabla 55, asociados a este contraste para Edad y Diabetes Mellitus son 0.212 y 0.172, respectivamente, observándose que estos valores son mayores al nivel de significancia que se ha tomado, que es del 10% es decir; que no se estaría rechazando la hipótesis de riesgos proporcionales para ninguna de las covariables.

Influencia de Individuos en la Estimación de los Coeficientes.

El supuesto de no influencia de los individuos sobre la estimación de cada coeficiente, puede estudiarse graficando los residuos tipo score versus el correspondiente valor de cada covariable. Para verificar este supuesto en la figura 2, se realiza la relación de las covariables Edad y Diabetes Mellitus con sus respectivas influencias, las cuales corresponden a los gráficos (c) y (d).

Figura 2. Verificación gráfica de los residuos de scores.

Gráfico (c)

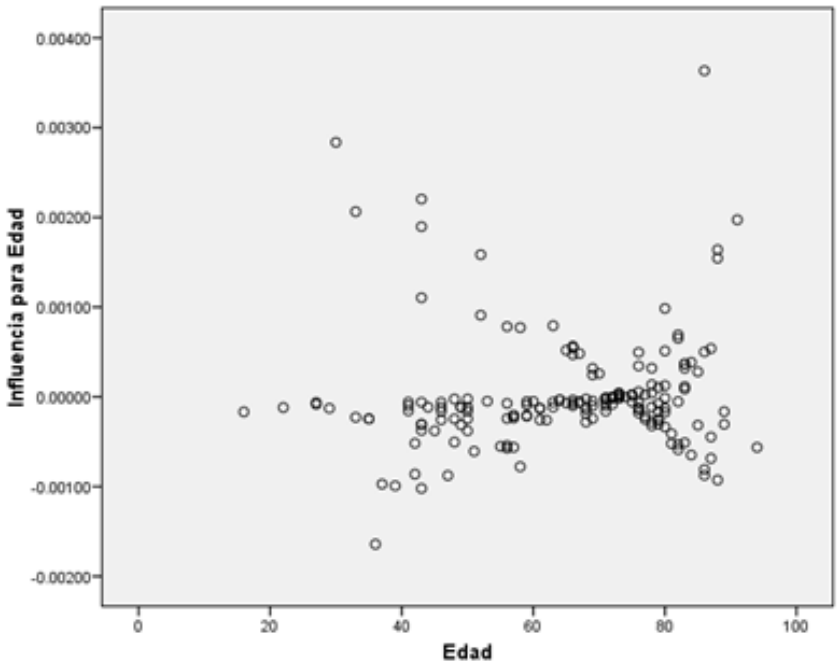
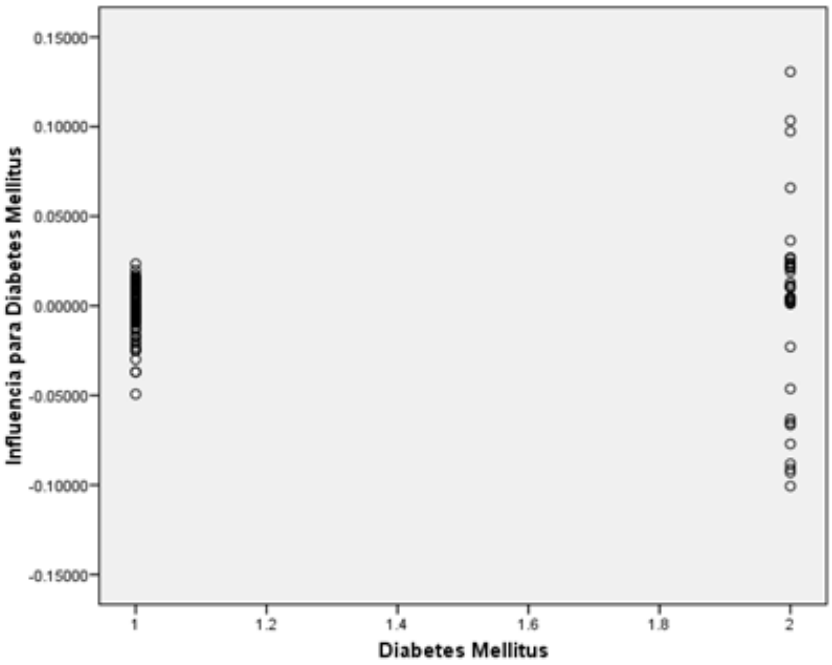


Gráfico (d)

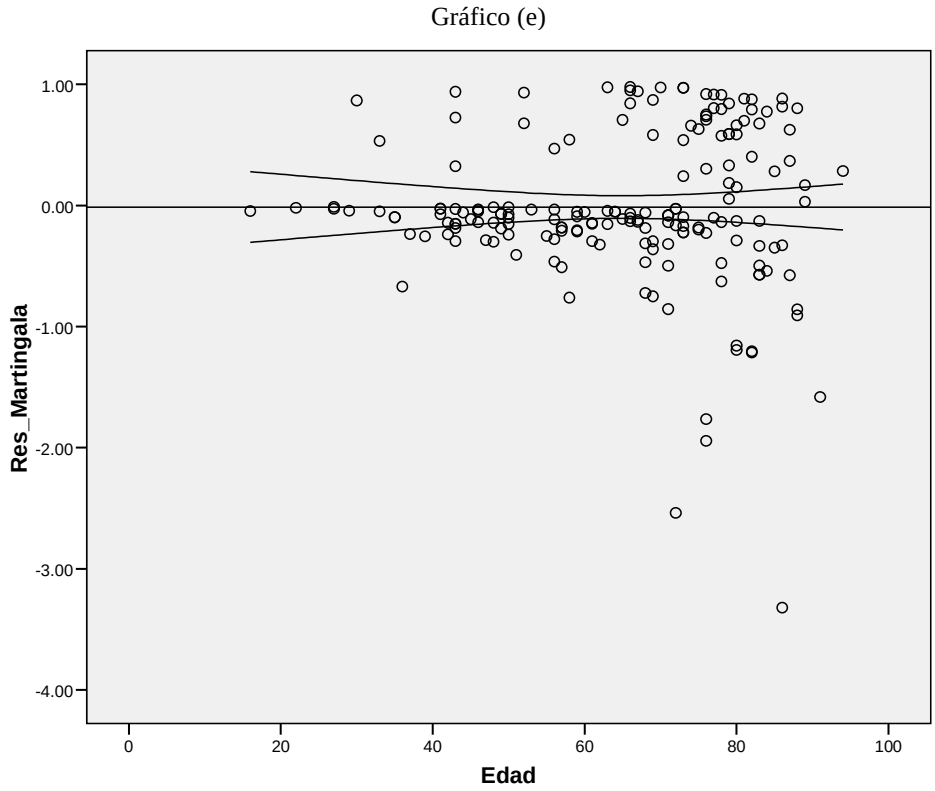


Al observar estos gráficos puede notarse que no existen individuos que estén influyendo en la estimación de sus respectivos coeficientes, ya que no se observan valores extremos respecto al eje Y, lo cual indica que no existe influencia alguna de los individuos en la estimación de cada coeficiente del modelo.

Forma Funcional de la Covariable Edad.

Por ser la covariable Edad continua la verificación del supuesto de la forma funcional de esta covariable que interviene en el modelo, se utilizará el gráfico de los residuos de martingala versus el valor correspondiente de la covariable y su respectiva curva suavizada.

Figura 3. Verificación gráfica de los residuos de martingala.



La forma funcional de la covariable Edad es correcta en el modelo; ya que se observa en la figura 3 que la nube de puntos de esta covariable tiende agruparse a una línea recta.

De los análisis realizados anteriormente, se puede concluir que el modelo encontrado con los datos de los pacientes con IR que consultan en el Hospital Militar Central que resultó ser:

$$\lambda(t/z) = \lambda_0 e^{(0.037\,Edad - 0.569\,Diabetes\ Mellitus)}$$

Cumple con todos los requerimientos exigidos para ser un modelo válido de Cox.

CONCLUSIONES.

- ✓ De acuerdo al análisis realizado de los datos, el perfil de los pacientes que adolecen más una insuficiencia renal son del sexo masculino con un intervalo de 76 a 86 años.

- ✓ La mayoría de los pacientes con IR proceden de la zona central, el cual corresponde a un 74.44% del total de la muestra, por lo tanto; se puede suponer que estos pacientes están más expuesto a la contaminación ambiental, industrial, etc, los cuales pueden ser considerados factores de riesgo

- ✓ Un 24.4% de los pacientes con insuficiencia renal ha utilizado productos agrícolas siendo el más utilizado el plaguicida Gramaxone con un 81.8%.

- ✓ El mayor tiempo en que los pacientes con IR han estado en contacto con plaguicidas fue de 5 a 10 años, el cual corresponde a un 29.5%, es decir; que estos pacientes no cumplen con las medidas de protección básica, por lo que existe una clara y estrecha relación entre el antecedente de uso de plaguicidas y el desarrollo de una insuficiencia renal.

- ✓ De los 180 pacientes con IR, 154 expresaron que la forma en que ellos se abastecían de agua para su consumo fue por medio del servicio de agua por cañería, es decir; mediante el abastecimiento publico ANDA, esto podría ser un factor de riesgo; ya que no posee los estándares de calidad en el tratamiento del agua para el consumo humano y podría estar contaminada a la hora de ser ingerida provocando diferentes enfermedades en el organismo.

- ✓ De los 180 pacientes encuestados, 138 han padecido Hipertensión Arterial, antes de haberle diagnosticado una insuficiencia renal, la cual obtuvo el primer lugar de las enfermedades que más han padecido los pacientes con IR; esto corresponde a un 76.7%, debido a que la Hipertensión Arterial mal controlada causa pérdida de la función renal, lo que genera una insuficiencia renal. Luego en segundo lugar se encuentra la Diabetes Mellitus con un total de 114 casos esto corresponde a un

63.3%. Entonces se tiene que ambas enfermedades son factores de riesgo de gran importancia para el deterioro de la función renal.

- ✓ De los 180 pacientes encuestados, los medicamentos que más han utilizado antes de haberle diagnosticado una insuficiencia renal es la Acetaminofén la que obtuvo el primer lugar con un total de 141 el cual corresponde a un 78.3%, este es suministrado como un antiinflamatorio su uso frecuente puede provocar daños hepáticos, es decir; que afecta las encimas de los órganos del cuerpo humano. Luego en segundo lugar se encuentra el medicamento Ibuprofeno con un total de 115 casos lo que corresponde a un 63.9%. Entonces tenemos que ambos medicamentos son factores de riesgo de gran importancia para el desarrollo de una función renal.

- ✓ De los 180 pacientes con IR existe un 30.6% que tiene dos años de padecer de dicha enfermedad, mientras que para un periodo entre diez y doce años, se tiene que solamente existe un paciente; el cual corresponde a un 0.6%, es decir; a medida que transcurre el tiempo el número de pacientes con la enfermedad va disminuyendo.

- ✓ En el Análisis de los datos para los 180 pacientes con IR que consultan en el Hospital Militar Central, se obtuvo que 76 de estos consumían de 3 a 6 vasos con agua, el cual corresponde a un 42.2%; mientras que solamente un 8.3% ingería la cantidad de agua recomendable, es decir; por lo menos 8 vasos con agua al día; para que el hígado, los riñones y el sistema digestivo e inmunológico cumplan muy bien con sus funciones.

- ✓ De los 180 pacientes encuestados 16 de estos tenían poco conocimiento de la enfermedad de insuficiencia renal, el cual representa un 8.89% y un 91.11% no sabe nada acerca de dicha enfermedad.

- ✓ Del los 114 pacientes que padecen de Diabetes Mellitus, existe un 48.2% que pertenecen al sexo femenino mientras; que un 51.8% corresponden al sexo masculino, por lo tanto; existe una diferencia mínima de 4% entre ambos sexos.

- ✓ De los 138 pacientes con Hipertensión Arterial, 54 de ellos pertenecen al sexo femenino; el cual corresponde a un 3.91%, mientras que un 60.9% pertenecen al sexo masculino. Obteniéndose así una diferencia porcentual aproximadamente del 30% de padecer de esta enfermedad en ambos sexos.

- ✓ De los 65 pacientes que han padecido de infección de vías urinarias, 40 de ellos pertenecen al sexo masculino; el cual corresponde a un 61.5%, mientras que un 38.5% pertenecen al sexo femenino.

- ✓ El sexo de los pacientes con IR no tiene ninguna influencia sobre el padecimiento de las enfermedades Diabetes Mellitus, Hipertensión Arterial e Infección de Vías Urinarias.

- ✓ Utilizando el Método de Kaplan & Meier se identificaron 65 muertes y 115 censuras de los 180 pacientes con IR.

- ✓ La probabilidad de vida para los pacientes con IR va disminuyendo a medida avanza el tiempo, es decir que la probabilidad de vida de los pacientes a los 8 meses de haber iniciado el estudio resulto ser de 0.978 esto significa que en media aproximadamente 98 de cada 100 pacientes con IR están vivos hasta este tiempo y al finalizar el estudio a los 82 meses se observó una probabilidad de vida de 0.144, es decir; que 14 de cada 100 pacientes se encuentran vivos hasta ese periodo de tiempo.

- ✓ En la gráfica de la función de supervivencia estimada obtenida mediante el método de Kaplan & Meier, se observó que el último valor de dicha función es un dato censurado el cual se ubica aproximadamente en el mes de estudio 82, es decir; que este paciente ha sobrevivido hasta la finalización del estudio.

- ✓ Del los 114 pacientes que padecen de Diabetes Mellitus existe un 64.0% de pacientes que sobreviven a la enfermedad de IR mientras que 66 no padecen de Diabetes Mellitus de los cuales un 63.6% han sobrevivido a la enfermedad de IR.

- ✓ Utilizando la prueba de Breslow se llegó a la conclusión que las curvas de supervivencia para el grupo de pacientes con Diabetes Mellitus y sin Diabetes Mellitus tienen el mismo comportamiento y por lo tanto coinciden en el mismo intervalo de tiempo observado.

- ✓ En el Modelo de Cox si la covariable continua Edad se aumentara en c unidades su coeficiente asociado, el riesgo de morir por causas relacionadas a la insuficiencia renal se incrementa, es decir; que un paciente 5 años antes de la edad actual el riesgo de morir es 1.2 veces menos que la que tiene actualmente y el riesgo de morir en un año se encuentra entre 1.019 y 1.056 con un nivel de confianza del 95%.

- ✓ En el Modelo de Cox se observó que la presencia de Diabetes Mellitus en un individuo con IR, el riesgo de muerte aumenta en 0.566 veces, que uno que no tenga Diabetes Mellitus y el riesgo de morir por causas asociadas a la Diabetes Mellitus se ubica entre el intervalo de 0.286 y 1.122 con un nivel de confianza del 95%.

- ✓ En el análisis de los datos para los pacientes con IR que consultan en el Hospital Militar Central, las covariables significativas mediante el análisis del modelo de Cox, fueron: Diabetes Mellitus y Edad, estas covariables son las que estarían modificando el riesgo de muerte en los pacientes con IR y pueden considerarse como factores importantes en el aumento de pacientes con esta enfermedad.

RECOMENDACIONES.

- Identificar tempranamente a los pacientes que están expuestos a factores de riesgo para desarrollar una insuficiencia renal, y darle el seguimiento adecuado para lograr disminuir la incidencia de la enfermedad.

- Realizar campañas a la población en general sobre los riesgos y efectos que conllevan a un insuficiencia renal

- Sugerir al Ministerio de Salud Pública ejerza vigilancia y control sobre la venta y comercialización indiscriminada de antiinflamatorios, analgésicos, etc, a la población en general sin prescripción médica, para evitar el consumo crónico de estos y el desarrollo a largo plazo de una insuficiencia renal.

- Sugerir al Ministerio del Medio Ambiente y Ministerio de Agricultura y Ganadería, que realice campañas educativas en la población agrícola, sobre el manejo y las medidas básicas de protección para el uso de plaguicidas, evitando así, complicaciones a corto y largo plazo en dicha población.

- Debido al impacto que tiene este tipo de trabajo a nivel nacional, se recomienda realizar un estudio similar incluyendo todos los pacientes con IR que consultan los centros públicos y privados, con la finalidad de obtener predictores válidos para todo el país. Esto sería factible a través de un gran proyecto nacional.

- En el caso en que el modelo de Cox no cumpla con uno de los requisitos de los supuestos de riesgos proporcionales, se tienen dos buenas alternativas que son los modelos de regresión de riesgos heterocedásticos (Hsieh, 2001) y los modelos de riesgos proporcionales generalizados (Bagdonavicius y Nikulin, 2001).

- La elección del software computacional SPSS se hizo en base a la comodidad de la herramienta. Sin embargo, existen otros disponibles en los sistemas estadísticos que se pueden utilizar para efectuar análisis de supervivencia tales como: S-PLUS, SAS y R. En tal sentido, se recomienda repetir este estudio o llevar a cabo otros estudios similares utilizando estos sistemas.

- En la medida de lo posible que las autoridades del Hospital Militar Central, implementen programas de prevención y atención temprana, para los pacientes que consultan en dicha institución; por enfermedades que podrían generar una insuficiencia renal como lo es la Diabetes Mellitus entre otras.

- Formar un centro de investigación en el Hospital Militar Central, que tenga como objetivo realizar estudios similares al presentado de otras enfermedades terminales o crónicas, en las que se muestre un incremento sustancial de pacientes.

- Que el estudio presentado se realice de manera periódica, para observar las diferencias de aumento o disminución de los resultados presentados en esta investigación; con los que se realicen posteriormente para tomar las medidas adecuadas respecto a las diferencias observadas.

ANEXO

Anexo 1. Encuesta pasada a los pacientes con IR del Hospital Militar Central

UNIVERSIDAD DE EL SALVADOR
FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICA
ESCUELA DE MATEMÁTICA

TEMA: ANÁLISIS ESTADÍSTICO QUE PREDIGA LAS CAUSAS O FACTORES QUE CONLLEVAN AL INCREMENTO DE PACIENES CON INSUFICIENCIA RENAL EN EL HOSPITAL MILITAR CENTRAL.

Objetivo: Determinar los factores de riesgo que conllevan al incremento de pacientes con Insuficiencia Renal en la población del Hospital Militar Central, con la finalidad de realizar un estudio estadístico.

Indicaciones: Conteste las siguientes preguntas complementando la información que se le solicita y marcando con una "x" la(s) respuesta(as) de su elección.

1. Sexo: F ☐ M ☐

2. Edad: _____

3. Zona de procedencia: ☐ Urbana ☐ Rural

4. Nivel educativo:

Educación básica ☐ Educación media ☐

Educación superior ☐ Sin estudios ☐

Otros: _____

5. Ocupación: _____

Cambio de trabajo: Si ☐ No ☐

Trabajo actual: _____

Trabajo antiguo: _____

6. ¿Cual es su ingreso económico?

< $100 al mes ☐

$100-$149 al mes ☐

$150- $300 al mes ☐

>$300 al mes ☐

7. ¿Ha utilizado productos agrícolas? Si ☐ No ☐

Si la respuesta es **SI,** diga cuales_____

8. ¿Si los ha utilizado, Por cuanto tiempo? _____

9. ¿De que forma se abastece usted de agua para su consumo?

- Rio ☐
- Pozo ☐
- Nacimiento ☐
- ANDA ☐
- Embotellada ☐

10. ¿Padece alguna de las siguientes enfermedades y cuanto tiempo tiene de padecerlas?

	SI	NO	TIEMPO
Diabetes Millitus	☐	☐	☐
Hipertensión Arterial	☐	☐	☐
Artritis Reumatoidea	☐	☐	☐
Infecciones de vías urinarias	☐	☐	☐
Otras_____			

11. ¿Que medicamentos a usado?

	SI	NO	DESDE CUANDO
a) Ibuprofeno	☐	☐	☐
b) Acetaminofen	☐	☐	☐
c) Diclofenac	☐	☐	☐
d) Indometacina	☐	☐	☐
e) Aspirina	☐	☐	☐
f) Enalapril	☐	☐	☐
g) Aldomet	☐	☐	☐
h) Glibenclamida	☐	☐	☐
i) Metformina	☐	☐	☐

j) Otros_____

12. ¿Desde hace cuanto tiempo padece usted de la enfermedad de la insuficiencia renal?_____

13. ¿Ha consumido productos enlatados?

☐ Mucho ☐ Poco ☐ Nada

14. ¿Conoce de la enfermedad de insuficiencia renal?

☐ Mucho ☐ Poco ☐ Nada

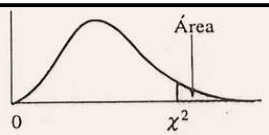
13. ¿Qué tipo de tratamiento se realiza?

- Farmacoterapia ☐
- Hemodiálisis ☐
- Hemodiafiltración ☐
- Dietoterapia ☐
- Diálisis peritoneal ☐
- Trasplante renal ☐

Desde cuando_____

Anexo 2. Distribución Chi Cuadrado χ^2.

La primera columna (gl) localiza cada distribución χ^2. Las otras columnas indican la proporción de área debajo de la distribución χ^2 que está más allá del valor indicado de χ^2. Los valores de χ^2 bajo los encabezados de 0.05 y 0.01 son los valores críticos de χ^2 para $\alpha = 0.05$ y 0.01. Para ser significativo, $\chi^2_{obt} \geq \chi^2_{crit}$.



Grados de libertad *gl*	*P* = .99	.98	.95	.90	.80	.70	.50	.30	.20	.10	.05	.02	.01
1	.000157	.000628	.00393	.0158	.0642	.148	.455	1.074	1.642	2.706	3.841	5.412	6.635
2	.0201	.0404	.103	.211	.446	.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210
3	.115	.185	.352	.584	1.005	1.424	2.366	3.665	4.642	6.251	7.815	9.837	11.341
4	.297	.429	.711	1.064	1.649	2.195	3.357	4.878	5.989	7.779	9.488	11.668	13.277
5	.554	.752	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086
6	.872	1.134	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	12.592	15.033	16.812
7	1.239	1.564	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475
8	1.646	2.032	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090
9	2.088	2.532	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578
16	5.812	6.614	7.962	9.312	11.152	12.624	15.338	18.418	20.465	23.542	26.296	29.633	32.000
17	6.408	7.255	8.672	10.085	12.002	13.531	16.338	19.511	21.615	24.769	27.587	30.995	33.409
18	7.015	7.906	9.390	10.865	12.857	14.440	17.338	20.601	22.760	25.989	28.869	32.346	34.805
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191
20	8.260	9.237	10.851	12.443	14.578	16.266	19.337	22.775	25.038	28.412	31.410	35.020	37.566
21	8.897	9.915	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	38.932
22	9.542	10.600	12.338	14.041	16.314	18.101	21.337	24.939	27.301	30.813	33.924	37.659	40.289
23	10.196	11.293	13.091	14.848	17.187	19.021	22.337	26.018	28.429	32.007	35.172	38.968	41.638
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980
25	11.524	12.697	14.611	16.473	18.940	20.867	24.337	28.172	30.675	34.382	37.652	41.566	44.314
26	12.198	13.409	15.379	17.292	19.820	21.792	25.336	29.246	31.795	35.563	38.885	42.856	45.642
27	12.879	14.125	16.151	18.114	20.703	22.719	26.336	30.319	32.912	36.741	40.113	44.140	46.963
28	13.565	14.847	16.928	18.939	21.588	23.647	27.336	31.391	34.027	37.916	41.337	45.419	48.278
29	14.256	15.574	17.708	19.768	22.475	24.577	28.336	32.461	35.139	39.087	42.557	46.693	49.588
30	14.953	16.306	18.493	20.599	23.364	25.508	29.336	33.530	36.250	40.256	43.773	47.962	50.892

Anexo 3. Pasos para encontrar el modelo definitivo de Cox.

Covariables		B	SE	Wald	df	Sig.	Exp(B)
Paso 1	sexo	,501	,341	2,159	1	,142	1,651
	EDAD1	,052	,013	15,836	1	,000	1,053
	area	-,734	,722	1,032	1	,310	,480
	zona	,019	,184	,010	1	,920	1,019
	Nivel_educativo	-,171	,236	,527	1	,468	,843
	Utilizado_productos	,396	2,231	,032	1	,859	1,486
	Pro_agricolas	-,339	,545	,387	1	,534	,712
	cuanto_tiempo	,377	,223	2,868	1	,090	1,459
	abastece_agua	-,286	,318	,809	1	,368	,751
	Diabetes_millitus	-1,100	,432	6,487	1	,011	,333
	Hipertencion	,569	,529	1,160	1	,281	1,767
	Artritis	,245	,468	,273	1	,601	1,277
	Vias_urinarias	-,561	,323	3,026	1	,082	,570
	Otras	2,353	1,429	2,711	1	,100	10,516
	Ibuprofeno	,538	,318	2,872	1	,090	1,713
	Acetaminofen	-,359	,446	,649	1	,420	,698
	Diclofenac	-,503	,385	1,705	1	,192	,605
	Indometacina	-1,805	1,295	1,941	1	,164	,164
	Aspirina	,096	,363	,071	1	,790	1,101
	Enalapril	-,438	,310	2,001	1	,157	,645
	Aldomet	,000	,781	,000	1	1,000	1,000
	Otros	-,245	,112	4,809	1	,028	,783
	vasos_agua	,026	,188	,018	1	,892	1,026
	consumo_enlatados	-,304	,226	1,811	1	,178	,738
	Conoce_enfermedad	-,563	,522	1,163	1	,281	,570

	Covariables	B	SE	Wald	df	Sig.	Exp(B)
Paso 2	sexo	,501	,341	2,159	1	,142	1,651
	EDAD1	,052	,013	15,860	1	,000	1,053
	area	-,734	,722	1,034	1	,309	,480
	zona	,019	,184	,010	1	,919	1,019
	Nivel_educativo	-,171	,236	,527	1	,468	,843
	Utilizado_productos	,396	2,220	,032	1	,858	1,486
	Pro_agricolas	-,339	,544	,389	1	,533	,712
	cuanto_tiempo	,377	,223	2,868	1	,090	1,459
	abastece_agua	-,286	,317	,812	1	,367	,751
	Diabetes_millitus	-1,100	,432	6,489	1	,011	,333
	Hipertencion	,569	,528	1,162	1	,281	1,767
	Artritis	,245	,467	,274	1	,601	1,277
	Vias_urinarias	-,561	,321	3,052	1	,081	,570
	Otras	2,353	1,425	2,726	1	,099	10,516
	Ibuprofeno	,538	,316	2,897	1	,089	1,713
	Acetaminofen	-,359	,445	,650	1	,420	,698
	Diclofenac	-,503	,384	1,718	1	,190	,605
	Indometacina	-1,805	1,289	1,960	1	,162	,164
	Aspirina	,096	,363	,071	1	,790	1,101
	Enalapril	-,438	,307	2,036	1	,154	,645
	Otros	-,245	,111	4,818	1	,028	,783
	vasos_agua	,026	,188	,019	1	,892	1,026
	consumo_enlatados	-,304	,225	1,829	1	,176	,738
	Conoce_enfermedad	-,563	,521	1,168	1	,280	,570

	Covariables	B	SE	Wald	df	Sig.	Exp(B)
Paso 3	sexo	,493	,330	2,227	1	,136	1,637
	EDAD1	,052	,013	16,281	1	,000	1,053
	area	-,732	,720	1,035	1	,309	,481
	Nivel_educativo	-,173	,236	,539	1	,463	,841
	Utilizado_productos	,386	2,215	,030	1	,862	1,471
	Pro_agricolas	-,335	,541	,382	1	,536	,716
	cuanto_tiempo	,377	,223	2,862	1	,091	1,458
	abastece_agua	-,284	,316	,809	1	,368	,753
	Diabetes_millitus	-1,093	,426	6,581	1	,010	,335
	Hipertencion	,571	,528	1,172	1	,279	1,770
	Artritis	,256	,454	,317	1	,573	1,291
	Vias_urinarias	-,561	,321	3,048	1	,081	,571
	Otras	2,382	1,397	2,907	1	,088	10,822
	Ibuprofeno	,536	,316	2,885	1	,089	1,710
	Acetaminofen	-,361	,445	,657	1	,418	,697
	Diclofenac	-,508	,380	1,791	1	,181	,601
	Indometacina	-1,797	1,287	1,952	1	,162	,166
	Aspirina	,094	,361	,067	1	,795	1,098
	Enalapril	-,434	,305	2,030	1	,154	,648
	Otros	-,246	,110	5,012	1	,025	,782
	vasos_agua	,027	,187	,022	1	,883	1,028
	consumo_enlatados	-,302	,224	1,821	1	,177	,739
	Conoce_enfermedad	-,562	,520	1,166	1	,280	,570

	Covariables	B	SE	Wald	df	Sig.	Exp(B)
Paso 4	sexo	,486	,327	2,208	1	,137	1,625
	EDAD1	,052	,013	16,526	1	,000	1,054
	area	-,723	,716	1,019	1	,313	,485
	Nivel_educativo	-,169	,233	,521	1	,470	,845
	Utilizado_productos	,362	2,219	,027	1	,871	1,436
	Pro_agricolas	-,334	,544	,378	1	,538	,716
	cuanto_tiempo	,377	,223	2,859	1	,091	1,458
	abastece_agua	-,282	,316	,799	1	,371	,754
	Diabetes_millitus	-1,100	,424	6,728	1	,009	,333
	Hipertencion	,565	,527	1,151	1	,283	1,760
	Artritis	,249	,451	,304	1	,582	1,282
	Vias_urinarias	-,563	,321	3,083	1	,079	,569
	Otras	2,362	1,391	2,883	1	,090	10,608
	Ibuprofeno	,538	,315	2,911	1	,088	1,713
	Acetaminofen	-,359	,445	,652	1	,419	,698
	Diclofenac	-,505	,379	1,778	1	,182	,603
	Indometacina	-1,850	1,237	2,238	1	,135	,157
	Aspirina	,101	,357	,080	1	,777	1,107
	Enalapril	-,433	,305	2,017	1	,156	,649
	Otros	-,245	,110	4,998	1	,025	,782
	consumo_enlatados	-,300	,223	1,807	1	,179	,741
	Conoce_enfermedad	-,572	,515	1,235	1	,266	,564

	Covariables	B	SE	Wald	df	Sig.	Exp(B)
Paso 5	sexo	,480	,325	2,182	1	,140	1,617
	EDAD1	,053	,012	19,191	1	,000	1,054
	area	-,723	,716	1,020	1	,313	,485
	Nivel_educativo	-,165	,232	,506	1	,477	,848
	Pro_agricolas	-,252	,191	1,746	1	,186	,777
	cuanto_tiempo	,392	,204	3,681	1	,055	1,479
	abastece_agua	-,284	,316	,806	1	,369	,753
	Diabetes_millitus	-1,095	,423	6,700	1	,010	,335
	Hipertencion	,566	,526	1,159	1	,282	1,762
	Artritis	,262	,443	,350	1	,554	1,300
	Vias_urinarias	-,560	,320	3,055	1	,080	,571
	Otras	2,396	1,376	3,033	1	,082	10,977
	Ibuprofeno	,537	,315	2,902	1	,088	1,711
	Acetaminofen	-,370	,439	,711	1	,399	,691
	Diclofenac	-,507	,379	1,793	1	,181	,602
	Indometacina	-1,866	1,233	2,290	1	,130	,155
	Aspirina	,109	,354	,095	1	,758	1,115
	Enalapril	-,427	,303	1,990	1	,158	,652
	Otros	-,246	,110	5,017	1	,025	,782
	consumo_enlatados	-,294	,220	1,781	1	,182	,745
	Conoce_enfermedad	-,564	,513	1,209	1	,272	,569

Covariables		B	SE	Wald	df	Sig.	Exp(B)
Paso 6	sexo	,468	,323	2,105	1	,147	1,597
	EDAD1	,052	,012	19,741	1	,000	1,053
	area	-,757	,712	1,130	1	,288	,469
	Nivel_educativo	-,154	,229	,454	1	,500	,857
	Pro_agricolas	-,261	,188	1,919	1	,166	,771
	cuanto_tiempo	,395	,203	3,789	1	,052	1,484
	abastece_agua	-,290	,318	,830	1	,362	,749
	Diabetes_millitus	-1,108	,421	6,941	1	,008	,330
	Hipertencion	,555	,524	1,120	1	,290	1,741
	Artritis	,237	,433	,299	1	,584	1,268
	Vias_urinarias	-,556	,320	3,012	1	,083	,574
	Otras	2,321	1,356	2,930	1	,087	10,187
	Ibuprofeno	,525	,313	2,819	1	,093	1,691
	Acetaminofen	-,356	,436	,668	1	,414	,700
	Diclofenac	-,495	,376	1,731	1	,188	,610
	Indometacina	-1,931	1,215	2,525	1	,112	,145
	Enalapril	-,433	,302	2,048	1	,152	,649
	Otros	-,255	,105	5,898	1	,015	,775
	consumo_enlatados	-,295	,219	1,803	1	,179	,745
	Conoce_enfermedad	-,562	,513	1,201	1	,273	,570

Covariables		B	SE	Wald	df	Sig.	Exp(B)
Paso 7	sexo	,515	,313	2,716	1	,099	1,674
	EDAD1	,051	,012	19,511	1	,000	1,052
	area	-,773	,716	1,166	1	,280	,462
	Nivel_educativo	-,156	,228	,467	1	,494	,855
	Pro_agricolas	-,257	,187	1,880	1	,170	,774
	cuanto_tiempo	,394	,201	3,852	1	,050	1,483
	abastece_agua	-,298	,319	,874	1	,350	,742
	Diabetes_millitus	-1,113	,421	6,985	1	,008	,329
	Hipertencion	,525	,526	,998	1	,318	1,690
	Vias_urinarias	-,593	,315	3,537	1	,060	,553
	Otras	2,324	1,361	2,917	1	,088	10,213
	Ibuprofeno	,531	,314	2,860	1	,091	1,700
	Acetaminofen	-,273	,409	,444	1	,505	,761
	Diclofenac	-,421	,352	1,431	1	,232	,656
	Indometacina	-1,950	1,215	2,577	1	,108	,142
	Enalapril	-,443	,303	2,134	1	,144	,642
	Otros	-,235	,099	5,683	1	,017	,790
	consumo_enlatados	-,331	,209	2,514	1	,113	,718
	Conoce_enfermedad	-,553	,512	1,166	1	,280	,575

Covariables		B	SE	Wald	df	Sig.	Exp(B)
Paso 8	sexo	,531	,311	2,914	1	,088	1,701
	EDAD1	,049	,011	19,290	1	,000	1,051
	area	-,747	,720	1,077	1	,299	,474
	Nivel_educativo	-,190	,222	,733	1	,392	,827
	Pro_agricolas	-,218	,178	1,502	1	,220	,805
	cuanto_tiempo	,356	,191	3,470	1	,062	1,428
	abastece_agua	-,300	,321	,872	1	,350	,741
	Diabetes_millitus	-1,161	,415	7,816	1	,005	,313
	Hipertencion	,503	,523	,924	1	,336	1,653
	Vias_urinarias	-,593	,315	3,531	1	,060	,553
	Otras	2,339	1,364	2,940	1	,086	10,370
	Ibuprofeno	,518	,314	2,721	1	,099	1,678
	Diclofenac	-,372	,344	1,168	1	,280	,689
	Indometacina	-2,001	1,212	2,725	1	,099	,135
	Enalapril	-,446	,301	2,202	1	,138	,640
	Otros	-,220	,096	5,238	1	,022	,803
	consumo_enlatados	-,326	,208	2,454	1	,117	,722
	Conoce_enfermedad	-,528	,513	1,060	1	,303	,590

Covariables		B	SE	Wald	df	Sig.	Exp(B)
Paso 9	sexo	,502	,307	2,664	1	,103	1,652
	EDAD1	,050	,011	20,293	1	,000	1,052
	area	-,750	,733	1,045	1	,307	,473
	Pro_agricolas	-,251	,174	2,077	1	,150	,778
	cuanto_tiempo	,396	,187	4,499	1	,034	1,486
	abastece_agua	-,342	,322	1,126	1	,289	,711
	Diabetes_millitus	-1,095	,407	7,239	1	,007	,334
	Hipertencion	,464	,514	,815	1	,367	1,591
	Vias_urinarias	-,527	,308	2,934	1	,087	,590
	Otras	2,106	1,346	2,450	1	,118	8,219
	Ibuprofeno	,447	,303	2,178	1	,140	1,563
	Diclofenac	-,372	,342	1,179	1	,277	,689
	Indometacina	-1,995	1,210	2,719	1	,099	,136
	Enalapril	-,432	,299	2,089	1	,148	,649
	Otros	-,190	,089	4,503	1	,034	,827
	consumo_enlatados	-,337	,205	2,695	1	,101	,714
	Conoce_enfermedad	-,465	,507	,840	1	,359	,628

Covariables		B	SE	Wald	df	Sig.	Exp(B)
Paso 10	sexo	,470	,305	2,379	1	,123	1,600
	EDAD1	,051	,011	20,475	1	,000	1,052
	area	-,772	,733	1,108	1	,292	,462
	Pro_agricolas	-,259	,174	2,218	1	,136	,772
	cuanto_tiempo	,394	,186	4,483	1	,034	1,483
	abastece_agua	-,327	,321	1,033	1	,309	,721
	Diabetes_millitus	-1,120	,405	7,656	1	,006	,326
	Vias_urinarias	-,515	,309	2,784	1	,095	,598
	Otras	2,073	1,346	2,372	1	,123	7,945
	Ibuprofeno	,427	,302	1,997	1	,158	1,532
	Diclofenac	-,311	,340	,837	1	,360	,733
	Indometacina	-1,931	1,205	2,569	1	,109	,145
	Enalapril	-,384	,292	1,730	1	,188	,681
	Otros	-,165	,085	3,798	1	,051	,848
	consumo_enlatados	-,338	,206	2,705	1	,100	,713
	Conoce_enfermedad	-,527	,505	1,088	1	,297	,590
Paso 11	sexo	,506	,303	2,789	1	,095	1,659
	EDAD1	,053	,011	21,746	1	,000	1,054
	area	-,735	,750	,961	1	,327	,479
	Pro_agricolas	-,265	,175	2,287	1	,130	,767
	cuanto_tiempo	,381	,186	4,207	1	,040	1,464
	abastece_agua	-,310	,331	,879	1	,349	,733
	Diabetes_millitus	-1,152	,403	8,151	1	,004	,316
	Vias_urinarias	-,568	,302	3,531	1	,060	,566
	Otras	2,028	1,348	2,263	1	,132	7,597
	Ibuprofeno	,388	,296	1,714	1	,190	1,474
	Indometacina	-2,108	1,193	3,123	1	,077	,121
	Enalapril	-,459	,279	2,703	1	,100	,632
	Otros	-,147	,083	3,177	1	,075	,863
	consumo_enlatados	-,295	,199	2,183	1	,140	,745
	Conoce_enfermedad	-,602	,495	1,482	1	,223	,548

Covariables		B	SE	Wald	df	Sig.	Exp(B)
Paso 12	sexo	,477	,301	2,523	1	,112	1,612
	EDAD1	,053	,011	21,746	1	,000	1,054
	area	-,212	,483	,192	1	,661	,809
	Pro_agricolas	-,258	,174	2,210	1	,137	,773
	cuanto_tiempo	,364	,183	3,980	1	,046	1,440
	Diabetes_millitus	-1,138	,404	7,933	1	,005	,321
	Vias_urinarias	-,610	,299	4,171	1	,041	,543
	Otras	1,512	1,214	1,550	1	,213	4,535
	Ibuprofeno	,392	,296	1,761	1	,185	1,480
	Indometacina	-2,068	1,192	3,009	1	,083	,126
	Enalapril	-,449	,280	2,584	1	,108	,638
	Otros	-,151	,083	3,304	1	,069	,860
	consumo_enlatados	-,316	,199	2,526	1	,112	,729
	Conoce_enfermedad	-,577	,493	1,366	1	,243	,562
Paso 13	sexo	,494	,299	2,730	1	,098	1,639
	EDAD1	,053	,011	22,396	1	,000	1,055
	Pro_agricolas	-,240	,171	1,978	1	,160	,787
	cuanto_tiempo	,367	,185	3,962	1	,047	1,444
	Diabetes_millitus	-1,142	,403	8,026	1	,005	,319
	Vias_urinarias	-,585	,293	3,985	1	,046	,557
	Otras	1,420	1,192	1,420	1	,233	4,138
	Ibuprofeno	,376	,293	1,641	1	,200	1,456
	Indometacina	-2,186	1,162	3,538	1	,060	,112
	Enalapril	-,468	,277	2,862	1	,091	,626
	Otros	-,152	,083	3,382	1	,066	,859
	consumo_enlatados	-,313	,199	2,491	1	,114	,731
	Conoce_enfermedad	-,584	,493	1,400	1	,237	,558

Covariables		B	SE	Wald	df	Sig.	Exp(B)
Paso 14	sexo	,478	,298	2,569	1	,109	1,613
	EDAD1	,050	,011	22,152	1	,000	1,051
	Pro_agricolas	-,258	,167	2,395	1	,122	,773
	cuanto_tiempo	,369	,181	4,173	1	,041	1,446
	Diabetes_millitus	-1,153	,402	8,226	1	,004	,316
	Vias_urinarias	-,548	,290	3,565	1	,059	,578
	Ibuprofeno	,412	,290	2,021	1	,155	1,510
	Indometacina	-2,062	1,154	3,194	1	,074	,127
	Enalapril	-,452	,275	2,696	1	,101	,637
	Otros	-,144	,083	3,046	1	,081	,866
	consumo_enlatados	-,313	,198	2,501	1	,114	,732
	Conoce_enfermedad	-,568	,493	1,327	1	,249	,567
Paso 15	sexo	,477	,299	2,540	1	,111	1,611
	EDAD1	,049	,011	21,642	1	,000	1,050
	Pro_agricolas	-,233	,166	1,976	1	,160	,792
	cuanto_tiempo	,347	,181	3,685	1	,055	1,415
	Diabetes_millitus	-1,171	,401	8,510	1	,004	,310
	Vias_urinarias	-,524	,287	3,322	1	,068	,592
	Ibuprofeno	,428	,290	2,188	1	,139	1,535
	Indometacina	-2,065	1,156	3,190	1	,074	,127
	Enalapril	-,429	,275	2,432	1	,119	,651
	Otros	-,137	,082	2,774	1	,096	,872
	consumo_enlatados	-,321	,197	2,661	1	,103	,726

Covariables		B	SE	Wald	df	Sig.	Exp(B)
Paso 16	sexo	,539	,290	3,449	1	,063	1,714
	EDAD1	,048	,011	20,782	1	,000	1,050
	cuanto_tiempo	,141	,094	2,240	1	,134	1,152
	Diabetes_millitus	-1,096	,397	7,636	1	,006	,334
	Vias_urinarias	-,587	,282	4,343	1	,037	,556
	Ibuprofeno	,422	,289	2,138	1	,144	1,525
	Indometacina	-2,139	1,158	3,409	1	,065	,118
	Enalapril	-,360	,270	1,786	1	,181	,697
	Otros	-,127	,082	2,396	1	,122	,881
	consumo_enlatados	-,271	,192	2,000	1	,157	,762
Paso 17	sexo	,480	,287	2,801	1	,094	1,616
	EDAD1	,049	,011	21,090	1	,000	1,050
	cuanto_tiempo	,129	,095	1,837	1	,175	1,138
	Diabetes_millitus	-,963	,381	6,376	1	,012	,382
	Vias_urinarias	-,524	,277	3,571	1	,059	,592
	Ibuprofeno	,433	,291	2,215	1	,137	1,541
	Indometacina	-1,926	1,147	2,818	1	,093	,146
	Otros	-,137	,082	2,810	1	,094	,872
	consumo_enlatados	-,256	,194	1,739	1	,187	,774

Covariables		B	SE	Wald	df	Sig.	Exp(B)
Paso 18	sexo	,505	,287	3,105	1	,078	1,657
	EDAD1	,047	,010	20,650	1	,000	1,048
	cuanto_tiempo	,161	,093	3,038	1	,081	1,175
	Diabetes_millitus	-,855	,373	5,250	1	,022	,425
	Vias_urinarias	-,470	,276	2,909	1	,088	,625
	Ibuprofeno	,342	,283	1,460	1	,227	1,408
	Indometacina	-1,750	1,136	2,370	1	,124	,174
	Otros	-,106	,078	1,853	1	,173	,899
Paso 19	sexo	,500	,287	3,037	1	,081	1,649
	EDAD1	,046	,010	20,159	1	,000	1,047
	cuanto_tiempo	,152	,092	2,733	1	,098	1,164
	Diabetes_millitus	-,792	,369	4,611	1	,032	,453
	Vias_urinarias	-,406	,270	2,258	1	,133	,666
	Indometacina	-1,655	1,132	2,137	1	,144	,191
	Otros	-,120	,077	2,443	1	,118	,887
Paso 20	sexo	,487	,286	2,902	1	,088	1,628
	EDAD1	,044	,010	19,571	1	,000	1,045
	cuanto_tiempo	,138	,090	2,327	1	,127	1,147
	Diabetes_millitus	-,712	,353	4,064	1	,044	,490
	Vias_urinarias	-,426	,267	2,551	1	,110	,653
	Otros	-,105	,076	1,945	1	,163	,900

Covariable		B	SE	Wald	df	Sig.	Exp(B)
Paso 21	sexo	,465	,285	2,673	1	,102	1,593
	EDAD1	,040	,009	17,804	1	,000	1,041
	cuanto_tiempo	,125	,091	1,899	1	,168	1,133
	Diabetes_millitus	-,703	,353	3,977	1	,046	,495
	Vias_urinarias	-,394	,267	2,184	1	,139	,674
Paso 22	sexo	,289	,263	1,212	1	,271	1,335
	EDAD1	,038	,009	17,881	1	,000	1,039
	Diabetes_millitus	-,675	,354	3,641	1	,056	,509
	Vias_urinarias	-,353	,264	1,781	1	,182	,703
Paso 23	EDAD1	,038	,009	17,158	1	,000	1,038
	Diabetes_millitus	-,634	,352	3,240	1	,072	,531
	Vias_urinarias	-,368	,264	1,937	1	,164	,692
Paso 24	EDAD1	,037	,009	17,004	1	,000	1,038
	Diabetes_millitus	-,569	,349	2,665	1	,086	,566

BIBLIOGRAFÍA.

- Sánchez de Rivera, Daniel Peña (2002). Regresión y Diseño de Experimentos. Alianza Editorial, S. A., Madrid.

- Cartín Brenes, Mayra (1990). Epidemiología y Demografía.

- Rotman, Kenneth J. (2008). Epidemiología Moderna. Editorial, Ediciones Díaz de Santos.

- Canavos, George C. (1988). Probabilidad y Estadística, Aplicaciones y Métodos. Editorial, McGRAW-HILL

- Orvezabal Morena, Mauro Javier. Díaz Rubio, Eduardo (2006). Factores Pronósticos y Predictivos de la Supervivencia Global y Libre de Progresión de Pacientes con Cáncer de Mama Metastasico en Tratamiento de Quimioterapia Intensiva. Editorial, Universidad Complutense de Madrid.

- Mendoza Reyes, Karina Elizabeth. Merche Peraza, Roxana Eugenia. Rodríguez Díaz, Ulises Josué (2003). Factores de Riesgo en Insuficiencia Renal Crónica, en Pacientes del Hospital Nacional Rosales, durante el mes de septiembre del 2003 (Tesis de Doctorado en Medicina).

- Galindo Martínez, Vilma Josefina. Molina Aguilar, Brenda Lizzette (2001). Estudio Bacteriológico del Líquido Peritoneal en Pacientes con Peritonitis Atendido en el Programa de Diálisis del Hospital Nacional San Juan de Dios del Departamento de San Miguel, durante los meses de mayo a julio del año 2001 (Tesis de Doctorado en Medicina).

- Aguilar Abrego, Manuel Enrique. Aguirre Quintanilla, Rafael Ernesto (2007) Presencia de Insuficiencia Renal Crónica y Factores Asociados en las Unidades de

Salud de Carolina y San Francisco del Monte, Cabañas en los meses de mayo a junio del 2007 (Tesis de Doctorado en Medicina).

- Estalin Ademir Mejía Hernández (2009). Análisis de Supervivencia y su Aplicación para Predecir la Calidad de Vida de los Nacidos Extremadamente Prematuros del Hospital de Maternidad. (Tesis de Licenciatura en Estadística).

- Dr. Pérez Rebasa (2006). Articulo especial "Conceptos Básicos del Análisis de Supervivencia" Corporación Sanitaria Parc Tauli. Subadell. Barcelona, España.

- Borges, Rafael Eduardo (2005) Análisis de supervivencia de pacientes con diálisis peritoneal". Revista Colombiana de Estadística, volumen 28 Nº 2. pp. 243 a 259. Diciembre 2005.
  http://www.ciencias.unal.edu.co/publicaciones/estadistica/rce/V28/V28_2_243Borges.pdf.

- Cuba, M., Barak, A., Pérez Rodríguez, M. (1996). Supervivencia de pacientes con insuficiencia renal crónica terminal en Holguín.

- Unidad de Epidemiología Clínica y Bioestadística. Complexo Hospitalario Universitario Juan Canalejo. A Coruña (España). Año 1995.

Páginas Web:

- http://www.medimexico.com.mx/info_e/info_e.html#tratamiento.
- http://www.tendencia%20central/medidas1.htm.
- htpp://www.tendencentral.htm.
- htpp://www.medidasdedipersion.htm.
- sisbib.unmsm.edu.pe/bibvirtualdata/monografias/basic/tineo_gf/cap1.pdf.
- http://www.sc.ehu.es/XIIIJoraede/Comunicaciones/Juan%20Gomez%20Garcia%20SanSebastian12004revision.pdf.
- http://thales.cica.es/rd/Recursos/rd97/UnidadesDidacticas/53-1-u-punt14.html.
- http://es.wikipedia.org/wiki/Mediana_(estad%C3%ADstica).
- http://www.monografias.com/trabajos27/datos-agrupados/datos-agrupados.shtml.