ESCUELA SUPERIOR POLITECNICA DEL LITORAL

Instituto de Ciencias Matemáticas

Ingeniería en Estadística Informática

“Factores predictorios de sobre vida en pacientes con melanoma, mediante el modelo regresión de Cox”

TESIS DE GRADO

Previa a la obtención del Título de:

INGENIERA EN ESTADISTICA INFORMATICA

Presentada por:

Leonor Alejandrina Zapata Aspiazu

GUAYAQUIL- ECUADOR

AÑO

2004

AGRADECIMIENTO

A todas las personas que de una u otra manera colaboraron en la realización de este trabajo y específicamente a la gran ayuda de mi director de tesis el Mat. John Ramírez Figueroa.

DEDICATORIA

A mis padres:

Walter Zapata y

Laura Aspiazu

A mi hermana Laura

TRIBUNAL DE GRADUACIÓN

MAT. JORGE MEDINA	**MAT. JOHN RAMIREZ**
DIRECTOR DEL ICM	**DIRECTOR DE TESIS**
ING. ROBERT TOLEDO	**ING. WASHINGTON ARMAS**
VOCAL	**VOCAL**

DECLARACIÓN EXPRESA

“La responsabilidad del contenido de esta tesis de grado, me corresponde exclusivamente; y el patrimonio intelectual de la misma a la ESCUELA SUPERIOR POLITECNICA DEL LITORAL”

(Reglamento de graduación de la ESPOL)

_____

Leonor Alejandrina Zapata Aspiazu

RESUMEN

El objetivo de esta investigación es estimar la probabilidad de sobrevida, en un pequeño intervalo de tiempo, de los pacientes de la institución SOLCA de la ciudad de Guayaquil que presentaron melanoma. Para obtener esta probabilidad se analizaron variables cuyo valor-p era significativo.

En el Capítulo I, se presenta todo acerca del cáncer de piel o melanoma, prevención, epidemiología, factores de riesgo, síntomas, etc. En el siguiente capítulo, se presentan las variables utilizadas para la investigación, además su respectiva codificación.

En el Capítulo III, se realiza el análisis descriptivo de todas las variables utilizadas y en el último capítulo se muestran los análisis de sobrevida empleando técnicas como Kaplan Meier y Regresión de Cox.

INDICE GENERAL

II. DETERMINACION DE LAS VARIABLES A SER INVESTIGADAS

ABREVIATURAS

ADN	Ácido desoxirribunocleico
(f/q)	Fallecimiento o quiebra
LDH	Lactato deshidrogenasa
RNM	Resonancia magnética nuclear
Rx	Rayos x
Sig	Significancia estadística
TAC	Tomografía axial computarizada
UV	Rayos ultravioleta
%	Porcentaje

SIMBOLOGIA

λ	Valor propio
$\bar{y}$	Media muestral
S^2	Varianza muestral
H_0	Hipótesis nula
H_1	Hipótesis alterna
n	Número de datos
y_i	Observación i
μ	Media poblacional
σ^2	Varianza poblacional
Ω	Espacio muestral omega
R	Rango
X_L	Valor máximo en un conjunto de datos
X_S	Valor mínimo en un conjunto de datos
γ_1	Coeficiente de sesgo
γ_2	Coeficiente de curtosis
$Cov(Y_1, Y_2)$	Covarianza entre Y_1 y Y_2
$E(Y_i)$	Valor esperado
p	Valor de significancia estadística
p	Cantidad de variables numéricas

x_{jk}	Valor particular de la k-ésima variable que es observado en el j-ésimo ítem
X	Arreglo rectangular de filas y columnas
e_i	Vectores propios
n_{ij}	Frecuencia observada en el renglón i y columna j
p_{ij}	Probabilidad en el renglón i y columna j
r_i	Número de observaciones en el renglón i
c_i	Número observaciones en la columna j
χ^2	Ji cuadrado
$f(t)$	Función densidad de probabilidad
$F(x)$	Función distribución de probabilidad
$S(x)$	Función de supervivencia
$H(t)$	Función de riesgo acumulado
t_i	Tiempo de observación en i
p_k	Probabilidad de sobrevivir en un período de tiempo k
d_k	# de fallecimientos en el instante k
R_k	# de individuos expuestos a riesgo en el instante t_k

$h_0(t)$	Función de riesgo sin considerar el efecto del conjunto de variables X
β	Parámetro desconocido beta
$S(t/x)$	Probabilidad de que el suceso final no ocurra hasta pasado un período de tiempo superior o igual a t
gl	Grados de libertad
ET	Error estándar

INDICE DE TABLAS

INDICE DE GRAFICOS

CAPITULO 1

CONCEPTOS BÁSICOS SOBRE EL CÁNCER

1.1. Células Normales Y Células Cancerosas

La célula es el elemento más simple, dotado de vida propia, que forma los tejidos organizados. Está compuesto por una masa rodeada de protoplasma que contiene un núcleo. Una pared celular rodea la célula y la separa de su ambiente.

Dentro del núcleo está el ADN, que contiene la información que programa la vida celular. El hombre está compuesto por millones de células.

La célula se divide y al hacerlo sus estructuras se dividen también en otras exactamente iguales a las anteriores, con los mismos componentes y funciones que la originaria.

Las células normales crecen a un ritmo limitado y permanecen dentro de sus zonas correspondientes. Las células musculares se forman y crecen en los músculos y no

en los huesos; las de los riñones no crecen en los pulmones, etc.

Estas funciones y este ritmo de crecimiento vienen determinados por el ADN. Algunas células tienen menos tiempo de vida que otras, como por ejemplo las células del intestino que tienen un período de vida de dos semanas, mientras que los hematíes viven durante unos tres meses.

Otras células pueden vivir el tiempo que viva la persona y solo se dividen para sustituirse a sí mismas, este sería el caso de las células óseas que actúan cuando hay que reparar una fractura. La sangre aporta el oxígeno y los nutrientes necesarios para la vida celular y recoge los productos de desecho producidos por las células y los transporta a los órganos de filtrado y limpieza (riñones, hígado, pulmones).

La Linfa es un líquido incoloro que se compone, en su gran mayoría, por linfocitos un tipo de glóbulos blancos, y que recorre todo el organismo a través de vasos linfáticos.

La célula normal pasa a convertirse en una célula cancerosa debido a un cambio o mutación en el ADN. A veces esas células, cuya carga genética ha cambiado mueren o son eliminadas en los ganglios linfáticos, pero otras veces siguen con vida y se reproducen.

Las células cancerosas tienen un aspecto diferente, bien porque su forma ha cambiado o porque contengan núcleos más grandes o más pequeños. Estas células son incapaces de realizar funciones que correspondan a las células pertenecientes a ese tejido. Generalmente se multiplican muy rápidamente, porque les falta un mecanismo de control del crecimiento.

Con frecuencia, son inmaduras debido a que se multiplican de una forma muy rápida y no tienen tiempo suficiente para crecer plenamente antes de dividirse. Al formarse un gran número de células cancerosas, se amontonan, presionan o bloquean a otros órganos y les impiden realizar sus trabajo.
Como no se limitan al espacio originario donde se forman, y se extienden a otras zonas, se dicen que son invasivas.

Tienden a emigrar a otros lugares, a través de la sangre o de la linfa. Las células que se encargan de la defensa del organismo suelen destruirlas, así separadas, pero si sobreviven pueden producir un nuevo crecimiento en un lugar diferente, metástasis, y a dañar a otros órganos.

1.2. ¿Qué es el Cáncer de Piel?

El cáncer de piel es una enfermedad producida por el desarrollo de células cancerosas en las capas exteriores de la piel. Existen dos tipos:

1.2.1. Tipo No Melanoma

El cáncer de tipo no melanoma es el más frecuente y se denomina no melanoma porque se forma a partir de otras células de la piel que no son los melanocitos. Dentro de este tipo se encuentran todos los cánceres de piel menos el melanoma maligno.

El cáncer de piel se da más en las personas de piel blanca y que han pasado mucho tiempo expuestas a los rayos solares. Aunque puede aparecer en cualquier parte de la piel, es más frecuente que se presente en la cara, cuello,

manos y brazos.

Se puede reconocer por un cambio en el aspecto de la piel, como una herida que no sana o una pequeña protuberancia, también puede aparecer una mancha roja, áspera o escamosa. Ante cualquier cambio o anormalidad de la piel, hay que acudir al médico. Éste puede extraer una muestra y analizarla (biopsia) para comprobar si es un tumor maligno o no.

1.2.2. Tipo Melanoma

El tipo melanoma es una enfermedad de la piel que se forma a partir de las células que coloran la piel los (melanocitos). Los melanocitos son encontrados en la epidermis, es decir en la capa superficial de la piel, y ellos contienen melanina que da color a la piel. El melanoma se lo conoce también como melanoma cutáneo o maligno.

1.3. ¿Factores que intervienen en el cáncer de piel?

Según los investigadores, la primera causa parece ser la intensidad de las radiaciones ultravioleta procedentes del sol,

que, debido a la degradación de la capa de ozono, el filtro natural, inciden con mayor vehemencia en la superficie de la tierra.

El espectro UVB de la radiación solar posee la mayor potencia de inducción de cáncer de piel, ya que induce daño estructural en el ADN celular, al mismo tiempo que estimula la proliferación de la epidermis. Estimaciones recientes han calculado que por cada reducción de un 1 % en la capa de ozono, la radiación UVB/UVC aumenta en un 2 % y el cáncer de piel en un 2 a 6 %. Siendo los rayos UVC más cortos que los rayos ultravioleta, éstos son potencialmente más dañinos, pero son totalmente filtrados por la parte superior de la capa de ozono. Sin embargo, conforme la capa de ozono es destruida el riesgo aumenta.

1.3.1 Factores del cáncer de células basales

Los casos de cáncer de piel de células basales son más comunes en personas de piel clara, ojos azules y pelo rubio. La gente albina correrá mayor riesgo. En el otro extremo, las personas de raza negra difícilmente adquieren la enfermedad.

Hay investigadores que comentan la posibilidad de la predisposición genética como factor de riesgo.

Las regiones situadas a gran altura reciben con mayor intensidad las radiaciones ultravioleta que las situadas al nivel del mar cubiertas por encima de ellas de un mayor espesor atmosférico. Esto explica una proporción mayor de casos ocurridos en poblaciones de los Andes.

La latitud geográfica. Se observa que en zonas tropicales los índices de radiación UVB son muy altos, debido a la verticalidad de la incidencia de la radiación. Así encontramos que Australia es un país con gran cantidad de casos de cáncer de piel y donde el gobierno realiza gran esfuerzo para prevenir esta enfermedad.

La nubosidad protege de las radiaciones ultravioletas; pero en mediciones experimentales se han encontrado resultados sorprendentes. Tal es así que, con nublados poco densos o discontinuos y que aparecen como un cielo cubierto, permiten la penetración de una gran proporción de radiaciones ultravioleta. Además, por la situación de la nube, se pueden producir fenómenos de reflexión que se puede producir en el

lateral de la masa nubosa, pudiendo sumarse en un punto de la superficie a la radiación directa en ese punto produciendo un incremento sorprendente de la intensidad ultravioleta.

Recientemente se está investigando que la exposición a campos magnéticos de 50 o 60 Hz, junto con la exposición a la radiación ultravioleta, incrementa la efectividad de ésta en la posibilidad de producir tumores en la piel. La exposición al arsénico, que puede presentarse en ciertos herbicidas.

Suele ocurrir en personas con edad por encima de los 40 años. Existe efecto acumulativo de la radiación ultravioleta adquirida durante toda la vida; por lo que, aunque en esas edades se tome poco el sol, puede aparecer cáncer de piel si anteriormente se ha tomando el sol en exceso.

1.4. Síntomas del Melanoma de Piel

El melanoma puede aparecer como un cambio en aquellas manchas de la piel. Cualquier llaga, protuberancia, marca, etc. que sea sospechosa pudiera ser un melanoma. La piel puede volverse áspera o escamosa o puede sangrar o

exudar. Se puede dar un melanoma a partir de un lunar, que cambie de aspecto, o textura.

Por lo general, un lunar es una mancha de color uniforme, de color café, canela o negro en la piel. Tienen menos de seis milímetros de diámetro y puede estar presente desde el nacimiento o puede aparecer en la infancia o juventud. La mayoría de las personas tienen lunares que son benignos.

Es importante reconocer sus cambios.

La regla del ABCD puede ayudar a reconocer las características de un melanoma:

Asimetría: la mitad del lunar no se corresponde con la otra mitad.

Bordes irregulares: Los bordes del lunar son desiguales.

Color: el color del lunar no es uniforme, sus tonalidades varían desde un marrón a un rojo, o azul.

Diámetro: el lunar tiene más de 6 milímetros de ancho. Aunque esta regla es útil para la mayoría de los melanomas, no todos se ajustan a estas características.

1.5. Tratamiento

Los estadios localizados del melanoma tienen unas posibilidades elevadas de ser curados con cirugía. El empleo de ésta para los diseminados se utilizará con intención paliativa, es decir, para disminuir los síntomas.

La lesión primaria debe ser extirpada, incluyendo piel, tejido celular subcutáneo y aponeurosis. Como para realizar el diagnóstico se habrá realizado una biopsia escisional, deberá cortarse por la cicatriz con un margen amplio, entre dos y tres centímetros.

Para los melanomas menores de 0,76 mm., será suficiente extirpar un margen de 1 cm. Cuando hay ganglios afectados, deberán ser extirpados. Esto se realizará cuando se evidencie una inflamación de los ganglios.

Biopsia de ganglios centinelas: esta técnica se encuentra en estudio y puede realizarse o no según el criterio médico. Consiste en averiguar qué ganglio es el que se drena fluido linfático a la zona del melanoma y analizarlo. Para ello lo que

se hace es inyectar una sustancia, coloreada o con un componente radioactivo, en la zona del melanoma.

Al cabo de un tiempo, se podrá observar coloreado o con el compuesto radioactivo aquel ganglio que haya absorbido la sustancia y que será el que pueda contener mayor número de células cancerosas, si el cáncer se ha extendido. Cuando se ha localizado el ganglio, se toma una muestra y se analiza al microscopio. Si presenta células cancerosas, se extirpará. También se extirparán los ganglios linfáticos restantes de esa zona.

Si se ha evidenciado la existencia de metástasis en otros órganos, puede realizarse una cirugía aunque no tenga como objetivo la curación. A veces la extirpación de metástasis en otros órganos, aumenta el tiempo de vida del paciente o, por lo menos, mejora los síntomas que éste presenta.

Quimioterapia

La quimioterapia sistémica se emplea como tratamiento (*)paliativo de los síntomas.
Se utiliza después del tratamiento de cirugía en algunas

metástasis dérmicas, cerebrales, intestinales u óseas. El tratamiento con un solo fármaco o con combinación de ellos es poco eficaz y las tasas de respuesta no superan el 30%. El tiempo de curación es poco. Aún así se siguen realizando investigaciones combinando varios fármacos.

Los medicamentos que se utilizan con más frecuencia son la dacarbacina (DTIC), la carmustina (BCNU), el taxol, el platino, la vinblastina y la vincristina.

Se pueden emplear distintas combinaciones de medicamentos, recientemente se han descrito resultados alentadores con la asociación de DTIC, platino, BCNU y tamoxifeno. Algunas combinaciones de quimioterápicos se pueden asociar a medicamentos de inmunoterapia como son el interferón, la interleukina-2 y los anticuerpos monoclonales.

*Paliativo: Tratamientos que en vez de centrarse en combatir la enfermedad, o tratar de prolongar la vida, van encaminados en que la existencia del paciente sea más cómoda y agradable mientras dure su lucha. Se conoce como mejor calidad de vida hasta que ocurre el deceso.

1.6. Prevención

El principal factor de riesgo en este cáncer es una exposición excesiva a la radiación ultravioleta. Evitar una intensa o prolongada exposición al sol, intentando no exponerse en horas en las que la radiación solar es mayor es la mejor medida de prevención que se puede utilizar.

Otras formas son la utilización de materiales que protejan aquellas zonas más delicadas como es el uso de sombreros, de gafas que absorban los rayos ultravioleta de un 99% a un 100%, o utilizar telas adecuadas para cubrir la piel. El uso de cremas protectoras solares reduce el peligro de la exposición. Deben utilizarse correctamente, hay distintos grados según sea el tipo de piel. Además deben de emplearse con un tiempo de antelación a la exposición solar. Como se ha visto, las cabinas y las lámparas bronceadoras resultan peligrosas, por lo que deben evitarse.

Cuando se observe un lunar que ha cambiado de aspecto, o que sangra, se debe acudir al médico. Éste lo puede extirpar y realizar una biopsia para comprobar si es maligno o no. Este tipo de cáncer puede prevenirse más que muchos otros.

Siguiendo las anteriores instrucciones pueden disminuirse o anularse gran parte de los factores de riesgo, con lo que las posibilidades de padecer un cáncer de piel también disminuyen.

1.6.1 Medidas antes y durante la exposición solar

Evitar el uso de productos que contengan alcohol y perfumes. Elegir el protector solar adecuado, en función del tipo de piel, del lugar de aplicación y de las condiciones ambientales.

Aplicar una buena cantidad del producto solar 30 minutos antes de tomar el sol, sobre la piel seca. Evitar tomar el sol entre las doce de la mañana y las cuatro de la tarde, en zonas de gran altitud, y en lugares próximos al ecuador.

El agua, la nieve y la arena actúan reflejando los rayos solares y aumentando su intensidad. Es por este motivo, por lo que pueden producirse quemaduras incluso en la sombra. Las primeras veces que se tome el sol, se deberá emplear un factor de protección más elevado.

Aún en los días nublados, hay que utilizar el protector solar. Hay que ingerir muchos líquidos para compensar la pérdida de líquidos debido a la exposición solar.

Después de un baño, cuando se haya sudado mucho, o tras pasar dos horas de la anterior aplicación, habrá que volver a aplicar la crema protectora.
Tras la exposición al sol, hay que ingerir muchos líquidos debido a la pérdida de éstos.

EPIDEMIOLOGIA

En los Estados Unidos, el melanoma es el octavo proceso maligno en orden de frecuencia y representa el 3% de todos los cánceres. En 1985 se diagnosticaron 11,2 casos por 100.000 habitantes de sexo masculino y 8,4 casos por 100.000 habitantes de sexo femenino, que determinaron 2,8 y 1,5 muertes, respectivamente. Las estimaciones indican que en 1993 se diagnosticarían 32.000 nuevos casos y habría 6.800 muertes.

Durante la última década la incidencia de melanoma ha aumentado con más rapidez que la de cualquier otro proceso maligno, alcanzando un promedio de 5-7% por año (que se duplica cada 10-15 años). Para el año 2000 1 de cada 75 individuos presentará

melanoma cutáneo. En todo el mundo la incidencia varía de una alta tasa de 28,4 por 100.000 habitantes en Australia a 0,2 por 100.000 en Japón.

FACTORES DE RIESGO

A. **EDAD.-** La incidencia del melanoma aumenta en forma gradual de menos de 1 por 100.000 antes de los 20 años a 26,8 por 100.000 hacia los 80 años de edad.

B. **SEXO.-** El riesgo es 1,3 veces mayor en los varones que en las mujeres.

C. **RAZA.-** El riesgo de los caucásicos es 17 veces mayor que el de los negros, y alcanza a un riesgo del 1% durante la vida. Las poblaciones asiática, hispánicas y de indios americanos están expuestos a un riesgo intermedio.

D. **Factores Genéticos**

   1. **Antecedentes Familiares.-** El riesgo de melanoma de los familiares directos de pacientes afectados es del cuádruple del de la población general.

   2. **Síndromes Familiares.-** Los síndromes de melanoma familiar, con herencia mendeliana verdadera, constituyen el 8-12% de los casos de melanoma cutáneo.

Típicamente estas lesiones aparecen antes que las lesiones no hereditarias y a menudo son múltiples.

3. **Anormalidades Genéticas**.- Por lo general, las células del melanoma son sumamente aneuploides y suelen presentar alteraciones cromosómicas.

E. Color de la Piel.- Los individuos con fenotipo rubio, sobretodo los que tienen cabello rubio o pelirrojo, piel clara y deficiente capacidad de broncearse tienen un riesgo sustancialmente más alto de presentar melanoma que la población general.

F. Patología dermatológica previa.

1. **Pecas y Nevos.-** El riesgo de melanoma maligno en individuos con pecas o más de 20 nevos es del triple del de aquellos que no presentan estos rasgos. El nevo es un crecimiento congénito (esto es, hereditario, de nacimiento, innato) o marca sobre la piel, como una mancha o lunar.

2. **Queratosis actínica.-** La existencia de Queratosis actínica se asocia con un riesgo más alto de melanoma maligno.

3. Lesiones pigmentadas precursoras

a. **Síndrome de nevo displásico**.- Este trastorno se caracteriza por un mayor riesgo de melanoma, nevos múltiples (<100) por lo menos un nevo mayor de 8 mm. y no menos de 1 nevo de la unión o compuesto con características atípicas, tanto macroscópicas (es decir, pigmentación moteada, irregularidad, asimetría y diámetro <6 mm.) y microscópicas (p. ej. Hiperplasia melanocítica y elongación del relieve de la red; núcleos melanocíticos hipercromáticos, agrandados; puentes de relieve de la red de melanocitos agregados; fibroplasia dérmica laminar y concéntrica y un infiltrado melanocítico).Una variante familiar común es el síndrome de lunar-melanoma múltiple atípico familiar.

b. **Nevos congénitos**.- Se observan nevos melanocíticos congénitos en el 1% de los recién nacidos y se clasifican en pequeños (<1,5cm), medianos (1,5- 20 cm.)y grandes (>20). Las características clínicas de estos hamartomas son:

superficie microscópicamente irregular, hiperpigmentación e hipertricosis. Los melanomas pueden surgir dentro de cualquier nevo congénito, independientemente del tamaño, aunque es probable que el riesgo sea proporcional al diámetro.

c. **Melanoma.**- El riesgo de melanoma en pacientes con diagnóstico previo de este cáncer es de 900 veces el de la población general, o un riesgo de por vida de un 3-5%.

d. **Xeroderma pigmentoso**.- Esta rara enfermedad autonómica recesiva se caracteriza por un defecto genético en la reparación de la lesión del ADN provocada por la radiación ultravioleta. Estos pacientes tienen un riesgo sumamente alto de melanoma cutáneo, carcinomas de células basales y escamosas, y sarcomas. La mayoría muere antes de los 25 años de edad.

e. **Exposición a la radiación**.- La alta incidencia de melanoma en los individuos que viven cerca del

ecuador, migran a climas soleados, presentan 3 o más quemaduras de sol con ampollas o realizan trabajos estivales al aire libre antes de los 20 años, y en las localizaciones anatómicas expuestas a la luz solar sugiere que la exposición acumulativa a la luz ultravioleta es un factor de riesgo importante de melanoma cutáneo. La reciente depleción de ozono pueden ser responsable, en parte, del aumento de la incidencia de este tumor. Sin embargo aún no se ha esclarecido la relación exacta entre la luz ultravioleta y el melanoma, dado que este puede afectar áreas relativamente no expuestas de la piel como ejemplo, las palmas, las plantas y las áreas del tronco cubiertas por un traje de baño y a pacientes jóvenes sin antecedentes de larga exposición al sol. Con suma probabilidad, el melanoma está relacionado con periodos de exposición al sol aguda, intensa e intermitente manifestada por quemaduras de sol ampollares.

f. **Hormonas.**- Algunos datos sugieren que las mujeres que toman anticonceptivos orales durante

más de 5 años y que tienen el primer hijo después de los 30 años presentan proliferación rápida y la diseminación del melanoma durante el embarazo.

g. **Dieta.**- No se han detectado factores dietéticos.

h. **Tabaquismo**.- El tabaco no ha sido implicado en la patogenia del melanoma cutáneo.

i. **Urbanización**.- El melanoma es más frecuente entre empleados urbanos que trabajan en ambientes interiores y practican actividades recreativas al aire libre, que en los individuos que desempeñan tareas de agricultura y en los obreros que trabajan fundamentalmente al aire libre.

ESTADIAJE O ESTADIOS Y CLASIFICACIÓN

Es preciso, ahora hablar de factores pronósticos del melanoma maligno cutáneo y realizar una revisión de los sistemas de clasificación:

Estadiaje del melanoma maligno
Según niveles de invasión (Clark)

Nivel I Limitado a epidermis.

Nivel II Atraviesa la membrana basal.

Nivel III Limita con dermis reticular.

Nivel IV Se extiende por dermis reticular.

ivel V Se extiende a la grasa subcutánea.

Profundidad en milímetros (Breslow)

Nivel I: 0,75 mm o menos

Nivel II: 0,76 mm - 1,50 mm

Nivel III: 1,51 mm - 4,0 mm

Nivel IV: 4,10 mm o más

Clasificación TNM

T0 Melanoma "in situ". Sin lesión invasiva. Nivel I de Clark.

T1 0,75 mm o menos. Nivel II de Clark.

T2 0,76 mm - 1,50 mm. Nivel III de Clark.

T3 1,51 mm - 4,0 mm. Nivel IV de Clark.

T4 4,10 mm o más. Nivel V de Clark y/o lesiones satélites a menos de 2 cm del tumor primario.

N1. Metástasis de diámetro máximo menor o igual a 3 cm en cualquier ganglio regional.

N2. Metástasis de diámetro máximo mayor a 3 cm en cualquier ganglio regional y/o metástasis en tránsito.

- N2a: metástasis de diámetro máximo mayor a 3 cm en cualquier ganglio regional.
- N2b: metástasis en tránsito.
- N2c: ambos

Las metástasis en tránsito afectan a la piel o al tejido subcutáneo a una distancia superior a 2 cm desde el tumor primario, pero no más allá de los ganglios linfáticos regionales.

M1a: metástasis en piel, tejido subcutáneo o ganglios linfáticos más allá de los ganglios linfáticos regionales.

M1b: metástasis viscerales.

CAPITULO 2

DETERMINACIÓN DE LAS VARIABLES A SER INVESTIGADAS

2.1. Descripción de las variables de estudio

Para poder realizar nuestra investigación acudimos a la Sociedad de Lucha Contra el Cáncer "SOLCA" de la ciudad de Guayaquil, donde se tomó los datos de los pacientes que ingresaron en el establecimiento anteriormente nombrado de salud, por presentar un cuadro clínico que se encuentra enmarcado dentro de las especificaciones de melanoma. Se trató de obtener los datos de todos los pacientes que cumplían con las características antes mencionadas, cabe recalcar que la información se logró recoger satisfactoriamente, ya que se encontraron todas las historias clínicas que van hacer objeto de este estudio.

Para poder iniciar con la recolección de la información, estuvimos asesorados de un experto en la materia, con la finalidad de poder establecer cuales deberían ser las variables de interés de las cuales obtendríamos los datos para el posterior análisis. Después de un análisis minucioso

junto con el experto en está área de la medicina tomando en cuenta los factores que inciden en esta enfermedad se estableció que el análisis debería basarse sobre veintidós variables las mismas que se describen a continuación.

***Variable # 1*: Sexo**

Esta variable nos indica el género de los pacientes de SOLCA de la ciudad de Guayaquil, la misma que sirve para determinar la proporción de hombres y mujeres con esta enfermedad.

***Variable # 2*: Edad**

La edad nos permitirá saber el tiempo transcurrido desde el nacimiento de un ser hasta el momento en que esta variable es medida. Con esta variable buscamos obtener información con respecto a la edad en años que tenia el paciente en el momento de contraer la enfermedad.

***Variable # 3*: Antecedentes**

Esta variable nos indica si el paciente con melanoma tuvo un familiar con la misma enfermedad o padeció de algún tipo de cáncer. Es importante porque el factor genético parece tener influencia en cualquier tipo de cáncer.

***Variable # 4*: Localización del tumor**

Esta variable nos indica en que parte de su cuerpo el paciente presenta la enfermedad, en este caso, en que sitio se presenta el melanoma. Las localizaciones comunes son las siguientes:

Localizaciones BANS por sus siglas en ingles.

Back: parte superior de la espalda

Arm: zona posterior de los brazos

Neck: parte posterior y lateral del cuello.

Scalp: zona posterior y superior del cuello cabelludo.

***Variable # 5*: Inicio de metástasis**

Esta variable nos indica si el paciente con melanoma presenta metástasis durante la enfermedad; es decir la diseminación del cáncer a otros órganos del cuerpo humano. La metástasis se refiere a la habilidad de las células cancerosas de penetrar en los vasos sanguíneos y linfáticos, circular por el torrente sanguíneo y luego invadir el tejido normal en otras partes del cuerpo.

***Variable # 6*: L.D.H.**

La enzima Lactato deshidrogenasa pertenece a la clase Óxido-reductasa. Es una enzima tetrámero que se encuentra

en corazón, hígado, músculo, eritrocitos, plaquetas y nódulos linfáticos. Se sintetiza desde dos genes individuales distintos, que originan polipéptidos estructuralmente diferentes pero con la misma actividad catalítica. Su función es la de reducir reversiblemente el piruvato a lactato. Está relacionada con el infarto de miocardio, hemolisis y enfermedades del parénquima hepático.

***Variable* # *7*: Gamma**

Para diagnosticar metástasis se realiza un tipo de estudio radiológico en este caso se utilizó rayos gamma. Esta variable sirve para conocer si el paciente recibió este tipo de tratamiento el cual le ayudará a mejorar su condición.

***Variable* # *8*:T.A.C**

El diagnóstico del melanoma maligno se hace mediante la extirpación del cáncer primario. Cuando se diagnostica el melanoma, es importante determinar si el cáncer se diseminó a otras partes del organismo. Los exámenes para determinar la extensión de la enfermedad son conocidos como procedimientos de clasificación por etapas y la última

extensión o etapa de la diseminación del melanoma determina el tratamiento y los resultados. La clasificación clínica se basa en si el cáncer se diseminó a los ganglios linfáticos regionales o a lugares distantes, lo que se determina mediante el examen médico, exploración con tomografía axial computarizada (T.A.C).

***Variable* # *9*:R.N.M**

Por sus siglas R.N.M. significa Resonancia Magnética Nuclear, también es un tratamiento que sirve para diagnosticar metástasis.

***Variable* # *10*: Cirugía**

Esta variable nos indica si el paciente con melanoma, ha sido sometido a cirugía para la extirpación del tumor.

***Variable* # *11*: Necrosis**

Esta variable nos indica si el paciente presentó gangrena, es decir, la putrefacción de la parte del cuerpo afectada por el melanoma.

***Variable # 12*: Estadiaje o Estadío**

Esta variable nos permite identificar en que fase de la enfermedad se encuentran mayormente los pacientes que presenta melanoma.

Etapa I: El melanoma maligno se encuentra en la capa externa de la piel (epidermis) y/ o sobre la capa interna de la piel (dermis), pero no se ha diseminado a los ganglios linfáticos.

Etapa II: El melanoma maligno se ha diseminado a la parte inferior de las capas internas de la piel (dermis), pero no se ha diseminado dentro del tejido bajo la dermis, o a los ganglios linfáticos cercanos.

Etapa III: Describe las tumoraciones que ocupan toda la dermis papilar expandiéndola. El melanoma maligno también puede tener cualquier tamaño con diseminación a los ganglios linfáticos regionales.

Etapa IV: El melanoma maligno primario es de cualquier tamaño, pero se ha diseminado a los ganglios linfáticos y/ o a lugares distantes.

***Variable # 13*: Quimioterapia Preoperatoria**

Nos indica si el paciente recibió tratamiento de quimioterapia antes de someterse a una cirugía.

***Variable # 14*: Quimioterapia Postoperatoria**

Nos indica si el paciente recibió tratamiento de quimioterapia después de someterse a una cirugía.

***Variable # 15*: Tratamiento completo**

Esta variable nos indica si el paciente con melanoma fue sometido a la quimioterapia preoperatoria; cirugía, quimioterapia postoperatoria y radioterapia, es decir si el paciente recibió los cuatro tratamientos.

***Variable # 16*: Metástasis durante la enfermedad**

Esta variable nos indica si el paciente con melanoma presentó diseminación de cáncer en algún órgano del cuerpo. La metástasis se refiere a la habilidad de las células cancerosas de penetrar en los vasos sanguíneos y linfáticos, circular por el torrente sanguíneo y luego invadir el tejido normal en otras partes del cuerpo.

***Variable # 17*: Localización de metástasis**

Esta variable nos indica específicamente, en qué órgano se diseminó el cáncer.

***Variable # 18:* Estado de ultima observación**

Esta variable nos muestra el número de pacientes vivos, muertos y los que abandonaron el tratamiento.

***Variable # 19*: Fecha de ingreso**

Esta variable, indica la fecha que ingresó el paciente a la institución SOLCA, de la ciudad de Guayaquil.

***Variable # 20*: Rx.**

El diagnóstico del melanoma maligno se hace mediante la extirpación del cáncer primario. Cuando se diagnostica el melanoma, es importante determinar si el cáncer se diseminó a otras partes del organismo. Los exámenes para determinar la extensión de la enfermedad son conocidos como procedimientos de clasificación por etapas y la última extensión o etapa de la diseminación del melanoma determina el tratamiento y los resultados. La clasificación clínica se basa en si el cáncer se diseminó a los ganglios

linfáticos regionales o a lugares distantes, lo que se determina mediante el examen médico, los rayos X. Es decir, para diagnosticar metástasis se realiza un tipo de estudio radiológico en este caso se utilizó los rayos x.

***Variable # 21*: Fecha de cirugía**

Variable que indica el día, mes y año en la cual el paciente con melanoma fue intervenido quirúrgicamente.

***Variable # 22*: Fecha de diagnóstico patológico**

Esta es una variable que nos indica la fecha exacta en que se le diagnosticó al paciente cual era su enfermedad. Esta variable es muy importante para realizar el análisis de sobrevida ya que a partir de ella conocemos el tiempo que el paciente permaneció vivo, es decir, para conocer el número de días que el paciente estuvo con la enfermedad tomamos la diferencia de fecha de diagnóstico patológico menos el estado de última observación.

2.2. Codificación de las variables a ser investigadas

Muchas de las variables de las cuales hemos obtenido la información son de tipo cualitativo, para poder realizar el

análisis estadístico de ellas debemos codificarlas mediante escalas de Lickert, con el objeto de convertir las variables cualitativas en variables cuantitativas y poderlas utilizar en el análisis de componentes principales y posteriormente para el análisis de la curva de sobrevivida mediante análisis de regresión de Cox.

***Variable* # *1*: Sexo**	**Codificación**
Masculino	0
Femenino	1

***Variable* # *3*: Antecedentes**	**Codificación**
No reporta	0
Si reporta	1

***Variable* # *4*: Localización del tumor**	**Codificación**
Dedo del pié	0
Párpado	1
Oído externo	2
Parte no específica de la cara	3
Cuero cabelludo	4
Tronco	5

Extremidades superiores y hombros	6
Extremidades inferiores	7

***Variable* # *5*: Inicio de metástasis**	**Codificación**
No reporta	0
Si reporta	1

***Variable* # *7*: Gamma**	**Codificación**
No recibió rayos gamma	0
Si recibió rayos gamma	1

***Variable* # *8*:T.A.C**	**Codificación**
No recibió	0
Si recibió	1

***Variable* # *9*:R.N.M**	**Codificación**
No recibió	0
Si recibió	1

***Variable* # *10*: Cirugía**	**Codificación**
No recibió cirugía	0
Si recibió cirugía	1

***Variable # 11*: Necrosis**	**Codificación**
No	0
Sí	1
No reporta	2

***Variable # 12*: Estadiaje**	**Codificación**
I	1
II	2
III	3
IV	4

***Variable # 13*: Quimioterapia Preoperatoria**	**Codificación**
No recibió	0
Si recibió	1

***Variable # 14*: Quimioterapia Postoperatoria**	**Codificación**
No recibió	0
Si recibió	1

***Variable # 15*:**

Tratamiento completo	**Codificación**
No recibió	0
Si recibió	1

***Variable # 16*:**

Metástasis durante la enfermedad	**Codificación**
No presentó metástasis	0
Si presentó metástasis	1

***Variable # 17*:**

Localización de metástasis	**Codificación**
Ninguna	0
Linfáticos regionales	1
Linfáticos a distancia	2
Hueso	3
Hígado	4
Pulmón, pleura o ambos	5
Cerebro	6
Ovario	7
Otra	8

Variable # 18:

Estado de ultima observación	**Codificación**
Vivo	0
Muerto	1
Abandono	2

***Variable # 20*: Rx.**	**Codificación**
No recibió rayos x	0
Si recibió rayos x	1

2.3. Técnicas Univariadas

El análisis estadístico se divide en tres grandes tipos: univariado, bivariado y multivariado. En el análisis univariado se describen las características de una variable por vez. También se lo llama estadística descriptiva.

En el análisis bivariado se investiga la influencia de una variable que es independiente, por vez, con respecto a la variable dependiente.

En el análisis multivariado se investiga la influencia de dos o más variables independientes, junto o no a una o más variables asociadas (covariables o cofactores) sobre una o más variables dependientes.

2.3.1. Variables en Bioestadística

Variable: es una característica o propiedad determinada del individuo, sea medible o no. Esta propiedad hace que las personas de un grupo puedan diferir de las de otro grupo en la muestra o población de estudio.

Muestra: es el grupo de pacientes u observaciones que se estudiará, la cual debe haberse elegido al azar y ser representativa de la población a la cual pertenece. En general la muestra es toda parte representativa de un conjunto, población o universo, cuyas características debe reproducir en pequeño lo más exacto posible. A partir del análisis de la muestra, obtenida correctamente y al azar, se pueden hallar conclusiones que sean extrapolables a la población de origen.

Población: conjunto de individuos, sujetos u observaciones con alguna característica en común. Conjunto de elementos de la misma especie que se pretende estudiar en una investigación científica y de la cual se obtiene una muestra.

Las poblaciones pueden ser clasificadas básicamente como sigue:
Población General o Madre: población real que se pretende estudiar y a la cual se extenderán las conclusiones de la muestra perteneciente a la misma.
Las variables se clasifican en:

Variable Cuantitativa: Es la que se puede medir. Habitualmente es llamada variable numérica o continua, o sea que posee una continuidad. Por ejemplo la edad, hematócrito, transporte de

oxigeno, altura, peso, frecuencia cardiaca o respiratoria, dosis de un medicamento.

Variable Cualitativa: Son variables que representan cualidades de la muestra, como por ejemplo la evolución del paciente hacia la mejoría o la muerte, color de ojos de un grupo de personas, sexo, etc. Estas variables también son llamadas categóricas o discretas, por dividirse en categorías.

Las variables cualitativas se clasifican en:

Variables Categóricas Dicotómicas: Son las que tienen dos valores fijos y excluyentes entre si como la evolución, presencia o ausencia de una enfermedad o característica en la muestra.

Variables Categóricas Nominales: Son variables cualitativas que no permiten establecer un orden, por ejemplo la raza, que puede ser blanca, negra, caucásica, etc., o los grupos sanguíneos A, B, AB o 0. También son excluyentes entre si, o sea que cada paciente pertenece a una u otra categoría pero no a dos al mismo tiempo.

Variables Categóricas Ordinales: Estas si permiten establecer un orden determinado.También son excluyentes entre sí.

Además de lo expuesto anteriormente, existe otra forma de clasificar a las variables que es también de suma importancia en estadística: en dependientes, independientes y asociadas.

Variable Dependiente: Es la variable motivo de nuestro interés, cuyos valores dependen de otras variables que pueden influir en ella. También se la llama variable de respuesta. Por ejemplo la sobrevida, respuesta al tratamiento, evolución, etc.

Variable Independiente: Es la que modifica de una u otra manera a la variable dependiente, llamándose también según el caso factor de riesgo, factor predictivo, etc.

Variable Asociada: Se denomina así a aquella variable independiente que no modifica por su sola presencia a la variable dependiente, pero que al combinarse con otra variable, si influye notoriamente a la anterior.

2.3.2. Medidas De Tendencia Central

La medida de tendencia central más común que se utiliza en Estadística es la media aritmética.

Definición.- La media de un conjunto de n mediciones $y_1, y_2, y_3, ..., y_n$ está dada por

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

el símbolo $\bar{y}$ se refiere a la media de una muestra. La media de todas las mediciones de una población se representara por el símbolo μ; más bien μ es una constante desconocida que se desea estimar a partir de la información de una muestra.

La media de un conjunto de mediciones solamente localiza el centro de la distribución de los datos; por si misma, no ofrece una descripción adecuada de un conjunto de mediciones.

Definición._ La mediana corresponde al valor central de la serie de datos observados.

Cuando se esta realizando el proceso de recolección de información se pueden presentar datos aberrantes, estos valores influyen sobre la media aritmética causando por consiguiente que exista una mayor diferencia entre la media de la población y la media aritmética, para evitar que esto ocurra se puede utilizar otra medida de tendencia central que es la mediana.

Para obtener el valor de la mediana se debe arreglar los datos en forma ascendente, el valor de la mediana es el valor que se encuentra en el centro de todas las observaciones. Si existen dos números en el centro se debe calcular el promedio de los dos, y ese será el valor de la mediana. La característica principal de esta medida es que al menos el 50% de las observaciones serán menores o iguales a ella.

Si ***n*** es impar la mediana es: $x\left(\frac{n+1}{2}\right)$

Si ***n*** es par la mediana es: $x\left(\frac{n}{2}\right)+x\left(\frac{n}{2}+1\right)$

Definición._ La moda es el valor que más se repite en una serie de datos.

La medida más común de variabilidad usada en la Estadística es la varianza que es una función de las desviaciones o distancias de las mediciones maestrales con respecto a su media.

Definición._ La varianza de un conjunto de mediciones $y_1, y_2, y_3, ..., y_n$ es la media del cuadrado de las desviaciones de las mediciones con respecto a su media. Simbólicamente la varianza de la muestra esta dada por:

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

La varianza de la población correspondiente se denota por el símbolo σ^2. A mayor varianza de un conjunto de mediciones, corresponde una mayor variación dentro del conjunto. La varianza es útil en la comparación de la variación relativa de dos conjuntos de mediciones, pero solo porta información con respecto a la variación en un solo conjunto cuando se interpreta en términos de la desviación estándar.

Definición._ La desviación estándar de un conjunto de mediciones es la raíz cuadrada positiva de la varianza; es decir,

$$s = \sqrt{s^2}$$

La desviación estándar correspondiente de la población se denota por σ.

Aunque esta relacionada estrechamente con la varianza, la desviación estándar puede utilizarse para dar una idea bastante exacta de la variación de los datos en un solo conjunto de mediciones.

Muchas distribuciones de datos de la vida real tienen la forma de una montaña. Es decir se pueden aproximar por una distribución de frecuencias con forma de una campana que se conoce como la curva normal. Los datos que tienen una distribución acampanada tienen características bien definidas con respecto a la variación, que se pueden expresar en el siguiente enunciado:

Regla Empírica: Para una distribución de mediciones que es aproximadamente normal, el intervalo

$\mu \pm \sigma$ Contiene aproximadamente 68% de las mediciones.

$\mu \pm 2\sigma$ Contiene aproximadamente 95% de las mediciones.

$\mu \pm 3\sigma$ Contiene casi todas las mediciones.

Definición de variable aleatoria.- Sea Ω un espacio muestral cualquiera, una variable aleatoria X es una función definida de Ω a R

$X : \Omega \to R$

Esto es, X es una función que a cada elemento del (*) espacio muestral le asigna un número real y solo uno.

* Espacio Muestral: es el conjunto Ω igual a los resultados posibles de un (**) experimento.

** Experimento: Es un proceso en el que se efectúa algún tipo de medida y que puede ser repetido.

Probabilidad: Número entre 0 y 1 inclusive, que mide la creencia que se tiene de llegue a ocurrir un evento especifico que sea resultado de un experimento.

2.3.3. Rango

Una de las medidas de dispersión que mayormente es utilizada es el rango, este es la diferencia que existe entre el mayor valor y el menor valor del conjunto de datos recolectados. Al rango se lo denota por R y se lo obtiene de la siguiente forma:

$$R = X_L - X_S$$

donde X_L es la observación de más alto valor y X_S es la observación de mas bajo valor.

2.3.4. Medidas de sesgo y kurtosis

A parte de las medidas de tendencia central y de las medidas de dispersión existen otras dos medidas que describen los datos estas medidas se las conoce como el coeficiente del sesgo y el coeficiente de la Kurtosis.

2.3.5. Coeficiente del sesgo

Este coeficiente describe la asimetría que existe en el conjunto de datos con respecto a la media, este coeficiente es calculado por la siguiente ecuación:

$$\gamma_1 = \left[\frac{n\left[\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^3\right]^2}{\left[\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right]^3} \right]^{1/2}$$

Si el coeficiente del sesgo es negativo entonces la mayoría de los datos se encuentran hacia la izquierda del valor de la media. Si el coeficiente del sesgo es positivo la mayoría de los datos se encuentran a la derecha del valor de la media. Cuando el coeficiente del sesgo es cero los datos se encuentran repartidos equitativamente tanto hacia la derecha como a la izquierda. Cuando el coeficiente del sesgo es positivo el valor de la media es mayor que el valor de la mediana, mientras que cuando el

coeficiente es negativo el valor de la mediana es mayor que el valor de la media, y cuando el coeficiente es cero entonces los valores de la media y la mediana son iguales. El coeficiente de sesgo indica el grado a la cual una distribución se desvía de su simetría. Esto es usado por conjuntos de datos unimodales (esto es, tienen una sola moda) y tienen un tamaño de muestra por lo menos de 100. Grandes magnitudes en el coeficiente de sesgo, descartan la noción de una distribución simétrica.

2.3.6. Coeficiente de Curtosis

El coeficiente de Kurtosis es una medida que nos permite observar la picudez de un conjunto de datos. Esta medida esta dada por la ecuación:

$$\gamma_2 = \frac{n\sum_{i=1}^{n}(X_i - \bar{X})^4}{\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2}$$

El coeficiente de curtosis es una medida relativa. Para una distribución normal, el coeficiente de curtosis es 3. Una distribución normal se denomina mesocúrtica, una distribución que es más picuda que la normal se denomina leptocúrtica, y una menos picuda que la normal es platicúrtica. Para una distribución leptocúrtica, el coeficiente de curtosis es más grande que 3. A

mayor pronunciamiento de la picudez de una distribución, más grande es el valor del coeficiente de curtosis. El coeficiente de curtosis se usa solamente para hacer inferencias de un conjunto de datos cuando el tamaño de la muestra es al menos de 100 y la distribución es unimodal, esto es, tiene una sola moda.

2.3.7. Covarianza de dos variables aleatorias

Intuitivamente pensamos en la dependencia de dos variables aleatorias $Y_1 y Y_2$ como el caso en el que una variable, digamos Y_1, crece o decrece cuando Y_2 cambia.

Definición._ La Covarianza de $Y_1 y Y_2$ se define como el valor esperado de $(Y_1 - \mu_1)(Y_2 - \mu_2)$. En la anotación de la esperanza la Covarianza será igual a: $Cov(Y_1, Y_2) = E[(Y_1 - \mu_1)(Y_2 - \mu_2)]$ en donde $y_1 = E(Y_1)$ y $y_2 = E(Y_2)$.

2.4. TECNICAS MULTIVARIADAS

2.4.1. Métodos multivariados aplicados

Los datos de variable múltiples se presentan en todas las ramas de la ciencia. Casi todos los datos reunidos actualmente por los investigadores se pueden clasificar como de variables múltiples. Por ejemplo un investigador de mercados podría querer identificar

las características de los individuos que le permitirían determinar si es probable que una persona compre un producto específico.

La medición y evaluación de unidades experimentales es una actividad principal de la mayoría de los investigadores, como ejemplo de unidades experimentales citamos; las personas, los insectos, los animales, los campos, los terrenos, las compañías, los árboles, los granos de trigo, los países. Se obtienen datos de variables múltiples siempre que un investigador mide o evalúa más de un atributo o característica de cada unidad experimental.

2.4.2. Un panorama general de los métodos multivariados.

Los métodos multivariados son extraordinariamente útiles para ayudar a los investigadores a hacer que tengan sentido conjuntos grandes, complicados y complejos de datos que constan de una gran cantidad de variables medidas en números grandes de unidades experimentales. La importancia y la utilidad de los métodos multivariados aumentan al incrementarse el número de variables que se están midiendo y el número de unidades experimentales que se están evaluando.

El objetivo primario de los análisis multivariados es resumir grandes cantidades de datos por medio de pocos parámetros.
El interés de los análisis multivariados es encontrar relaciones entre las variables de respuesta, las unidades experimentales y tanto las variables respuesta como las unidades experimentales.

Muchas técnicas multivariadas tienden a ser de naturaleza exploratoria en lugar de confirmatoria, es decir tienden a motivar hipótesis en lugar de probarlas.

2.4.3. Técnicas dirigidas por las variables y por los individuos

Las técnicas dirigidas por las variables son aquellas que se enfocan en las relaciones que podrían existir entre las variables respuesta que se están midiendo. Algunos ejemplos de este tipo de técnicas se encuentran en los análisis realizados sobre las matrices de correlación, el análisis de componentes principales, el análisis por factores, el análisis de regresión y el análisis de correlación canónica.

Las técnicas dirigidas por los individuos son las que se interesan en las relaciones que podrían existir entre las unidades experimentales o individuos que se están midiendo, o en ambos.

Ejemplos de este tipo de técnicas se encuentra el análisis discriminante, el análisis por agrupación, y el análisis multivariado de la varianza (MANOVA).

Muchos métodos multivariados ayudan a los investigadores a crear nuevas variables que tengan propiedades deseables.

Algunas técnicas multivariadas que crean nuevas variables son el análisis de componentes principales, el análisis por factores, el análisis de correlación canónica, el análisis discriminante canónico y el análisis de variables canónicas.

2.5. LA ORGANIZACIÓN DE LOS DATOS

2.5.1. Matriz de Datos

Se toman p datos de cada uno de los individuos o unidades experimentales en una muestra, es decir, características que son de interés para el investigador, de esto se origina una matriz de datos. Se usará p para representar la cantidad de variables numéricas de respuesta que se están midiendo u n representará el número de unidades experimentales sobre las cuales se están midiendo las variables. Usaremos la notación x_{jk} para indicar el valor particular de la k-ésima variable que es observada en el j-ésimo ítem.

Los x_{jk} se pueden disponer en una matriz, llamada matriz de datos.

Se pueden disponer esos datos en un arreglo rectangular, llamado $\mathbf{X}$ de n filas y p columnas.

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} \cdots & x_{1k} \cdots & x_{1p} \\ x_{21} & x_{22} \cdots & x_{2k} \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{j1} & x_{j2} \cdots & x_{jk} \cdots & x_{jp} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} \cdots & x_{nk} \cdots & x_{np} \end{bmatrix}$$

2.5.2. Estadística Descriptiva

Sea $x_{11}, x_{21}, \ldots, x_{n1}$ son n medidas sobre la primera variable. Entonces el promedio aritmético de esta medida es:

$$\bar{x}_1 = \frac{1}{n} \sum_{j=1}^{n} x_{j1}$$

En general

$$\bar{x}_k = \frac{1}{n} \sum_{j=1}^{n} x_{jk}$$

k =1,2,……p

Una medida de dispersión es provisto por la varianza muestral, definida por n medidas sobre las primeras variables como:

$$s_1^{\,2} = \frac{1}{n}\sum_{j=1}^{n}\left(x_{j1} - \overline{x}_1\right)^2$$

Donde $\overline{x}_1$ es la media muestral de x_{j1}. En general para *p* variables tenemos:

$$s_k^{\,2} = \frac{1}{n}\sum_{j=1}^{n}\left(x_{jk} - \overline{x}_k\right)^2$$

Para $k = 1,2,\ldots p$

La estadística descriptiva de n observaciones o medidas en p variables aleatorias representada en forma de arreglo tiene la siguiente estructura.

Media Muestral

$$\overline{\mathrm{X}} = \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ \overline{x}_p \end{bmatrix}$$

Varianza y Covarianza Muestral

$$\mathrm{R} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$$

Correlación Muestral

$$\mathrm{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

2.6. Componentes Principales

El análisis de componentes principales se encuentra estrechamente relacionado con la estructura de la varianza y Covarianza de un conjunto de variables a través de una combinación lineal de esas variables.

Los objetivos generales de estos son: la reducción de datos, y la interpretación, esto se logra a través de p componentes que son requeridas para reproducir el sistema total de la variabilidad, mucha de esta variabilidad puede ser contenida en un pequeño número k que llamaremos componentes principales. Si existe información en la k componentes como existe en las p variables

originales, y el conjunto original de datos, compuesto por n medidas sobre p variables, es reducido a un conjunto de datos que consiste de n medidas sobre k componentes principales.

Un análisis de componentes principales frecuentemente da a conocer relaciones que previamente no se esperaban, permitiendo realizar interpretaciones que puede tener un resultado no familiar u ordinario , en investigaciones de tamaño grande resulta muy útil trabajar con esta técnica multivariada.

2.6.1. Fundamentos

Las componentes principales son combinaciones lineales de las p variables aleatorias originales $\mathbf{X}_1,\mathbf{X}_2,\ldots\ldots\mathbf{X}_{p.}$ Estas combinaciones lineales geométricamente representan un nuevo sistema coordenado, obtenido por la rotación del sistema original, en el cual los ejes coordenados serán $\mathbf{X}_1,\mathbf{X}_2,\ldots\ldots\mathbf{X}_{p.}$ Los nuevos ejes representan la dirección que maximizan la variabilidad y provee una simple descripción de la estructura de la Covarianza.

Como ya se mencionó las componentes principales dependen solamente de la matriz de Covarianza Σ o de la matriz de

correlación ρ de $\mathbf{X}_1,\mathbf{X}_2,\ldots\ldots\mathbf{X}_p$. No es necesario que las variables originales sean normales multivariadas.

Sea el vector aleatorio $X^{'} = \left[X_1, \quad X_2, \quad \cdots, \quad X_P\right]$ con su matriz de Covarianza Σ con los valores propios $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

Consideremos las combinaciones lineales:

$$
\begin{aligned}
Y_1 &= a'_1 X = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\
Y_2 &= a'_2 X = a_{21}X_1 + a2_2X_2 + \cdots + a_{2p}X_p \\
&\vdots \\
Y_p &= a'_p X = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p
\end{aligned}
$$

De aquí podemos obtener lo siguiente:

$$Var(Y_i) = a_i^{'} \sum a_i \qquad i = 1,2,...., p$$

$$Cov(Y_i, Y_K) = a_i^{'} \sum a_k \qquad k = 1,2,..., p$$

Las componentes principales son variables artificiales, que no están relacionadas entre sí.

De este modo la primera componente principal es la combinación lineal $a_1'X$ de máxima varianza, $Var(Y_1) = a_1' \sum a_1$ sujeto a la restricción de que $Var(Y_1) = a_1' a_1 = 1$

La segunda componente principal es la combinación lineal $a_2'X$ que maximiza $Var(a_2'X)$ sujeto a $a_2' a_2 = 1$ y $Cov(a_1'X, a_2'X) = 0$.

De este modo la i-ésima componente es la combinación lineal $a_i'X$ que maximiza la $Var(a_i'X)$ sujeto a $a_i' a_i = 1$ y $Cov(a_i'X, a_k'X) = 0$ para k < i.

2.6.2. Obtención de las componentes principales

Para la obtención de las componentes principales, consideremos que $\sum$ es la matriz de varianza y covarianza obtenida a partir del vector $X^t = [X_1, X_2, ..., X_p]$, y además que de la matriz $\sum$ obtenemos los pares de valores y vectores propios $(\lambda_1, e_1), (\lambda_2, e_2), ... (\lambda_p, e_p)$, donde $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0.$ de aquí que la i-ésima componente principal esta dada por:

$$Y_t = e_1{}^t X = e_{1i} X_1 + e_{2i} X_2 + ... + e_{pi} X_p$$

$$para \quad i = 1,2,..., p$$

Además

$$Var(Y_i) = e_i{}^t \sum e_i = \lambda_i \qquad i = 1,2,..., p$$

$$Cov(Y_i) = e_i{}^t \sum e_k = 0 \qquad i \neq k$$

Se debe considerar que existen algunos de los λ_i iguales entonces los coeficientes del respectivo vector propio son iguales y por lo tanto la componente principal correspondiente a ese valor propio no es único.

El total de la varianza de la población esta dado por:

$$Total\ de\ la\ \text{var}ianza = \sigma_{11} + \sigma_{22} + ... + \sigma_{pp}$$
$$= \lambda_1 + \lambda_2 + ... + \lambda_p$$

Consecuentemente, la proporción del total de la varianza de explicación determinada por la k-ésima componente principal es:

$$\begin{bmatrix} \text{Proporción del total} \\ \text{de la varianza} \\ \text{explicada por la} \\ k-\text{ésima componente} \end{bmatrix} = \frac{\lambda_k}{\lambda_1 + \lambda_2 + ... + \lambda_p}$$

$$para \quad k = 1,2,..., p$$

2.6.3. Obtención de componentes principales a partir de datos estandarizados.

Cuando trabajamos con variables cualitativas al mismo tiempo que con variables cuantitativas, es recomendable estandarizar las variables originales.

Con lo que obtenemos un conjunto de variables Z, de la siguiente forma:

$$Z_1 = \frac{(X_1 - \mu_1)}{\sqrt{\sigma_{11}}}$$
$$Z_2 = \frac{(X_2 - \mu_2)}{\sqrt{\sigma_{22}}}$$
$$\vdots$$
$$Z_p = \frac{(Xp - \mu_p)}{\sqrt{\sigma_{pp}}}$$

En notación matricial.

$$Z = (V^{1/2})^{-1}(x - \mu)$$

De aquí, la matriz de Covarianza se puede determinar por la siguiente ecuación:

$$Cov(Z) = (V^{1/2})^{-1} \sum (V^{1/2})^{-1} = \rho$$

La matriz diagonal de la desviación estándar $V^{1/2}$ esta establecida por:

$$V^{1/2} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{\sigma_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\sigma_{pp}} \end{bmatrix}$$

Las componentes principales pueden ser obtenidas a partir de los valores propios de la matriz de correlación de la matriz X, que es la matriz de los datos observados.

Continuaremos utilizando la notación anteriormente empleada, así tenemos que la i-ésima componente esta determinada por:

$$Y_i = e_i{}^t Z = e_i{}^t (V^{1/2})^{-1} (X - \mu), \qquad i = 1,2,..., p$$

Además que

$$\sum_{i=1}^{p} Var(Y_i) = \sum_{i=1}^{p} Var(Z_i) = p$$

y

$$\rho_{Y_i, Z_k} = e_{ik}\sqrt{\lambda_i}, \qquad i,k = 1,2,..., p$$

En este caso $(\lambda_1, e_1), (\lambda_2, e_2), ...(\lambda_p, e_p),$ son los pares de valores y vectores propios de la matriz de correlación ρ, con $\lambda_1 \lambda_2, ..., \lambda_p \geq 0.$

Entonces obtenemos:

$$\begin{bmatrix} \text{Pr}oporción \quad de\ la \\ \text{var}ianza \quad de \\ \text{exp}licación \quad para \quad la \\ i-ésima \quad componente \end{bmatrix} = \frac{\lambda_i}{p} \qquad k = 1,2,...., p$$

2.6.4. Determinación del número óptimo de Componentes Principales

Para determinar cual es el número óptimo de componentes principales con las que se debe trabajar existen cuatro métodos:

1. Quizás el método mas utilizado sea el implantado por Káiser en 1960. Este criterio consiste en retener solo aquellas componentes cuyos valores sean mayores que 1. Para

establecer la acuracidad de este método se han realizado estudios, en los cuales se han considerado entre 10 y 40 variables. En estos estudios, el criterio tuvo acuracidad cuando el número de variables era alto. La acuracidad de un grupo de datos o instrumento de medida se refiere al grado de uniformidad de las observaciones alrededor de un valor deseado tal como el promedio, o un valor objetivo deseado.

2. El método gráfico llamado Prueba Scree, el cual fue propuesto por Castell en 1966. En este método la magnitud de los valores son graficados en el orden en el que fueron obtenidos, generalmente los sucesivos valores propios descienden rápidamente , se recomienda trabajar con las componentes principales correspondientes a los valores propios hasta observar el descenso más pronunciado.

3. Este método fue desarrollado por Lawlwv en 1940, consiste en realizar una prueba estadística significativa para el número de factores que se deben de retener, sin embargo, como todas las pruebas estadísticas, se ve influenciado por el tamaño de la muestra, y un tamaño de muestra grande

producirá la retención de un número alto de componentes principales.

4. El último método consiste en retener tantas componentes principales como para contener al menos entre el 80% y el 90% de la varianza total explicada, mediante este método se retienen sólo las variables que son esenciales para las variables especificadas.

2.7. TABLAS DE CONTINGENCIA

Un problema común en el análisis de datos enumerativos se refiere a la independencia de dos métodos de clasificación de eventos observados. Por ejemplo, podríamos clasificar una muestra de individuos de según el sexo y según su opinión con respecto a una cuestión política para probar la hipótesis de que las opiniones con respecto a esta cuestión son independientes del sexo, o podríamos clasificar a los pacientes que padecen cierta enfermedad según el tipo de medicamento y según el porcentaje de recuperación para ver si el porcentaje de recuperación depende del tipo de medicamento. En cada uno de estos ejemplos queremos investigar la *dependencia o contingencia* entre dos criterios de clasificación.

Para probar la hipótesis de independencia de dos variables de clasificación se utiliza el procedimiento de la prueba ji-cuadrada.
Las frecuencias observadas de las variables de clasificación se las conoce como una tabla de contingencia.

A una tabla de contingencia con r renglones y c columnas se le conoce como una tabla r x c (“r x c” se lee r x c). A los totales de renglones y columnas se les denomina frecuencias marginales.

Sea n_{ij} la frecuencia observada en el renglón ***i*** y la columna ***j*** de la tabla de contingencia, y sea p_{ij} la probabilidad de que una observación caiga en esta celda. Si se seleccionan las observaciones independientemente, entonces las frecuencias de las celdas tienen una distribución multinomial, y el estimador de máxima verosimilitud de p_{ij} es simplemente la frecuencia relativa observada para esta celda. Es decir,

$$\hat{p}_{ij} = \frac{n_{ij}}{n}, i = 1,.....,c$$

Así mismo al considerar el renglón ***i*** como una sola celda, la probabilidad para el renglón ***i*** está dada por p_i y por lo tanto

$\hat{p}_i = \frac{r_i}{n}$, (en donde r_i denota el número de observaciones en el renglón ***i)*** es el estimador de máxima verosimilitud de p_i .

Aplicando razonamientos análogos encontramos que el estimador de máxima verosimilitud de la j-ésima probabilidad de la columna $\frac{c_j}{n}$, en donde c_j denota el número de observaciones en la columna ***j***.

Según al hipótesis nula, la estimación de máxima verosimilitud del valor esperado de n_{11} es

$$\hat{E}(n_{11}) = n(\hat{p}_{1\cdot}\ \hat{p}_{A}) = n\left(\frac{r_1}{n}\right)\left(\frac{c_1}{n}\right) = \frac{r_1 \cdot c_1}{n}$$

En otras palabras, vemos que la estimación del valor esperado de la frecuencia de la celda observada n_{ij} para una tabla de contingencia, es igual al producto de sus respectivos totales de renglón y de columna, dividido entre la frecuencia total. Es decir,

$$\hat{E}(n_{ij}) = \frac{r_i c_j}{n}$$

Ahora, podemos utilizar las frecuencias esperadas y observadas de las celdas, para calcular el valor del estadístico de la prueba:

$$\chi^2 = \sum_{ij} \frac{\left[n_{ij} - \hat{E}(n_{ij})\right]^2}{\hat{E}(n_{ij})}$$

Donde la sumatoria se extiende a todas las celdas rc en la tabla de contingencia r x c. Si $\chi^2 > \chi_\alpha^{\ 2}$ con **v = (r-1) (c-1)** grados de libertad, se rechaza la hipótesis nula de independencia en ele nivel de significancia α; de lo contrario, se acepta la hipótesis nula.

Columnas					
Filas	A	B	C	D	**Total**
1	n_{11}	n_{12}	n_{13}	n_{1j}	$\sum n_{1j}$
2	n_{21}	n_{22}	n_{23}	n_{2j}	$\sum n_{2j}$
.					
.					
.	n_{i1}	n_{i2}	n_{i3}	n_{ij}	$\sum n_{ij}$
Total	$\sum n_{i1}$	$\sum r$	$\sum n_{i3}$	$\sum n_{ij}$	$\sum n_{ij}$

2.8. ANÁLISIS DE SUPERVIVENCIA

Se denomina análisis de supervivencia al conjunto de técnicas que permiten estudiar la variable “tiempo hasta que ocurre un evento” y su dependencia de otras posibles variables explicatorias. Por ejemplo, en el estudio de enfermedades crónicas o tratamientos muy agresivos, el tiempo hasta que ocurre la muerte del enfermo (tiempo de supervivencia) y su dependencia de la aplicación de distintos tratamientos, pero en otras enfermedades, el tiempo hasta la curación, o el tiempo hasta la aparición de la enfermedad. En procesos de control de calidad se estudia el tiempo hasta que un cierto producto falla (tiempo de fallo), o el tiempo de espera hasta recibir un servicio (tiempo de espera), etc.

Debido a que la variable tiempo es una variable continua podría ser, en principio, estudiada mediante las técnicas de análisis de la varianza o los modelos de regresión. Hay, sin embargo, dos dificultades importantes para este planteamiento. En primer lugar, en la mayor parte de los estudios citados la variable tiempo no tiene una distribución normal, más bien suele tener una distribución asimétrica y aunque podrían intentarse transformaciones que la normalizaran, existe una segunda dificultad que justifica un planteamiento específico para estas variables, y es que para

observarlas se tiene que prolongar el estudio durante un período de tiempo suficientemente largo, en el cual suelen ocurrir pérdidas, que imposibilitan la observación del evento.

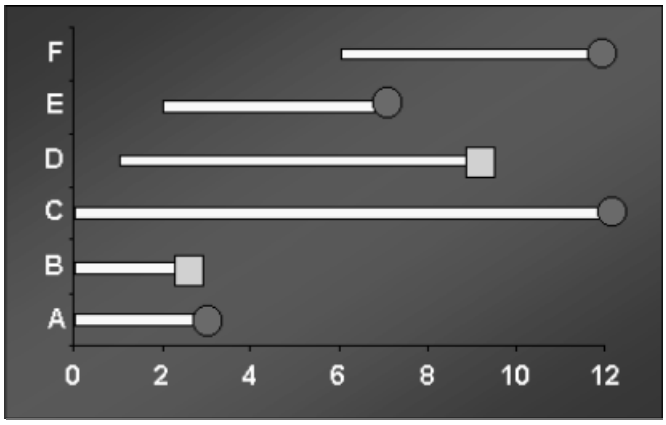
Existen tres motivos por los que pueden aparecer estas pérdidas, en primer lugar por fin del estudio. Supóngase, por ejemplo, que para evaluar una intervención quirúrgica arriesgada se sigue en el tiempo, durante un año, a dos grupos de pacientes.

A los de un grupo se les practicó la intervención y a los de otro no, y se registró la duración del intervalo de tiempo entre la intervención (o la entrada en el estudio, para el grupo no intervenido) y la muerte. Al final del estudio puede haber individuos que no hayan muerto. Otra causa es la pérdida propiamente dicha, por ejemplo se quiere evaluar la eficacia de un tratamiento preventivo para el SIDA, y se sigue durante cinco años a individuos VIH+. Algunos de los individuos, y puede ser un número importante, desaparecerán del estudio en algún momento del mismo por diversos motivos: cambio de domicilio, falta de interés, etc. Una última causa de pérdida es la ocurrencia de un evento competitivo, en los ejemplos anteriores puede ser muerte por alguna otra causa ajena al estudio. Aunque los ejemplos anteriores

son del ámbito de Ciencias de la Salud, estos mismos problemas aparecen en cualquier estudio que necesite un largo tiempo de observación.

Hay que tener en cuenta también que la variable 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 los ejemplos citados, la observación no comienza en el mismo instante para todos los individuos. En algunos textos se denomina pérdida por la izquierda a esta no coincidencia de los tiempos en que comienza la observación, ya que, si el estudio está diseñado para acabar en un tiempo determinado, el efecto de esta no coincidencia es reducir, para los que empiezan más tarde, el tiempo de observación. En el esquema de la figura se detallan todas las posibles pérdidas. Evidentemente, se pueden evitar las pérdidas por la izquierda diseñando el estudio para que acabe, no en un tiempo establecido con carácter general, sino, para cada individuo, en un tiempo determinado después del inicio de la observación.



Esquema temporal de un estudio para observar tiempos de espera para un evento, por ejemplo supervivencia en una intervención quirúrgica. Con el círculo se representan las pérdidas y con el cuadrado las muertes (ocurrencia del evento).

El individuo **A** desaparece del estudio 3 meses después de la intervención (sería una pérdida en sentido estricto). El **B** fallece a los 2,5 meses. El **C** sigue vivo al acabar el estudio (sería una pérdida a los 12 meses por fin del estudio). El **D**, al que se le interviene en el mes 1, fallece en el 9, el tiempo de supervivencia sería 8 meses (hay 1 mes de pérdida por la izquierda). El **E**, al que se le interviene en el mes 2, se pierde en el 7 (sería una pérdida a los 5 meses, ya que hay pérdida en sentido estricto y pérdida por la izquierda). El **F**, al que se le interviene en el mes 6, sigue vivo al

acabar el estudio, sería una pérdida a los 6 meses (existe pérdida por fin del estudio y pérdida por la izquierda).

Si se quisiera aplicar un modelo de regresión lineal a un estudio de este tipo, habría que eliminar del mismo las observaciones perdidas, ya que para ellas no se conoce el valor de la variable; sin embargo sí se tiene alguna información útil sobre la misma: se sabe que es mayor que el tiempo en el que se produjo la pérdida.

2.8.1. Distribución de la variable tiempo de espera

La variable tiempo de espera es una variable aleatoria continua y no negativa, cuya función de probabilidad puede especificarse de varias maneras. La primera es la habitual función densidad de probabilidad f(t), y relacionadas con ella, la función de supervivencia S(t) y la función de riesgo h(t).

La función densidad de probabilidad f(t) para una variable continua se define como una función que permite calcular la probabilidad de que la variable tome valores en un intervalo a través de la fórmula:

$$P(a < T < b) = \int_a^b f(t)\,dt \qquad 0 < t < \infty$$

La función de supervivencia S(t) se define como:

Denotamos por $x$la edad en periodos anuales, de un ente (individuo, empresa), donde x podrá tomar cualquier valor de 0 (cero) al límite superior de supervivencia.

Consideremos un recién nacido (empresa recién creada) y asociemos la variable aleatoria ξ a la edad de f/q del ente considerado. Sea $F(x)$la función de distribución de ξ ,

$$F(x) = P(\xi \leq x), x \geq 0$$

y establezcamos

$S(x) = 1 - F(x) = P(\xi > x), x \geq 0$

donde $F(0) = 0$, lo cual implica que $S(0) = 1$.

La función $S(x)$se denomina “función de supervivencia” ya que para cualquier valor positivo de x la $S(x)$nos da la probabilidad de que un recién nacido (o empresa recién creada) alcance la edad x.

La distribución de ξ se puede definir por la función $F(x)$, o bien por la $S(x)$. Por tanto $P(x < \xi \leq z) = F(z) - F(x) = S(x) - S(z)$.

La probabilidad de que un recién nacido fallezca/quiebre (f/q) entre x y y superviviendo a la edad x, sería:

$$P(x < \xi \leq y / \xi > x) = \frac{F(y) - F(x)}{1 - F(x)} = \frac{S(x) - S(y)}{S(x)}.$$

$$S(t) = P(T \geq t) = \int_{t}^{\infty} f(u)\,du$$

Por lo tanto, la función de supervivencia da la probabilidad complementaria de la habitual función de distribución acumulativa

F(t) = P(T / t), es decir, S(t) = 1 - F(t).

Otro modo de expresar la probabilidad para la variable tiempo de espera es por medio de la función de riesgo h(t) que es la función de densidad de probabilidad de T, condicionada a que T t. Por ejemplo, para la supervivencia a una intervención quirúrgica, la función de riesgo a los 2 años es la de densidad de probabilidad de morir a los 2 años de la intervención, condicionada a que ya se ha sobrevivido hasta entonces. Esta probabilidad sería, realmente, la que en cada momento le importa al enfermo intervenido.

Se puede demostrar que

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

A veces se usa también la función de riesgo acumulada H(t), más difícil de interpretar, que se define como

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

y que verifica

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

Es decir, las cuatro funciones están relacionadas; si se conoce una cualquiera de ellas, se pueden obtener las demás.

A pesar de que el tiempo es una variable continua, un observador sólo tiene acceso a valores discretos de la misma. Los datos observados para cualquiera de las experiencias descritas en la introducción son una serie de valores discretos. Conviene, por lo tanto, definir las funciones anteriores en el caso (práctico) de considerar a la variable tiempo como discreta, es decir, como un conjunto discreto de valores $t_1 < t_2 < \ldots$

El suponerlos ordenados de menor a mayor no representa ninguna pérdida de generalidad, de hecho es así como se observa el tiempo.

Para una variable discreta, la función densidad de probabilidad f (t) se define como:

$$f(t_i) = P(T = t_i), \quad i = 1, 2, \ldots$$

y la función de supervivencia:

$$S(t_i) = \sum_{j \mid t_j \geq t_i} f(t_j)$$

La función de supervivencia da, por lo tanto, para cada valor t_i de T, la probabilidad de que la variable T sea mayor o igual que t_i (en este caso **no** es la complementaria de la función de distribución puesto que la probabilidad de que T sea igual a t_i, que en las variables discretas en general no es cero, está incluida en ambas funciones), aunque otros textos, justamente para que siga siendo la complementaria de la función de distribución la definen sin incluir el igual.

Las funciones de riesgo y riesgo acumulado para una variable

discreta también son:

$$h(t_i) = \frac{f(t_i)}{S(t_i)} \qquad H(t_i) = -Ln(S(t_i))$$

2.8.2. Algunos conceptos mas sobre el análisis de supervivencia

El objetivo del Análisis de Sobrevida es estimar, en función del tiempo, la probabilidad de que ocurra un determinado suceso final.

Debido a que la probabilidad de supervivencia está ligada a un conjunto de aspectos relacionados con los hábitos de vida del paciente, para estimar la probabilidad de reaparición de los síntomas en función del tiempo transcurrido desde el tratamiento, se aplicara en modelo de regresión de Cox.

2.8.3. Método de Kaplan-Meier

Dada una variable cuyos valores corresponden al tiempo que transcurre hasta que ocurre un determinado suceso final, el objetivo del análisis es estimar, en función del tiempo, la probabilidad de que ocurra dicho suceso.

2.8.3.1. Formulación del método de Kaplan- Meier

Sea una muestra de N individuos de los que se conoce si ha ocurrido o no un suceso final en un periodo de observación *t* (para cada individuo el periodo de observación puede ser distinto). El objetivo que persigue el análisis es obtener una función de *t* cuyos valores proporcionen la probabilidad de que el suceso final no ocurra hasta pasado un periodo de tiempo superior o igual a *t* o equivalente, si denominamos <<fallecer =0>> a la ocurrencia del suceso final y <<sobrevivir = 1>> a la no ocurrencia del mismo, se trata de obtener una función cuyos valores proporcionen la probabilidad de sobrevivir, al menos, hasta el instante *t*.

Para cada individuo *i* de la muestra, *i* = 1,.....,N, se dispone de observaciones en dos variables *T* y δ, siendo *T* el tiempo de observación y δ una variable dicotómica cuyos valores indican la ocurrencia o no de el suceso final, donde t_i es el tiempo a que ha sido sometido a observación del i-ésimo individuo.

La probabilidad de que el suceso final ocurra hasta pasado un determinado período de tiempo superior o igual a t_k podría estimarse como, del total de individuos cuyo tiempo de

observación ha sido superior o igual a t_k, la proporción de supervivientes se la obtiene de la siguiente forma:

$$\frac{N.^o\, de \quad individuos\, tales \quad que\left(T \geq t_k, \delta = 1\right)}{N.^o\, de \quad individuos \quad tales \quad que\left(T \geq t_k\right)}$$

Este método de estimación de la supervivencia, puntual en el tiempo o directo, aunque sencillo es poco preciso. El objetivo del método de Kaplan-Meier es obtener, mediante probabilidades condicionadas, una descripción más precisa de la evolución de la supervivencia a lo largo del tiempo.

En términos generales, la probabilidad de sobrevivir un periodo de tiempo t_k, P_k, se estimará como producto de la probabilidad estimada de sobrevivir un periodo de tiempo t_{k-1}, P_{k-1}, por la probabilidad estimada de sobrevivir un periodo de tiempo t_k habiendo sobrevivido un periodo de tiempo $t_{k-1}, P_{k,k-1}$:

$$P_k = p_{k,k-1}P_{k-1} = p_{k,k-1}p_{k-1,k-2}P_{k-2} = \ldots = p_{k,k-1\ldots}$$

El valor de la función de supervivencia en el instante t_k, la probabilidad de sobrevivir hasta, al menos el instante t_k,se

estima como la proporción de supervivencia acumulada hasta t_k:

$$P_k = P_{k-1}\left(1-\frac{d_k}{R_k}\right) \quad \text{y} \quad P_1 = 1-\frac{d_1}{N}$$

donde:

d_k es el número de fallecimientos en el instante t_k.

R_k es el número de individuos expuestos a riesgo en el instante t_k.

2.8.4. Regresión de Cox

Dada una variable cuyos valores corresponden al tiempo que transcurre hasta que ocurre un determinado suceso final y un conjunto de una o más variables independientes cuantitativas o cualitativas, 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.

2.8.4.1. Formulación de la regresión de Cox

En la regresión de Cox , se supone que existe un conjunto de variables independientes, $X_1,\ldots,X_p,$ cuyos valores influyen en el tiempo que transcurre hasta que ocurre el suceso final. Si se

define la función de riesgo, $h(t)$, como el limite, cuando $\Delta t \to 0$, de la probabilidad de que el suceso final ocurra en un pequeño intervalo $(t, t+\Delta t)$, supuesto que no ha ocurrido antes del instante t, el modelo que se postula es:

$$h(t/X) = h_0(t)g(X) = h_0(t)e^z$$

donde:

$h(t/X)$ Es la función de riesgo, considerando la información del conjunto de variables $X = \{X_1, \ldots, X_p\}$.

$h_0(t)$ Es la función de riesgo sin considerar el efecto del conjunto de variables $X = \{X_1, \ldots, X_p\}$.

La función de riego se puede expresar como el producto de una función de t y otra función que únicamente depende de $X_1, \ldots, X_p$, en particular si:

$$g(X) = e^z$$

siendo Z la combinación lineal:

$$Z = \sum_{i=1}^{p} \beta_i X_i = \beta_1 X_1 + \ldots + \beta_p X_p$$

tenemos el modelo de regresión de Cox.

El análisis consistirá entonces en estimar los parámetros desconocidos $\beta_1,\ldots,\beta_p$. Observemos que, si las estimaciones de todos los parámetros fueran nulas, significaría que las variables $X_1,\ldots,X_p$, no influyen en el tiempo transcurrido hasta que ocurre el suceso final. En dicho caso, la función $g(X)$ será igual a 1 y, en consecuencia, $h(t/X)$ coincidiría con $h_0(t)$.

La función de sobrevida, $S(t/X)$, probabilidad de que el suceso final no ocurra hasta pasado un periodo de tiempo superior o igual a t, puede obtenerse, mediante una relación matemática, directamente a partir de la función de riesgo:

$$S(t/X)=\exp\left\{-\int_0^t h(s/X)ds\right\}$$

Es por esto que una vez estimados los parámetros del modelo, además de la estimación de la función de riesgo se obtendrá la estimación del valor de la función de supervivencia para cada instante t.

2.8.4.2. Variables cualitativas en la regresión de Cox.

Si, entre las variables independientes, se encuentra alguna variable cualitativa, sus valores serán recodificados, mediante una pequeña manipulación de sus valores, creando variables con valores numéricos que correspondan en algún sentido con su valor original.

En el caso de variables con dos categorías, sus valores sería 0 y 1. Siendo el valor 1 que indica la presencia de la cualidad correspondiente a una de las dos categorías y el 0 la ausencia de dicha cualidad. Cuando una variable presente más de dos categorías, se generarán tantas variables como el número de categorías existentes menos uno. Cada nueva variable tomará el valor de 1 para una determinada categoría y 0 en el resto.

Mediante este esquema de codificación, los coeficientes de las nuevas variables reflejarán el efecto de las categorías representadas respecto al efecto de la categoría de referencia.

2.8.4.3. Selección de las variables.

En la construcción de la función *Z* para el modelo de regresión logística se selecciona un subconjunto de variables independientes que mas información aportaba sobre las

probabilidades de pertenecer a cualquiera de los dos grupos establecidos por los valores de la variable dependiente.

De la misma forma, en la construcción de la función *Z* para el modelo de regresión de Cox, podrá seleccionarse aquel subconjunto de las variables independientes que mas información aporte sobre la probabilidad de que, para cada posible valor de *t,* el suceso final no ocurra hasta pasado un periodo de tiempo $t + \Delta t$, supuesto que no ha ocurrido antes de *t.*

Tanto el método de Forward como los criterios basados en la Puntuación eficiente de Rao y en el estadístico de Wald para la selección y eliminación de variables, podrán ser utilizados para el método de regresión de Cox.

2.8.4.4. Estadístico de Wald.

El estadístico de Wald, para las variables incluidas en la ecuación de regresión de Cox, juega exactamente el mismo papel que en la regresión logística. Es decir, para cualquier variables independiente X_j seleccionada, si β_j es el parámetro asociado en la ecuación de regresión, el estadístico de Wald permite contrastar la hipótesis nula:

$$H_0 : \beta_j = 0$$

La interpretación de dicha hipótesis es que la información que se perdería al eliminar la variable X_j 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. Será eliminada si dicho máximo es mayor que un determinado valor critico prefijado (si no se indica lo contrario , 0.1).

2.8.4.5. Puntuación eficiente de Rao

Se supone que β_j es el parámetro asociado a la variable X_j , supuesto que entrara en la ecuación de regresión en el siguiente paso; este estadístico nos permite contrastar la siguiente hipótesis nula:

$$Ho: \beta_j = 0$$

Cuya interpretación es que si la variable X_j fuera seleccionada, la

información que aportaría no sería significativa. Si el *p*-valor asociado es menor que α se rechazará la hipótesis nula al nivel de significancia α.

Bajo este punto de vista, en cada una de la selección de las variables, la candidata a ser seleccionada será la que tenga el mínimo *p*-valor asociado al estadístico Puntuación eficiente de Rao. Será eliminada si dicho mínimo es menor que un determinado valor crítico prefijado (si no se indica lo contrario, 0.05).

2.9. Método forward para la selección de variables.

Si el proceso comienza sin ninguna variable seleccionada, entonces:

1. En el primer paso se introduce la variable que presente el mínimo *p*-valor asociado al estadístico Puntuación eficiente de Rao, siempre y cuando se verifique el criterio de selección. En caso contrario, el proceso finalizará sin que ninguna variable sea seleccionada.

2. En este paso se introduce la variable que presente el

mínimo *p*-valor asociado al estadístico Puntuación eficiente de Rao, siempre que verifique el criterio de selección. En caso contrario, el proceso finalizará y la función Z se construirá a partir de la información de la variable independiente introducida en el primer paso.

3. En el siguiente paso se introduce la variable que presente el mínimo *p*-valor asociado al estadístico Puntuación eficiente de Rao, siempre que verifique el criterio de selección. Si, al introducir una variable, el máximo *p*-valor asociado al estadístico de Wald para las variables previamente incluidas verifica el criterio de eliminación, antes de proceder a la selección de una nueva variable, se eliminará la variable correspondiente.

4. Cuando ninguna variable verifique el criterio de eliminación, se vuelve a la etapa 3. La etapa 3 se repite hasta que ninguna variable no seleccionada satisfaga el criterio de selección y ninguna de las seleccionadas satisfaga el de eliminación.

Si el proceso comienza con una o más variables seleccionadas,

en el primer paso se analizará la posibilidad de seleccionar a las que no están.

2.9.1. Estimación de los parámetros

Recordemos que, a partir del modelo de regresión de Cox, dado el conjunto de variables independientes $X = \{ X_1,..,X_p \}$, el límite cuando Δt tiende a cero, de la probabilidad de que el suceso final ocurra en un pequeño intervalo $(t, t + \Delta t)$, supuesto que no haya ocurrido antes del instante t, vendrá dado por:

$h(t/X) = h_0(t)g(X) = h_0(t)\ e^z$.

Siendo la Z la combinación lineal:

$$Z = \beta_1 X_1 + ... + \beta_p X_p.$$

Y, $\beta_1,...,\beta_p$. parámetros desconocidos a estimar.

El criterio para obtener los coeficientes $B_1,...,B_p$, estimaciones de los parámetros desconocidos de $\beta_1,...,\beta_p$ es el de máxima verosimilitud. A partir de $B_1,...,B_p$, la estimación de la función Z sería:

$$Z = B_1 X_1 + ... + B_p X_p.$$

Y, en consecuencia, la estimación de *g(X)* será:

$$g(X) = e^{z} = (e^{B1})^{X1}....(e^{Bp})^{Xp}.$$

Luego para los valores fijos de los restantes términos, cuanto mayor sea el coeficiente B_i mayor será la estimación de *g(X)* o, la de *h(t/X).* Lo que se quiere decir es que mayor será la probabilidad estimada de que el suceso final ocurra en un pequeño intervalo (*t, t* + Δt), supuesto que no haya ocurrido antes del instante t.

2.9.2. Bondad de Ajuste

La bondad de ajuste se aplica en las situaciones cuando se desea analizar cuan probables son los resultados muestrales a partir de un modelo ajustado. La probabilidad de los resultados obtenidos se denomina verosimilitud y para comprobar si ésta difiere de 1, en el cual el modelo se ajusta perfectamente a los datos, se utiliza el estadístico:

*-2LL = -2 * Logaritmo de la verosimilitud*

Cuanto más próximo a cero sea el valor del estadístico *–2LL*, más próxima a 1 será la verosimilitud.

Para todas las variables utilizadas en la función Z tenemos garantía de que, por el criterio de eliminación en el proceso de selección, el *p*-valor asociado al estadístico de Wald es menor que 0.1. En este sentido, para comprobar que el modelo es adecuado, una alternativa es contrastar en una única hipótesis nula, que todos los parámetros correspondiente al conjunto de variables incluidas en el modelo son iguales a cero.

Para contrastar la hipótesis nula se utilizará el estadístico *Ji*-cuadrado global para el modelo y evaluaremos el cambio que se produce en el estadístico *–2LL*.

CAPITULO 3

ANALISIS UNIVARIADO

En este análisis, para las variables consideradas se presentan las correspondientes medidas de tendencia central, dispersión, sesgo y curtosis; a las variables continuas se les realiza además una prueba de bondad de ajuste, utilizando el método de Kolmogorov-Smirnov.

3.1. Información Personal

En esta sección se analizan las características de orden personal de los pacientes de SOLCA en la ciudad de Guayaquil desde al año 1997 al año 2001, que presentaron melanoma maligno.

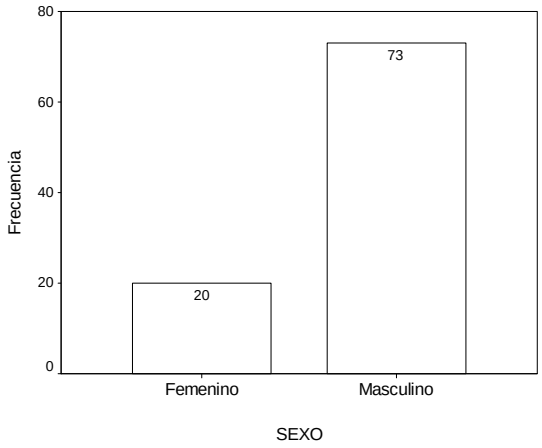
Género

En cuanto al género de los pacientes que presentaron melanoma maligno de piel en SOLCA, de un total de 93 pacientes el 78.50% afecta a hombres y el 21.50% a mujeres; de modo que la razón mujeres / varones es de 1:3.65; esto lo podemos apreciar en la Tabla I y en el Gráfico 1 respectivamente. El género masculino es el que presenta mayor número de casos en cuanto a pacientes con melanoma.

Tabla I
Género de los pacientes

Género	N de pacientes con melanoma	Frecuencia Relativa
Masculino	73	0.7850
Femenino	20	0.2150
Total	***93***	***1.000***

Gráfico 1
Histograma del género de pacientes



Edad

La Tabla II muestra los parámetros correspondientes a la edad de los pacientes de la institución SOLCA de la ciudad de Guayaquil; la edad promedio en años de pacientes que presentaron melanoma es 44.5054 $\pm$ 23.6862 años, mientras que la mediana nos indica que el 50% de los pacientes tiene una edad menor o igual a 50 años; y la dispersión de la variable edad respecto a la media,

medida por la desviación estándar de los datos, es de 23.6862 años.

Existe al menos un paciente que tiene 9 años de edad y alguien con edad de 86 años; la distribución es segada a la izquierda, la edad que más se repite en el grupo de los pacientes es de 66 años. Además se puede apreciar que el 25% de los pacientes tienen edades menores o iguales a 18 años; el 50% tiene una edad menor o igual a 50 y el 75% del conjunto de los pacientes cuentan con edad menor o igual a 66 años.

Tabla II
Parámetro de la Edad de Pacientes

Total		93
Media		44.5054
Mediana		50
Moda		66
Desviación Estándar		23.6862
Varianza		561.0353
Sesgo		-0.166
Curtosis		-1.493
Mínimo		9
Máximo		86
Cuartiles:	25	18
	50	50
	75	66

Ahora nos referimos a la misma variable edad de los pacientes con melanoma maligno, pero esta vez particionada por intervalos que

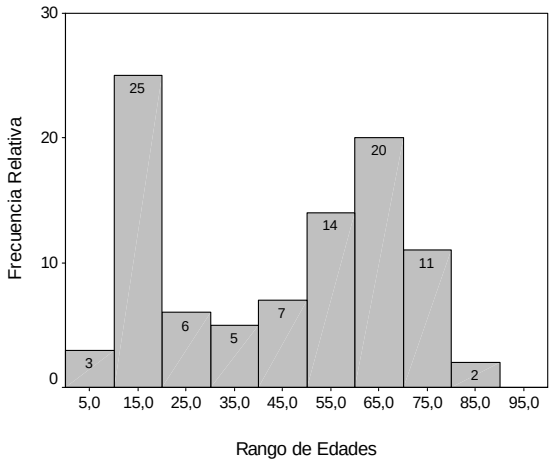
van desde 0 años hasta los 90. Como es posibles observar en la Tabla III, treinta y dos de cada mil pacientes tienen edades entre 0 y 10 años, trescientos uno de cada mil entre 10 y 20 años y así sucesivamente. Es de notar que los intervalos donde se distribuye la mayor cantidad de pacientes son los siguientes: de 10 a 20 con el 26.9% del total; de 50 a 60 con el 15.1%; de 60 a 70 con el 21.5%; de 70 a 80 años con el 11.8%. El detalle de esta información se condensa en la Tabla III.

Tabla III
Distribución De Frecuencias De La Edad De Los Pacientes Con Melanoma

Edad	Nª de Pacientes	Frecuencia Relativa	Frecuencia Relativa Acumulada
[0 a 10)	3	0,032	0,032
[10 a 20)	25	0,269	0,301
[20 a 30)	6	0,065	0,366
[30 a 40)	5	0,054	0,419
[40 a 50)	7	0,075	0,495
[50 a 60)	14	0,151	0,645
[60 a 70)	20	0,215	0,860
[70 a 80)	11	0,118	0,979
[80 a 90)	2	0,022	1,000
Total	*93*	*1,000*	

El Gráfico 2 muestra un histograma de la variable edad correspondiente a los pacientes con melanoma de la institución SOLCA.

Gráfico 2
Distribución de la Edad de Pacientes con Melanoma



El Cuadro 1, nos da información respecto a una propuesta de bondad de ajuste para esta misma variable y como puede observarse la hipótesis nula debe ser rechazada, pues con tres decimales de precisión el valor p de la prueba es de 0.016, es decir como el valor-p es menor a 0.10 por lo tanto se rechaza la hipótesis nula.

Cuadro 1

Bondad de Ajuste (K-S): Edad de los Pacientes

$\mathbf{H}_0$: La Edad de los pacientes tiene una distribución que es N ($\overline{X}, \sigma^2$) N (44.5054, 561.0353) **Vs.** $\mathbf{H}_1$: No es verdad $\mathbf{H}_0$ $Sup_x \left\| \hat{F}(x) - F_0(x) \right\| = 0.161$ Valor p =0.016

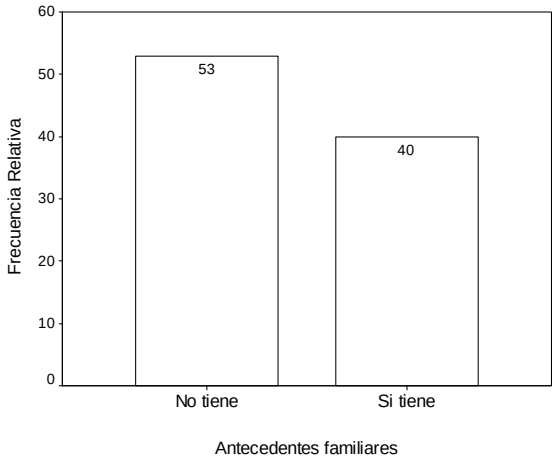
Antecedentes

El riesgo de contraer melanoma de piel de familiares directos afectados por esta enfermedad según nos muestra la Tabla IV es del 43%; mientras que el 57% de los pacientes no presentan antecedentes familiares.

Tabla IV
Antecedentes familiares

Antecedentes	N pacientes con antecedentes familiares	Frecuencia Relativa
Si	40	0.430
No	53	0.570
Total	***93***	***1.000***

Gráfico 3
Antecedentes familiares



3.2. Datos Clínicos

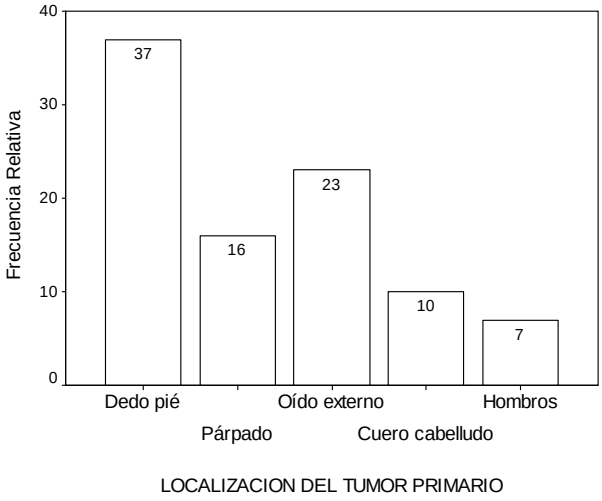
Localización del tumor primario

Los pacientes que fueron diagnosticados con melanoma de piel presentan la lesión de la enfermedad en las siguientes partes del cuerpo, como se detalla en la Tabla V y Gráfico 4.

Tabla V
Localización del tumor primario

Localización	N de pacientes	Frecuencia Relativa
Dedo Pié = 0	**37**	**0.3978**
Párpado = 1	**16**	**0.1720**
Oído externo = 2	**23**	**0.2475**
Cuero Cabelludo = 4	**10**	**0.1075**
Hombros = 6	**7**	**0.0752**
Total	***93***	***1.000***

Gráfico 4
Localización del tumor primario

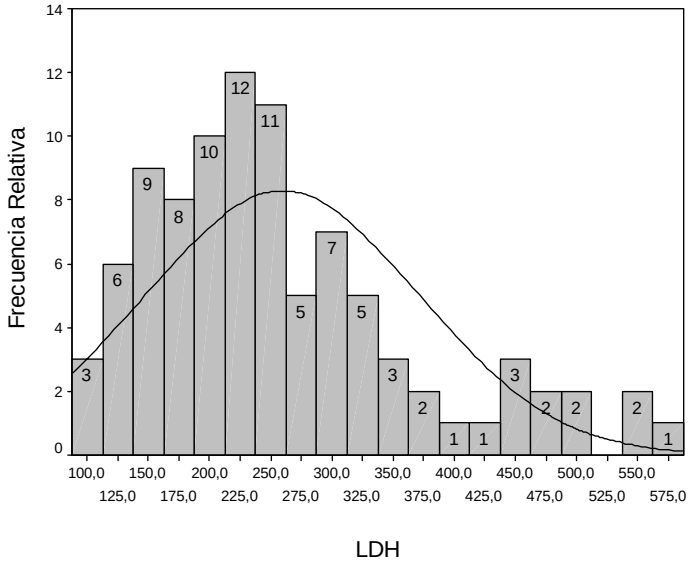


Lactato deshidrogenasa (LDH)

Tabla VI
LDH de pacientes con Melanoma

Total		93
Media		265.5269
Mediana		240
Moda		186
Desviación Estándar		110.9457
Varianza		12308.95
Sesgo		1.012
Curtosis		0.529
Mínimo		109
Máximo		579
Cuartiles:	25	184
	50	240
	75	319

Gráfico 5
Distribución del LDH de Pacientes con Melanoma



El Cuadro 2, nos da información respecto a una propuesta de bondad de ajuste para esta misma variable y como puede observarse la hipótesis nula debe ser rechazada, pues con tres decimales de precisión el valor p de la prueba es de 0.106, es decir como el valor-p no es menor a 0.10 podemos decir que se rechaza la hipótesis nula de que LDH tiene una distribución normal.

Cuadro 2
Bondad de Ajuste (K-S): LDH de los Pacientes

$\mathbf{H}_0$: El LDH en de los pacientes tiene una distribución que es N ($\overline{\mathbf{X}}, \sigma^2$)
N (265.5269, 12308.95)
Vs.
$\mathbf{H}_1$: No es verdad $\mathbf{H}_0$
$\underset{x}{Sup}\left\|\hat{F}(x) - F_0(x)\right\| = 0.126$
Valor p =0.106

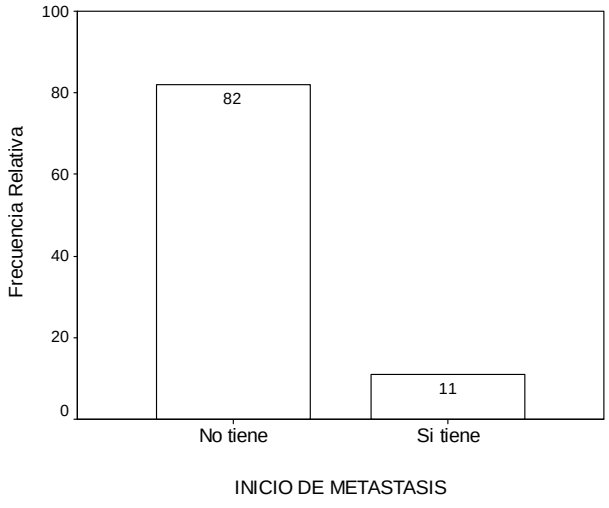
Inicio de Metástasis

De la tabla que se muestra a continuación observamos que el 88.17% de los pacientes afectados con melanoma no presentaron metástasis, mientras que el 11.83% si fueron diagnosticados con la metástasis.

Tabla VII
Inicio de Metástasis

Inicio Metástasis	N de pacientes	Frecuencia Relativa
No Metástasis = 0	82	0.8817
Sí Metástasis = 1	11	0.1183
Total	***93***	***1.000***

Gráfico 6
Inicio de Metástasis



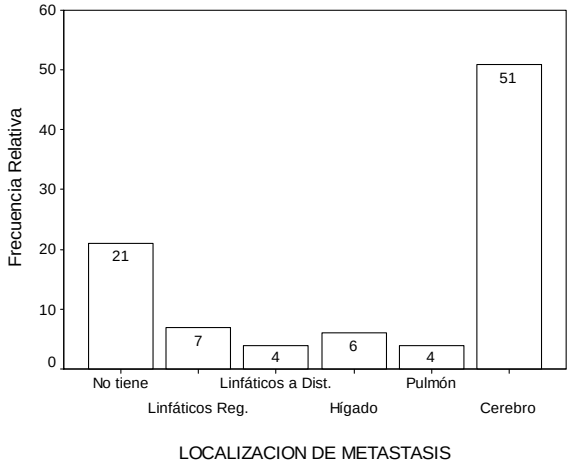
Localización de Metástasis

En la tabla siguiente se detalla el sitio en el cual se presenta la metástasis, por ejemplo el 22.58% de los pacientes no lo presenta, el 7.52% lo presenta en los linfáticos regionales, el 4.30% en los linfáticos a distancia, el 6.45% en hígado, el 4.30% en el pulmón, pleura o ambos y el 54.85% en el cerebro.

Tabla XVIII
Localización de Metástasis

Localización de Metástasis	**N pacientes**	**Frecuencia Relativa**
NO METASTASIS =0	21	0.2258
LINFATICOS REGIONALES = 1	7	0.0752
LINFATICOS A DISTANCIA = 2	4	0.0430
HIGADO = 4	6	0.0645
PULMON = 5	4	0.0430
CEREBRO = 6	51	0.5485
Total	***93***	***1.000***

Gráfico 7
Localización de Metástasis



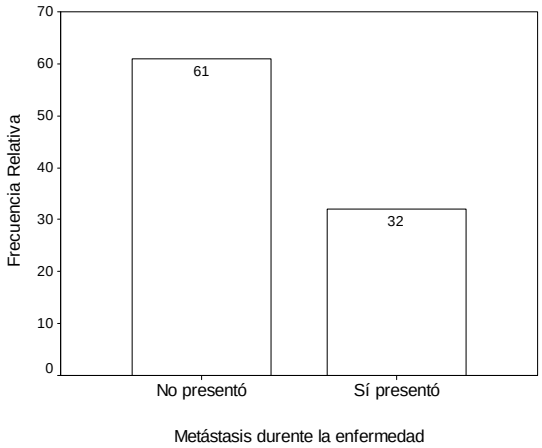
Metástasis Durante la Enfermedad

En la tabla siguiente se muestra que el 65.60% de los pacientes no presenta metástasis es decir aún no se le disemina el cáncer a otras partes del cuerpo, mientras que el 34.40% restante si presenta diseminación de las células cancerosas a otras regiones.

Tabla IX
Metástasis durante la enfermedad

Metástasis durante la enfermedad	N de pacientes	Frecuencia Relativa
NO METASTASIS = 0	61	0.6560
SI METASTASIS = 1	32	0.3440
Total	***93***	***1.000***

Gráfico 8
Metástasis durante la enfermedad



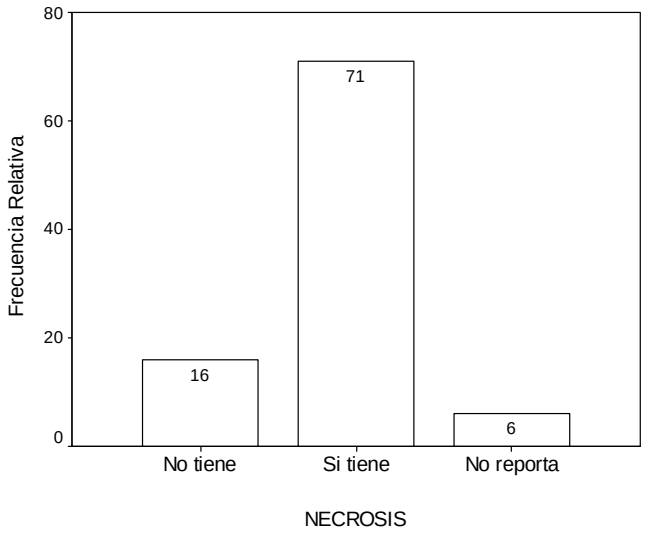
NECROSIS

El 17.20% de los pacientes con melanoma no presentan necrosis (o gangrena), el 76.34% si presenta necrosis en algún lugar de su cuerpo, en tanto que el restante 6.46% no lo reporta.

Tabla X
Necrosis en pacientes con Melanoma

Necrosis	N de pacientes	Frecuencia Relativa
No necrosis = 0	16	0.1720
Sí necrosis = 1	71	0.7634
No reporta = 2	6	0.0646
Total	***93***	***1.000***

Gráfico 9
Necrosis en pacientes con Melanoma



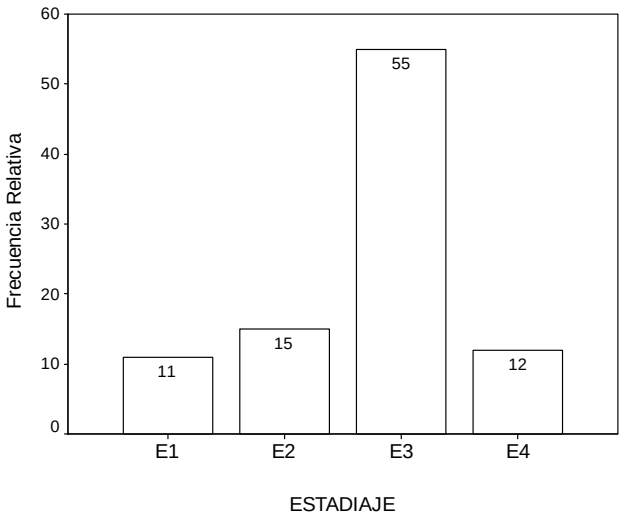
Estadiaje o Estadíos

El 11.82% de los 93 pacientes con melanoma se encuentran en la fase del estadío I, el 16.12% en el estadío II, el 59.13% en el estadío III y el 12.93% en el estadío IV.

Tabla XI
Estadiaje

Estadiaje	N de pacientes	Frecuencia Relativa
I	11	0.1182
II	15	0.1612
III	55	0.5913
IV	12	0.1293
Total	***93***	***1.000***

Gráfico 10
Estadiaje o Estadíos



3.3. Tratamientos

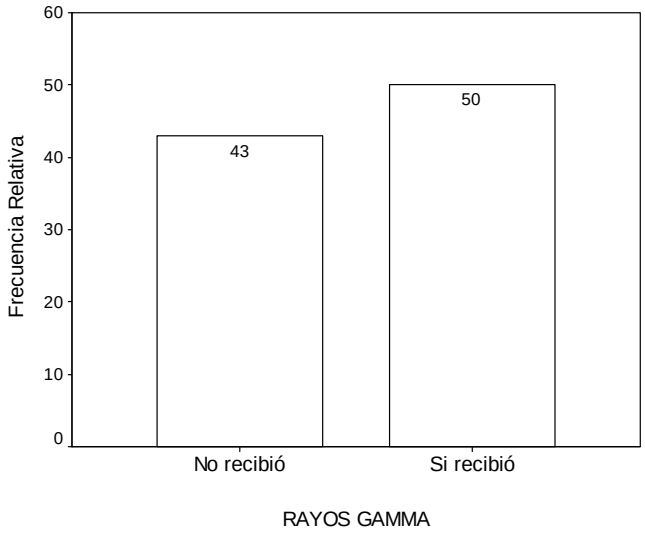
Gamma

La tabla siguiente nos muestra que el 46.24% de los pacientes con melanoma no recibieron rayos gamma, mientras que el 53.76% si lo tomaron.

Tabla XII
Pacientes que recibieron rayos Gamma

Rayos Gamma	N de pacientes	Frecuencia Relativa
No rayos gamma = 0	43	0.4624
Sí rayos gamma = 1	50	0.5376
Total	***93***	***1.000***

Gráfico 11
Pacientes que recibieron rayos Gamma



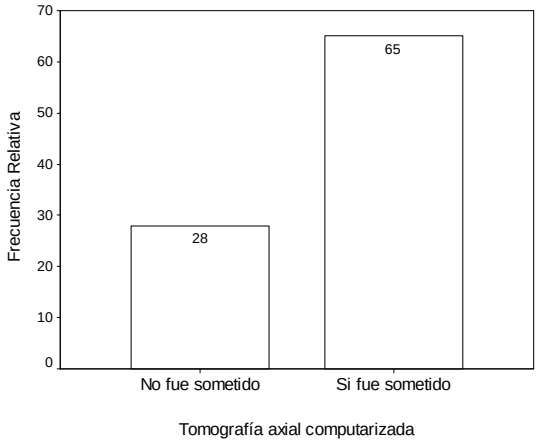
TOMOGRAFIA AXIAL COMPUTARIZADA (T.A.C)

Para determinar si el cáncer se diseminó a los ganglios linfáticos regionales o a lugares distantes se lo hace a través del examen médico, exploración con tomografía computarizada (T.A.C.), del mismo que presentamos la Tabla siguiente que muestra a 28 pacientes que no se sometieron a este examen, es decir el 30.10% del total de los pacientes; mientras que 65 pacientes si fueron sometidos a este examen esto es el 69.90% de ellos.

Tabla XIII
T.A.C en pacientes con Melanoma

T.A.C	N de pacientes	Frecuencia Relativa
0	28	0.3010
1	65	0.6990
Total	***93***	***1.000***

Gráfico 12
T.A.C en pacientes con Melanoma



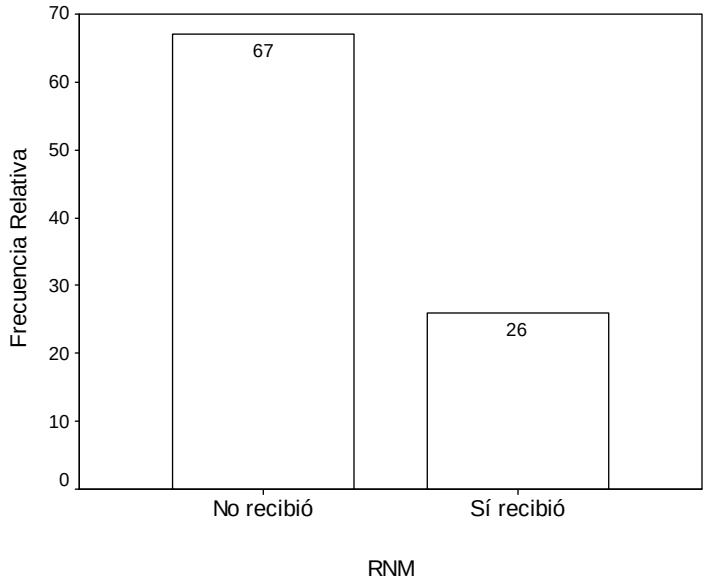
RESONANCIA MAGNETICA NUCLEAR (R.N.M.)

De un total de 93 pacientes con melanoma el 72.04% fueron sometidos a la resonancia magnética nuclear, en tanto que el 27.96% restante no fue sometido a dicho tratamiento.

Tabla XIV
R.N.M en pacientes con melanoma

R.N.M	N de pacientes	Frecuencia Relativa
No R.M.N.= 0	67	0.7204
Sí R.M.N = 1	26	0.2796
Total	***93***	***1.000***

Gráfico 13
R.N.M en pacientes con melanoma



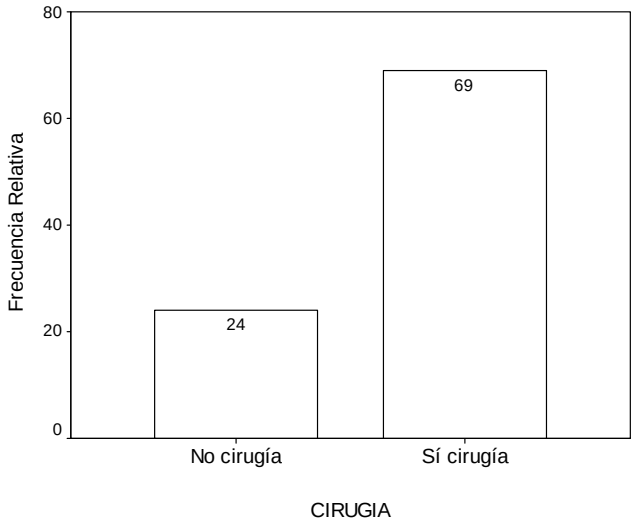
CIRUGIA

El 25.80% de los pacientes con melanoma no se sometieron a cirugía, mientras que el restante 74.20% si fue sometido a alguna intervención quirúrgica.

Tabla XV
Pacientes con melanoma que fueron sometidos a cirugía

Cirugía	N de pacientes	Frecuencia Relativa
No cirugía = 0	24	0.2580
Sí cirugía = 1	69	0.7420
Total	***93***	***1.000***

Gráfico 14
Pacientes con melanoma que fueron sometidos a cirugía



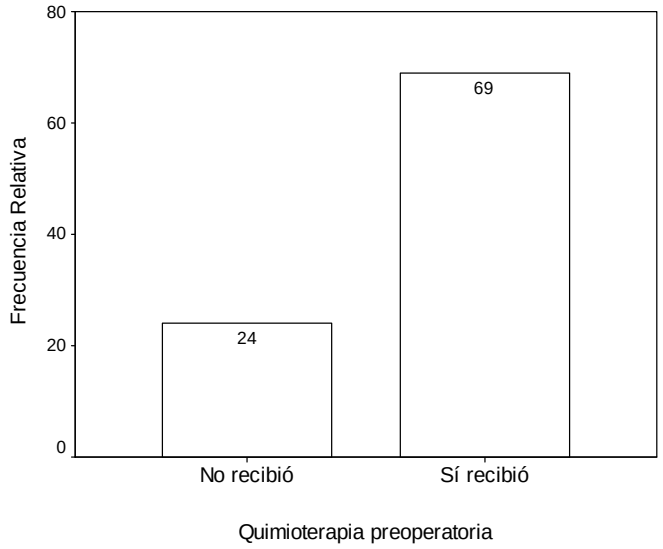
Quimioterapia Preoperatoria

El 25.80% de los pacientes no fueron sometidos a la quimioterapia preoperatoria, mientras que el 74.2% si fue sometido a este tratamiento.

Tabla XVI
Pacientes con melanoma que fueron sometidos a Quimioterapia Preoperatorio

Quimioterapia Preoperatoria	N de pacientes	Frecuencia Relativa
NO QUI PRE = 0	24	0.2580
SI QUI PRE = 1	69	0.742
Total	***93***	***1.000***

Gráfico 15
Pacientes con melanoma que fueron sometidos a Quimioterapia Preoperatorio



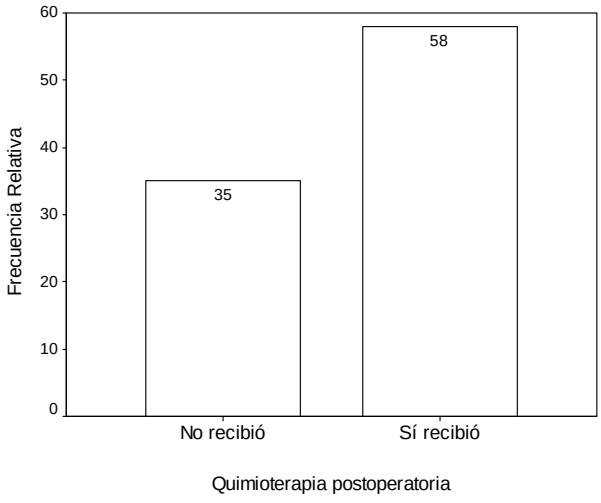
Quimioterapia Postoperatoria

El 37.63% de los pacientes no fueron sometidos a la quimioterapia postoperatoria, mientras que el 62.37% si fue sometido a este tratamiento.

Tabla XVII
Pacientes con melanoma que fueron sometidos a Quimioterapia Postoperatorio

Quimioterapia Postoperatoria	N pacientes	Frecuencia Relativa
NO QUI POS = 0	35	0.3763
SI QUI POS = 1	58	0.6237
Total	***93***	***1.000***

Gráfico 16
Pacientes con melanoma que fueron sometidos a Quimioterapia Postoperatorio



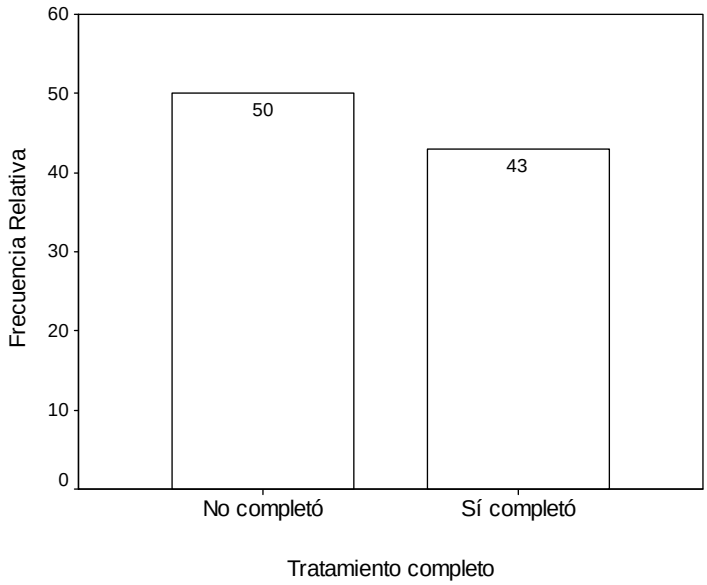
Tratamiento Completo
(Quimioterapia Preoperatoria + Quimioterapia Postoperatoria +Cirugía)

El 53.76% de los pacientes de SOLCA no recibió el tratamiento completo, mientras tanto el 46.24 si lo recibió.

Tabla XVIII
Pacientes con Melanoma que recibieron tratamiento completo

Tratamiento Completo	N de pacientes	Frecuencia Relativa
NO TRA COM = 0	50	0.5376
SI TRA COM = 1	43	0.4624
Total	***93***	***1.000***

Gráfico 17
Pacientes con Melanoma que recibieron tratamiento completo



3.4. TABLAS DE CONTINGENCIA

En cada uno de los casos que se analizarán a continuación queremos investigar la dependencia o contingencia entre dos criterios de clasificación. Además, el objetivo de este análisis consiste en determinar las variables que serán seleccionadas para el estudio de la Regresión de Cox, ya que se tomarán las variables que tienen un valor p significativo, es decir un valor p menor a 0,05.

H_0: La Edad es Independiente del Estadio

Vs.

H_1: La Edad si depende del estadío

Tabla XIX
Tabla De Contingencia De Las Variables Edad Vs. Estadio

EDAD (años)	ESTADÍO				TOTAL
	E1	E2	E3	E4	
≤ 30	3	3	24	4	34
> 30	8	13	30	8	59
TOTAL	**11**	**16**	**54**	**12**	**93**

Tabla XX
Prueba Chi-Cuadrado De Las Variables Edad Vs. Estadio

	Valor	gl	Nivel de significancia
Chi-cuadrado de Pearson	4,098	3	0,251
Números de casos válidos	93		

Al realizar este análisis de un total de 93 pacientes con melanoma la Tabla XXVI nos muestra un nivel de significancia de 0.251 con 3 grados de libertad , por lo tanto se acepta la hipótesis nula, es decir la edad es independiente del estadío.

Ho: La Edad es Independiente del estado de la última observación

Vs.

H_1: La Edad si depende del estado de última observación

Tabla XXI
Tabla De Contingencia De Las Variables Edad Vs. Estado De Última Observación

EDAD (años)	ESTADO DE ULTIMA OBSERVACION		TOTAL
	Muertos y Abandonos	Vivos	
≤ 30	24	20	44
> 30	3	46	49
TOTAL	**27**	**66**	**93**

Tabla XXII
Prueba Chi-Cuadrado De Las Variables Edad Vs. Estado De Última Observación

	Valor	gl	Nivel de significancia
Chi-cuadrado de Pearson	21,621	1	0,000
Números de casos válidos	93		

La Tabla XXII presenta un nivel de significancia de 0,000 con 1 grado de libertad, por lo tanto se rechaza la hipótesis nula

Ho: La Edad es Independiente del estado de la última observación, por lo tanto, a edad es un factor dependiente asociado al estado de última observación.

Ho: El Tiempo de Enfermedad Es Independiente del Estado de última observación

Vs.

H_1: El Tiempo de Enfermedad sí depende del Estado de última observación

Tabla XXIII
Tabla De Contingencia De Las Variables
Tiempo De Enfermedad Vs. Estado De Última Observación

Tiempo de enfermedad (años)	ESTADO DE ULTIMA OBSERVACION		TOTAL
	Muertos y Abandonos	Vivos	
[0 a 1)	13	66	79
[1 a 2)	2	10	12
[2 a 3)	1	1	2
TOTAL	**16**	**77**	**93**

Tabla XXIV
Prueba Chi-Cuadrado De Las Variables
Tiempo De Enfermedad Vs. Estado De Última Observación

	Valor	gl	Nivel de significancia
Chi-cuadrado de Pearson	1,544	2	0,462
Números de casos válidos	93		

Al realizar la prueba Chi-cuadrado obtenemos un nivel de significancia de 0.462 con 2 grados de libertad, por lo tanto no se rechaza la hipótesis nula, es decir, el tiempo de la enfermedad es independiente del estado de la última observación.

Ho: El Tiempo de enfermedad es Independiente del estadío

Vs.

H_1: El Tiempo de enfermedad depende del estadío

Tabla XXV
Tabla De Contingencia De Las Variables
Tiempo De Enfermedad Vs. Estadio

Tiempo de enfermedad (Años)	ESTADÍO				TOTAL
	E1	E2	E3	E4	
[0 a 1)	11	14	42	12	79
[1 a 2)		2	10		12
[2 a 3)			2		2
TOTAL	11	16	54	12	93

Tabla XXVI
Prueba Chi-Cuadrado De Las Variables
Tiempo De Enfermedad Vs. Estadio

	Valor	gl	Nivel de significancia
Chi-cuadrado de Pearson	6,686	6	0,351
Números de casos válidos	93		

Esta tabla nos muestra un nivel de significancia de 0.351 con 6 grados de libertad, por lo tanto concluimos que no se rechaza la hipótesis nula, es decir, el tiempo de la enfermedad es independiente del estadío.

CAPITULO 4

ANÁLISIS ESTADÍSTICO MULTIVARIADO

4.1. Introducción

Las Componentes Principales explican la estructura de varianza y Covarianza de un conjunto de p variables observables a través de unas pocas combinaciones lineales de ellas, el objetivo de las componentes principales es reducir el número de variables de trabajo y simplificar la interpretación.

Se consideraron las siguientes variables:

- Variable # 2: Edad
- Variable # 7: L.D.H
- Variable # 15: Estadío
- Variable # 33: Tiempo de Enfermedad

4.2. Análisis de Componentes Principales de las variables observadas

Se tienen p variables observables $\mathbf{X}_1,\mathbf{X}_2,\ldots\ldots\mathbf{X}_p$, en nuestro caso tendremos las siguientes variables:

$\mathbf{X}_1$ = Edad

$\mathbf{X}_2$ = L.D.H

$\mathbf{X}_3$ = Estadío

$\mathbf{X}_4$ = Tiempo de enfermedad

Que lo representamos como un vector

$$X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \in \Re^4$$

En
tonces se definen en principio p variables aleatorias no observables de la siguiente manera:

$$Y_1 = a'_1 X = a_{11}X_1 + a_{12}X_2 + a_{13}X_3 + a_{14}X_4$$
$$Y_2 = a'_2 X = a_{21}X_1 + a_{22}X_2 + a_{23}X_3 + a_{24}X_4$$
$$Y_3 = a'_3 X = a_{31}X_1 + a_{32}X_2 + a_{33}X_3 + a_{34}X_4$$
$$Y_4 = a'_3 X = a_{41}X_1 + a_{42}X_2 + a_{43}X_3 + a_{44}X_4$$

Al realizar el análisis de componentes principales se presenta en la Tabla XXIV la matriz de varianza y Covarianza de las variables antes mencionadas.

TABLA XXVII
Matriz de varianza y Covarianza

	X_1 = Edad	X_2 = L.D.H	X_3 = Estadío	X_4 = Tiempo enfermedad
X_1 = Edad	56,729	-166,052	0,635	562,208
X_2 = L.D.H	-166,052	9829,905	-7,166	-13729,377
X_3 = Estadío	0,635	-7,166	0,468	12,856
X_4 = Tiempo enfermedad	562,208	-12,856	12,856	369583,87

Los valores propios o eigenvalores (λ_i) representan las varianzas de las componentes principales, el porcentaje de variabilidad contenida en las cuatro componentes principales, se muestran en la Tabla XXVIII.

TABLA XXVIII

Eigenvalores

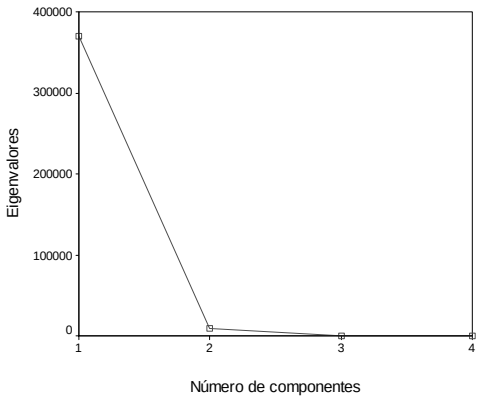
λ_1	Varianza de Y_i	% de la varianza	% Acumulado
λ_1	370107,9	97,533	97,533
λ_2	9308,971	2,543	99,986
λ_3	53,605	0,01413	100,00
λ_4	0,458	0,0001206	100,00

Una manera visual para determinar el número apropiado de componentes principales es el gráfico de sedimentación con los eigenvalores λ_i ordenados de mayor a menor versus i, es decir la magnitud del eigenvalor versus este número.

Para determinar el número apropiado de componentes principales observamos el codo en el gráfico, en este caso, es claro ver que debemos escoger dos componentes principales, porque con éstas logramos explicar el 99,986% de la variación total.

A continuación se presenta el gráfico de sedimentación en el cual se puede observar el número óptimo de componentes principales que se deben retener.

Gráfico 18
Gráfico de sedimentación



A continuación se presentan las componentes principales:

$$Y_1 = 0.056X_1 - 0.113X_2 + 0.011X_3 + 0.992X_4$$

$$Y_2 = 0.060X_1 - 0.050X_2 + 0.997X_3 + 0.011X_4$$

$$Y_3 = 0.991X_1 - 0.108X_2 + 0.059X_3 + 0.055X_4$$

$$Y_4 = -0.107X_1 + 0.986X_2 - 0.049X_3 - 0.111X_4$$

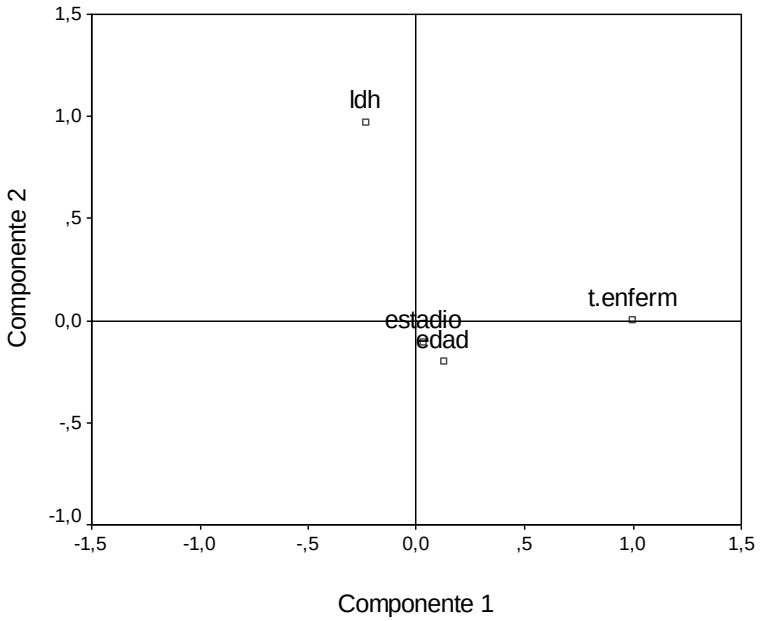
en donde;

$\mathbf{X}_1$ = Edad, $\mathbf{X}_2$ = L.D.H, $\mathbf{X}_3$ = Estadío y $\mathbf{X}_4$ = Tiempo de enfermedad.

En la primera componente principal observamos que la variable

X_4 = Tiempo de enfermedad, posee el peso más significativo, en la segunda componente la más significativa es la variable $\mathbf{X}_3$ = Estadío.

Gráfico 19
COMPONENTE 1 VS COMPONENTE 2



Como se puede observar en el Gráfico 19, la variable X_4 = Tiempo de enfermedad tiene el peso más significativo en el modelo, esto significa que a medida que el paciente tenga muchos años con la enfermedad, la recuperación de éste va ser tardía.

Al efectuar el análisis de Componentes Principales con los datos originales, resulta que con solo dos componentes principales se explica el 99.986% de la varianza total, más, las variables cuyas escalas son en promedio más altas

dominan estas dos componentes siendo estas X_4 = Tiempo de enfermedad, ya que posee el peso más significativo, en la segunda componente la más significativa es la variable $\mathbf{X}_3$ = Estadío por lo tanto deberían considerarse primero a la hora de realizar un análisis.

4.3. Análisis de sobrevida de las variables observadas

4.3.1. Método de Kaplan-Meier

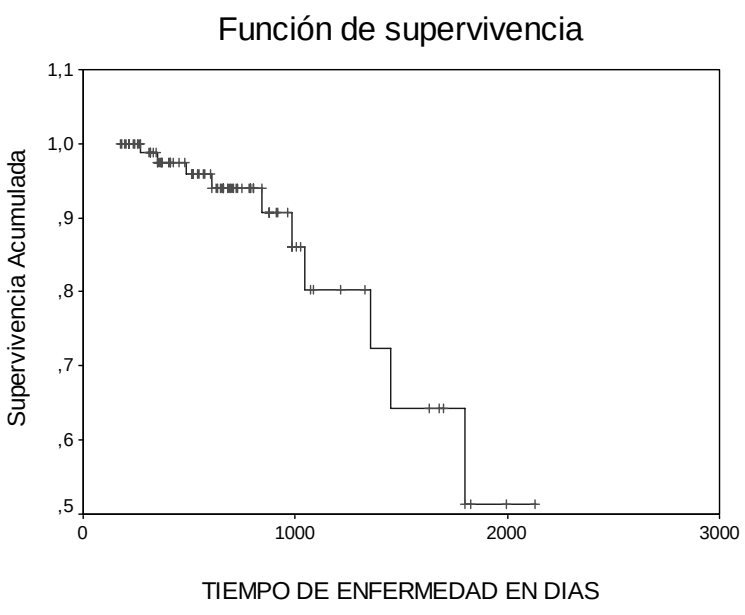
Dada una variable cuyos valores corresponden al tiempo que transcurre hasta que ocurre un determinado suceso final, el objetivo del análisis es estimar, en función del tiempo, la probabilidad de que ocurra dicho suceso.

Tabla XXIX
Kaplan Meier Para El Tiempo De Sobrevida

Tiempo en días	Estado	Sobrevida Acumulada	Error Estándar	Eventos Acumulados
271	1	0.9878	0.0121	1
350	1	0.9748	0.0176	2
485	1	0.9588	0.0235	3
604	1	0.9400	0.0296	4
846	1	0.9065	0.0436	5
987	1	0.8611	0.0606	6
1045	1	0.8037	0.0792	7
1356	1	0.7234	0.1044	8
1149	1	0.6430	0.1198	9
1798	1	0.5144	0.1497	10

La Tabla XXIX, muestra la probabilidad de que ocurra el suceso muerte en pacientes con melanoma, a los 271 días es de 0.9688, la probabilidad de que ocurra la muerte de los pacientes con melanoma a los 350 días es de 0.9748, de igual manera, la probabilidad de que ocurra el evento muerte a los 485 días es de 0.9588. A medida que avanza el tiempo, la probabilidad de sobrevida de los pacientes decrece, para entender mejor, lo podemos apreciar en el gráfico siguiente.

Gráfico 20
Función de supervivencia



4.3.2. ANALISIS DE REGRESION DE COX

El análisis de supervivencia esta constituido por un grupo de técnicas estadísticas que permiten valorar la probabilidad de desarrollar un evento o suceso (resultado de una exposición, evolución de una enfermedad o beneficio de un tratamiento) con el paso del tiempo. Inicialmente se desarrollaron para estimar la probabilidad de sobrevivir y de ahí la denominación de técnicas de análisis de supervivencia.

El análisis de supervivencia incluye una variable dependiente que, es el tiempo que tarda en producirse un suceso.

Recordemos que el objetivo del análisis de Regresión de Cox, dada una variable cuyos valores corresponden al tiempo que transcurre hasta que ocurre un determinado suceso final y un conjunto de una o más variables independientes cuantitativas o cualitativas, 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. Para realizar

este análisis se utilizaron las siguientes variables que fueron recomendadas por el especialista de SOLCA.

Cabe recalcar que no todas las variables que se utilizan para este estudio tienen un valor pronóstico estadísticamente significativo porque el tamaño de la muestra es pequeño; esto no quiere decir que ninguna de las variables sea útil, una de ellas podría ser un factor pronóstico importante. La selección de variables hay que hacerla paso a paso, recalculando los valores-p después de eliminar (o introducir) cada variable y además con la experiencia del experto de SOLCA. Las variables que se utilizarán para este análisis son las siguientes:

X1= Edad

X2= L.D.H.

X3= Sexo

X4= Estadío

X5= Tiempo de Enfermedad

Cabe recalcar, que las variables Edad, L.D.H., y Tiempo de Enfermedad son variables cuantitativas, por lo tanto las utilizaremos en su forma natural, en cambio la variable

Estadío, es una variable cualitativa sus valores numéricos corresponden a alguna categoría, por lo tanto procedemos a recodificar esta variable.

En el caso de variables con dos categorías, sus valores se recodificarán a valores 0 y 1. El valor 1 indicará la presencia de la cualidad correspondiente a una de las dos categorías, y el 0, la ausencia de dicha cualidad. Cuando una variable presente más de dos categorías, se generarán tantas variables como el total de la categoría menos uno. Cada nueva variables tomará valor 1 para una determinada categoría y 0 el resto, de tal forma que los individuos en una misma categoría tomarán valor 1 en una misma variable y 0 en el resto. La categoría no considerada, o categoría referencia, estará representada por el valor 0 en todas las nuevas variables. Mediante este esquema de codificación, los coeficientes de las nuevas variables reflejarán el efecto de las categorías representadas respecto al efecto de la categoría referencia. Por lo tanto la variable Estadío quedará recodificada de la siguiente manera:

	E1	E2	E3
Estadío 1	0	0	0
Estadío 2	1	0	0
Estadío 3	0	1	0
Estadío 4	0	0	1

La tabla XXXIII, muestra un resumen completo del proceso de casos, el mismo que tiene como variable dependiente el Tiempo de Sobrevida (fecha de diagnostico – fecha de última observación).

Los casos disponibles en el análisis para el evento(muerte) fue de 10 es decir el 10.8% de los datos analizados, debido que durante el tiempo de observación de los pacientes se encontraron que fueron 10 los decesos, los datos censurados (es decir los vivos mas los abandonos) fueron 72 que equivale al 77,4% de los datos analizados, además podemos observar que existen casos que fueron excluidos , los casos con valores perdidos fue igual a 0 es decir no existieron datos perdidos, los casos con tiempo no positivo fue igual 0 es decir no existieron datos con tiempo negativo, y los Casos censurados antes del evento más temprano en un estrato fue igual a 11 es decir el 11.8% de los datos analizados, el total de datos analizados fue de 93 pacientes que equivale al 100% de los datos analizados.

Tabla XXX
Resumen Del proceso de casos

		N	Porcentaje
Casos disponibles en el análisis	Evento[a]	10	10,8%
	Censurado	72	77,4%
	Total	82	88,2%
Casos excluidos	Casos con valores perdidos	0	,0%
	Casos con tiempo no positivo	0	,0%
	Casos censurados antes del evento más temprano en un estrato	11	11,8%
	Total	11	11,8%
Total		93	100,0%

a. Variable dependiente: TIEMPO DE ENFERMEDAD

4.3.3. BONDAD DE AJUSTE DEL MODELO

Comprobar la bondad del ajuste es analizar cuán probables son los resultados muestrales a partir del modelo ajustado. La probabilidad de los resultados obtenidos se denomina verosimilitud. Para comprobar si la verosimilitud difiere de 1 (que el modelo se ajusta perfectamente a los datos) se utiliza el estadístico:

-2LL= -2 x Logaritmo de la verosimilitud

Cuanto más próximo a cero sea el valor del estadístico -2LL, más próxima a 1 será la verosimilitud y mejor será el modelo.

El objetivo de esta prueba es contrastar una única hipótesis nula en la que todos los parámetros correspondientes al conjunto de variables incluidas en el modelo son iguales a cero, es decir:

$$\mathrm{H}_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$$
vs.
$$\mathrm{H}_1 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 \neq 0$$

Para contrastar la hipótesis nula de que todos los parámetros asociados son nulos, se utiliza el estadístico Chi-cuadrado global para el modelo. Ahora, si el valor-p, asociado al estadístico para esta prueba Chi-cuadrado es menor que 0.1, entonces se rechaza la H_0 de que los parámetros β sean cero, es decir, éstos no serán nulos, lo cual significa que en la regresión de Cox, estos parámetros tendrán un valor que aún no conocemos, y por lo tanto al estimar la probabilidad de sobrevida de los pacientes lo haremos en base a éstas variables. Cuando todos son iguales a cero, al ser la función Z igual a cero, $g(x)$ será igual a 1, recordemos que Z es igual a $\hat{Z} = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6$ ($h(t/X) = h_0(t)g(X) = h_0(t)e^z$) y en consecuencia $h(t/X)$ coincidirá con $h_0(t)$.

En la Tabla XXI el resultado que se obtiene muestra, primero una valoración global del modelo: El valor -2 Log verosimilitud es una medida del ajuste del modelo a los datos; cuanto menor sea ese número, mejor será el ajuste, es decir, cuanto más próximo a cero sea el valor del estadístico -2LL más próxima a 1 será la verosimilitud y mejor será el modelo, en este caso al aplicar -2 Log de 59.661, se obtiene -3.551, esto quiere decir que el modelo no es perfecto, porque lo ideal es que el valor del estadístico -2LL se acerque a cero para que el modelo sea perfecto, esto no significa que el modelo que se empleó no sirve clínicamente, pues, el tamaño de la muestra es la que hace que el modelo no sea perfecto, además, según la experiencia del experto de SOLCA, estos valores que arrojó el modelo son importantes para posteriores diagnósticos.

El test estadístico adjunto (cambio desde el paso anterior) compara el modelo sin ninguna variable, en la hipótesis nula, con el modelo que incorpora esas tres variables. El resultado no es significativo, p = 0.634, puesto que el tamaño de muestra es pequeño, si lo fuera, diríamos que al menos una de las variables es útil como factor pronóstico de la

supervivencia, sin embargo, con la ayuda del experto de SOLCA, decidimos realizar el análisis.

Tabla XXXI
Prueba Bondad De Ajuste Sobre Los Coeficientes Del Modelo

-2 log de la verosimilitud	Global (puntuación)			Cambio desde el paso anterior			Cambio desde el bloque anterior		
	Chi-cuadrado	gl	Sig.	Chi-cuadrado	gl	Sig.	Chi-cuadrado	gl	Sig.
59,661	4,315	6	,634	4,189	6	,651	4,189	6	,651

En la tabla XXII, podemos observar la Tabla de Supervivencia la misma que nos muestra el tiempo de todos los eventos(muertes) que han ocurrido en el análisis los mismos que son iguales a 10 decesos, los valores estimados de la función de Sobrevida, evaluada sobre las medias de las variables independientes.

Dado que hasta el instante *t* = 271, cuyo valor estimado de la función de Sobrevida es 0.991 el mismo que es próximo a 1, y podemos observar también que a partir del tiempo *t* = 350, cuyo valor estimado de la función de sobrevida es 0.981, hasta llegar al tiempo *t* =1798 en el cual la probabilidad de sobrevida es 0.332.

Tabla XXXII

Tabla de supervivencia

Tiempo	Impacto acum. línea base	En la media de las covariables		
		Supervivencia	ET	Impacto acum.
271,000	,026	,991	,009	,009
350,000	,055	,981	,014	,019
485,000	,094	,968	,020	,032
604,000	,140	,953	,026	,048
846,000	,255	,916	,045	,087
987,000	,437	,861	,068	,150
1045,000	,724	,780	,102	,249
1356,000	1,271	,646	,164	,436
1449,000	1,947	,513	,213	,668
1798,000	3,215	,332	,239	1,104

En la tabla XXIII podemos observar que en la columna B aparecen los estimadores de los coeficientes de las variables, y en la columna ET los correspondientes errores estándar.

Si un coeficiente valiese 0, esa variable no influiría en el modelo, por ello es importante contrastar las hipótesis nulas $B_i = 0$: En la columna Sig se presentan los valores- p de esos contrastes. Por último, la columna Exp(B) muestra el riesgo relativo de aumentar una unidad en la covariable correspondiente. Así por ejemplo, esto indica que el paciente que tiene una edad mayor de 44.5054 años (la media

respectiva de la edad de los pacientes) tiene 0.967 veces más posibilidades de morir (una posibilidad más de morir) que los pacientes menores de 44.5054 años. Los pacientes que presentan la coenzima LDH por encima de 259.5269, tienen aproximadamente 1 vez más de posibilidad de morir que los pacientes con un valor inferior a ésta.

De manera análoga, los pacientes de sexo masculino tienen 1.831 veces posibilidades más de morir que los de sexo femenino con esta enfermedad. Así mismo, pacientes en que la etapa de su enfermedad se encuentra en el estadío 2, tienen 3.126 veces mas posibilidades de morir que los pacientes que se encuentran en el estadío 1; y los pacientes que se hallan estadío 3, tienen 8.831 veces más posibilidades de morir que otros pacientes en los que la etapa de su enfermedad no esta muy avanzada.

Tabla XXXIII

Variables en la ecuación

							95,0% IC para Exp(B)	
	B	ET	Wald	gl	Sig.	Exp(B)	Inferior	Superior
EDAD	-,033	,022	2,258	1	,133	,967	,926	1,010
LDH	-,003	,003	,839	1	,360	,997	,990	1,004
SEXO	,605	,871	,482	1	,487	1,831	,332	10,092
E1	-,829	,943	,773	1	,379	,437	,069	2,770
E2	1,140	1,177	,938	1	,333	3,126	,311	31,393
E3	2,178	1,697	1,648	1	,199	8,831	,317	245,724

Como podremos recordar, para determinar si la información proporcionada por la variable X es redundante, se utiliza el valor p asociado al estadístico de Wald, si una variable es candidata a ser seleccionada en un paso, el criterio de entrada se basaba en el valor p, si este es menor que 0.1 la variable debe ser incluida en el modelo, y si el valor p es mayor que 0.1 la variable no aporta significativamente en el modelo y por lo tanto debe ser excluida del mismo.

Recordemos que, a partir del modelo de regresión de Cox , dado el conjunto de variables independientes $X = \{X_1, ..., X_p\}$, el limite, cuando Δt tiende a cero, de la probabilidad de que el suceso final ocurra en un pequeño intervalo $(t, t + \Delta t)$, supuesto que no ha ocurrido antes del instante t, vendrá dado por:

$$h(t / X) = h_0(t) g(X) = h_0(t) e^z$$

$$\hat{Z} = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6$$

En donde

$X_1 = EDAD$

$X_2 = LDH$

$X_3 = SEXO$

$$X_4 = E_1$$
$$X_5 = E_2$$
$$X_6 = E_3$$

Y en consecuencia, la estimación de $\hat{g}(x)$ será:

$$\hat{g}(x) = e^{\hat{z}}$$

$$\hat{g}(x) = \left[0.967^{(EDAD)} 0.997^{(L.D.H)} 1.831^{(Sexo)} 0.437^{(E1)} 3.126^{(E2)} 8.831^{(E3)}\right]$$

Esto es para cualquier valor que puedan tomar las variables que se encuentran dentro del modelo.

Por ejemplo para un paciente cuya edad es 65 años, de sexo masculino, su nivel L.D.H. es de 364, y con un Estadio 3, tendremos el siguiente resultado:

$$X_1 = EDAD = 65 años$$
$$X_2 = LDH = 364$$
$$X_3 = SEXO = 1$$
$$X_4 = E_1 = 0$$
$$X_5 = E_2 = 0$$
$$X_6 = E_3 = 1$$

$$\hat{g}(x) = \left[0.967^{(65)} 0.997^{(364)} 1.831^{(1)} 0.437^{(0)} 3.126^{(0)} 8.831^{(1)}\right]$$

$\hat{g}(X)$ = 0.6115

$$h(t/x) = h(t_0)\mathbf{0}.6115$$

El paciente con estos factores tiene una probabilidad del 61.15% de morirse.

Cabe recalcar que la función $h(t_0)$ clínicamente no tiene ningún significado.

Ahora la probabilidad de fallecimiento que tiene otro paciente con 65 años de edad, de sexo masculino y LDH de 364 pero que su enfermedad se encuentra en el estadio 2 es el siguiente:

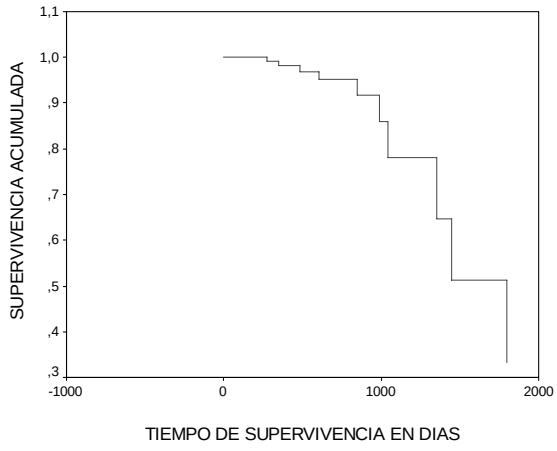
$$\hat{g}(x) = \left[0.967^{(65)} 0.997^{(364)} 1.831^{(1)} 0.437^{(0)} 3.126^{(1)} 8.831^{(0)}\right]$$

$\hat{g}(X)$ = 0.21648

El paciente con estos factores tiene una probabilidad de 21.64% de morirse.

La figura siguiente nos muestra la representación gráfica de los valores de la función de Sobrevida frente al tiempo. La probabilidad de sobrevida de los 93 pacientes con melanoma es 0.991 a los 271 días, es decir a los 9 meses aproximadamente, luego esta probabilidad disminuye poco a poco, hasta que a los 32.9 meses (2.74 dos años y medio) la probabilidad de sobrevida ha disminuido al 0.861, luego a los 4.025 años (1449 días) esta probabilidad es de 0.513, es decir, esta probabilidad se redujo a la mitad, y por último observamos que a los 5 años (1798 días) la probabilidad de sobrevida es de 0.332.

Gráfico 21
Función de supervivencia



CONCLUSIONES

1. La población objetivo investigada fue de 93 pacientes de SOLCA de la ciudad de Guayaquil, datos recabados desde 1997 al año 2001. En cuanto al género de los pacientes que presentaron melanoma maligno de piel, de un total de 93 pacientes el 78.50% afecta a hombres y el 21.50% a mujeres, es decir más de la mitad de los pacientes con esta enfermedad son varones; de modo que la razón mujeres/varones es de 1:3.65.

2. La edad promedio en años de los pacientes que presentaron melanoma es 44.5054 años, mientras que la mediana nos indica que el 50% de los pacientes tiene una edad menor o igual a 50 años; y la dispersión de la variable edad respecto a la media, medida por la desviación estándar de los datos, es de 23.6862 años. Existe al menos un paciente que tiene 9 años de edad y alguien con edad de 86 años; la distribución es segada a la izquierda, la edad que más se repite en el grupo de los pacientes es de 66 años. Además se puede apreciar que el 25% de los pacientes tienen edades menores o

iguales a 18 años; el 50% tiene una edad menor o igual a 50 y el 75% del conjunto de los pacientes cuentan con edad menor o igual a 66 años.

3. De los 93 pacientes con melanoma el 43% presentó antecedentes familiares, es decir, por lo menos un familiar presentó algún tipo de cáncer, y el 57% restante no lo presentó.

4. Con respecto a la localización del tumor el 39.78% de los pacientes presentaron el tumor en el dedo del pié, el 24.75% de ellos en el oído externo, el 17.20% en el párpado, el 10.75% en el cuero cabelludo y por último el 7.52% en los hombros.

5. El 88.17% del total de los pacientes con melanoma no presentaron metástasis mientras que el 11.83% mostraron diseminación del cáncer en alguna parte del cuerpo.

6. El mínimo valor de la enzima Lactato deshidrogenasa L.D.H. que se encuentra en el corazón, hígado, músculos, eritrocitos, plaquetas y nódulos linfáticos es de 102 y el máximo valor es de 579. El valor promedio de esta sustancia es de 259.5269, mientras que el 50% de los pacientes tiene un L.D.H. menor o igual a 235. Además se puede observar que el 25% de los pacientes presentan un L.D.H. menor o igual a 184, y el 75% de los pacientes tienen un L.D.H. menor o igual a 307.50.

7. En cuanto al tratamiento que recibieron los pacientes con melanoma en SOLCA, el 53.76% recibieron rayos gamma, es decir, más de la mitad de los pacientes son sometidos a estos rayos, el 46.24% no recibieron este tratamiento, el 27.96% de los pacientes fueron sometidos a la resonancia magnética nuclear, el 25.80% de ellos fueron intervenidos quirúrgicamente y el 46.24% recibió el tratamiento completo, es decir, quimioterapia preoperatoria, cirugía y quimioterapia postoperatoria.

8. Setenta y uno de los noventa y tres pacientes con melanoma presentaron necrosis, es decir el 76.34% del total de los pacientes.

9. Con respecto al Estadiaje, el 11.82% de los pacientes se encuentran en la fase del estadío 1, es decir la enfermedad se encuentra en la capa externa de la piel pero aún no se ha diseminado a los ganglios linfáticos, el 16.12% en el estadío 2, aquí el melanoma ya se diseminó a las capas internas de la piel, el 59.13% se encuentra en el estadío 3, en esta etapa el melanoma invade el tejido subcutáneo y el 12.93% en el estadío 4, es decir, la enfermedad ya se ha diseminado a los ganglios linfáticos, los pacientes se encuentran en la última fase de este mal.

10. Con respecto a metástasis durante la enfermedad el 65.60% de los pacientes con melanoma no reportaron metástasis mientras que el 34.40% si la presentaron.

11. Más de la mitad de los pacientes presentó metástasis en el cerebro, es decir, el 54.85% de ellos, el 22.58% no la presentaron.

12. Al realizar las tablas de contingencia observamos que las variables edad y estadío son independientes, es decir, la una no depende de la otra. La variable edad es un factor independiente asociado al estado de última observación. Además, la variable tiempo de enfermedad es un factor dependiente asociado al estado de última observación. Podemos afirmar que el tiempo de enfermedad es un factor independiente asociado al estadío.

13. Efectuando el análisis de Componentes Principales con los datos originales, resulta que con solo dos componentes principales se explica el 99.986% de la varianza total, más, las variables cuyas escalas son en promedio más altas dominan estas dos componentes siendo estas X_4 = Tiempo de enfermedad, ya que posee el peso más significativo, en la segunda componente la más significativa es la variable $\mathbf{X}_3$ = Estadío.

14. La variable X_4 = Tiempo de enfermedad es la variable que más pesa dentro del modelo y debería ser la primera en considerarse a la hora de realizar un análisis de cáncer de melanoma.

15. Al realizar el análisis de sobrevida utilizando el método de Kaplan-Meier, se utilizaron las variables Tiempo de Enfermedad medido en días y Estado de última observación, la misma que nos indica si el paciente estuvo vivo, muerto o abandonó el tratamiento, de las cuales se obtuvo que a los 9 meses (271 días) la probabilidad de sobrevida es de 0.9878, a los 11.6 meses (350 días) es de 0.9748, mientras que a los 33 meses (987 días) esta probabilidad decrece a 0.8611 y por último ésta se reduce a la mitad, es decir, a 0.5144 aproximadamente a los 4.92 años.

16. Al realizar el análisis de sobrevida mediante la técnica de Regresión de Cox se emplearon las siguientes variables: X1= Edad, X2= L.D.H., X3= Sexo, X4= Estadío, como variable dependiente del tiempo X5= Tiempo de

Enfermedad y como variable de estado X6= Estado de última observación, lo que lanzó el siguiente modelo:

$$h(t/X) = h_0(t)g(x)$$

$$\hat{g}(x) = \left[0.967^{(EDAD)}0.997^{(L.D.H)}1.831^{(Sexo)}0.437^{(E1)}3.126^{(E2)}8.831^{(E3)}\right]$$

Por ejemplo para un paciente cuya edad es 65 años, de sexo masculino, su nivel L.D.H. es de 364, y con un Estadio 3, tendremos el siguiente resultado:

$\hat{g}(X)$ = 0.6115

El paciente con estos factores tiene una probabilidad del 61.15% de morirse.

RECOMENDACIONES

1. Para que la recolección de los datos sea ágil y eficiente se recomienda que la institución SOLCA se provea de un sistema de base de datos en donde se pueda encontrar con facilidad toda la información actualizada acerca de los pacientes y la variedad de enfermedades que se tratan en este establecimiento.

2. Es elemental que el ICM establezca convenios para realizar investigaciones con instituciones de este tipo, que aportan a la calidad de vida de las personas, ya que con los resultados obtenidos los médicos especialistas están capacitados para dar un mejor diagnóstico a los futuros pacientes y a su vez los estudiantes del ICM pondrán en práctica todos los conocimientos obtenidos a lo largo de su vida académica.

3. Diseñar un buen registro de información en donde los médicos puedan llenar con facilidad todos los datos de

los pacientes, y así, facilitar el ingreso de esta información a la base de datos de la institución.

4. En el análisis de Componentes Principales se recomienda basarse principalmente en el método que consiste en retener aquellas componentes que nos proporcione más del 70% del total de la información.

5. Para obtener mejores resultados al realizar el análisis de Regresión de Cox se recomienda hacerlo en un período de tiempo de 5 años como mínimo.

BIBLIOGRAFÍA

1 Mendenhall, Wackerly, Scheaffer, Estadística Matemática con Aplicaciones, Grupo Editorial Iberoamérica S.A. México, Segunda Edición, páginas 606 – 610.

2 Johnson, R. Wichern, D. Applied Multivariate Statistical Analysis, Prentice Hall. New Jersey United States, Fourth Edition, páginas 458 – 472.

3 Johnson, D. Métodos multivariados aplicados al análisis de datos, International Thomson Editores, México. México, páginas 77 – 99.

4 Walpole, E. & Freíd, J. Estadística Matemática con Aplicaciones Editor Hugo Acevedo Espinosa, Cuarta edición, México D.F. México.

5 Cochram, W.G. & Cox, G.M. 1991. Diseños experimentales, 2ª edición. Trillas, México DF.

6 Amitava Mitra, Fundamentals of Quality Control and Improvement, Second Edition, páginas 154-155.