ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

INSTITUTO DE CIENCIAS MATEMÁTICAS

INGENIERÍA EN ESTADÍSTICA INFORMÁTICA

"Factores predictores de sobrevida en pacientes con diferentes patologías cancerosas, mediante el modelo de regresión de Cox. Caso: Estómago"

TESIS DE GRADO

PREVIO A LA OBTENCIÓN DEL TÍTULO DE:

INGENIERA EN ESTADÍSTICA E INFORMÁTICA

PRESENTADA POR:

Andrea Malizza Abarca Pérez

GUAYAQUIL - ECUADOR

2004

DECLARACIÓN EXPRESA

La responsabilidad del contenido de ésta Tesis de Grado corresponde exclusivamente a la autora y el patrimonio intelectual de la misma a la Escuela Superior Politécnica del Litoral.

Andrea Abarca Pérez

A Jehová Dios, nuestro creador.

A mis padres por su amor, comprensión y apoyo.

A mi familia .

RESUMEN

La población objetivo ha ser analizada son los pacientes de SOLCA pero solo aquellos que fueron diagnosticados con Cáncer de Estómago en el año de 1999, el presente estudio muestra un análisis estadístico de algunas características de la población.

El conjunto correspondiente a la población investigada lo forman 115 pacientes que recibieron atención en la institución ya mencionada. El género masculino tuvo mayor ocurrencia con el 65% que el femenino con el 35%, las edades están entre los 31 a 88 años, con un promedio de 65 años.

El tipo morfológico del cáncer, el 80% fue de tipo intestinal y el 54% de los pacientes tenían metástasis en otros órganos en el cual

el 25% fue en el hígado. La mayor parte de los pacientes se encontraban en la fase o estadio IV con el 62%.

En lo relacionado al Análisis Multivariado debemos retener las dos primeras componentes principales, ya que las dos primeras componentes principales explican el 99,998% de la información total.

En lo relacionado al análisis de Sobrevivida para el modelo de Regresión de Cox, se analizó dos modelos para contrastar y verificar la información obtenida, en los cuales no variaron los resultados obtenidos en gran diferencia.

ÍNDICE GENERAL

SIMBOLOGÍA

Σ	sumatoria
$\bar{X}$	media aritmética
R	rango
S^2	varianza
S	desviación estándar
Y_1	coeficiente del sesgo
Y_2	coeficiente de kurtosis
$\sum$	matriz de varianzas y covarianzas
ρ	matriz de correlación
λ	valor propio
Z	matriz de datos estandarizados
μ	media
$V^{1/2}$	matriz diagonal de la desviación estándar

$h(t/X)$	función de riesgo
$h_0(t)$	Es la función de riesgo sin considerar el efecto del conjunto de variables
g(X)	Es la función de riesgo considerando el efecto del conjunto de variables
$S(t/X)$	función de supervivencia
X	conjunto de variables independientes
Z	combinación lineal del conjunto de variables independientes

ÍNDICE DE TABLAS

ÍNDICE DE GRÁFICOS

INTRODUCCIÓN

El término "cáncer" se refiere a un grupo de enfermedades en las cuales las células crecen y se diseminan libremente por el cuerpo. Es difícil imaginarse a alguien que no haya escuchado acerca de esta enfermedad. Muchas personas de una forma u otra han sido afectadas por el cáncer; por eso, es importante que todas las personas tengan un conocimiento básico sobre el cáncer.

El cáncer gástrico, es el cáncer que comienza en cualquier parte del estómago. El estómago es uno de los muchos órganos ubicados en el abdomen, el área del cuerpo que se encuentra entre el tórax y la pelvis. Otros órganos que se encuentran en el abdomen son el

hígado, el páncreas, la vesícula biliar y el colon. Es importante diferenciar entre estos órganos, debido a que los cánceres y otras enfermedades que los afectan presentan diferentes síntomas y se tratan de forma diferente.

La causa exacta del cáncer de estómago se desconoce, aunque se cree que hay muchos factores de riesgo que contribuyen a que las células del estómago se vuelvan cancerosas. Los síntomas del cáncer de estómago pueden parecerse a los de otras condiciones o problemas médicos. Siempre consulte a su médico para el diagnóstico.

Por medio de la información obtenida a través de SOLCA, se realizó un análisis de los Factores Predictores de Sobrevida por medio de la Regresión de Cox. Al principio de esta investigación, se presenta una introducción sobre este tipo de cáncer, seguido por el Análisis Univariado, Multivariado y de Supervivencia, seguido por las correspondientes conclusiones y recomendaciones para finalizar el estudio..

I. IDENTIFICACIÓN DEL CÁNCER DE ESTÓMAGO

1.1 DESCRIPCIÓN GENERAL DEL CÁNCER

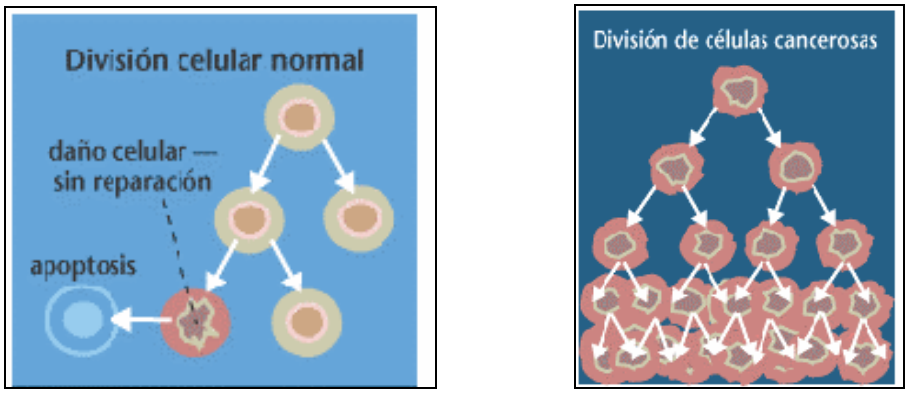
Al referirse al término cáncer, también relacionado como tumor maligno, por lo general se lo relaciona con una seria condición médica con resultados frecuentes en la muerte de la persona que lo contrajo. Por eso es necesario tener un conocimiento básico del cáncer, su origen, causas, diagnóstico, tratamientos y su prevención.

Este término, cáncer, se refiere a un grupo de enfermedades en las cuales las células crecen y se esparcen libremente por el cuerpo sin control. Es muy difícil actualmente imaginarse a alguien que no haya escuchado acerca de esta enfermedad. Muchas personas de una forma u otra han sido afectadas por el cáncer; puede ser que un ser querido, una amistad, o quizá hasta ellos mismos sean sobrevivientes de esta enfermedad.

1.2 ORIGEN DEL CÁNCER

El cáncer ocurre por un descontrol en el crecimiento normal de las células. En los tejidos normales, las tasas relacionadas con el crecimiento de células nuevas y con la muerte de las células viejas se mantienen en balance. En el cáncer se altera este balance. Esta alteración puede ser el resultado del crecimiento descontrolado de células o la incapacidad de las células a someterse a la "apoptosis". La apoptosis, o "el suicidio de las células", es el proceso en el cual las células viejas o dañadas se autodestruyen normalmente.

FIGURA 1-1
DIFERENCIA DE LA DIVISIÓN CELULAR ENTRE CELULAS NORMALES Y CANCEROSAS



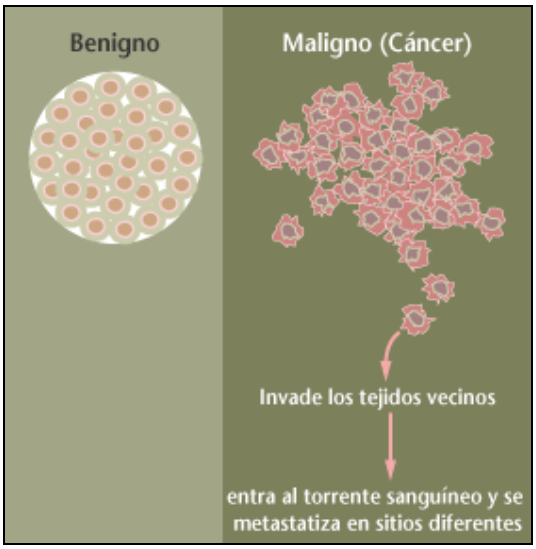
www.press2.nci.nhi.gov

El aumento gradual en el número de células con capacidad para dividirse crea una masa creciente de tejido que se conoce como "tumor" o "neoplasma"[1]. El tumor crecerá rápidamente de tamaño si la división de las células es relativamente rápida y no hay señales "suicidas" que provoquen la muerte de las células. Si las células se dividen más lentamente, el crecimiento del tumor será más lento. Sin importar la rapidez del crecimiento, los tumores crecen en tamaño porque las nuevas células se producen en cantidades mayores de lo que es necesario. La formación normal del tejido se alterará gradualmente, entre más y más se acumulen las células que se dividen.

[1] Neoplasma: Tejido celular anormal de nueva formación.

Los tumores se clasifican en benignos y malignos, lo cual depende de si se esparcen por invasión o por metástasis[2]. Los tumores benignos son tumores que no se pueden esparcir por invasión o por metástasis; por lo tanto, sólo crecen localmente. Los tumores malignos son tumores que se pueden esparcir por invasión y por metástasis. Por definición, el término "cáncer" se aplica sólo a los tumores malignos.

FIGURA 1-2
DIFERENCIA ENTRE UN TUMOR BENIGNO Y MALIGNO



www.press2.nci.nhi.gov

La invasión se refiere a la migración y a la penetración directas de las células cancerosas en el tejido vecino. La metástasis se refiere a la habilidad de las células cancerosas

[2] Metástasis: Reproducción de una enfermedad en órganos distintos de aquel en el que se presentó primero.

de penetrar en los vasos sanguíneos y linfáticos, circular por el torrente sanguíneo y luego invadir el tejido normal en otras partes del cuerpo.

Un tumor maligno, un "cáncer", es un problema más serio para la salud que un tumor benigno, porque las células cancerosas se pueden diseminar a otras partes distantes del cuerpo y pueden alterar las funciones de estos órganos vitales y por lo tanto poner la vida en peligro.

1.3 CAUSAS DEL ORIGEN DEL CÁNCER

Hoy se considera que la causa que origina el cáncer es la mutación[3] de los genes que controlan la proliferación de células normales; éstas mutaciones pueden ser producidas por los carcinógenos dañinos del ADN, como los subproductos derivados del tabaco y la radiación. Sin embargo, algunas mutaciones que causan el cáncer son simplemente errores espontáneos que aparecen en las moléculas normales de ADN cuando las células duplican su ADN antes de que se dividan.

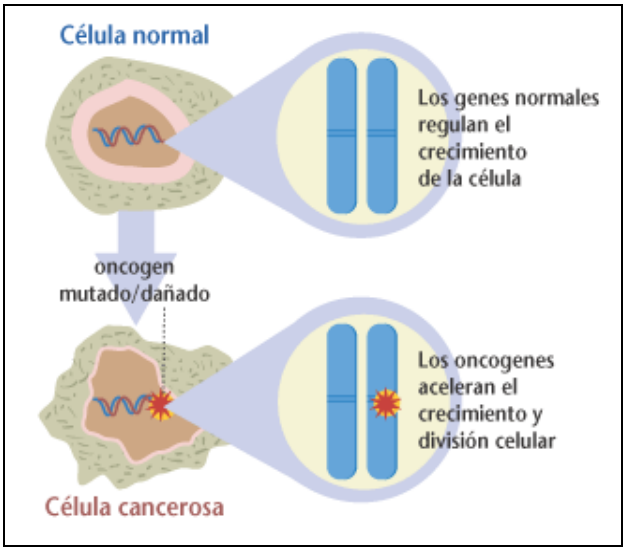
[3] Mutación: cambio en el material genético, que aparece bruscamente y no es debido a recombinación genética.

Las mutaciones que contribuyen al desarrollo del cáncer afectan a tres clases de genes: oncogenes, genes supresores de tumor y genes de reparación del ADN.

1.3.1 Los oncogenes

Los oncogenes, o genes dañados, son genes que se encuentran en cada célula y puede causar que una célula sana se desarrolle maligna bajo condiciones particulares; ya que contribuyen dando instrucciones a las células para que produzcan proteínas que estimulan la división y el crecimiento; por lo cual es el primer grupo de genes implicados en el desarrollo del cáncer, ya que su actividad o sobreactividad puede acelerar el crecimiento excesivo de células.

FIGURA 1-3
LOS ONCOGENES



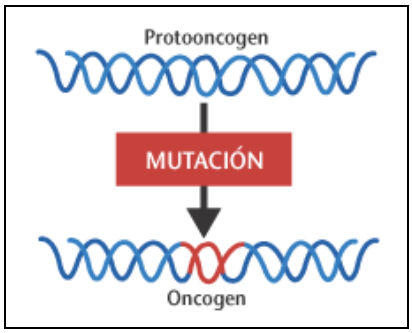
www.press2.nci.nhi.gov

Los oncogenes están relacionados con los genes normales llamados protooncogenes. Los protooncogenes son una familia de genes normales que codifican principalmente a las proteínas involucradas en el mecanismo de control del crecimiento normal de las células.

Los oncogenes resultan de la mutación de protooncogenes. Por ser formas mutantes de los protooncogenes, los oncogenes se asemejan a los protooncogenos en el sentido que codifican la producción de proteínas involucradas en el control de

crecimiento. Sin embargo, los oncogenes codifican versiones alteradas (o cantidades excesivas) de estas proteínas de control de crecimiento, alterando de esta manera el mecanismo de señalamiento de crecimiento de las células.

FIGURA 1-4
MUTACIÓN DE UN PROTOONCOGEN EN UN ONCOGEN



www.press2.nci.nhi.gov

Al producir versiones o cantidades anormales de proteínas de control de crecimiento, los oncogenes hacen que el mecanismo de señalamiento de crecimiento de la célula sea hiperactivo; es decir entre más activo esté el mecanismo, más rápido se dividen y crecen las células. Una célula cancerosa puede contener uno o más oncogenes, lo cual indica que existen uno o más componentes anormales en este mecanismo.

1.3.2 Los genes supresores de tumor

El segundo grupo de genes implicados en el cáncer son los "genes supresores de tumor". Los genes supresores de tumor son genes normales cuya AUSENCIA puede conducir al cáncer. En otras palabras, si una célula pierde un par de genes supresores de tumor o si son inactivados por mutación, su falta de funcionamiento puede causar el cáncer. Los individuos que heredan un alto riesgo de desarrollar cáncer frecuentemente nacen con una copia defectuosa del gen supresor de tumor.

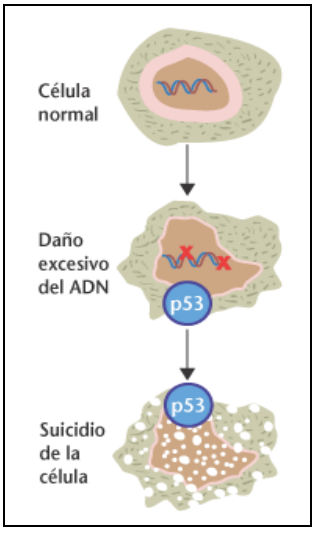
Debido a que los genes ocurren naturalmente en pares (uno heredado de cada uno de los padres), un defecto heredado en una copia no causará el cáncer debido a que la otra copia normal aún funciona. Pero si la segunda copia se somete a la mutación, la persona entonces puede desarrollar cáncer porque ya no existe alguna copia del gen que funcione.

Los genes supresores de tumor son una familia de genes normales que ordenan a las células a producir proteínas que restringen el crecimiento y la división de

células. Ya que los genes supresores de tumor codifican las proteínas que disminuyen el crecimiento y la división de células, la pérdida de estas proteínas permite que las células crezcan y se dividan de forma incontrolada.

Un gen supresor de tumor en particular codifica la proteína llamada "p53" la cual puede provocar el suicidio de células (apoptosis). La proteína p53 actúa como un "freno" que detiene el crecimiento y la división de las células que han sufrido daño en su ADN. Si no se puede reparar el daño, la proteína p53, con el tiempo, iniciará el suicidio celular, previniendo así el crecimiento descontrolado de las células genéticamente dañadas.

FIGURA 1-5
FUNCIÓN DE LOS GENES SUPRESORES DE TUMOR

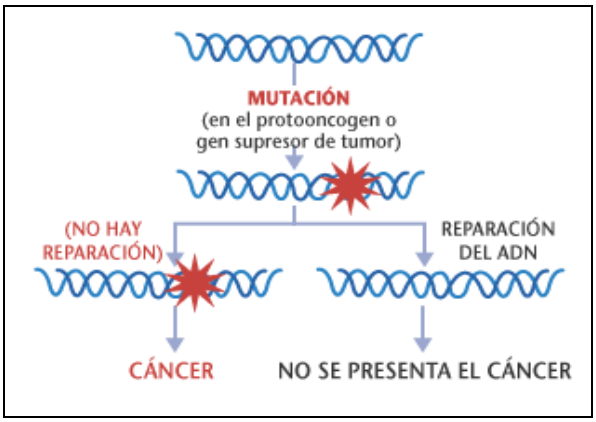


www.press2.nci.nhi.gov

1.3.3 Genes de reparación del ADN

Los "genes reparadores de ADN" son la tercera clase de genes implicados en el cáncer. Estos genes codifican proteínas cuya función normal es corregir errores que surgen cuando las células duplican su ADN antes de dividirse. Las mutaciones en los genes reparadores de ADN pueden conducir al fracaso en la reparación de ADN, lo cual a su vez permite mutaciones subsecuentes en los genes supresores de tumor y que los protooncogenes se acumulen.

FIGURA 1-6
FUNCIÓN DE LOS GENES DE REPARACIÓN DEL ADN



www.press2.nci.nhi.gov

1.4 LOS DIFERENTES TIPOS DE CÁNCER

El cáncer puede surgir casi en cualquier parte del cuerpo.

El carcinoma, el más común entre los diferentes tipos de cáncer, proviene de las células que cubren las superficies externas e internas del cuerpo.

Los sarcomas son cánceres que surgen de las células que se encuentran en los tejidos que sostienen el cuerpo como el hueso, el cartílago, el tejido conectivo, el músculo y la grasa.

Los linfomas son cánceres que se originan en los ganglios linfáticos y en los tejidos del sistema inmune del cuerpo.

Las leucemias son cánceres de células inmaduras de la sangre producidas en la médula ósea y que tienden a acumularse en grandes cantidades dentro del torrente sanguíneo.

Los científicos utilizan una variedad de nombres técnicos para distinguir las diferentes clases de carcinomas, sarcomas, linfomas y leucemias. En general, estos nombres usan diferentes prefijos que representan el nombre de la célula afectada. Por ejemplo, el prefijo "osteo" significa hueso,

entonces un cáncer que surge del hueso se llama osteosarcoma. Similarmente, el prefijo "adeno" significa glándula; así pues, un cáncer que se presenta en las células de una glándula se llama adenocarcinoma.

1.5 DESCRIPCIÓN DEL CÁNCER GÁSTRICO

El cáncer gástrico, generalmente conocido como cáncer de estómago, es una enfermedad en la cual se encuentran células cancerosas (malignas) en los tejidos del estómago y puede comenzar en cualquier parte de este.

El estómago es un órgano en forma de J que se encuentra en la parte superior del abdomen entre el tórax y la pelvis, donde los alimentos se digieren. Los alimentos llegan al estómago a través del esófago que conecta la boca con el estómago. Después de pasar por el estómago, los alimentos parcialmente digeridos pasan al intestino delgado y luego al intestino grueso o colon.

Debido que el estómago es uno de los órganos ubicados en el abdomen, junto con el hígado, el páncreas, la vesícula biliar y el colon, es importante diferenciar entre estos órganos, debido a que los cánceres y otras enfermedades que los afectan presentan diferentes síntomas y se tratan de forma diferente.

Como todo cáncer en su etapa inicial casi no produce síntomas e incluso puede encontrarse en el estómago durante mucho tiempo y crecer considerablemente antes de causar síntomas. La causa exacta del cáncer de estómago se desconoce, aunque se cree que hay muchos factores de riesgo que contribuyen a que las células del estómago se vuelvan cancerosas.

1.6 FACTORES DE RIESGOS.

Antes de enunciar los factores de riesgos es importante saber que significa. Un factor de riesgo es cualquier cosa que pueda aumentar las probabilidades de una persona de desarrollar

una enfermedad. Puede ser una actividad como fumar, la dieta, su historia familiar o muchas otras cosas. Distintas enfermedades, incluyendo los cánceres, tienen factores de riesgo diferentes.

Aun cuando estos factores pueden aumentar los riesgos de una persona, éstos no necesariamente causan la enfermedad. Algunas personas con uno o más factores de riesgo nunca contraen la enfermedad, mientras otras la desarrollan sin tener factores de riesgo conocidos.

Pero el saber los factores de riesgo de cualquier enfermedad puede ayuda a guiar a las personas en las acciones apropiadas, incluyendo el cambio de comportamiento y el ser monitoreado clínicamente.

Entre los factores de riesgo del cáncer de estómago se incluyen los siguientes:

- Infección por Helicobacter pylori.
- Una dieta que incluye lo siguiente:

  - Cantidades elevadas de alimentos ahumados.
  - Carnes y pescados curados con sal.
  - Alimentos con alto contenido de almidón y con poca fibra.
  - Vegetales en vinagre.
  - Alimentos y bebidas que contienen nitratos y nitritos.

- El abuso del cigarrillo.
- El abuso del alcohol.
- Cirugía previa del estómago.
- Anemia megaloblástica o perniciosa (causada por la deficiencia de vitamina B12).
- Enfermedad de Ménétrier.
- La edad de 55 años o más (la mayoría de los pacientes tienen entre 60 y 70 años).
- Sexo masculino (la enfermedad se les diagnostica a más hombres que a mujeres).
- Tener sangre de tipo A.
- Antecedentes familiares de lo siguiente:

  - Cáncer de colon no polipósico.
  - Poliposis familiar adenomatosa.

- Cáncer de estómago.

- Antecedentes de pólipos en el estómago.

También la exposición a factores ambientales como polvos y vapores en el lugar de trabajo.

1.7 SÍNTOMAS Y DIAGNÓSTICO DEL CÁNCER GÁSTRICO.

Los síntomas incluyen un dolor impreciso en el abdomen superior, que puede asociarse con mal apetito y perdida de peso. Muchos individuos se ponen anémicos pero, por lo demás, *no muestran ningún síntoma antes que ocurra la diseminación metastática*. Otros síntomas pueden incluir nausea, vómitos, variación en los hábitos intestinales, mal apetito y debilidad e infección con Helicobacter pylori. Para obtener un diagnóstico se requiere una historia médica completa y exacta y un examen físico, puede que algunos pacientes necesiten que se les practiquen evaluaciones de diagnóstico más extensas, que pueden ser:

- Examen de sangre oculta en las heces: el cual busca indicios de sangre escondida (oculta) en las heces. Se coloca una cantidad muy pequeña de heces en una tarjeta especial, y luego se examina en el consultorio del médico o se envía al laboratorio.
- Serie gastrointestinal (GI) superior: También llamada esofagografía, examen de diagnóstico que examina los órganos de la parte superior del sistema digestivo: el esófago, el estómago y el duodeno (la primera sección del intestino delgado). Se ingiere un líquido denominado bario (un producto químico metálico [líquido yesoso] utilizado para recubrir el interior de los órganos de forma que aparezcan en una placa de rayos X). Después se toman los rayos X para evaluar los órganos digestivos.
- Esofagogastroduodenoscopia: También llamada EGD o endoscopia superior. Es un procedimiento que le permite al médico examinar el interior del esófago, el estómago y el duodeno. Un tubo con luz, delgado y flexible, llamado endoscopio, se pasa por la boca y la garganta, y luego por el esófago, el estómago y el duodeno. El endoscopio le permite al médico ver dentro de este área del cuerpo, así como introducir instrumentos a través del endoscopio para

tomar muestras de tejido y realizar una biopsia (si es necesario).

- Ecografía endoscópica: esta técnica de imagen utiliza ondas sonoras para crear una imagen computarizada del interior del esófago y del estómago. El endoscopio se pasa por la boca y la garganta, y luego por el esófago y el estómago. Como en una endoscopia normal, esto le permite al médico ver el interior de este área del cuerpo, así como introducir instrumentos para tomar muestras de tejido (biopsia).

La probabilidad de recuperación (pronostico) y la selección del tratamiento dependerán de la etapa en la que se encuentre el cáncer (si se encuentra en el estomago o si se ha diseminado a otras partes del cuerpo) y del estado de salud general del paciente.

1.8 ETAPAS DEL CÁNCER GÁSTRICO

Una vez que se encuentra cáncer en el estómago, se hacen otras pruebas para determinar si las células cancerosas se han diseminado a otras partes del cuerpo. Este proceso se

denomina clasificación por etapas. El médico necesita saber la etapa en la que se encuentra la enfermedad para poder planear el tratamiento adecuado. Las siguientes etapas se emplean en la clasificación del cáncer del estómago:

- **Etapa 0:** El cáncer del estómago en etapa 0 es un cáncer en su etapa inicial. El cáncer sólo se encuentra en la capa más interior de la pared estomacal.
- **Etapa I:** El cáncer se encuentra en la segunda o tercera capa de la pared estomacal y no se ha diseminado a los ganglios linfáticos cercanos al cáncer, o se encuentra en la segunda capa de la pared estomacal y se ha diseminado a los ganglios linfáticos que se encuentran muy cerca del tumor. (Los ganglios linfáticos son estructuras pequeñas en forma de fríjol que se encuentran en todo el cuerpo y cuya función es producir y almacenar células que combaten la infección.)
- **Etapa II:** Se puede presentar cualquiera de las siguientes situaciones:

1. El cáncer se encuentra en la segunda capa de la pared estomacal y se ha diseminado a los ganglios linfáticos que se encuentran lejos del tumor.

2. El cáncer sólo se encuentra en la capa muscular (la tercera capa) del estómago y se ha diseminado a los ganglios linfáticos muy cercanos al tumor.

3. El cáncer se encuentra en las cuatro capas de la pared estomacal pero no se ha diseminado a los ganglios linfáticos ni a otros órganos.

- **Etapa III:** Se puede presentar cualquiera de las siguientes situaciones:

1. El cáncer se encuentra en la tercera capa de la pared estomacal y se ha diseminado a los ganglios linfáticos que se encuentran lejos del tumor.

2. El cáncer se encuentra en las cuatro capas de la pared estomacal y se ha diseminado a los ganglios linfáticos que están muy cerca del tumor o lejos del tumor.

3. El cáncer se encuentra en las cuatro capas de la pared estomacal y se ha diseminado a tejidos cercanos. El cáncer puede haberse diseminado o no a los ganglios linfáticos muy cercanos al tumor.

- **Etapa IV:** El cáncer se ha diseminado a los tejidos cercanos y a los ganglios linfáticos que se encuentran lejos del tumor o se ha diseminado a otras partes del cuerpo.

1.9 TRATAMIENTO DEL CÁNCER DE ESTÓMAGO

El tratamiento específico del cáncer de estómago será determinado por el médico basándose en lo siguiente:

- Su edad, su estado general de salud y su historia médica.
- Qué tan avanzada está la enfermedad.
- Su tolerancia a determinados medicamentos, procedimientos o terapias.
- Sus expectativas para la trayectoria de la enfermedad
- Su opinión o preferencia.

El tratamiento del cáncer de estómago puede incluir:

- Cirugía

  La cirugía puede ser necesaria para extirpar el tejido canceroso así como tejido contiguo no canceroso. La operación más común se llama gastrectomía. Si se extirpa parte del estómago, se llama gastrectomía parcial o subtotal. Si se extirpa todo el estómago, se llama gastrectomía total.

- Radioterapia

  La radioterapia utiliza rayos de alta energía para eliminar o reducir las células cancerosas.

- Quimioterapia

  La quimioterapia usa medicamentos para eliminar las células cancerosas .

II. DETERMINACIÓN DE LAS VARIABLES A SER INVETIGADAS

2.1. Objetivo

En la ciudad de Guayaquil, actualmente, funciona desde hace 50 años el Instituto Oncológico de La Sociedad de Lucha Contra el Cáncer, conocida como SOLCA, la cual provee de servicios de ayuda para combatir ésta enfermedad para personas de limitados recursos económicos.

El estudio que se hace a continuación está enfocado en los pacientes que contrajeron ésta enfermedad y su sobrevida.

2.2. Población Objetivo y población investigada

La población objetivo ha ser analizada son los pacientes de SOLCA pero solo aquellos que fueron diagnosticados con Cáncer de Estómago.

La población investigada incluye solo a los pacientes que fueron diagnosticados en el año de 1999.

2.3. Marco

El marco lo conforman todas las historias clínicas de los pacientes diagnosticados con Cáncer Gástrico en el año de 1999, conformando su listado 115 personas.

2.4. Instrumento de medida

El instrumento de medida que se utilizó para recoger la información fue un cuestionario que consta de información acerca del paciente como la edad, sexo, nivel de instrucción, lugar de residencia y antecedentes familiares; además de preguntas que contienen nuestras variables de interés.

Cabe recalcar que el cuestionario que se utilizó fue orientado por un médico conocido en el tema.

2.5. Descripción de las variables a ser investigadas

En el cuestionario que se utilizó para la recolección de la información de las historias clínicas de los pacientes, se utilizaron veinticuatro variables las cuales se describen a continuación.

Variable # 1: Sexo

El sexo de los pacientes nos brindará información de la ocurrencia de éste tipo de cáncer en cada uno de ellos.

Variable # 2: Edad

En esta variable nos referimos a la edad en años que tenía el paciente cuando se lo diagnosticó con Cáncer Gástrico.

Variable # 3: Lugar de Residencia

El Lugar de Residencia, se refiere al lugar donde la persona usualmente habita, a través de esta variable se busca conocer de donde proviene y para futuras mediciones de seguimiento.

Variable # 4: Nivel de Instrucción

Nivel de Instrucción, se refiere al grado de educación que tiene el paciente.

Variable # 5: Institución que Envía

Con ésta variable se busca obtener información si el paciente ya fue atendido por otra institución médica antes de ir a SOLCA.

Variable # 6: Diagnostico Previo

Esta variable está ligada a la anterior y nos da información acerca si se diagnosticó con esta enfermedad al paciente antes de acudir a SOLCA.

Variable # 7: Tratamiento Previo

Con ésta variable se sabrá si el paciente recibió algún tipo de tratamiento en la institución anterior antes de acudir a SOLCA.

Variable # 8: HPV

El HPV, o Hicto Bacto Pilori, con esta variable se desea conocer si el paciente contenía ésta bacteria en su organismo al momento de ser diagnosticado.

Variable # 9: Antecedentes Familiares.

Con ésta variable se busca información si existe algún familiar que haya sido diagnosticado con ésta enfermedad.

Variable # 10: Clasificación de la lesión.

Esta variable nos permitirá saber el tipo del lugar de la lesión que tenía el paciente al momento de ser diagnosticado. Se clasifica en los siguientes tipos: El 1/3 proximal cardias, 1/3 medio cuerpo, 1/3 distal antro, solapada (en dos o más lugares) y SAI.

Variable # 11: Tipo Morfológico

Esta variable nos permitirá a saber cual es el diagnostico según el tipo de morfología que presenta que presenta el paciente después de los resultados.

Variable # 12: M

Esta variable, Metástasis, da información si el cáncer ha invadido a otros órganos o se encuentra en otra parte del cuerpo que fuera del estómago.

Variables # 13 Lugar de Metástasis

Con ésta variable permite conocer el tipo de metástasis según el lugar donde el cáncer se encuentra ubicado.

Variable # 14: Estadio.

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

Variable # 15: Tiempo de enfermedad.

Con ésta variable se desea conocer el tiempo transcurrido en años desde que el paciente fue diagnosticado con cáncer gástrico.

Variable # 16: Estado de la Ultima Observación.

Esta variable nos permitirá saber el estado físico en que se encontraba el paciente en la última cita.

Variable # 17: Tratamiento Cronológico

Esta variable nos permitirá conocer el tipo de tratamiento o tratamientos que recibió el paciente los cuales pueden ser cirugía, radioterapia, o quimioterapia.

Variable # 18: Tipo de Cirugía

En ésta variable se obtiene información del tipo de cirugía que se le practicó al paciente, estás pueden ser gastrectomía total, gastrectomía subtotal, esplenectomía, pancreatectomía, omentectomía y paliativa.

Variable # 19: Recibió Radioterapia

Esta variable brinda información acerca del paciente que presenta cáncer de estómago recibió como parte de su tratamiento Radioterapia.

Variable # 20: Completo Radioterapia

Esta variable nos permitirá conocer si el paciente completó las dosis de radioterapia las cuales fueron como parte de su tratamiento.

Variable # 21: Recibió Quimioterapia

Esta variable brinda información acerca del paciente que presenta cáncer de estómago recibió como parte de su tratamiento Quimioterapia.

Variable # 22: Completo Quimioterapia

Esta variable nos permitirá conocer si el paciente acudió a todas las sesiones de quimioterapia las cuales fueron programadas como parte de su tratamiento.

2.6. Codificación de las variables a ser investigadas.

Debido a que la mayoría de la variable descritas anteriormente son cualitativas y para poder analizarlas estadísticamente se las decidió codificarlas en escala nominal con el propósito de poder realizar más adelante el análisis de componentes principales y el análisis de la curva de sobrevivida mediante análisis de regresión de Cox.

Variable # 1: Sexo

Masculino	0
Femenino	1

Variable # 3: Lugar de Residencia

Guayaquil	0
Fuera de Guayaquil	1

Variable # 4: Nivel de Instrucción

Ninguna	0
Primaria	1
Secundaria	2
Superior	3

Variable # 5: Institución que Envía

Voluntaria	0
Medico Particular	1
H. Luis Vernaza	2

Clínica Kennedy	3
H. Guayaquil	4
H. IESS	5
Clínica Alcívar	6
SOLCA Machala	7

Variable # 6: Diagnostico Previo

No	0
Si	1

Variables # 7: Tratamiento previo

No	0
Si	1

Variable # 8: HPV

No	0
Si	1

Variables # 9: Antecedentes familiares

No	0
Si	1

Variable # 10: Clasificación de la lesión

1/3 proximal cardias	0
1/3 medio cuerpo	1
1/3 distal antro	2
Solapada	3
SAI	4

Variable # 11: Tipo Morfológico

Intestinal	0
Difuso	1

Variable # 12: M

No	0
Si	1

Variables # 13: Lugar de Metástasis

Ninguno	0
Abdominal y/o Epiplón	1
Huesos	2
Hígado	3
Pulmón, pleura o ambos	4
Bazo	5
Ovario	6
Carcinomatosis	7
Vesícula Biliar	8
Hepática y Pulmonar	9

Variable # 14: Estadio

No reporte	0
I	1
II	2

III	3
IV	4

Variables # 15: Tiempo de enfermedad

No especifica	0
Menos de 1 año	1
De 1 a 2 años	2
De 2 a 3 años	3
Mayor 3 años	4

Variable # 16: Estado de la Ultima Observación

Paciente vivo	0
Paciente Fallecido	1
Paciente Abandono Tratamiento	2

Variable # 17: Tratamiento Cronológico

Ninguno	0
Cirugía	1

Radioterapia	2
1 y 2	3
Quimioterapia	4
1 y 4	5
2 y 4	6
1, 2 y 4	7

Variable # 18: Tipo de Cirugía

Ninguna	0
Gastrectomía total	1
Gastrectomía subtotal	2
Gastrectomía total y Esplenectomía	3
Gastrectomía subtotal y Esplenectomía	4
Gastrectomía total y Pancreatectomía	5
Gastrectomía subtotal Omentectomía	6
Pancreatectomía y Omentectomía	7

Variable # 19: Radioterapia

No	0
Si	1

Variable # 20: Completo Radioterapia

No	0
Si	1

Variable # 21: Quimioterapia

No	0
Si	1

Variable # 22: Completo Quimioterapia

No	0
Si	1

2.7. Análisis Univariado

El análisis univariado consiste en realizar el análisis descriptivo de cada una de las variables objeto de estudio. Al realizar este tipo de análisis se considera tres tipos de medidas que son: las medidas de tendencia central, las medidas de dispersión y las medidas de sesgo y kurtosis.

En lo posterior de éste capitulo se detallará en que consiste cada una de éstas medidas.

2.7.1. Medidas de tendencia central

Los datos que fueron recopilados necesitan ser descritos para realizar una evaluación mas objetiva de los mismos, para ello existen algunas medidas numéricas que serán usadas para resumir la información de los valores observados. Las medidas de tendencia central permiten saber la localización de las observaciones y el valor alrededor del cual se encuentran. Las medidas de tendencia central que se analizaran en el Capitulo 3 serán las siguientes:

- Media Aritmética

La media aritmética es un estimador de la media de la población. Esta medida es una de las más utilizadas cuando se requiere evaluar un conjunto de medidas de una característica determinada. La media aritmética es el promedio del conjunto de observaciones y se la denota por $\bar{X}$, la ecuación para cálculo es la siguiente:

$$\bar{X} = \frac{\sum_{i=1}^{n} Xi}{n}$$

Donde n es el número total de observaciones y Xi es el valor que toma cada observación.

- Mediana

Cuando se esta realizando el proceso de recolección de información se puede 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.

2.7.2. Medidas de Dispersión

Es importante no solo conocer los valores de tendencia central que tienen las variables observadas sino también es importante la variabilidad a que están sujetas. Algunas de las medidas que nos proporcionan información acerca de la variabilidad se detallan a continuación.

- Rango

Es una de las medidas de dispersión que mayormente se utilizada, este valor se lo obtiene por 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 mas alto valor y X_S es la observación de mas bajo valor.

- Varianza

La varianza mide las oscilaciones de las observaciones alrededor de la media. La varianza muestral se la determina por medio de la siguiente ecuación:

$$S^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}$$

Si comparamos la varianza con el rango llegamos a la conclusión de que la varianza nos proporciona una mejor explicación de la variabilidad, por cuanto el rango solo considera al mayor y menor valor en cambio la varianza considera todas las observaciones. Cabe indicar además, que dos conjuntos

diferentes de datos pueden tener el mismo rango pero tener una variabilidad diferente.

- Desviación Estándar

La desviación estándar también mide la variabilidad de las observaciones con respecto a la media, es igual a la raíz cuadrada de la varianza. Está medida de dispersión siempre es positiva y se la denota por S. La ecuación que se utiliza para su calculo es:

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

2.7.3. 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.

- 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.

- Coeficiente de Kurtosis

El coeficiente de Kurtosis es una medida que nos permite observar el grado en el cual las observaciones forman una cresta. Esta medida esta dada por la ecuación:

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

2.7.4. Covarianzas

Al igual que las medidas anteriores, es importante conocer la relación existente o no entre dos variables por lo cual se utiliza la covarianza. La forma en la cual se calcula éste valor es:

$$Cov(X_i, Y_i) = E\left[\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)\right]$$

donde X_i y Y_i representan los diferentes valores que pueden tomar las variables X & Y. Si obtenemos un valor de la covarianza positivo significa que, bajo las mismas condiciones, si una variable se incrementa la otra también se incrementará. En cambio si obtenemos un valor negativo significa que ha medida que la una variable se incrementa la otra variable decrece.

2.8. Análisis Multivariado

Los objetivos relacionados con la explicación de un fenómeno físico o social pueden lograrse recogiendo y analizando los datos. Así al realizar la investigación de algún fenómeno, se debe recoger observaciones de diferentes variables; el método por medio del cual se realiza el análisis de observaciones simultaneas sobre muchas variables es llamado Análisis Multivariado.

Los objetivos al aplicar la técnica multivariada son los siguientes:

- Reducir los datos tanto como sea posible, mediante el sacrificio de una pequeña cantidad de información, esto facilita la interpretación del mismo.
- Crear variables que agrupen objetos o variables similares, esto se debe hacer basado en las características medidas.
- Investigar la dependencia entre las variables, resulta muy interesante determinar si una variable depende o no de otra.

2.8.1. Análisis de Componentes Principales

El análisis de componentes principales esta relacionado con las matrices de varianza y covarianza de un conjunto de variables a través de algunas combinaciones lineales de esas variables.

Las componentes principales son un conjunto de combinaciones lineales de las p variables aleatorias observadas $X_1, X_2, ..., X_p$, en términos geométricos dichas combinaciones lineales constituyen un nuevo sistema coordenado, el cual se obtiene a partir de las p variables originales $X_1, X_2, ..., X_p$.

2.8.2. Procedimiento para la obtención de las componentes principales.

Al realizar un análisis estadístico utilizando el método de componentes principales no se requiere asumir normalidad de las variables aleatorias observadas.

Las componentes principales dependen únicamente de la matriz de covarianzas $\sum$ o de la matriz de correlación p de $X_1, X_2, ..., X_p$.

Ahora bien, si tenemos el vector aleatorio $X^t = [X_1, X_2, ..., X_p]$, constituido por las p variables originalmente observada y obtenemos a partir de estos datos la matriz de covarianza $\sum$ de la cual obtenemos los valores propios $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0.$

Consideremos las combinaciones lineales:

$$Y_1 = a_1' X = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p$$

$$Y_2 = a_2' X = a_{21}X_1 + a_{22}X_2 + ... + a_{2p}X_p$$

$$\vdots$$

$$Y_p = a_p' X = a_{p1}X_1 + a_{p2}X_2 + ... + a_{pp}X_p$$

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.

2.8.3. Obtención de las Componentes principal

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

$(\lambda_1, e_1), (\lambda_2, e_2), \ldots (\lambda_p, e_p),$ donde $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0.$ de aquí que la i-ésima componente principal esta dada por:

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

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

Además

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

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

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:

$$Total\ de\ la\ \text{var}ianza = \sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp}$$

$$= \lambda_1 + \lambda_2 + \ldots + \lambda_p$$

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

$$\begin{bmatrix} \text{Pr}\,oporción\ \ del\ \ total \\ de\ \ la\ \ \text{var}\,ianza \\ \text{exp}\,licada\ \ por\ \ la \\ k-ésima\ \ componente \end{bmatrix} = \frac{\lambda_k}{\lambda_1 + \lambda_2 + ... + \lambda_p}$$

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

2.8.4. Obtención de Componentes Principales a partir 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:

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

2.8.5. Determinación del número optimo 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 tubo acuracidad cuando el número de variables era alto.
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.

2.9. Análisis de Supervivencia

El objetivo del Análisis de Supervivencia 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.

2.9.1. 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.

2.9.1.1. Formulación del Problema

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 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, el modelo que se postula es:

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

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 $X = \{ X_1,\ldots,X_p \}$.

Es decir, se supone que la función de riesgo 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 la Z la combinación lineal:

$$Z = \sum \beta_j X_j = \beta_1 X_1 + \ldots + \beta_p X_p.$$

Se tiene el modelo de regresión de Cox.

El análisis consiste en estimar los parámetros desconocidos $\beta_1,\ldots,\beta_p$. Notemos que, si las estimaciones de 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 tal caso, la función de g(X) sería 1 y por lo tanto $h(t/X)$ coincidiría con $h_0(t)$.

La función de supervivencia, $S(t/X)$, probabilidad de que el suceso final no ocurra hasta pasado un periodo de tiempo superior o igual a t, se la obtiene a partir de la función de riesgo:

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

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.

2.9.1.2. 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.

2.9.1.3. Selección de las variables

En la construcción de la función Z, podrá seleccionarse aquel subconjunto de las variables independientes que más información aporte sobre la probabilidad de que, para cada

posible valor de *t*, el suceso final no ocurra hasta pasado un periodo de tiempo. Tanto el método como los criterios para la selección y eliminación de variables serán el método Forward y los criterios basados en la Puntuación eficiente de Rao y el estadístico de Wald.

Estadístico de Wald

El estadístico de Wald, para cualquier variable independiente X_j seleccionada, si β_j es el parámetro asociado en la ecuación de regresión, permite contrastar la hipótesis nula:

Ho: $\beta_j = 0$

La interpretación de la 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 significancia α.

Bajo esto, en cada una de la selección de las variables, la candidata a ser eliminada será la que tenga el máximo valor *p*-valor asociado al estadístico de Wald. Será eliminada si dicho máximo es mayor que un determinado valor crítico prefijado (si no se indica lo contrario, 0.1).

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).

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.

2.9.1.4. 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:

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

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

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

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

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

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

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 $(t,\ t + \Delta t)$, 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*.

III. ANÁLISIS ESTADÍSTICO UNIVARIADO

3.1. Introducción

En este capítulo se realizará el análisis univariado de las variables, lo cual consiste en un análisis descriptivo de cada una de ellas. Cabe recalcar que las variables objeto de estudio fueron previamente seleccionadas con la ayuda de un médico relacionado con el tema que se está analizando.

3.2. Estadística descriptiva de las variables.

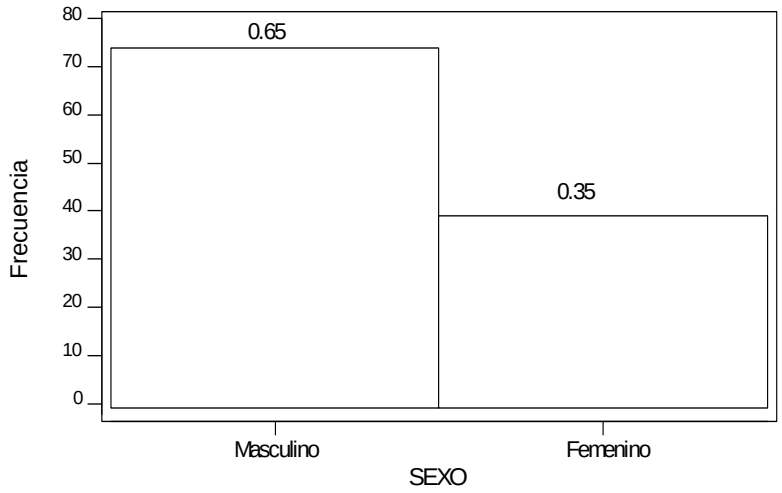
Para la elaboración de la estadística descriptiva de las variables, se utilizó el conjunto de datos observados,

debemos considerar que el número de casos sobre los que obtuvimos los datos es ciento quince. A continuación mostraremos el análisis Univariado para cada una de las variables.

Variable # 1: Sexo

Esta variable indica la ocurrencia de éste tipo de cáncer en cada género de los pacientes SOLCA. Al observar el histograma de frecuencia de esta variable notamos que, en el género masculino la ocurrencia es del 65% mientras que en el femenino es tan solo del 35% de los pacientes con Cáncer Gástrico registrados en el año 1999.

FIGURA 3-1
HISTOGRAMA DE FRECUENCIA PARA EL SEXO

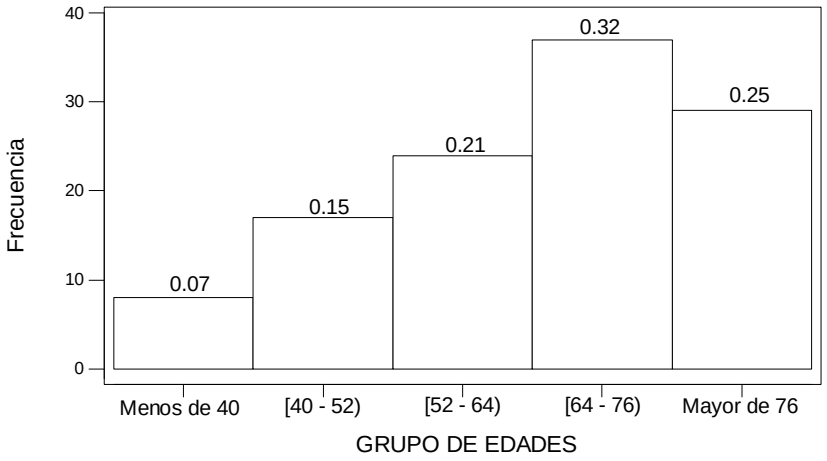


Variable # 2: Edad

Con esta variable se conoce la edad que tenían los pacientes cuando fueron diagnosticados con cáncer de estómago en SOLCA. Observando el histograma de frecuencia podemos notar que la edad de los pacientes varia entre los 31 y 88 años de edad.

Debido al rango de la variable edad, que es 57, se dispuso utilizar grupos de edades para la construcción del histograma de frecuencia como se puede observar en la figura 3-2. El 32% de los pacientes se encuentran dentro del intervalo [64–76), el 25% tenían más de 76 años, un 21% está comprendido entre los [52-64) y el restante 22% tenían menos de 52 años de edad.

FIGURA 3-2
HISTOGRAMA DE FRECUENCIA PARA LA EDAD



Como mínimo valor observado tenemos la edad de 31 años y como máximo la edad de 88 años; la media de los datos observados es de 64.69 años con una varianza de 213.182 mostrando una alta variabilidad.

La distribución de esta variable está sesgada negativamente, refiriéndonos con esto a que la mayor parte de los datos se encuentran al lado derecha de la media, esto se puede observar a través del coeficiente del sesgo que es –0.46 y comparando la media con la mediana que es 67, siendo mayor ésta; el coeficiente de Kurtosis también tiene un valor negativo de –0.54 lo cual quiere decir que la distribución tiene una cresta pronunciada hacia la derecha.

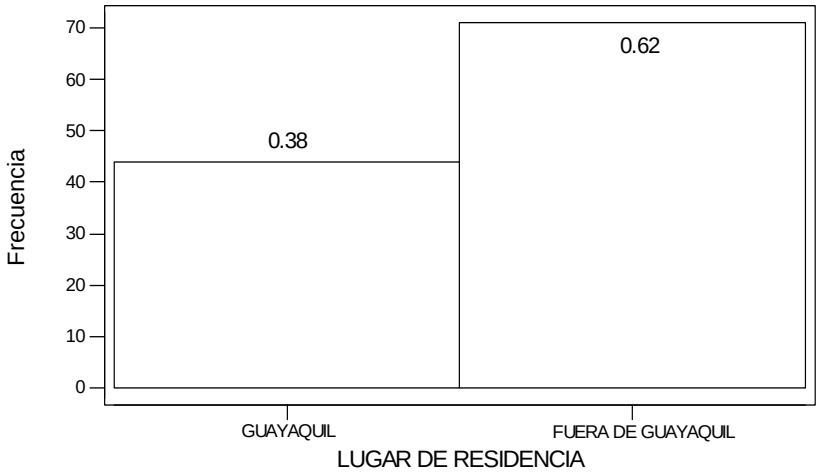
TABLA 3-1
ESTADÍSTICA DESCRIPTIVA DE LA EDAD

	Edad
Mínimo	31
Máximo	88
Rango	57
Mediana	67
Media	64.687
Desviación estándar	14.6007
Varianza	213.182
Sesgo	-0,46
Kurtosis	-0.54
Total	115

Variable # 3: Lugar de residencia

Esta variable nos permite saber el lugar de donde reside habitualmente el paciente, si este lugar se encuentra fuera o dentro de la ciudad de Guayaquil donde funciona SOLCA. Analizando el histograma de frecuencia se observa que el 62% de los pacientes provienen fuera de la ciudad de Guayaquil mientras que solo el 38% reside en esta ciudad.

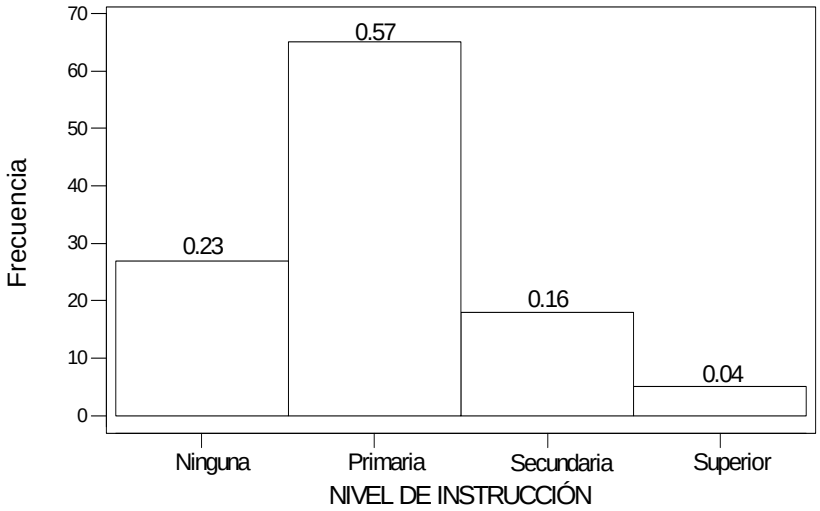
FIGURA 3-3
HISTOGRAMA DE FRECUENCIA PARA EL LUGAR DE RESIDENCIA



Variable # 4: Nivel de Instrucción

La finalidad de esta variable es conocer el grado de educación que tiene el paciente con este tipo de patología. Observando las correspondientes frecuencias en el histograma, se tiene que la mayor parte de los pacientes, el 57%, tan solo posee educación primaria, el 23% no posee ningún tipo de educación, el 16% posee educación secundaria y sólo el 4% educación superior.

FIGURA 3-4
HISTOGRAMA DE FRECUENCIA PARA EL NIVEL DE INSTRUCCIÓN



Como podemos notar la media de los datos observados es 1.008700 lo cual indica que el promedio del nivel de educación obtenido es primario con una moderada variabilidad debido a su varianza que es 0.570099; también se observa por el coeficiente del sesgo, 0.607748, la mayor parte de los datos se ubican al lado izquierdo de la media, además el coeficiente de Kurtosis, 0.445586, muestra que la distribución de los datos es puntiaguda hacia la izquierda.

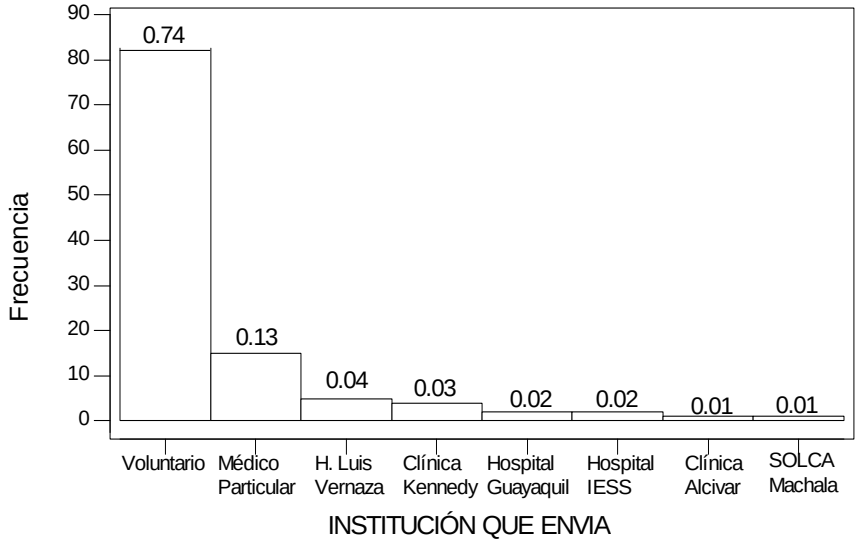
TABLA 3-2
ESTADÍSTICA DESCRIPTIVA DEL NIVEL DE INSTRUCCIÓN

	Nivel
Media	1.00870
Desviación estándar	0.75505
Varianza	0.570099
Sesgo	0.607748
Kurtosis	0.445586
Total	115

Variable # 5: Institución que Envía

Esta variable nos permitirá conocer si el pacientes antes de acudir a SOLCA fue a otra institución de salud. Al ver el histograma de frecuencia podemos observar que el 74% de los pacientes acudieron a SOLCA de forma voluntaria, el 13% de los pacientes fueron remitidos por un médico particular y el restante 13% se encuentra repartido entre el Hospital Vernaza, Clínica Kennedy, Hospital Guayaquil, Hospital del IESS, clínica Alcívar y SOLCA de Machala con 4%, 3%, 2%, 2%, 1% y 1% correspondientemente.

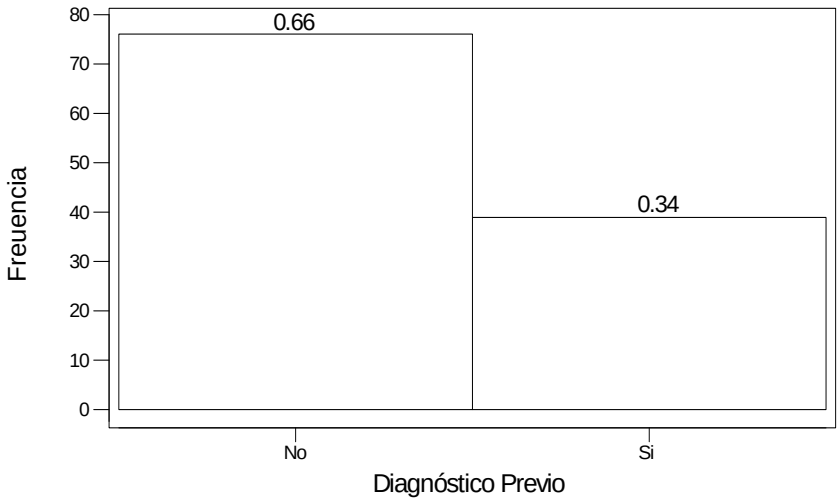
FIGURA 3-5
HISTOGRAMA DE FRECUENCIA DE LA INSTITUCIÓN QUE ENVIA



Variable # 6: Diagnostico previo

Esta variable nos proporcionará información acerca de que si el paciente antes de acudir a SOLCA fue diagnosticado previamente con cáncer de estómago. Por medio del histograma de frecuencia observamos que el 66% de los pacientes acudieron a ésta institución sin ningún diagnóstico previo mientras que el restante de los paciente, 34%, si fueron diagnosticados previamente con éste tipo de cáncer.

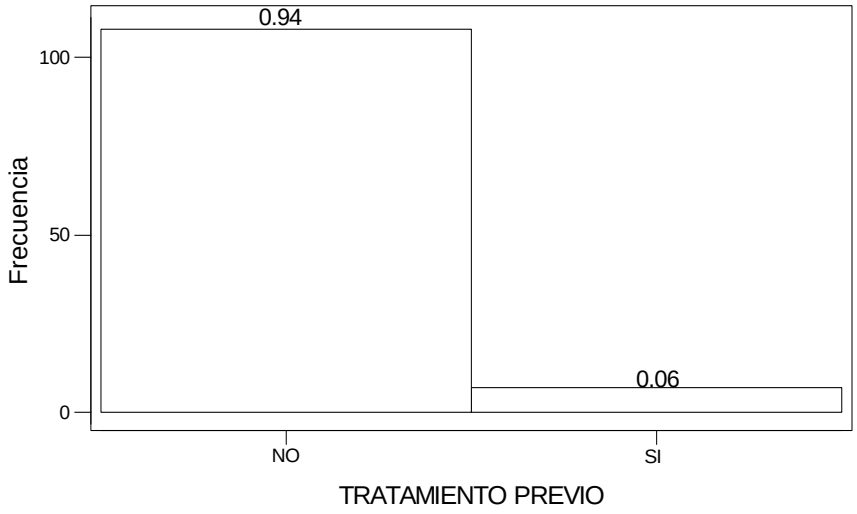
FIGURA 3-6
HISTOGRAMA DE FRECUENCIA DE DIAGNOSTICO PREVIO



Variable # 7: Tratamiento Previo

Esta variable nos permitirá saber si es que el paciente que presenta los síntomas de esta enfermedad recibió un tratamiento previo en alguna institución de salud antes de acudir a SOLCA. Al ver el histograma de frecuencia de esta variable, podemos observar que el 94% de los pacientes que acudieron a SOLCA no tuvieron un tratamiento previo, mientras que el 6% de los pacientes que acudieron a SOLCA manifestaron que si habían recibido un tratamiento previo.

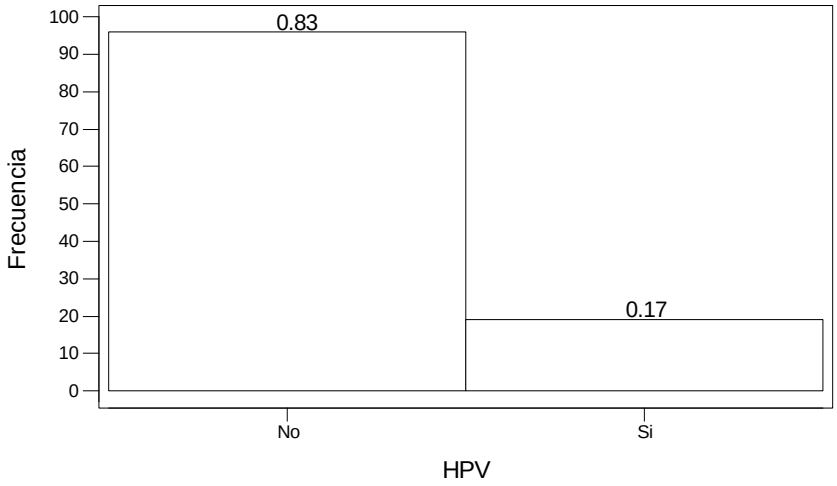
FIGURA 3-7
HISTOGRAMA DE FRECUENCIA DE TRATAMIENTO PREVIO



Variable # 8: HPV

Con esta variable se desea conocer si el paciente contenía ésta bacteria, Hicto Bacto Pílori, en su organismo al momento de ser diagnosticado. Al observar el histograma de frecuencia de esta variable notamos que, el 83% de los pacientes no poseían ésta bacteria en su organismo mientras que el 17% si poseían.

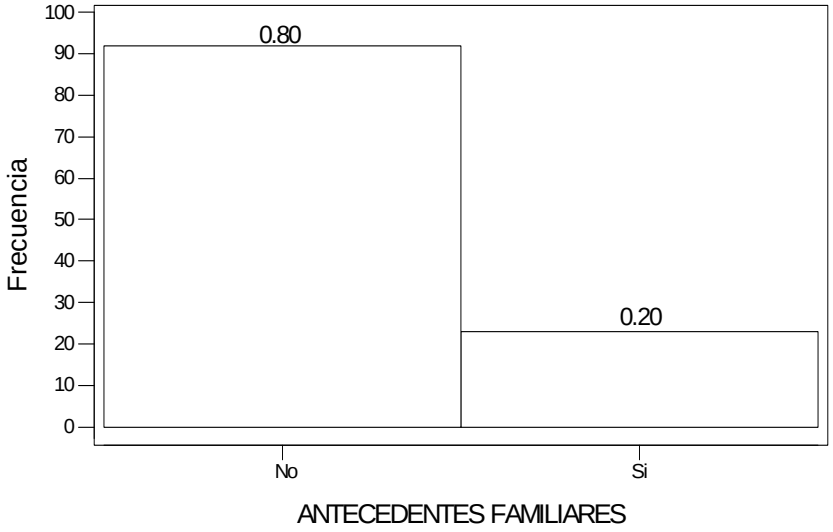
FIGURA 3-8
HISTOGRAMA DE FRECUENCIA DEL HPV



Variable # 9: Antecedentes Familiares.

Con ésta variable se busca información si existe algún familiar que haya sido diagnosticado con ésta enfermedad. Observando el histograma de frecuencias notamos que el 80% de los diagnosticados no tenían ningún familiar que haya padecido de ésta enfermedad a diferencia del 20% que declaró que si había antecedentes de este tipo en su familia.

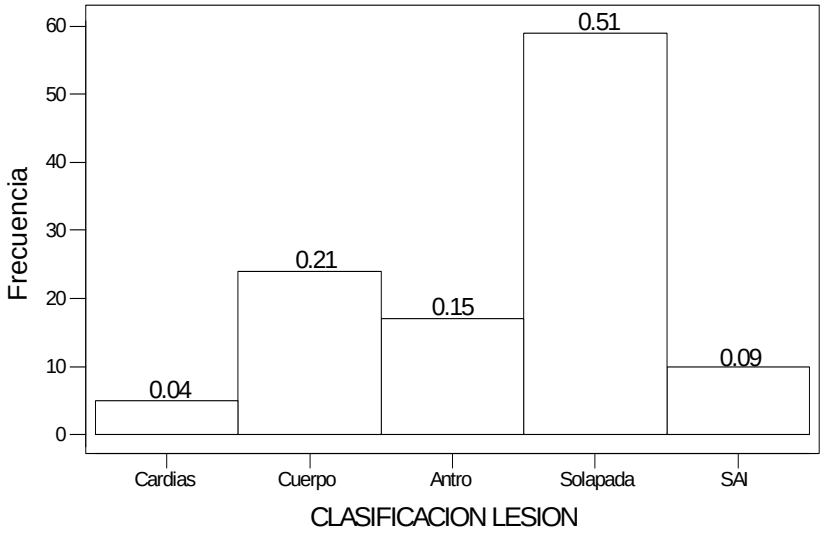
FIGURA 3-9
HISTOGRAMA DE FRECUENCIA DE ANTECEDENTES FAMILIARES



Variable # 10: Clasificación de la lesión.

Esta variable nos permitirá saber el tipo de lugar de la lesión en el estómago que tenía el paciente al momento de ser diagnosticado. Al analizar el histograma de frecuencias observamos que el 4% tiene la lesión en el 1/3 proximal cardias, 21% en el 1/3 medio cuerpo, el 15% en el 1/3 distal antro, 51% fue una lesión solapada, es decir en dos o más lugares y 9% fue clasificada como SAI.

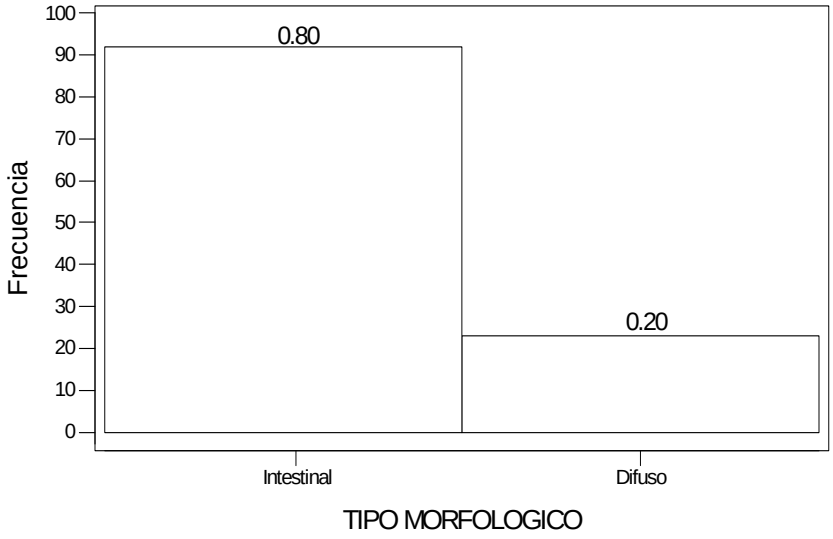
FIGURA 3-10
HISTOGRAMA DE FRECUENCIA DE CLASIFICACIÓN DE LA LESION



Variable # 11: Tipo Morfológico

Esta variable nos permitirá a saber cual es el diagnostico según el tipo de morfología que presenta que presenta el paciente después de los resultados. Observando el histograma de frecuencias notamos que el 80% de los pacientes tiene del tipo morfológico Intestinal mientras que el 20% de los pacientes del tipo Difuso; lo que quiere decir que en su histología constan con Adenocarcinoma de Tipo Intestinal y Adenocarcinoma de Tipo Difuso correspondientemente.

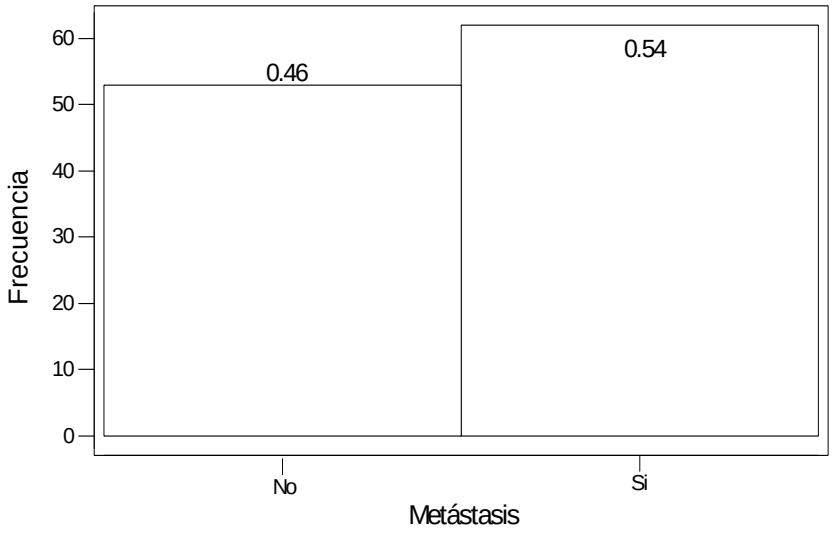
FIGURA 3-11
HISTOGRAMA DE FRECUENCIA DE TIPO MORFOLÓGICO



Variable # 12: M

Esta variable, Metástasis, da información si el cáncer ha invadido a otros órganos o se encuentra en otra parte del cuerpo que fuera del estómago. Al observar su correspondiente histograma de frecuencia se tiene que en el 46% de los pacientes el cáncer no se había esparcido por otras partes del cuerpo mientras que, en el 54% si se había esparcido.

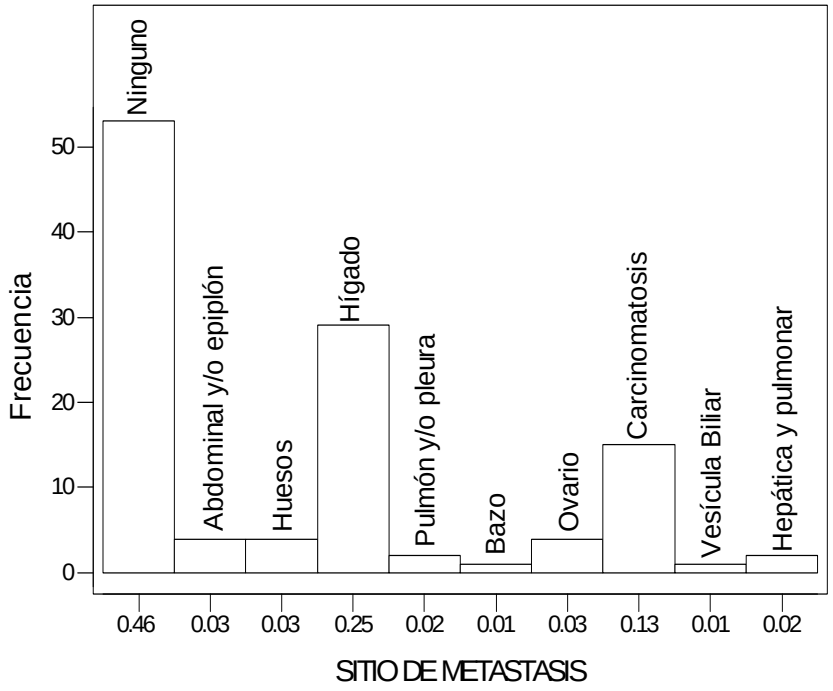
FIGURA 3-12
HISTOGRAMA DE FRECUENCIA DE METÁSTASIS



Variable # 13: Lugar de la metástasis.

Con ésta variable permite conocer el lugar donde el cáncer se ha esparcido. Al analizar el histograma de frecuencias observamos que en el 25% de los pacientes el cáncer se ha esparcido al hígado, el 13% posee metástasis carcinomatosis, 3% en los ovarios, huesos, abdomen y/o epiplón cada uno; pulmón y/o pleura, pulmonar y hepática con el 2% cada una y en el 1% se encontró en la Vesícula Biliar, también en este diagrama se confirma que el 46% de los pacientes no poseían metástasis.

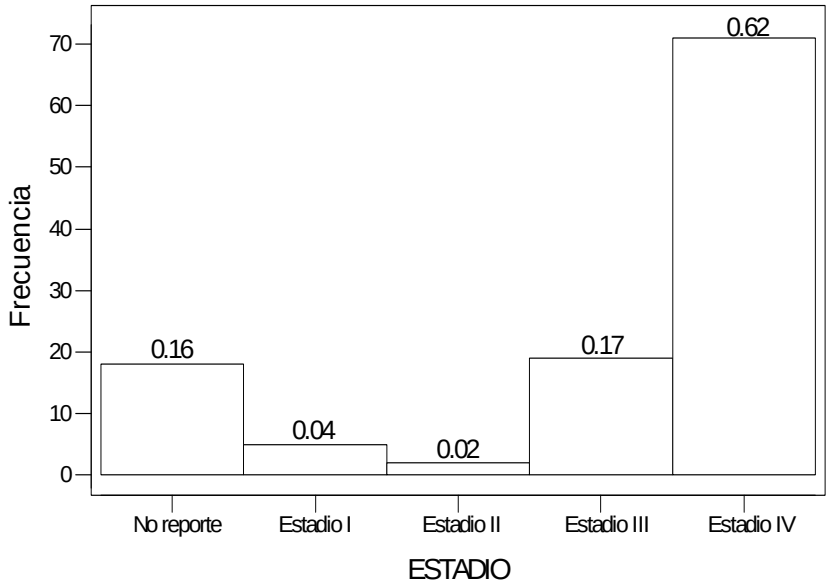
FIGURA 3-13
HISTOGRAMA DE FRECUENCIA DE LUGAR DE LA METÁSTASIS



Variable # 14: Estadio.

Esta variable nos permite identificar en que fase de la enfermedad se encuentran los pacientes. Los resultados que muestra el histograma de frecuencias fueron que en la mayor parte de los pacientes se encontraban en la fase o etapa IV del cáncer con el 62%, en la etapa III el 17%, en la etapa I y en la etapa II el 2%, mientras que el 16% de los pacientes no poseían en el reporte la etapa en que se encontraba su cáncer.

FIGURA 3-14
HISTOGRAMA DE FRECUENCIA DEL ESTADIO



Como podemos ver el promedio de los datos observados es de 3,04348, la varianza de los datos de esta variable es de 2,23494, lo cual indica que existe una moderada variabilidad de los datos.

El coeficiente del sesgo es -1,30987, lo cual significa que la distribución de esta variable esta sesgada negativamente lo cual indica que una gran parte de las observaciones se encuentran hacia la derecha de la media, como podemos observar el valor de la kurtosis es

de 0,054, con lo que podemos afirmar que la distribución de los datos de esta variable es ligeramente puntiaguda a la izquierda.

TABLA 3-3
ESTADÍSTICA DESCRIPTIVA DE LOS ESTADIOS

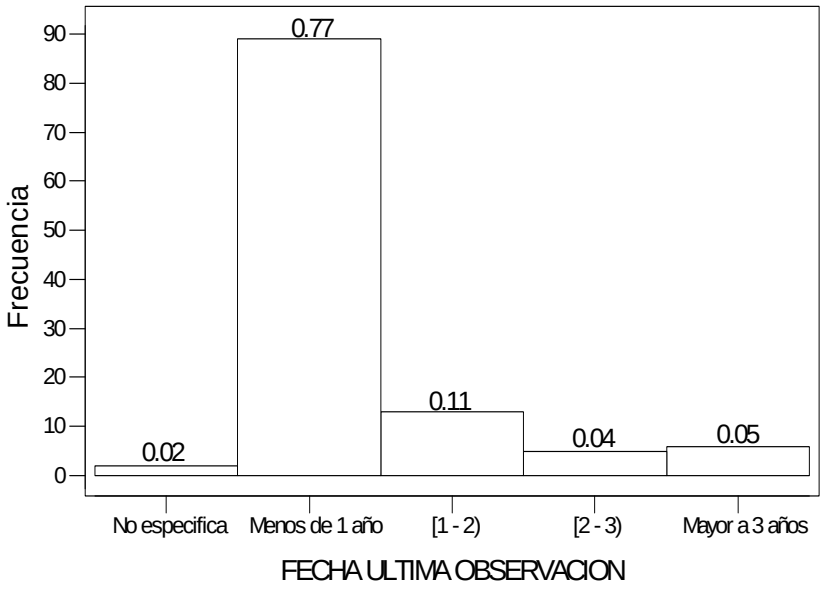
	Estadio
Media	3.04348
Desviación estándar	1.49497
Varianza	2.23494
Sesgo	-1.30987
Kurtosis	0.054
Total	115

Variable # 15: Tiempo de enfermedad.

Con ésta variable se desea conocer el tiempo transcurrido que asistió al hospital de SOLCA desde que el paciente fue diagnosticado con cáncer gástrico hasta el registro de su última observación que puede ser abandono, muerto o vivo en el momento de la toma de los datos. En este caso también se agrupo el tiempo transcurrido. Al analizar el histograma de frecuencias observamos que el 77% de los pacientes asistió a consulta un periodo menor de un año, mientras que el 11% asistió entre un año y dos, el 4% entre dos y tres

años y el 5% asistió más de tres años, y un 2% que solo asistió el día que se le diagnostico.

FIGURA 3-15
HISTOGRAMA DE FRECUENCIA DE TIEMPO DE ENFERMEDAD



Como podemos leer en la tabla esta variable, en días, tiene como mínimo menos de un día y como máximo de asistencia 1415 días por lo cual su rango coincide con el máximo que es de 1415 días, una mediana de 117 días con una media de 234.34 días con varianza de 100,896 días lo cual muestra que esta variable posee una alta variabilidad.

El coeficiente del sesgo es 2.071, lo cual significa que la distribución de esta variable esta sesgada positivamente lo cual indica que una gran parte de las observaciones se encuentran hacia la derecha de la media, como podemos observar el valor de la kurtosis es de 3.83394, con lo que podemos afirmar que la distribución de los datos de esta variable es puntiaguda a la izquierda.

TABLA 3-4
ESTADÍSTICA DESCRIPTIVA DE TIEMPO DE ENFERMEDAD

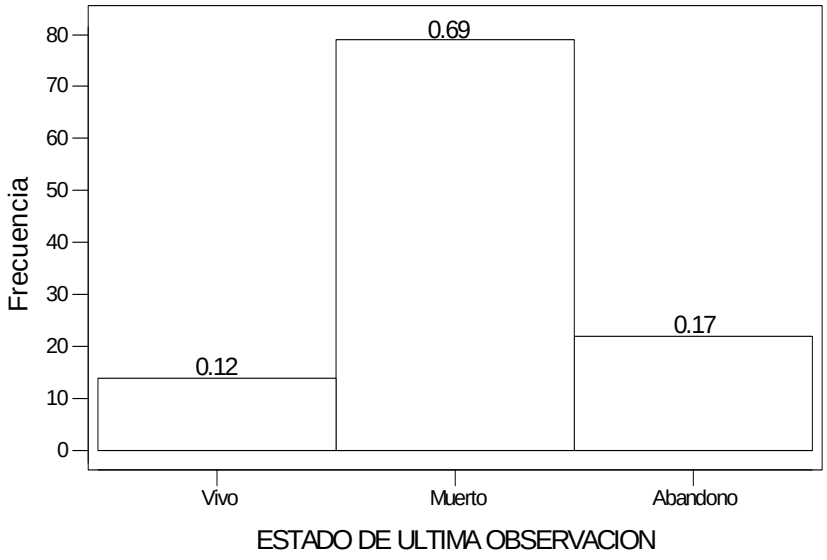
	Fecha
Mínimo	0
Máximo	1415
Rango	1415
Mediana	117
Media	234.348
Desviación estándar	317.641
Varianza	100896
Sesgo	2.07100
Kurtosis	3.83394
Total	115

Variable # 16: Estado de la Ultima Observación.

Esta variable nos permitirá saber el estado físico en que se encontraba el paciente en la última cita. Al analizar el histograma de frecuencias observamos que la mayoría de los pacientes, el 69%, ya habían fallecidos,

el 17% había abandonado el tratamiento o dejo de concurrir al hospital y el 12% se encontraba vivo.

FIGURA 3-16
HISTOGRAMA DE FRECUENCIA DE ESTADO DE LA ÚLTIMA OBSERVACIÓN

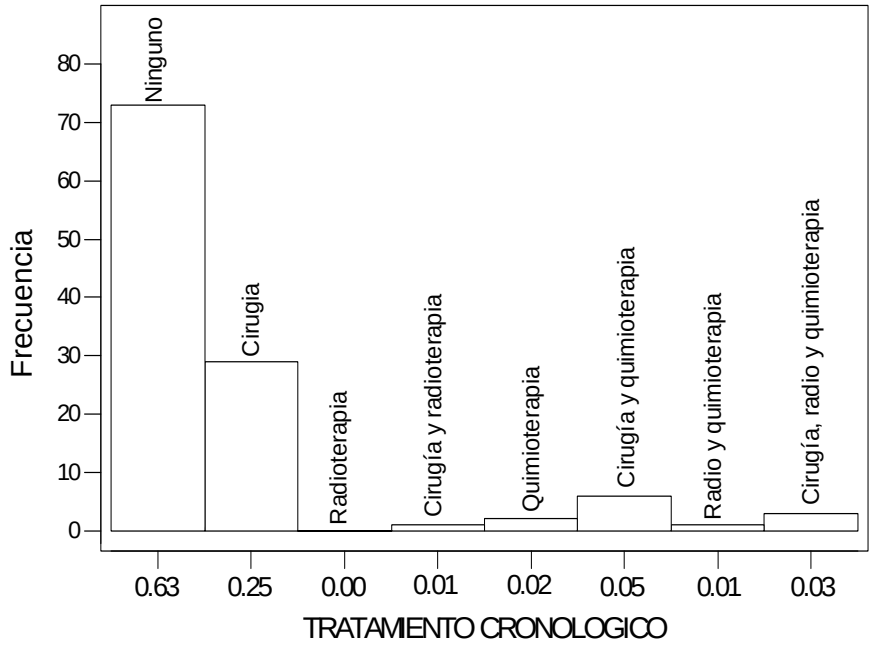


Variable # 17: Tratamiento Cronológico

Esta variable nos permitirá conocer el tipo de tratamiento o tratamientos que recibió el paciente mientras concurrió a SOLCA. Analizando el histograma de frecuencia observamos que el 63% no recibió ningún tipo de tratamiento, el 25% se le realizó cirugía, el 5%

se le realizó cirugía asociada con quimioterapia y el restante se encuentra distribuido entre cirugía, radio y quimioterapia asociadas, quimioterapia, cirugía y radioterapia asociadas y, radio y quimioterapia asociadas.

FIGURA 3-17
HISTOGRAMA DE FRECUENCIA DE TRATAMIENTO CRONOLÓGICO

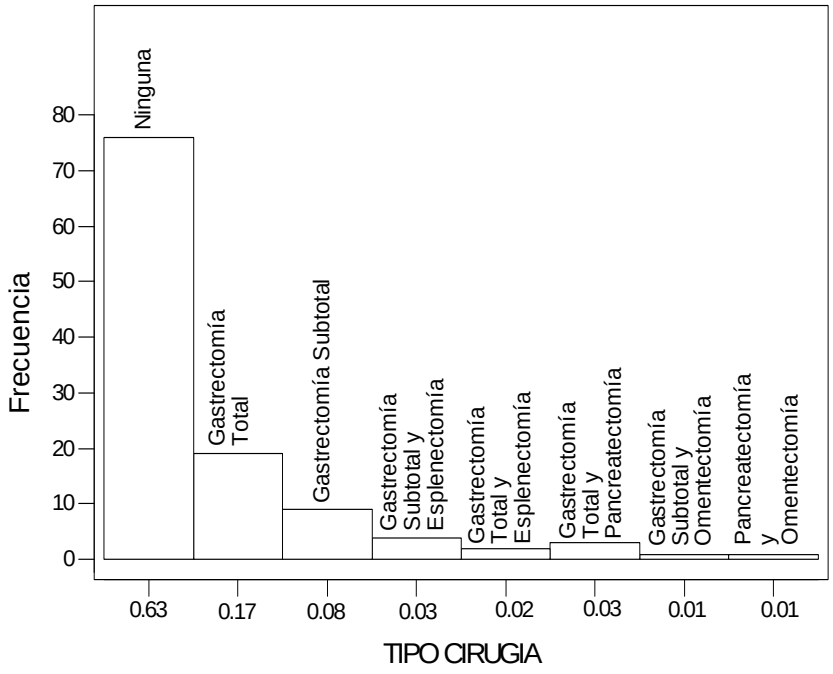


Variable # 18: Tipo de Cirugía

En ésta variable se obtiene información del tipo de cirugía que se le practicó al paciente; los resultados

fueron los siguientes, 17% se le realizo una gastrectomía total, el 8% se le practicó gastrectomía subtotal, con 3% gastrectomía subtotal y esplenectomía asociadas, gastrectomía total y pancreatectomía asociadas cada una, el 2% se le realizó gastrectomía total con esplenectomía y cada intervención quirúrgica con el 1% gastrectomía subtotal asociada con una omentectomía y una pancreatectomía con omentectomía asociada.

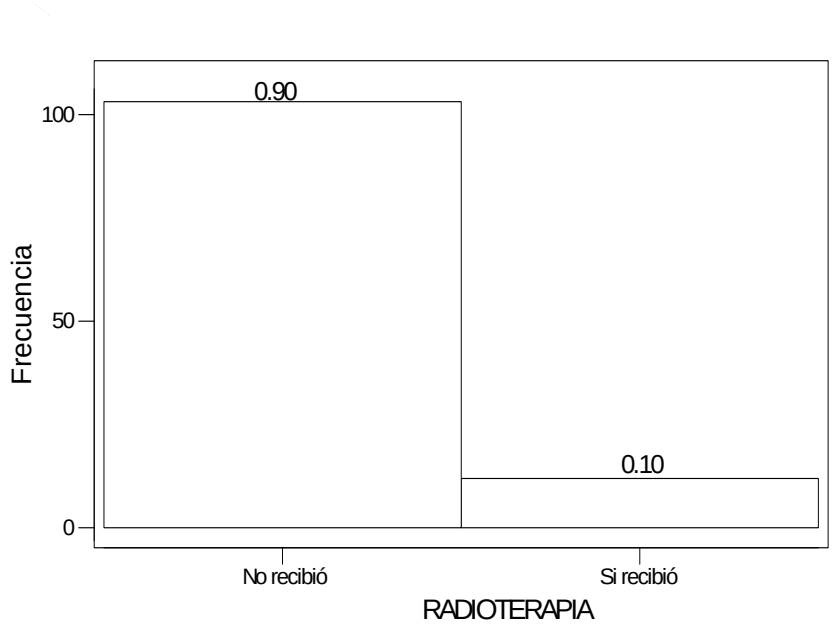
FIGURA 3-18
HISTOGRAMA DE FRECUENCIA DE TIPO DE CIRUGIA



Variable # 19: Recibió Radioterapia

Esta variable brinda información acerca del paciente que presenta cáncer de estómago si recibió como parte de su tratamiento Radioterapia. Al analizar el histograma de frecuencias tenemos que el 90% no recibió este tipo de tratamiento a diferencia del 10% que si recibió Radioterapia para combatir este cáncer.

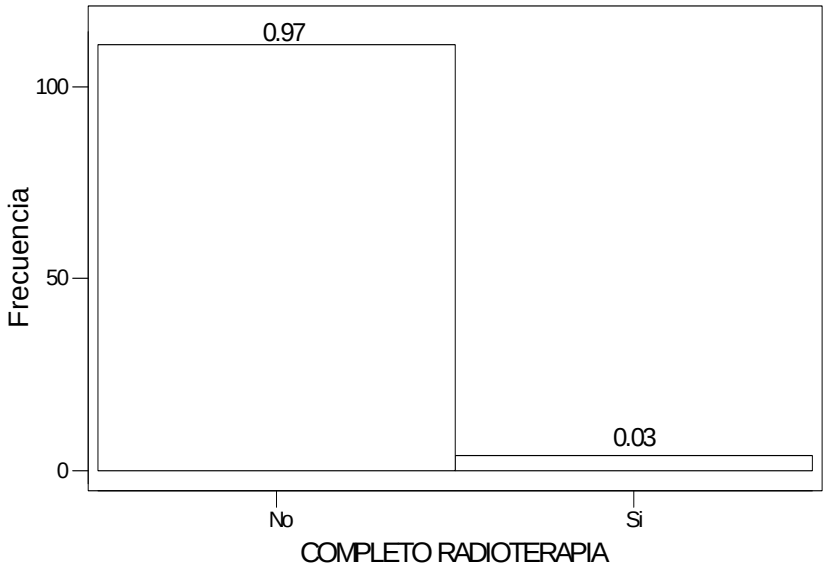
FIGURA 3-19
HISTOGRAMA DE FRECUENCIA DE RECIBIO



Variable # 20: Completo Radioterapia

Esta variable nos permitirá conocer si el paciente completó las dosis de radioterapia las cuales fueron designadas como parte de su tratamiento. Como se puede observar en el histograma de frecuencias el 97% de los pacientes no completó la dosis de radioterapia mientras que sólo el 3% si completó con las dosis de Radioterapia.

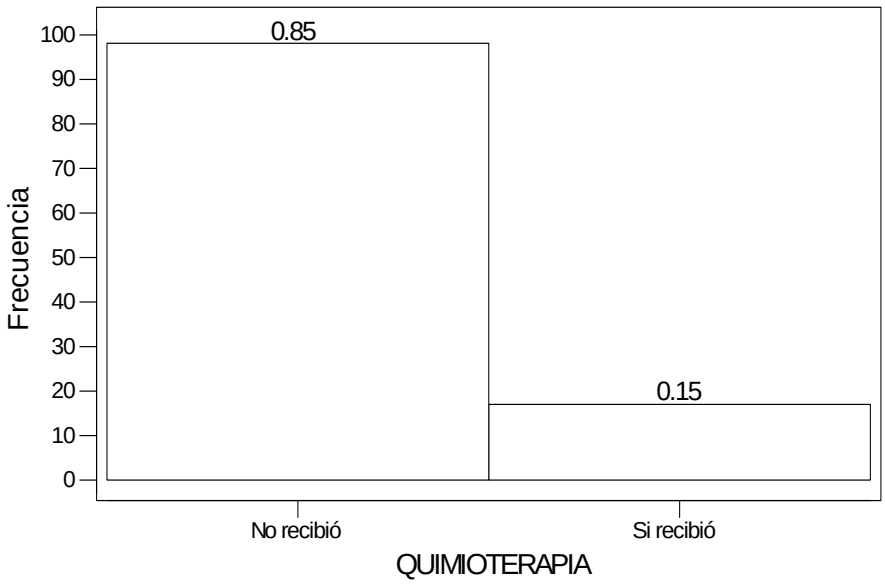
FIGURA 3-20
HISTOGRAMA DE FRECUENCIA DE COMPLETO RADIOTERAPIA



Variable # 21: Recibió Quimioterapia

Esta variable brinda información acerca del paciente que presenta cáncer de estómago recibió como parte de su tratamiento Quimioterapia. Al ver el histograma de frecuencias notamos que la mayoría de los pacientes, el 85%, no recibió ciclos de quimioterapia como parte del tratamiento a diferencias del 15% que si recibió.

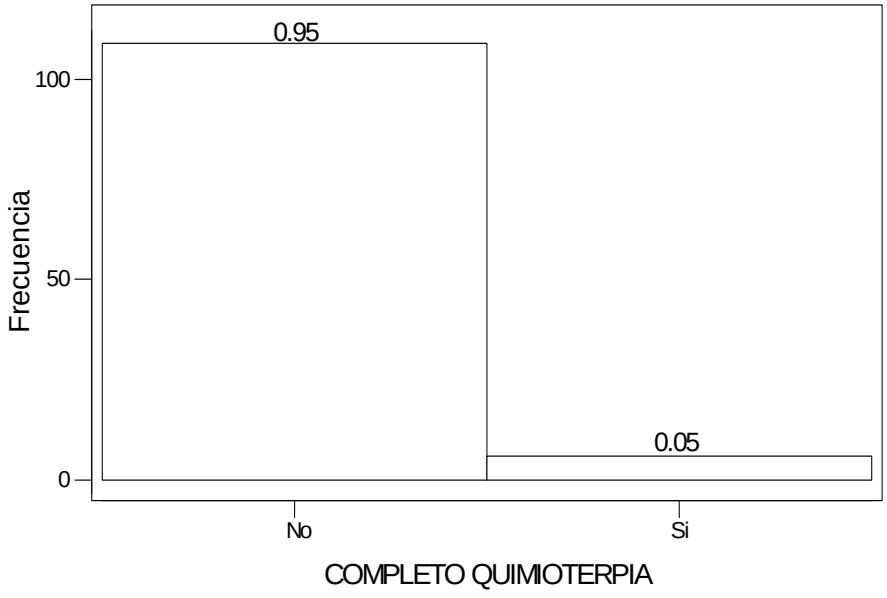
FIGURA 3-21
HISTOGRAMA DE FRECUENCIA DE RECIBIO QUIMIOTERAPIA



Variable # 22: Completo Quimioterapia

Esta variable nos permitirá conocer si el paciente acudió a todas las sesiones de quimioterapia las cuales fueron programadas como parte de su tratamiento. Al analizar el histograma de frecuencias observamos que el 95% de los pacientes no completó todas las sesiones programadas mientras que el 5% si completó los ciclos de quimioterapia programados.

FIGURA 3-22
HISTOGRAMA DE FRECUENCIA DE COMPLETO QUIMIOTERAPIA



3.3 Tablas de contingencias de las variables mas importantes

Para realizar el contraste de las tablas de contingencias, se utilizó las variables que en conjunto el médico guía son las más significativas en el presente estudio.

Las tablas de contingencias sirven para poder analizar la independencia existente entre una variable y otra. A continuación se postula y presenta los resultados de las siguientes hipótesis de las variables escogidas.

Ho: El sexo es independiente del estadio

La prueba Chi-cuadrado, con 4 grados de libertad y un nivel de significancia de 0.881, acepta la hipótesis nula por no haber suficiente evidencia estadística para su rechazo, lo que quiere decir que el sexo y el estadio son variables independientes mutuamente.

TABLA 3-5
TABLA DE CONTINGENCIA DE LAS VARIABLES
SEXO * ESTADIO

Count

	ESTADIO					
	N.R.	1	2	3	4	Total
Masculino	12	3	2	12	46	75
Femenino	6	2		7	25	40
Total	18	5	2	19	71	115

TABLA 3-6
PRUEBA CHI-CUADRADO DE LAS VARIABLES
SEXO * ESTADIO

	Valor	gl	Nivel de significancia
Pearson Chi-Square	1,185	4	,881
Número de casos válidos	115		

Ho: El nivel de instrucción es independiente del estado de la última observación

La prueba Chi-cuadrado, con 6 grados de libertad y un nivel de significancia (2 colas) de 0.418, acepta la hipótesis nula por no haber suficiente evidencia estadística para su rechazo, lo que quiere decir que el nivel de instrucción y el estado de la última observación son variables independientes una de la otra.

TABLA 3-7
TABLA DE CONTINGENCIA DE LAS VARIABLES NIVEL DE INSTRUCCIÓN * ESTADO ÚLTIMA OBSERVACIÓN

Count

	Vivo	Muerto	Abandono	Total
Ninguno	5	15	7	27
Primaria	8	46	11	65
secundaria	1	15	2	18
Superior		3	2	5
Total	14	79	22	115

TABLA 3-8
PRUEBA CHI-CUADRADO DE LAS VARIABLES NIVEL DE INSTRUCCIÓN * ESTADO ÚLTIMA OBSERVACIÓN

	Valor	gl	Nivel de significancia
Pearson Chi-Square	6,047	6	,418
Número de casos válidos	115		

Ho: Tiempo de enfermedad es independiente del estadio

La prueba Chi-cuadrado, con 16 grados de libertad y un nivel de significancia (2 colas) de 0.004, se rechaza la hipótesis nula por haber suficiente evidencia estadística, lo que quiere decir que el tiempo de enfermedad y es estadio son variables dependientes mutuamente.

TABLA 3-9
TABLA DE CONTINGENCIA DE LAS VARIABLES
TIEMPO DE ENFERMEDAD * ESTADIO

	ESTADIO					
	N.R.	1	2	3	4	Total
No especifica					2	2
Menos de 1 año	16	3	1	10	59	89
[1 - 2)	1			5	7	13
[2 - 3)	1			2	2	5
Mayor de 3 años		2	1	2	1	6
Total	18	5	2	19	71	115

TABLA 3-10
PRUEBA CHI-CUADRADO DE LAS VARIABLES
TIEMPO DE ENFERMEDAD * ESTADIO

	Valor	gl	Nivel de significanci
Pearson Chi-Square	34,614	16	,004
Número de casos válidos	115		

Ho: Tiempo de enfermedad es independiente del estado de la última observación

La prueba Chi-cuadrado, con 8 grados de libertad y un nivel de significancia (2 colas) de 0.006, rechaza la hipótesis nula por encontrar suficiente evidencia estadística, lo que quiere decir que el tiempo de enfermedad y el estado de la última observación son variables dependientes una de la otra.

TABLA 3-11
TABLA DE CONTINGENCIA DE LAS VARIABLES TIEMPO DE ENFERMEDAD * ESTADO ÚLTIMA OBSERVACIÓN

Count

	Vivo	Muerto	Abandono	Total
No especifica			2	2
Menos de 1 año	9	63	17	89
[1 - 2)	1	9	3	13
[2 - 3)	3	2		5
Mayor de 3 años	1	5		6
Total	14	79	22	115

TABLA 3-12
PRUEBA CHI-CUADRADO DE LAS VARIABLES TIEMPO DE ENFERMEDAD * ESTADO ÚLTIMA OBSERVACIÓN

	Valor	gl	Nivel de significancia
Pearson Chi-Squar	21,528	8	,006
Número de casos válidos	115		

Ho: Tratamiento cronológico es independiente del estadio

La prueba Chi-cuadrado, con 24 grados de libertad y un nivel de significancia (2 colas) de 0.000, rechaza la hipótesis nula por haber suficiente evidencia estadística, lo que quiere decir que el tratamiento cronológico y el estadio son variables dependientes.

TABLA 3-13
TABLA DE CONTINGENCIA DE LAS VARIABLES TRATAMIENTO * ESTADIO

	No repote	1	2	3	4	Total
Ninguno	18	2			53	73
Cirugía		3	2	14	10	29
Radioterapia					1	1
Cirugía/radioterapia					2	2
Quimioterapia				4	2	6
Cirugía/quimioterapia					1	1
Cirugía/radio/quimio				1	2	3
Total	18	5	2	19	71	115

TABLA 3-14
PRUEBA CHI-CUADRADO DE LAS VARIABLES TRATAMIENTO * ESTADIO

	Valor	gl	Nivel de significancia
Pearson Chi-Square	66,380	24	,000
Número de casos válido	115		

Ho: Edad es independiente del estadio

La prueba Chi-cuadrado, con 16 grados de libertad y un nivel de significancia (2 colas) de 0.000, rechaza la hipótesis nula por haber suficiente evidencia estadística, lo que quiere decir que el tratamiento edad y el estadio son variables dependientes.

TABLA 3-15
TABLA DE CONTINGENCIA DE LAS VARIABLES
EDAD * ESTADIO

	No reporte	1	2	3	4	Total
Menor de 40 años		3	2		3	8
[40 -52)					17	17
[52 - 64)					24	24
[64 - 76)				10	27	37
Mayor de 76 años	18	2		9		29
Total	18	5	2	19	71	115

TABLA 3-16
PRUEBA CHI-CUADRADO DE LAS VARIABLES
EDAD * ESTADIO

	Valor	gl	Nivel de significancia
Pearson Chi-Square	147,584	16	,000
Número de casos válido	115		

3.4 Bondad de Ajuste para las variables más importantes.

Ho: El sexo sigue una Distribución de Poisson

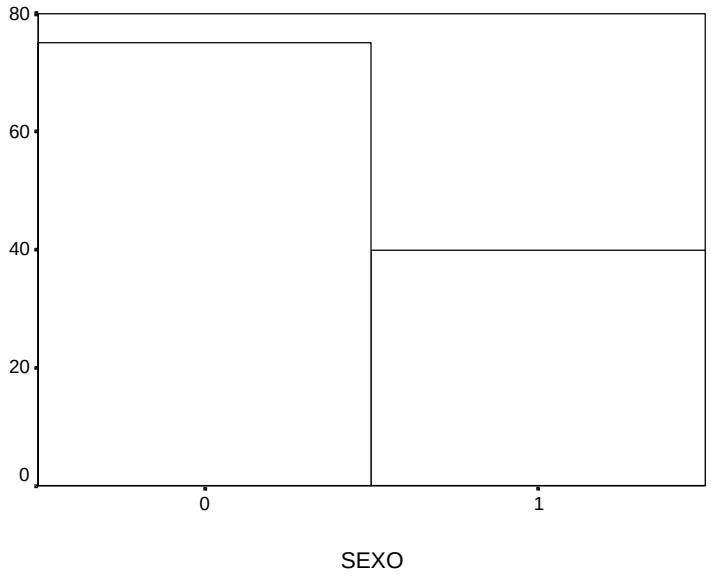
En la tabla 3-17, podemos observar el análisis de bondad de ajuste para la variable Sexo, para lo cual se contrastó la hipótesis nula (Ho: Sexo sigue una

Distribución de Poisson), para la cual se realizo la prueba de Kolmogorov Smirnov la que dio como resultado 0.826, con un nivel de significancia de 0.508 la misma que es mayor que $\alpha = 0.05$ por lo tanto se acepta la hipótesis nula (Ho). Por lo tanto la Edad si sigue una Distribución Normal.

TABLA 3-17
BONDAD DE AJUSTE PARA EL SEXO

		SEXO
N		115
Parametros de Poisson	Media	,35
Diferencias más extremas	Absoluta	,054
	Positiva	,048
	Negativa	-,054
Kolmogorov-Smirnov Z		,580
Nivel de significancia (2 colas)		,890

FIGURA 3-23
HISTOGRAMA DEL SEXO



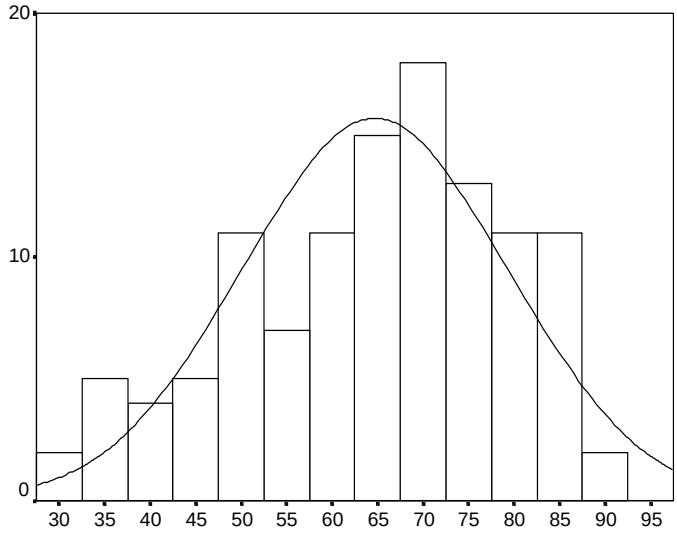
Ho: Edad sigue una Distribución Normal

En la tabla 3-18, podemos observar el análisis de bondad de ajuste para la variable Edad, para lo cual se contrastó la hipótesis nula (Ho: Edad sigue una Distribución Norma), para la cual se realizo la prueba de Kolmogorov Smirnov la que dio como resultado 0.826, con un nivel de significancia de 0.508 la misma que es mayor que $\alpha = 0.05$ por lo tanto se acepta la hipótesis nula (Ho). Por lo tanto la Edad si sigue una Distribución Normal.

TABLA 3-18
BONDAD DE AJUSTE PARA LA EDAD

		EDAD
N		115
Parametros normales	Media	64,69
	Desviación estandar	14,60
Diferencias más extremas	Absoluta	,077
	Positiva	,055
	Negativa	-,077
Kolmogorov-Smirnov Z		,826
Nivel de significancia (2 colas)		,502

FIGURA 3-24
HISTOGRAMA DE LA EDAD



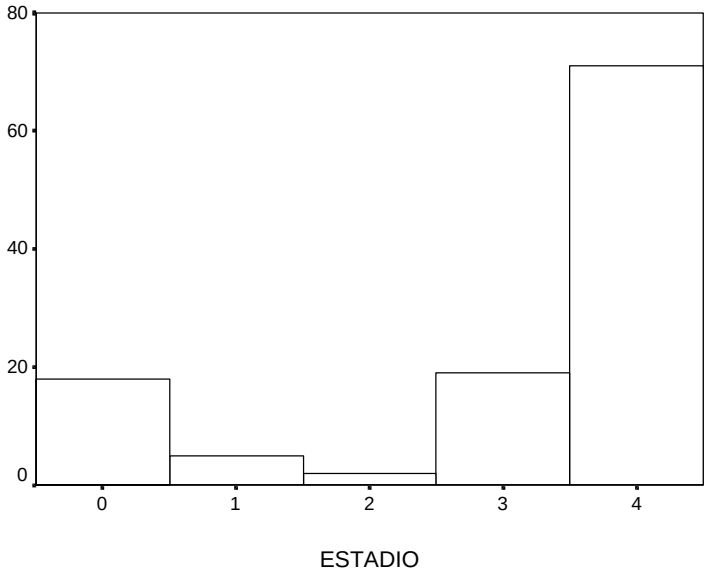
Ho: El Estadio sigue una Distribución Poisson

En la tabla 3-19, podemos observar el análisis de bondad de ajuste para la variable Edad, para lo cual se contrastó la hipótesis nula (Ho: El Estadio sigue una Distribución Poisson), para la cual se realizo la prueba de Kolmogorov Smirnov la que dio como resultado 2.733, con un nivel de significancia de 0.000 la misma que es menor que $\alpha = 0.05$ por lo tanto se rechaza la hipótesis nula (Ho). Por lo tanto el Estadio no si sigue una Distribución de Poisson.

TABLA 3-19
BONDAD DE AJUSTE PARA EL ESTADIO

		ESTADIO
N		115
Parametros de Poisson	Media	3,04
Diferencias más extremas	Absoluta	,255
	Positiva	,192
	Negativa	-,255
Kolmogorov-Smirnov Z		2,733
Nivel de significancia (2 colas)		,000

FIGURA 3-25
HISTOGRAMA DEL ESTADIO



IV. ANÁLISIS ESTADÍSTICO MULTIVARIADO

4.1. Introducción

En este capítulo presentaremos el Análisis Estadístico Multivariado, para ello empleamos la técnica de reducción de datos, denominada *componentes principales,* el mismo que tiene como objetivo principal conocer cuales son las variables que tienen un mayor peso dentro del análisis es decir contribuyen con mayor información en el mismo. Luego de realizar el análisis anteriormente mencionado procederemos a realizar el denominado *análisis de sobrevivida*. El método de componente principales consiste en generar variables artificiales en términos o partir de variables aleatorias

originales, para conocer cuales son las variables que influyen más en el análisis. Para lo cual podemos utilizar la matriz de varianzas y Covarianzas (denotada por Σ) o la matriz de correlación (denotada por ρ) de los datos observados.

Para realizar el Análisis Estadístico Multivariado nos ayudamos del paquete estadístico SPSS versión 10 para Windows. Valiéndonos de este paquete obtenemos en primer lugar la matriz de varianzas y covarianzas, posteriormente utilizamos el método de reducción de datos. Para generar las componentes principales emplearemos solamente las variables cuantitativas que se describen a continuación.

Para el análisis de componentes principales, consideramos que las variables en las que debemos basar nuestra investigación son:

- *Variable # 2*: Edad
- *Variable # 4*: Nivel de Instrucción
- *Variable # 14: Estadio*
- *Variable # 15: Tiempo de enfermedad*

Para determinar el número óptimo de componentes principales que debemos retener consideraremos los métodos 1, 2, y 4 descritos en el capitulo 2[1] del presente trabajo.

Para el análisis de sobrevivida consideramos junto con el experto que las variables, en las que debemos basar nuestro análisis son:

- Sexo
- Edad
- Nivel de instrucción
- Tiempo de Enfermedad
- Tratamiento cronológico
- Estado de Última Observación (fallecido, vivo o abandono)
- Estadio

[1] Páginas 61 -62

Cabe destacar que el Tiempo de enfermedad se lo obtiene de la siguiente forma:

(Fecha de última observación – Fecha de diagnóstico)

4.2. Análisis Estadístico Multivariado de las variables observadas

La tabla 4-1, contiene los valores de la matriz de varianzas y covarianzas, de los datos originales, estos valores representan la relación lineal entre las variables; como se puede observar existe una alta relación inversa entre las variables Edad y Tiempo de Enfermedad también observamos que la variable Tiempo de Enfermedad obtuvo un alto valor de su varianza así como la Edad, por lo tanto estas variables tienen un alto peso en esta matriz.

TABLA 4-1
MATRIZ DE VARIANZAS Y COVARIANZAS

Variables	Edad	Nivel de Instr.	Estadio	Tiempo
Edad	213,182	-0,806	-5,291	-134,289
Nivel de Instr.	-0,806	0,565	0,191	-26,256
Estadio	-5,291	0,191	2,216	-35,773
Tiempo	-134,289	-26,256	-35,773	100002,236

La tabla 4-2, contiene los valores propios de la matriz de varianzas y covarianzas, de los datos originales, estos valores representan la varianza de las componentes principales. También consta el porcentaje de explicación que le corresponde a las componentes principales. Adicionalmente encontramos el total del porcentaje de explicación.

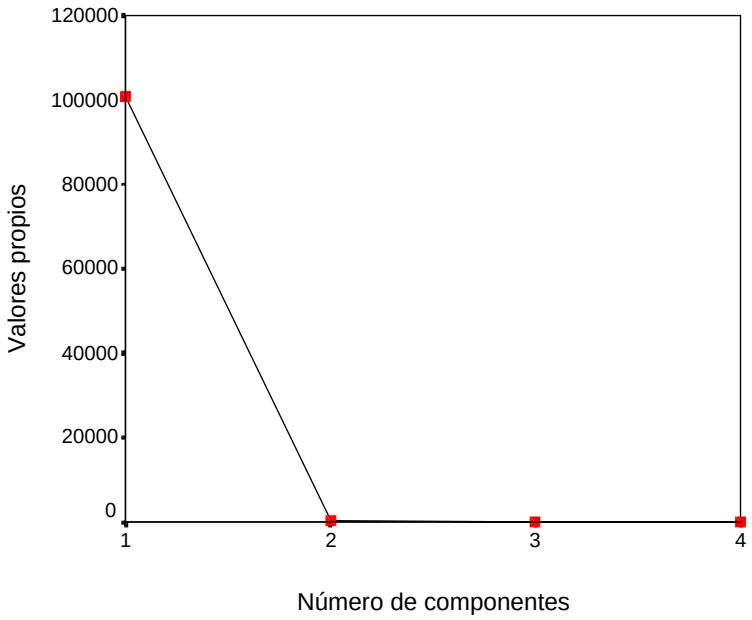
Notamos que la primera componente principal tiene 99,767% del total de la varianza, lo que nos indica que utilizar la primera componente obtendremos el 99,767% del total de la información. La segunda componente principal nos proporciona 0,211% de la información total la misma que junto con la primera componente principal nos explica el 99,998% del total de la información.

TABLA 4-2
PORCENTAJE DE EXPLICACION DE LAS COMPONENTES

λ_i	**Varianza**	**% de explicación**	**Total % de explicación**
λ_1	100896,1	99,787	99,767
λ_2	213,139	0,211	99,998
λ_3	2,101	2,078E-03	99,999
λ_4	0,543	5,368E-04	100,000

La figura 4-1 muestra los valores propios de la matriz de varianzas y covarianzas. Esta figura nos permite determinar el número óptimo de componentes por medio del criterio de las raíces latentes, al utilizar este criterio resulta que es óptimo trabajar solamente con la primera componente, ya que luego del primer valor propio, es decir, el que corresponde a la primera componente, se observa un descenso bien pronunciado.

FIGURA 4-1
CRITERIO DE LAS RAICES LATENTES



Al trabajar con el primer método para la retención óptima de componentes principales, debemos trabajar con 3 componentes, ya que según este método se debe retener aquellas componentes cuyas varianzas sean mayores que 1, sin embargo al trabajar con las dos primeras componentes se obtiene 99,998% del total de la información, y logramos disminuir significativamente el número de variables.

Por consiguiente el número óptimo de componentes que debemos retener será dos, es decir, retener solamente las dos primeras componentes principales.

La primera componente principal será:

Y_1 = -0,029Edad - 0,111Nivel de instrucción –

0,076Estadio + 1Tiempode enfermedad

La variable que predomina en esta componente es la variable 15, es decir ,el Tiempo de enfermedad.

La segunda componente principal será:

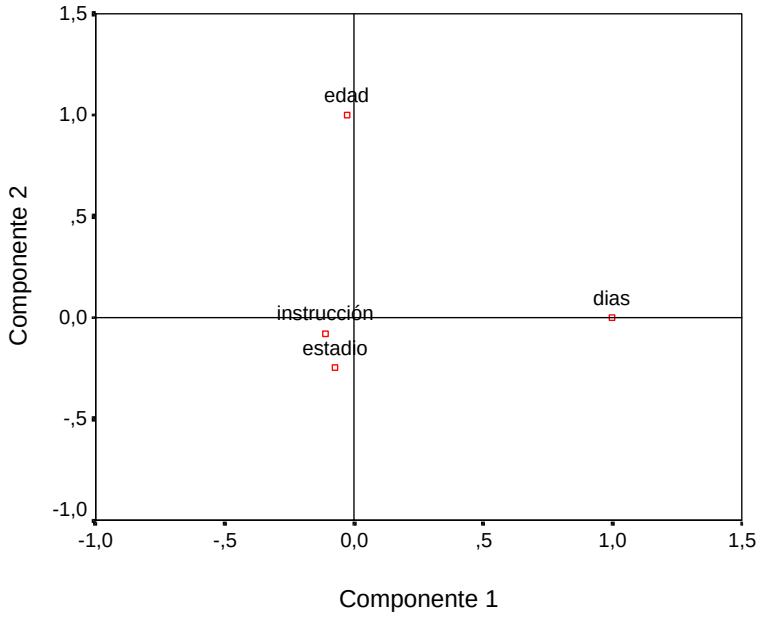
Y_2 = 1Edad – 0,078Nivel de instrucción –0,249Estadio +

0,000Tiempo de enfermedad

La variable que predomina en esta componente es la variable 2, es decir , La Edad del paciente.

Por lo tanto la primera componente principal será llamada Tiempo de Enfermedad dado que esta variable es la que aporta con mayor información a esta componente y la segunda componente principal será la Edad por la misma razón.

FIGURA 4-2
COMPONENTE 1 VS COMPONENTE 2

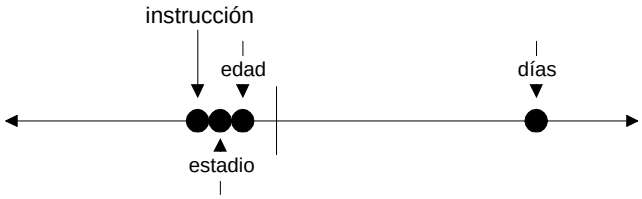


En la figura 4-2, podemos observar que la variable Tiempo de enfermedad (o días en la figura 4-2) y la variable Edad son las variables que mas pesan dentro del modelo y deberían ser la primeras en considerarse a la hora de realizar un análisis, también podemos

observar que éstas dos variables no están relacionadas pues casi son ortogonales entre sí.

Puesto que el segundo eje (o componente principal) tiene apenas un 0.2% de información, podemos considerar sólo el primer eje (o la primera componente principal). Así vamos a proyectar todas las variables en el primer eje, obteniéndose:

FIGURA 4-3
COMPONENTE 1



Vemos que se forman dos grupos; en el primero formado por las variables Nivel de Instrucción, Estadio y Edad del paciente y el segundo solo por el Tiempo de Enfermedad (o días en el gráfico 4-3).

Debido a que la variable Tiempo de Enfermedad (o días) es la que más pesa en el primer eje, se puede utilizar en diferentes modelos estadísticos ésta variable

en representación de las cuatro variables cuantitativas con una pérdida de información despreciable.

4.3. Análisis de sobrevida de las variables observadas por medio del modelo de Regresión de Cox.

Para realizar el modelo de Regresión de Cox plantearemos dos enfoques. En el primero se recodifican las variables Nivel de Instrucción, Tratamiento Cronológico y Estadio en variables dicotómicas y en el segundo enfoque se recodifican las variables Nivel de Instrucción y Estadio usando una escala lickert y Tratamiento Cronológico en una variable dicotómica, mientras que el resto de las variables dentro del modelo permanecen en su codificación original. Al final se hace una comparativa entre los resultados obtenidos mediante estos dos enfoques.

4.3.1. Regresión de Cox utilizando variables dicotómicas

Para realizar el modelo de Regresión de Cox se recodificó las variables en variables dicotómicas como lo recomienda el modelo, el mismo que expresa lo siguiente: Si entre las variables independientes, se encuentran alguna variable cualitativa, sus valores numéricos que corresponden en algún sentido a las categorías originales. 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.

Como podemos observar las variables Edad y Tiempo de enfermedad son variables cuantitativas, por lo tanto estas son utilizadas en su forma original, la variable Sexo es una variable de dos categorías, mientras que las variables Nivel de Instrucción, Tratamiento cronológico y Estadio tienen mas de dos categorías, por lo tanto estas deben ser recodificadas mediante variables dicotómicas.

1. Nivel de Instrucción

	N1	N2	N3
Ninguno	0	0	0
Primario	1	0	0
Secundario	0	1	0
Superior	0	0	1

2. Tratamiento cronológico

	T1	T2	T3	T4	T5	T6	T7
Ninguno	0	0	0	0	0	0	0

Cirugía	1	0	0	0	0	0	0
Radioterapia	0	1	0	0	0	0	0
Cirugía / radioterapia	0	0	1	0	0	0	0
Quimioterapia	0	0	0	1	0	0	0
Cirugía / quimioterapia	0	0	0	0	1	0	0
Quimio / Radioterapia	0	0	0	0	0	1	0
Cirugía/Radio/Quimio	0	0	0	0	0	0	1

3. Estadio

	E1	E2	E3	E4
No Reporte	0	0	0	0
Estadio 1	1	0	0	0
Estadio 2	0	1	0	0
Estadio 3	0	0	1	0
Estadio 4	0	0	0	1

La tabla 4-3, nos 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) requerido fue de 79 que equivale al 68.7% de los datos

analizados, debido que durante el tiempo de observación de los pacientes se encontraron que fueron 79 los decesos, los datos censurados fueron 36 que equivale al 31.34% de los datos analizados estos son aquellos que seguían vivo o habían abandonado el tratamiento durante el periodo de observación, además podemos observar que no existen casos que fueron excluidos como casos con valores perdidos, casos con tiempo no positivo o casos censurados antes del Evento más temprano, el total de datos analizados fue de 115 lo que corresponde al número de pacientes en el estudio.

TABLA 4-3
RESUMEN DEL PROCESO DE CASOS

		N	Porcentaje
Casos disponibles en el análisis	Evento[a]	79	68,7%
	Censurados	36	31,3%
	Total	115	100,0%
Casos excluidos	Casos con valores perdidos	0	,0%
	Casos con tiempo no positivo	0	,0%
	Casos censurados antes del evento más temprano en el estrato	0	,0%
	Total	0	,0%
Total		115	100,0%

a. Variable Dependiente: TIEMPO

Como se puede observar en el bloque 0 de la regresión, el cual se encuentra la tabla 4-4, se obtuvo mediante la Prueba Bondad de Ajuste el coeficiente del modelo, el mismo que se realiza mediante el estadístico -2LL (-2 Log de verosimilitud), el cual dio como resultado el valor de 592.100, la razón de la existencia de este valor radica en que el modelo se está validando es el correspondiente a la función h(t / X).

TABLA 4-4
PRUEBA BONDAD DE AJUSTE SOBRE EL COEFICIENTE DEL MODELO

-2 Log de verosimilitud
592,100

Analizando el bloque 1, se obtiene la tabla 4-5, en la cual constan, el valor del coeficiente del modelo que se lo obtiene mediante el estadístico –2LL, el cual en este paso dio como resultado 550.072, como se puede comparar con el resultado mostrado en la tabla 4-3 este presenta un valor diferente, la razón a este cambio es que en este paso se evalúa la mejora del modelo al incorporar el Tiempo de Sobrevida. También la tabla nos muestra la comparación entre el análisis global, en

el mismo paso y con el bloque anterior o paso 0. En el análisis global se obtuvo el valor del estadístico Chi-cuadrado de 40.164, con 15 grados de libertad y con un nivel de significancia de 0.00 el cual es menor a 0.05, queriendo decir que el cambio fue estadísticamente ligeramente significativo. Para el Cambio desde el paso anterior se obtiene que el estadístico Chi-cuadrado es igual a 42.028, con 15 grados de libertad y con un nivel de significancia de 0.00 el mismo que es menor a 0.05, entonces se puede decir que el cambio fue estadísticamente significativo, y para el Cambio desde el bloque anterior se obtuvo los mismo valores tanto para el estadístico Chi-cuadrado, como para los grados de libertad y el nivel de significancia, lo que refleja lo dicho anteriormente.

TABLA 4-5
PRUEBA BONDAD DE AJUSTE SOBRE LOS COEFICIENTES DEL MODELO

	Global (puntuación)			Cambios desde el paso anterior			Cambios desde el bloque anterior		
-2 Log de verosimilitud	Chi-cuadrado	gl	Sig.	Chi-cuadrado	gl	Sig.	Chi-cuadrado	gl	Sig.
550,072	40,164	15	,000	42,028	15	,000	42,028	15	,000

Si bien el valor de –2LL no es cercano a cero (550,072) como se desease que fuera, en cual el modelo se ajusta perfectamente a los datos; sin embargo, el modelo pasa la prueba χ^2.

La tabla 4-6a y 4-6b, muestra la Tabla de Sobrevida en la cual consta el tiempo de todos los Eventos (muerte) que han ocurrido en el análisis los mismos que son iguales a 79 decesos, los valores estimados de la función de Sobrevida, evaluada sobre las medias de las variables independientes. Dado que en el instante t = 1, cuyo valor estimado de la función de Sobrevida es 0.994 el mismo que es próximo a 1 se produce el primer deceso, es decir que la probabilidad de no fallecer al día 1 después de ser diagnosticado es de 99.4%, así mismo siguiendo analizando la tabla, la probabilidad de no fallecer a los 110 días después del diagnostico es del 69.3%, a los 202 días es de 49.3%, a los 611 días es de 12.9% y a los 1210 días es del 0.1%.

TABLA 4-6a
TABLA DE SOBREVIDA

Tiempo	Evento acumulado	Impacto acumulado línea base	En la media de las covariables		
			Sobrevida	Error Estándar	Impacto Acumulado
1	1	0,005	0,994	0,018	0,006
5	2	0,011	0,988	0,036	0,013
6	3	0,017	0,981	0,054	0,019
7	5	0,029	0,968	0,090	0,032
8	6	0,035	0,961	0,108	0,039
9	7	0,041	0,955	0,126	0,046
11	8	0,048	0,948	0,145	0,054
16	9	0,054	0,941	0,164	0,061
18	10	0,061	0,933	0,182	0,069
21	11	0,068	0,926	0,201	0,077
22	12	0,075	0,919	0,220	0,085
23	13	0,082	0,911	0,239	0,093
24	14	0,090	0,904	0,259	0,101
26	15	0,097	0,896	0,278	0,110
27	16	0,105	0,888	0,299	0,119
33	18	0,122	0,871	0,399	0,138
34	20	0,139	0,855	0,379	0,157
36	21	0,148	0,846	0,399	0,167
37	22	0,157	0,837	0,419	0,178
38	23	0,167	0,828	0,440	0,189
39	24	0,176	0,819	0,460	0,199
40	25	0,186	0,810	0,480	0,210
44	26	0,196	0,801	0,500	0,221
47	27	0,206	0,792	0,520	0,233
50	28	0,216	0,783	0,540	0,245
51	29	0,227	0,773	0,560	0,257
63	30	0,239	0,764	0,580	0,270
67	31	0,250	0,754	0,600	0,282
70	32	0,261	0,744	0,619	0,295
74	33	0,273	0,734	0,639	0,309
83	34	0,286	0,724	0,658	0,323
89	35	0,298	0,714	0,677	0,337
94	36	0,311	0,704	0,697	0,352
110	37	0,325	0,693	0,716	0,367
117	38	0,339	0,682	0,735	0,383
121	39	0,353	0,671	0,753	0,399
123	40	0,368	0,660	0,772	0,416
145	41	0,384	0,648	0,791	0,434
149	42	0,401	0,636	0,810	0,453
150	43	0,419	0,623	0,828	0,474

TABLA 4-6b
TABLA DE SOBREVIDA

Tiempo	Evento acumulado	Impacto acumulado línea base	En la media de las covariables		
			Sobrevida	Error Estándar	Impacto Acumulado
162	44	0,439	0,609	0,847	0,495
164	45	0,459	0,596	0,866	0,518
169	47	0,502	0,567	0,901	0,567
184	48	0,525	0,553	0,917	0,593
188	49	0,549	0,538	0,933	0,621
194	50	0,574	0,523	0,949	0,649
198	51	0,600	0,508	0,963	0,678
202	52	0,626	0,493	0,976	0,708
221	53	0,653	0,478	0,988	0,738
245	54	0,680	0,464	0,998	0,769
258	55	0,709	0,449	1,007	0,801
261	56	0,740	0,434	1,014	0,836
262	57	0,772	0,418	1,019	0,872
283	58	0,808	0,402	1,023	0,912
296	59	0,846	0,384	1,025	0,956
300	60	0,890	0,366	1,024	1,005
312	61	0,938	0,346	1,020	1,060
314	62	0,990	0,327	1,016	1,119
324	63	1,043	0,308	1,009	1,179
355	64	1,101	0,288	0,996	1,243
369	65	1,163	0,269	0,981	1,314
393	66	1,231	0,249	0,962	1,391
442	67	1,329	0,223	0,929	1,501
479	68	1,434	0,198	0,890	1,620
491	69	1,548	0,174	0,843	1,749
509	70	1,672	0,151	0,791	1,889
611	71	1,810	0,129	0,732	2,044
620	72	1,959	0,109	0,669	2,213
845	73	2,161	0,087	0,578	2,441
867	74	2,464	0,062	0,461	2,784
1083	75	2,986	0,034	0,308	3,373
1159	76	3,682	0,016	0,170	4,160
1175	77	4,640	0,005	0,072	5,242
1210	78	5,853	0,001	0,023	6,612
1349	79	7,730	0,000	0,003	8,733

En la tabla 4-7 se muestra todas las variables que participan en la ecuación, las cuales son: Sexo, cuyo estimación del coeficiente β = 0.243, con un Error estándar de 0.277,el estadístico de Wald es de 0.769 con 1 grado de libertad y con un nivel de significancia de 0.380 y el Exp(B=0.243)=1.275; la variable Edad con una estimación de su coeficiente β = 0.006, con un Error estándar de 0.009 y el estadístico de Wald es igual a 0.426, con 1 grado de libertad y con un nivel de significancia de 0.161 y el Exp(B=0.006)=1.006. Además tenemos el nivel de instrucción la cual está codificada de la siguiente forma, N1, N2 y N3; para N1 la estimación del coeficiente fue β = -0.512, con un Error estándar de 0.365 el estadístico de Wald es de 1.966, con 1 grado de libertad y con un nivel de significancia de 0.161, además tenemos que Exp(-0.512) = 0.599, para N2 la estimación del coeficiente fue β = -0.270, con un Error estándar de 0.438 el estadístico de Wald es de 0.381, con 1 grado de libertad y con un nivel de significancia de 0.537, además tenemos que Exp(-0.270) = 0.763 y, para N3 la

estimación del coeficiente fue β = -0.459, con un Error estándar de 0.755 su estadístico de Wald es de 0.369 con 1 grado de libertad y con un nivel de significancia de 0.543, además tenemos que Exp(-0.459) = 0.632.

El Estadio que se encuentra subdividido en E1, E2, E3 y E4. La estimación para E1 de β = 0.141, con un Error estándar de 0.848, el estadístico de Wald es igual a 0.028 con 1 grado de libertad y con un nivel de significancia de 0.868, además tenemos que Exp(0.141) = 1.152, la estimación para E2 de β = -0.172, con un Error estándar de 1.213, el estadístico de Wald es 0.020 con 1 grado de libertad y con un nivel de significancia de 0.887, además tenemos que Exp(-0.172) = 0.842; la estimación para E3 de β = 0.582, con un Error estándar de 0.647 su estadístico de Wald de 0.809, con 1 grado de libertad y con un nivel de significancia de 0.368, además tenemos que Exp(0.582) = 1.789 y finalmente para E4 la estimación del coeficiente fue β = 0.866, con un Error estándar de 0.407 el estadístico de Wald es de 4.524 con 1 grado de

libertad y con un nivel de significancia de 0.033, además tenemos que Exp(0.866) = 2.377.

Y finalmente la variable Tratamiento que fue codificada de la siguiente forma: T1, T2, T3, T4, T5, T6 y T7. Para para T1 la estimación del coeficiente fue β = -1.414, con un Error estándar de 0.486 su estadístico de Wald es de 8.452 con 1 grado de libertad y con un nivel de significancia de 0.004, además tenemos que Exp(-1.414) = 0.243, para T2 no se presentó valores estimados debido a que ningún paciente del estudio recibió este tipo de tratamiento; para T3 la estimación del coeficiente β = -0.983 con un Error estándar de 1.038 el estadístico de Wald es 0.897 con 1 grado de libertad y con un nivel de significancia de 0.344, además tenemos que Exp(-0.983) = 0.374; para T4 la estimación del coeficiente fue β = 0.263 con un Error estándar de 0.834 el estadístico de Wald es de 0.099 con 1 grado de libertad y con un nivel de significancia de 0.753 además tenemos que Exp(0.263) = 1.300; para T5 la estimación del coeficiente fue β = -1.712 con un Error estándar de 0.646 su estadístico de Wald

de 7.016 con 1 grado de libertad y con un nivel de significancia de 0.008 además tenemos que Exp(-1.712) = 0.180; para T6 la estimación del coeficiente β = -13.10 con un Error estándar de 326.369 su estadístico de Wald de 0.002 con 1 grado de libertad y con un nivel de significancia de 0.968 además tenemos que Exp(-13.10) = 0.000; y para T7 la estimación del coeficiente fue β = -2.326 con un Error estándar de 1.051 su estadístico Wald de 4.898 con 1 grado de libertad y con un nivel de significancia de 0.027 y su Exp(-2.326) = 0.098.

TABLA 4-7
VARIABLES EN LA ECUACIÓN

	B	Error estándar	Wald	gl	Sig.	Exp(B)
SEXO	,243	,277	,769	1	,380	1,275
EDAD	,006	,009	,426	1	,514	1,006
N1	-,512	,365	1,966	1	,161	,599
N2	-,270	,438	,381	1	,537	,763
N3	-,459	,755	,369	1	,543	,632
E1	,141	,848	,028	1	,868	1,152
E2	-,172	1,213	,020	1	,887	,842
E3	,582	,647	,809	1	,368	1,789
E4	,866	,407	4,524	1	,033	2,377
T1	-1,414	,486	8,452	1	,004	,243
T2			,	0	,	
T3	-,983	1,038	,897	1	,344	,374
T4	,263	,834	,099	1	,753	1,300
T5	-1,712	,646	7,016	1	,008	,180
T6	-13,10	326,369	,002	1	,968	,000
T7	-2,326	1,051	4,898	1	,027	,098

Además la tabla 4-7 nos indica que los hombres tienen 1.3 veces más probabilidad de contraer éste tipo de cáncer que las mujeres; las personas mayores de 65 años tienen 1.006 veces más que las personas menores de 65 años; de las personas que se encuentran en la Etapa 4 del cáncer tienen 2.3 veces más probabilidad de no superar ésta patología, en la Etapa 3 del cáncer 1.8 veces más mientras que el Etapa 1 es 1.2 veces más.

Como podremos recordar, para determinar si la información proporcionada por la variable X es redundante, se utiliza el p-valor asociado al estadístico de Wald, si una variable es candidata a ser seleccionada en un paso, el criterio de entrada se basa en el p-valor, si este es menor que 0.15 la variable debe ser incluida en el modelo, y si el p-valor es mayor que 0.15 la variable no aporta significativamente en el modelo y por lo tanto debe ser excluida del mismo. Como podemos observar en la Tabla 4-7, las variables Sexo, Edad, Nivel de instrucción (N1,N2,N3), Estadio (E1,E2,E3) y Tratamiento (T2,T3,T4,T6) no deberían incluirse en el

modelo debido a que no contribuyen significativamente mientras que la variable Estadio (E4) y Tratamiento (T1,T5,T7) son seleccionadas debido a que si aportan significativamente en el modelo.

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:

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

$\hat{Z}$ = 0.243Sexo + 0.006Edad – 0.512N1 – 0.270N2 – 0.459N3 + 0.141E1 – 0.172E2 + 0.582E3 + 0.866E4 – 1.414T1 – 0.983T3 + 0.263T4 – 1.712T5 – 13.10T6 – 2.326T7

siendo $\hat{Z}$ la combinación lineal de las variables

Luego, la estimación de $\hat{g}(X)$ será:

$$\hat{g}(X) = [e^{0.243(Sexo)} e^{0.006(Edad)}] * [e^{-0.512(N1)} e^{-0.270(N2)} e^{-0.459(N3)}] * [e^{0.141(E1)} e^{-0.172(E2)} e^{0.582(E3)} e^{0.866(E4)}] * [e^{-1.414(T1)} e^{-0.983(T3)} e^{0.263(T4)} e^{-1.712(T5)} e^{-13.10(T6)} e^{-2.326(T7)}]$$

O, lo que es equivalente $\hat{g}(X)$ será:

$$\hat{g}(X) = [(1.275)^{Sexo}(1.006)^{Edad}] * [(0.599)^{N1}(0.763)^{N2}(0.632)^{N3}] * [(1.152)^{E1}(0.842)^{E2}(1.789)^{E3}(2.377)^{E4}] * [(0.243)^{T1}(0.374)^{T3}(1.3)^{T4}(0.18)^{T5}(0.0)^{T6}(0.098)^{T7}]$$

Esto es para cualquier valor que pueden tomar las variables que se encuentran dentro del modelo. Por ejemplo para un paciente cuyo sexo es masculino, edad es 65 años, su nivel de instrucción primario, con un Estadio 3 y con tratamiento cirugía, tendremos el siguiente resultado:

$\hat{g}(X)$ =(1.275)^{0}(1.006)^{65}(0.599)^{1}(0.763)^{0}(0.632)^{0}(1.152)^{0}(0.842)^{0}(1.789)^{1}(2.377)^{0}(0.243)^{1}(0.374)^{0}(1.3)^{0}(0.18)^{0}(0.0)^{0}(0.098)^{0}

$\hat{g}(X)$ =0,38416

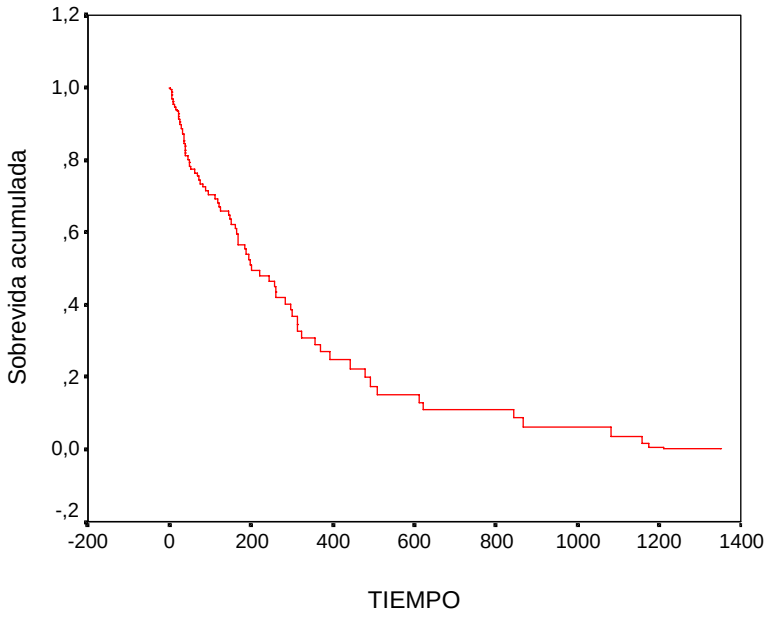
$h(t/X)=h_0(t)$0,38416

En el ejemplo dado, el valor obtenido nos dice que el paciente con las características dadas tiene el 38.416% de probabilidad de sobrevida.

Lo que nos indica que para cualquier valor de la función de riesgo $h(t_0)$ (en función del tiempo), obtendremos el valor de la función de riesgo considerando la información de las variables $h(t/X)$. El valor de ésta función $h_0(t)$ no es tan relevante en el modelo a excepción que el valor de la función $\hat{g}(X)$ sea cero.

La figura 4-4, nos muestra la representación gráfica de los valores de la función de Sobrevida frente al tiempo. La curva se mantiene un descenso progresivo suave hasta el t = 442, a partir del cual la curva se tiende a mantener plana con pequeños periodos constantes en su disminución hasta el instante t = 1349 donde tiene una altura de 0.000.

FIGURA 4-4
FUNCIÓN DE SOBREVIDA EN MEDIA DE COVARIABLES



4.3.2. Regresión de Cox utilizando variables dicotómicas y variables en escala lickert

Para realizar el modelo de Regresión de Cox, en este enfoque se recodificó las variable Tratamiento Cronológico en una variable dicotómica y las variables Nivel de Instrucción y Estadio en variables en escala licker, además de que en el proceso se eliminó 18 datos de la variable Estadio correspondiente a la categoría No Reporte para una mejor interpretación.

Y como se mencionó en el anterior enfoque las variables Edad y Tiempo de enfermedad son variables cuantitativas, por lo tanto estas son utilizadas en su forma original, la variable Sexo es una variable de dos categorías. A continuación se muestra las tres variables recodificadas:

1. Nivel de Instrucción

Ninguno	0
Primario	1
Secundario	2
Superior	3

2. Tratamiento cronológico

	T1	T2	T3	T3	T5	T6	T7
Ninguno	0	0	0	0	0	0	0
Cirugía	1	0	0	0	0	0	0
Radioterapia	0	1	0	0	0	0	0
Cirugía / radioterapia	0	0	1	0	0	0	0
Quimioterapia	0	0	0	1	0	0	0
Cirugía / quimioterapia	0	0	0	0	1	0	0
Quimio / Radioterapia	0	0	0	0	0	1	0
Cirugía/Radio/Quimio	0	0	0	0	0	0	1

3. Estadio

Estadio 1	0
Estadio 2	1
Estadio 3	2
Estadio 4	3

La tabla 4-8, nos 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

70 que equivale al 72.2% de los datos analizados, debido que durante el tiempo de observación de los pacientes y datos omitidos se encontraron que fueron 70 los decesos, los datos censurados fueron 27 que equivale al 27.8% de los datos analizados estos son aquellos que seguían vivo o habían abandonado el tratamiento durante el periodo de observación, además podemos observar que no existen casos que fueron excluidos como casos con valores perdidos, casos con tiempo no positivo o casos censurados antes del Evento más temprano, el total de datos analizados fue de 97 lo que corresponde al número de pacientes en el estudio con los datos omitidos.

TABLA 4-8
RESUMEN DEL PROCESO DE CASOS

		N	Porcentaje
Casos disponibles en el análisis	Evento[a]	70	72,2%
	Censurados	27	27,8%
	Total	97	100,0%
Casos excluidos	Casos con valores perdidos	0	,0%
	Casos con tiempo no positivo	0	,0%
	Casos censurados antes del evento más temprano en el estrato	0	,0%
	Total	0	,0%
Total		97	100,0%

a. Variable dependiente: TIEMPO

Como se puede observar en el bloque 0 de la regresión, el cual se encuentra la tabla 4-9, se obtuvo mediante la Prueba Bondad de Ajuste el coeficiente del modelo, el mismo que se realiza mediante el estadístico -2LL (-2 Log de verosimilitud), el cual dio como resultado el valor de 506.250, la razón de la existencia de este valor radica en que el modelo se está validando es el correspondiente a la función h(t / X).

TABLA 4-9
PRUEBA BONDAD DE AJUSTE SOBRE EL COEFICIENTE DEL MODELO

-2 Log de Verosimilitud
506,250

Analizando el bloque 1, se obtiene la tabla 4-10, en la cual constan, el valor del coeficiente del modelo que se lo obtiene mediante el estadístico –2LL, el cual en este paso dio como resultado 466.248, como se puede comparar con el resultado mostrado en la tabla 4-8 este presenta un valor diferente, la razón a este cambio es que en este paso se evalúa la mejora del modelo al incorporar el Tiempo de Sobrevida. También la tabla

nos muestra la comparación entre el análisis global, en el mismo paso y con el bloque anterior o paso 0. En el análisis global se obtuvo el valor del estadístico Chi-cuadrado de 38.591, con 10 grados de libertad y con un nivel de significancia de 0.00 el cual es menor a 0.05, queriendo decir que el cambio fue estadísticamente ligeramente significativo. Para el Cambio desde el paso anterior se obtiene que el estadístico Chi-cuadrado es igual a 40.002, con 10 grados de libertad y con un nivel de significancia de 0.00 el mismo que es menor a 0.05, entonces se puede decir que el cambio fue estadísticamente significativo, y para el Cambio desde el bloque anterior se obtuvo los mismo valores tanto para el estadístico Chi-cuadrado, como para los grados de libertad y el nivel de significancia, lo que refleja lo dicho anteriormente.

TABLA 4-10
PRUEBA BONDAD DE AJUSTE SOBRE LOS COEFICIENTES DEL MODELO

-2 Log de verosimilitud	Global (puntuación)			Cambios desde el paso anterior			Cambios desde el bloque anterior		
	Chi-cuadrado	gl	Sig.	Chi-cuadrado	gl	Sig.	Chi-cuadrado	gl	Sig.
466,248	38,591	10	,000	40,002	10	,000	40,002	10	,000

Si bien el valor de –2LL no es cercano a cero (466,248) como se desease que fuera, en cual el modelo se ajusta perfectamente a los datos; sin embargo, el modelo pasa la prueba χ^2.

La tabla 4-11a y 4-11b, muestra la Tabla de Sobrevida en la cual consta el tiempo de todos los Eventos (muerte) que han ocurrido en el análisis los mismos que son iguales a 70 decesos, los valores estimados de la función de Sobrevida, evaluada sobre las medias de las variables independientes. Dado que en el instante t = 1, cuyo valor estimado de la función de Sobrevida es 0.993 el mismo que es próximo a 1 se produce el primer deceso, es decir que la probabilidad de no fallecer al día 1 después de ser diagnosticado es de 99.3%, así mismo siguiendo analizando la tabla, la probabilidad de no fallecer a los 110 días después del diagnostico es del 70.2%, a los 202 días es de 50.9%, a los 611 días es de 12.4% y a los 1210 días es del 0.3%.

TABLA 4-11a
TABLA DE SOBREVIDA

Tiempo	Evento acumulado	Impacto acumulado línea base	En la media de las covariables		
			Sobrevida	Error Estándar	Impacto Acumulado
1	1	0,005	0,993	0,024	0,007
5	2	0,010	0,986	0,049	0,014
6	3	0,015	0,979	0,073	0,022
7	4	0,020	0,971	0,098	0,029
8	5	0,025	0,964	0,123	0,037
9	6	0,031	0,956	0,148	0,045
11	7	0,036	0,949	0,173	0,053
16	8	0,042	0,941	0,199	0,061
18	9	0,048	0,933	0,224	0,070
21	10	0,054	0,925	0,250	0,078
22	11	0,060	0,917	0,275	0,087
24	12	0,066	0,908	0,303	0,097
26	13	0,073	0,899	0,330	0,107
27	14	0,080	0,890	0,358	0,117
33	16	0,094	0,871	0,414	0,138
34	18	0,110	0,852	0,468	0,160
36	19	0,118	0,842	0,496	0,172
37	20	0,126	0,832	0,524	0,184
38	21	0,134	0,822	0,552	0,196
40	22	0,142	0,812	0,579	0,208
44	23	0,151	0,802	0,606	0,220
47	24	0,159	0,792	0,633	0,233
50	25	0,168	0,782	0,660	0,246
51	26	0,178	0,771	0,688	0,260
63	27	0,188	0,760	0,716	0,275
70	28	0,198	0,749	0,743	0,289

TABLA 4-11b
TABLA DE SOBREVIDA

Tiempo	Evento acumulado	Impacto acumulado línea base	En la media de las covariables		
			Sobrevida	Error Estándar	Impacto Acumulado
74	29	0,208	0,737	0,770	0,305
89	30	0,219	0,726	0,797	0,320
94	31	0,230	0,714	0,824	0,337
110	32	0,242	0,702	0,852	0,354
117	33	0,254	0,690	0,878	0,371
121	34	0,267	0,677	0,905	0,390
123	35	0,280	0,664	0,931	0,410
145	36	0,294	0,650	0,958	0,430
149	37	0,309	0,636	0,984	0,452
150	38	0,325	0,622	1,009	0,475
162	39	0,342	0,607	1,035	0,499
164	40	0,360	0,591	1,061	0,526
169	42	0,398	0,559	1,108	0,582
184	43	0,418	0,543	1,130	0,611
198	44	0,440	0,526	1,152	0,643
202	45	0,461	0,509	1,171	0,675
221	46	0,484	0,493	1,189	0,707
245	47	0,506	0,477	1,205	0,740
261	48	0,530	0,460	1,217	0,776
262	49	0,557	0,443	1,227	0,815
283	50	0,587	0,424	1,236	0,858
296	51	0,619	0,405	1,243	0,905
300	52	0,654	0,384	1,245	0,957
312	53	0,694	0,362	1,244	1,015
314	54	0,737	0,340	1,241	1,077
324	55	0,781	0,319	1,234	1,142
355	56	0,829	0,298	1,218	1,212
369	57	0,882	0,275	1,199	1,290
393	58	0,937	0,254	1,177	1,370
442	59	1,018	0,226	1,134	1,489
479	60	1,107	0,198	1,082	1,618
491	61	1,203	0,172	1,020	1,759
509	62	1,309	0,147	0,950	1,914
611	63	1,428	0,124	0,871	2,087
620	64	1,554	0,103	0,788	2,273
845	65	1,734	0,079	0,662	2,535
1083	66	2,049	0,050	0,491	2,996
1159	67	2,477	0,027	0,310	3,622
1175	68	3,089	0,011	0,157	4,516
1210	69	3,950	0,003	0,056	5,774
1349	70	5,569	0,000	0,007	8,141

En la tabla 4-12 se muestra todas las variables que participan en la ecuación, las cuales son: Sexo, cuyo estimación del coeficiente β = 0.448, con un Error estándar de 0..288,el estadístico de Wald es de 2.418 con 1 grado de libertad y con un nivel de significancia de 0.12 y el Exp(B=0.448)=1.565; la variable Edad con una estimación de su coeficiente β = 0.007, con un Error estándar de 0.009 y el estadístico de Wald es igual a 0.519, con 1 grado de libertad y con un nivel de significancia de 0.471 y el Exp(B=0.007)=1.007; la variable Nivel de Instrucción, la estimación del coeficiente fue β = -0.050, con un Error estándar de 0.211 el estadístico de Wald es de 0.056, con 1 grado de libertad y con un nivel de significancia de 0.813, además tenemos que Exp(-0.050) = 0.951; para la variable Estadio la estimación del coeficiente fue β = 0.169, con un Error estándar de 0.237 el estadístico de Wald es de 0.512, con 1 grado de libertad y con un nivel de significancia de 0.474, además tenemos que Exp(0.169) = 01.185.

Y finalmente la variable Tratamiento que fue codificada de la siguiente forma: T1, T2, T3, T4, T5, T6 y T7. Para T1 la estimación del coeficiente fue β = -1.436, con un Error estándar de 0.421 su estadístico de Wald es de 11.661 con 1 grado de libertad y con un nivel de significancia de 0.001, además tenemos que Exp(-1.436) = 0.238, para T2 no se presentó valores estimados debido a que ningún paciente del estudio recibió este tipo de tratamiento; para T3 la estimación del coeficiente β = -10.26 con un Error estándar de 1.037 el estadístico de Wald es 0.980 con 1 grado de libertad y con un nivel de significancia de 0.322, además tenemos que Exp(-10.26) = 0.358; para T4 la estimación del coeficiente fue β = 0.210 con un Error estándar de 0.773 el estadístico de Wald es de 0.074 con 1 grado de libertad y con un nivel de significancia de 0.786 además tenemos que Exp(0.210) = 1.234; para T5 la estimación del coeficiente fue β = -1.633 con un Error estándar de 0.580 su estadístico de Wald de 7.916 con 1 grado de libertad y con un nivel de significancia de 0.005 además tenemos que Exp(-

1.633) = 0.195; para T6 la estimación del coeficiente β = -13.126 con un Error estándar de 336.134 su estadístico de Wald de 0.002 con 1 grado de libertad y con un nivel de significancia de 0.969 además tenemos que Exp(-13.126) = 0.000; y para T7 la estimación del coeficiente fue β = -2.408 con un Error estándar de 1.045 su estadístico Wald de 5.304 con 1 grado de libertad y con un nivel de significancia de 0.021 y su Exp(-2.408) = 0.090.

TABLA 4-12
VARIABLES EN LA ECUACIÓN

	B	Error estándar	Wald	gl	Sig.	Exp(B)
SEXO	,448	,288	2,418	1	,120	1,565
EDAD	,007	,009	,519	1	,471	1,007
INSTRUCC	-,050	,211	,056	1	,813	,951
ESTADIO	,169	,237	,512	1	,474	1,185
T1	-1,436	,421	11,661	1	,001	,238
T2			,	0	,	
T3	-1,026	1,037	,980	1	,322	,358
T4	,210	,773	,074	1	,786	1,234
T5	-1,633	,580	7,916	1	,005	,195
T6	-13,126	336,134	,002	1	,969	,000
T7	-2,408	1,045	5,304	1	,021	,090

También la tabla 4-12 nos indica que los hombres tiene 1.6 veces más probabilidad de contraer cáncer de

estómago, las personas mayores de 65 años de contraer es 1.007 veces más que las personas menores de 65 años y las personas con Estadio 4 tienen 1.18 veces más probabilidad de no superar ésta enfermedad.

Como podremos recordar, para determinar si la información proporcionada por la variable X es redundante, se utiliza el p-valor asociado al estadístico de Wald, si una variable es candidata a ser seleccionada en un paso, el criterio de entrada se basa en el p-valor, si este es menor que 0.15 la variable debe ser incluida en el modelo, y si el p-valor es mayor que 0.15 la variable no aporta significativamente en el modelo y por lo tanto debe ser excluida del mismo. Como podemos observar en la Tabla 4-12, las variables Edad, Nivel de instrucción, Estadio y Tratamiento (T2,T3,T4,T6) no deberían incluirse en el modelo debido a que no contribuyen significativamente mientras que las variables Sexo y Tratamiento (T1,T5,T7) son seleccionadas debido a que si aportan significativamente en el modelo.

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:

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

$\hat{Z}$ = 0.448Sexo + 0.007Edad – 0.050Nivel de Instrucción + 0.169Estadio – 1.436T1 – 1.026T3 + 0.210T4 – 1.633T5 – 13.126T6 – 2.408T7

siendo $\hat{Z}$ la combinación lineal de las variables

Luego, la estimación de $\hat{g}(X)$ será:

$\hat{g}(X) = [e^{0.448(Sexo)} e^{0.007(Edad)}] * [e^{-0.050(Nivel\_de\_instrucción)}] * [e^{0.169(Estadio)}] * [e^{-1.436(T1)} e^{-1.026(T3)} e^{0.210(T4)} e^{-1.633(T5)} e^{-13.126(T6)} e^{-2.408(T7)}]$

O, lo que es equivalente $\hat{g}(X)$ será:

$\hat{g}(X)$ =[(1.565)^{Sexo}(1.007)^{Edad}(0.951)$^{Nivel\_instrucción}$ (1.185)^{Estadio}]*[(0.238)^{T1}(0.358)^{T3}(1.234)^{T4}(0.195)^{T5}(0.0 0)^{T6}(0.090)^{T7}]

Esto es para cualquier valor que pueden tomar las variables que se encuentran dentro del modelo. Usando el mismo ejemplo anterior, es decir para un paciente cuyo sexo es masculino, edad es 65 años, su nivel de instrucción primario, con un Estadio 3 y con tratamiento cirugía, tendremos el siguiente resultado:

$\hat{g}(X)$ =(1.565)^{0}(1.007)^{65}(0.951)^{1}(1.185)^{2}(0.238)^{1}(0.358)^{0}(1.234)^{0}(0.195)^{0}(0.0)^{0}(0.090)^{0}

$\hat{g}(X)$ =0,50016

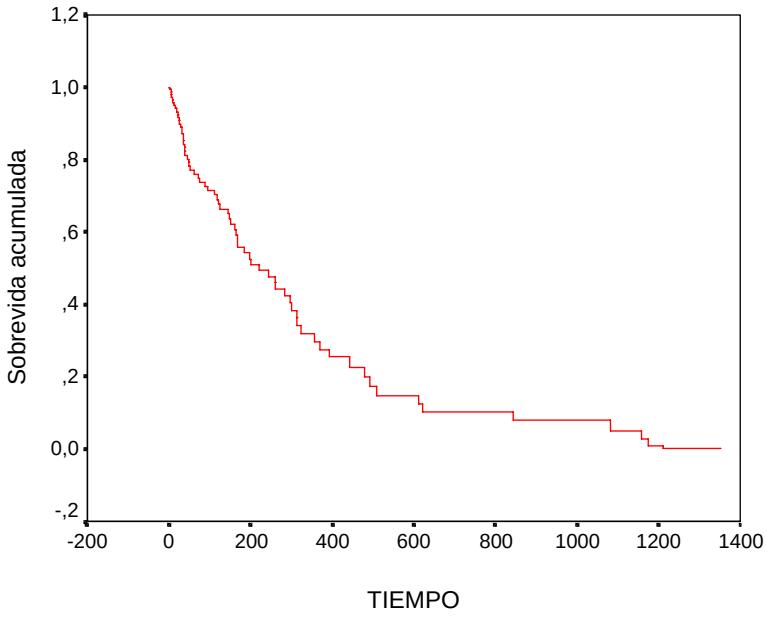
$h(t/X)= h_0(t)$ 0,50016

Este valor nos indica que el paciente con las características dadas tiene el 50.016% de probabilidad de sobrevida.

Lo que nos indica que para cualquier valor de la función de riesgo $h(t_0)$ (en función del tiempo), obtendremos el valor de la función de riesgo considerando la información de las variables $h(t/X)$.

La figura 4-5, nos muestra la representación gráfica de los valores de la función de Sobrevida frente al tiempo. La curva se mantiene un descenso progresivo suave hasta el t = 442, a partir del cual la curva se tiende a mantener plana con pequeños periodos constantes en su disminución hasta el instante t = 1349 donde tiene una altura de 0.000.

FIGURA 4-5
FUNCIÓN DE SOBREVIDA EN MEDIA DE COVARIABLES



4.3.3. Comparación de los modelos

Comparando los dos enfoques vemos primeramente que en la Prueba de Bondad de Ajuste, el coeficiente del estadístico –2LL no varió significativamente y ambos modelos pasaron la prueba χ^2. En la tabla de sobrevida aumento levemente la probabilidad de no fallecer en un tiempo específico.

TABLA 4-13
VARIABLES EN LA ECUACIÓN

	Modelo 1			Modelo 2		
Variables	**Sig.**	**B**	**Exp(B)**	**Sig.**	**B**	**Exp(B)**
sexo	0,380	0,243	1,275	0,120	0,448	1,565
Edad	0,514	0,006	1,006	0,471	0,007	1,007
Nivel de Inst.	-	-	-	0,813	-0,050	0,951
N1	0,161	-0,512	0,599	-	-	-
N2	0,537	-0,270	0,763	-	-	-
N3	0,543	-0,459	0,632	-	-	-
Estadio	-	-	-	0,474	0,169	1,185
E1	0,868	0,141	1,152	-	-	-
E2	0,887	-0,172	0,842	-	-	-
E3	0,368	0,582	1,789	-	-	-
E4	0,033	0,866	2,377	-	-	-
T1	0,004	-1,414	0,243	0,001	-1,436	0,238
T2	-	-	-	-	-	-
T3	0,344	-0,983	0,374	0,322	-1,026	0,358
T4	0,753	0,263	1,300	0,786	0,210	1,234
T5	0,008	-1,712	0,180	0,005	-1,633	0,195
T6	0,968	-13,100	0,000	0,969	-13,126	0,000
T7	0,027	-2,326	0,098	0,021	-2,408	0,090

Como se puede observar en la tabla 4-13, en ambos modelos se excluyen las variables Edad, Nivel de Instrucción, Estadio y Tratamiento (T2,T3,T4,T6) y se incluyen en el modelo las variable Tratamiento (T1,T5,T7), en lo que se diferencian es en que en el primero excluye la variable Sexo del modelo mientras que en el segundo no, aparte de que en el primero si incluye la variable dicotómica Estadio (E4).

En el ejemplo realizado con las mismas características de un paciente se tiene que la función $\hat{g}(X)$ varía de 0.38416 a 0.50016, lo que nos indica que un mismo paciente en el segundo modelo tiene mayor probabilidad de sobrevida que en el primer modelo.

Por lo tanto la variable Nivel de Instrucción no influye en el modelo mientras que la variable Estadio dicotómica si influye.

V. CONCLUSIONES Y RECOMENDACIONES

5.1 CONCLUSIONES

1. El género masculino tuvo mayor ocurrencia con el 65% que el femenino con el 35%, en el diagnostico registrado en el año 1999 y la edad mínima observada fue de 31 años, mientras que la edad máxima fue de 88 años; en el intervalo [64–76) de edad, fue donde se encontraron la mayor cantidad de pacientes con el 32% del total de toda la población analizada con cáncer gástrico.

2. Se pudo observar que la mayor cantidad de los pacientes que acudieron, para ser atendidos en la Sociedad de Lucha Contra

el Cáncer(SOLCA) con el diagnóstico cáncer de estómago, residían en su mayoría de la ciudad de Guayaquil con el 62% del total de la población, mientras que el restante 38% residían de otras partes del País.

3. La mayoría de los paciente, 57%, tan solo poseía educación primaria, el 23% no poseía ningún tipo de educación, el 16% de los pacientes tenían educación superior y sólo el 4% tenían educación superior.

4. Se pudo también observar que la mayoría de los pacientes, 74%, acudió voluntariamente a SOLCA; el 66%, sin ningún diagnóstico previo y 94% sin ningún tipo de tratamiento previo.

5. El 83% de los pacientes no poseían en su organismo la bacteria Hicto Bacto Pílori o HPV y el 80% no poseía historia médica en su familia que haya sido detectado con cáncer gástrico.

6. El momento de ser diagnosticado, el 51% de los pacientes tenía una lesión solapada y el 21% en el 1/3 medio cuerpo

del estómago. El tipo morfológico del cáncer, el 80% fue de tipo intestinal y el 54% de los pacientes tenían metástasis en otros órganos en el cual el 25% fue en el hígado.

7. La mayor parte de los pacientes se encontraban en la fase o estadio IV con el 62%, en el estadio III el 17%, en el estadio I el 2%; mientras que el 16% no poseían reporte en el estadio en que se encontraba.

8. En la variable Tiempo de Enfermedad, pudimos observar que el 77% de los pacientes que presentan cáncer de estómago asistió a consulta en un periodo menor de un año, el 11% asistió entre un año y dos, el 4% entre dos y tres años y el 5% concurrió más de tres años a consulta.

9. En la variable Estado de Última Observación, pudimos observar que el 69% de los pacientes con cáncer de estómago se encuentran fallecidos, el 12% de los pacientes se encuentran Vivos, y el 17% restante de los pacientes abandonaron el tratamiento para esta enfermedad.

10. El 63% de los pacientes no recibió ningún tipo de tratamiento cronológico para combatir éste tipo de cáncer, mientras que el 25% recibió cirugía como tratamiento de la cual gastrectomía total fue la que más se utilizó.

11. Mediante Tablas de contingencia pudimos observar que las variables Sexo y Estadio son independientes, es decir no tienen influencia la una con la otra.

12. Entre las variables Tiempo de Enfermedad y Estadio; Nivel de Instrucción y Estado de Última Observación; Tratamiento cronológico y Estadio; Edad y Estadio, pudimos concluir que son variables dependientes es decir si existe dependencia entre estas variables.

En lo relacionado al Análisis Multivariado de los datos obtenidos mediante las variables: Edad, Nivel de instrucción, Estadio y Tiempo de Enfermedad, tenemos:

13. Al trabajar con la matriz de varianzas y covarianzas de los datos reales debemos retener las dos primeras componentes principales, ya que las dos primeras

componentes principales explican el 99,998% de la información total. De esta manera tenemos que la primera componente principal será:

Y_1 = -0,029(Edad) -0.111(Nivel de instrucción) -0.076 (Estadio) +1(Tiempo de Enfermedad)

14. El Tiempo de Enfermedad es la variable que mas pesa dentro del modelo y debería ser la primera en considerarse a la hora de realizar un análisis de cáncer gástrico.

15. Puesto que la segunda componente sólo aporta con el 0.2% de la información, se proyecto todas las variables en el eje de la primera componente y vemos que se forma un grupo de variables entre el Nivel de Instrucción, Estadio, y Edad.

16. La variable Tiempo de Enfermedad tiene un comportamiento diferente al resto de variables.

En lo relacionado al análisis de Sobrevivida de los datos obtenidos mediante las variables: Sexo, Edad, Nivel de

Instrucción, Estadio, Tratamiento Cronológico, Tiempo de Enfermedad, y Estado de Última Observación, tenemos:

17. Para el modelo de Regresión de Cox, se utilizaron todas las variables anteriormente mencionadas, y se utilizó como variable dependiente el Tiempo de sobrevida, y como variable de estado se utilizó la variable Estado de Última Observación para el evento "fallecido" se obtuvo el siguiente modelo:

$$\hat{Z} = 0.243\text{Sexo} + 0.006\text{Edad} - 0.512\text{N1} - 0.270\text{N2} - 0.459\text{N3} + 0.141\text{E1} - 0.172\text{E2} + 0.582\text{E3} + 0.866\text{E4} - 1.414\text{T1} - 0.983\text{T3} + 0.263\text{T4} - 1.712\text{T5} - 13.10\text{T6} - 2.326\text{T7}$$

18. Al comparar ambos modelos, la variable que es significativa en el modelo es el Tipo de Tratamiento (T3,T5,T7).

19. La variable Estadio (E4) es significativa en el primer modelo mientras que la variable Sexo lo es en el segundo modelo, esta discrepancia se debe a que en el primer modelo el

número de casos correspondiente al evento Muerte fue de 79 mientras que en el segundo fue de 70.

5.2 RECOMENDACIONES

1. Para poder tener un mejor acceso a la información para estudios posteriores se recomienda que la institución de SOLCA, se provea de un sistema de base de datos donde la información acerca de cada una de las enfermedades y de los pacientes que se tratan en esta institución pueda ser recabada de forma ágil y eficiente, para las personas que requieran de las mismas.

2. Diseñar un elemento de captura de información, para que pueda ayudarse a los doctores a la hora de llenar las historias clínicas de los pacientes, para luego esta información pueda ser colocada en la base de datos.

3. Capacitar o contratar personal encargado de ingresar debidamente la información que se necesitará, tanto para estudios posteriores como para el control de los pacientes mediante los doctores.

4. Considerando los resultados obtenidos de acuerdo al método para retener el número optimo de componentes principales, se recomienda basarse principalmente en el método que consiste en retener aquellas componentes que nos proporcione mas del 70% del total de la información.

5. Para el análisis de Regresión de Cox, se recomienda que se realice este análisis para un intervalo de tres años como mínimo para poder obtener mayor cantidad de datos y así poder obtener mejores resultados que nos ayuden conocer cuales son las variables mas significativas a la hora de tratar esta enfermedad.

6. Puesto que la primera componente principal explica un 99,987% de la información total de la muestra, se recomienda usar esta componente para futuros estudios, puesto que contiene las cuatro variables métricas en el modelo, para cualquier modelo matemático, en particular en la Regresión de Cox.

BIBLIOGRAFÍA

1. **JOHNSON, R AND WICHERN, D** (1998). "*Applied Multivariate Statistical Analysis*", Prentice Hall, Upper Saddle River, New Yersey, USA.

2. **JOHNSON, D.** (1998). "*Métodos Multivariados Aplicados al Análisis de Datos*", International Thompson Editores, México, México.

3. **VISAUTA, V.** (1997). "*Análisis Estadístico con SPSS para Windows.* Estadística Básica", McGraw – Hill / Interamericana S.A. Madrid, España.

4. **MENDENHALL, W.** (1994). "*Estadística Matemática con Aplicaciones*", Grupo Editorial Iberoamérica, México, México.

5. **FREUND J., WALPOLE E.** (1990) "*Estadística Matemática con Aplicaciones*", Cuarta edición, Prentice Hall / Hispanoamericana S.A. México, México.

6. **CANAVOS, G.** "*Probabilidad y Estadística Aplicaciones y Métodos*", McGraw-Hill/ Interamericana S.A. México, México.

7. **LA CIENCIA DETRÁS DE LAS NOTICIAS**. Junio 2003. Entendiendo el cáncer,
www.press2.nci.nhi.gov/sciencebehind/index.htm