ASPECTOS BÁSICOS DE ESTIMACIÓN NO PARAMÉTRICA EN ANÁLISIS DE SOBREVIVENCIA

Una aplicación a un estudio de deserción estudiantil.

DIEGO ALEJANDRO CARDONA HURTADO

JENNY CAROLINA TRUJILLO BONILLA

UNIVERSIDAD DEL TOLIMA

FACULTAD DE CIENCIAS

MTEMÁTICAS CON ÉNFASIS EN ESTADÍSTICA

IBAGUÉ

2013

ASPECTOS BÁSICOS DE ESTIMACIÓN NO PARAMÉTRICA EN ANÁLISIS DE SOBREVIVENCIA

Una aplicación a un estudio de deserción estudiantil.

DIEGO ALEJANDRO CARDONA HURTADO
JENNY CAROLINA TRUJILLO BONILLA

Director
JAIRO ALFONSO CLAVIJO MÉNDEZ
Profesor adscrito al departamento de matemáticas y estadística

Trabajo para optar al título de
PROFESIONAL EN MATEMÁTICAS CON ÉNFASIS EN ESTADÍSTICA

UNIVERSIDAD DEL TOLIMA
FACULTAD DE CIENCIAS
MATEMÁTICAS CON ÉNFASIS EN ESTADÍSTICA
IBAGUÉ
2013

A nuestras madres quienes nos ensañor el valor e importacia de la vida, y a nuestros padres por no dejar que lo olvidemos y siempre brindarnos ánimo para no desistir a pesar de las adversidades.

AGRADECIMIENTOS

Al profesor Jairo Alfonso Clavijo, director de éste trabajo de grado, quien con su conocimiento, apoyo, dedicación e infinita paciencia nos orientó de la mejor manera durante la elaboración de éste.

Al profesor Rodrigo Vásquez quien con sugerencias y consejos logró mejorar la fundamentación teórica de este trabajo.

A todas aquellas personas que durante este tiempo transcurrido nos apoyaron.

RESUMEN

En el presente trabajo se exponen las principales técnicas de estimación no paramétrica de sobrevivencia como son las de Kaplan Meier y la de Nelson Aalen como ejemplo de su aplicación se mostrará el caso de deserción de estudiantes de la carrera de matemáticas que iniciaron estudios en el año 2002 los cuales fueron seguidos hasta el año 2011, „época en la que finalizó totalmente esa cohorte. Se aplicaron estas técnicas en forma general y luego por géneros para hacer sus respectivas comparaciones utilizando test estadísticos como el de Log – Rank, Tarone – Ware y Breslow.

ABSTRACT

This research will display the main non-parametric estimation techniques for survival such as the Kaplan-Meier and Nelson Aalen, as an example of its application is shown the case of students dropping out of math program with emphasis on statistical studies, they began in 2002 and were followed until 2011, 'time when fully concluded that cohort. We applied these techniques in general, making respective comparisons between genders and using the statistical tests such as Log - Rank, Tarone - Ware and Breslow.

CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

INTRODUCCIÓN

El análisis de sobrevivencia constituye una serie de procedimientos y técnicas estadísticas en las cuales la variable de interés es el tiempo que transcurre hasta que sucede un determinado evento, como puede ser tiempo de recurrencia, tiempo que dura la eficacia de una intervención, tiempo de un aprendizaje determinado, en este caso el tiempo que dura un estudiante en el programa de matemáticas con énfasis en estadística desde que se matriculó en el año 2002 hasta que finalizó sus estudios ya sea porque se graduó o porque se retiró de estos estudios sin haberse graduado. El término sobrevivencia se debe a que en las primeras aplicaciones de este método de análisis se utilizaba como evento la muerte de un paciente.

En el presente trabajo nos enfocaremos en los estimador no paramétricos como el de Kaplan–Meier, el cual es el encargado de indicar la probabilidad de que el grupo analizado sobreviva más allá de un determinado intervalo, y el método no paramétrico de Nelson Aalen el cual es de gran utilidad para el caso que se presente truncamiento a la izquierda y censura a la derecha, ya que corrige el sesgo producido por la subestimación de sobrevivencia.

Además de los estimadores no paramétricos también estudiaremos las comparaciones entre la deserción de los hombres y las mujeres para encontrar sus diferencias si existen, para ello se utilizaran test estadísticos como el de Log – Rank, Breslow (Wilcoxon generalizado) y Tarone – Ware.

Desde que se creó el programa de matemáticas con énfasis en estadística de la universidad del Tolima se ha presentado con preocupación el problema de la deserción de los estudiantes, esto ha llevado que al finalizar cada año se gradué una parte muy pequeña en comparación con los que entran al primer semestre, lo cual ha mostrado que los estudiantes ingresan al programa no con el interés de terminar su estudio y graduarse sino con la intención de cambiarse de carrera en otro semestre o simplemente por no quedarse haciendo nada, lo cual provoca que más adelante se

retiren de esta ya que no le colma las expectativas generadas. Por lo tanto aplicaremos los aspectos básicos de estimación no paramétrica de sobrevivencia principalmente el de Kaplan – Meier y Nelson Aalen a esta problemática que se presenta no solamente en nuestra carrera sino también en todos los programas de la universidad y en general en el país.

1. OBJETIVOS

1.1 GENERAL

- Presentar los aspectos teóricos básicos para la estimación de los modelos de sobrevivencia no paramétricos de Kaplan y Meier y Nelson Aalen

1.2 ESPECÍFICOS

- Aplicar los conceptos básicos en la estimación de los modelos de sobrevivencia no paramétrica de Kaplan – Meier y Nelson Aalen a la deserción de las estudiantes del programa matemáticas con énfasis en estadística.

- Pronosticar el tiempo de duración en el programa de matemáticas con énfasis en estadística para los estudiantes que ingresan a este.

- Comparar las tasas de deserción entre estudiantes hombres y mujeres en el programa de matemáticas con énfasis en estadística

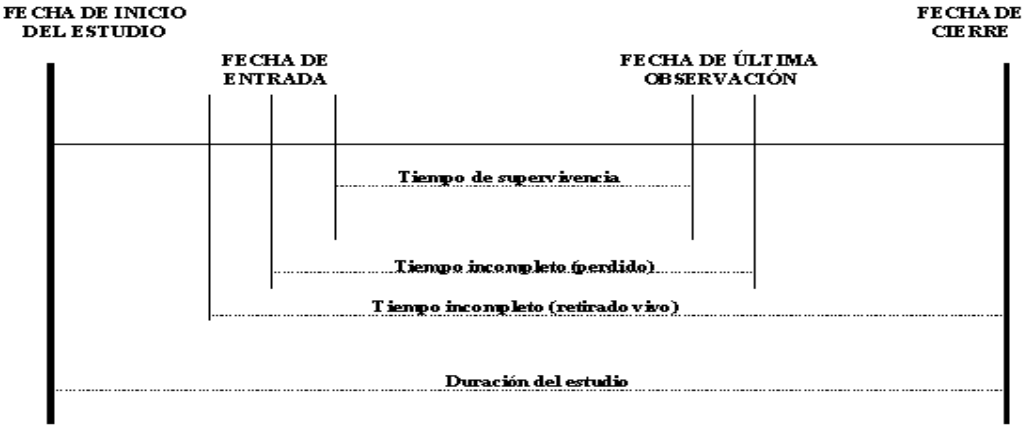
2. TEORIA DEL ANÁLISIS DE SOBREVIVENCIA

2.1 CONCEPTOS BÁSICOS

El análisis de sobrevivencia es una de las áreas de la estadística que más ha crecido en los últimos años, la cual utiliza una serie de métodos estadísticos en donde se estudia la distribución de tiempo hasta la ocurrencia de un evento de interés, en el momento que ocurre este evento se denomina tiempo de fallo. En nuestro caso el evento de interés es el retiro del estudiante por causas diferentes a la graduación.

La observación de cada estudiante se inicia el día que entra a estudiar (tiempo = 0) y continua hasta que se gradué o hasta que el tiempo de seguimiento se interrumpa por otro motivo. Cuando el tiempo de seguimiento termina el día del grado hablaremos de una observación censurada. A continuación se presenta un esquema general de un estudio de sobrevivencia.

Figura 1 Esquema general de un estudio de sobrevivencia



Fuente (Fernandez, 1995)

El periodo de seguimiento puede terminar por las siguientes razones:

a. El estudiante decide no seguir en el estudio y lo abandona.
b. El estudio termina antes de aparecer el evento de interés, es decir que no
c. se retira sino que se gradúa.
d. El estudiante se gradúa o se retira después de los 10 semestres.

En el análisis es fundamental definir los tiempos de sobrevivencia con exactitud, para determinar las censuras, el seguimiento viene definido por una fecha de inicio y una fecha de cierre que determinan el tiempo de seguimiento. Las fechas de inicio y cierre en algunos casos son diferentes para cada individuo, pues los estudiantes o personas incluidas en el estudio se incorporan en momentos diferentes, pero en nuestro caso el inicio es igual para los del semestre A 2002 y los del semestre B 2002, pero la fecha de cierre son diferentes casi en su mayoría.

El tiempo de sobrevivencia se define como el tiempo transcurrido desde el acontecimiento o estado inicial hasta el estado final ya sea porque el estudiante se graduó (censurado) o se retiró por otra causa. El estado inicial debe ser definido de manera que la fecha en que se produjo el evento pueda ser conocida exactamente (fecha de inicio de clases). Teniendo en cuenta que en el estado inicial las fechas pueden ser diferentes. Pero en nuestro caso no se presenta ya que todos los estudiantes ingresan en el mismo año (2002)

El acontecimiento o suceso estudiado también debe estar perfectamente definido para poder determinar exactamente la fecha del mismo. Este evento está casi siempre asociado al retiro del estudiante pero no tiene por qué ser así, ya que puede hacer referencia también a la fecha de graduación o por causas diferentes a las estudiadas.

En la última observación que se le hace a cada estudiante se deben registrar dos variables fundamentales, la primera es su estado, retirado o censurado y la segunda es la fecha de la información de dicho estado. El período de tiempo transcurrido entre la fecha de entrada y la fecha de la última observación o contacto se conoce como tiempo de participación en el estudio. Si el estudiante se retiró podremos con la fecha

de retiro calcular el tiempo de supervivencia. Si el estudiante está activo a la fecha de la última observación se podrá calcular el tiempo incompleto o censurado aportado por dicho estudiante.

Los requisitos necesarios para disponer de datos adecuados para un análisis de supervivencia son:

a. Definir apropiadamente el origen o inicio del seguimiento.
b. Definir apropiadamente la escala del tiempo.
c. Definir apropiadamente el evento.

(Fernandez, 1995)

2.1.1 Tipos de observaciones**:** Los datos de nuestro estudio pueden estar sesgados por las censuras o los truncamientos.

- Génesis de censuras: Pérdidas de seguimiento o fin del estudio.
- Génesis de truncamientos: Entrada en el estudio después del hecho que define el origen.

Censuras: la censura se presenta cuando:

- No se observan los eventos en todos los individuos, es decir que no todos los estudiantes se retiran por causas diferentes a la graduación.
- No se espera lo suficiente... a que aparezca el evento. En nuestro caso el tiempo de espera es 10 semestres.

La censura se puede presentar de dos maneras:

- Censura por la derecha o izquierda
- Censura por intervalos

En nuestro trabajo nos enfocaremos a datos censurados por la derecha.

Censura por la derecha: Es el caso más común de los datos incompletos y se caracteriza por que al realizar la última observación al estudiante aún no se ha

presentado el evento de interés y no se conoce cuándo ocurrirá; en nuestro caso el retiro del mismo. Algunas razones para que se presente este tipo de censura son:

- Que haya ocurrido en el individuo otro evento que no sea el de interés en el estudio (graduación)
- Que se determine el tiempo final del estudio y hasta ese momento no haya ocurrido el evento de interés (el estudiante siguió estudiando hasta la última observación), pero en este caso no se presenta.

Los tipos de censura por la derecha son:

- Censura tipo I: El periodo de seguimiento presenta una fecha de inicio y un tiempo final, que es el tiempo máximo de observación para que ocurra el evento de interés. En este tipo el número de censuras es aleatorio y la duración del estudio es fija, en nuestro caso aplica para aquellos estudiantes que al realizar la última observación en el semestre B de 2011 no han presentado el evento de interés (retiro)

- Censura tipo II: Se termina la investigación cuando se ha retirado un determinado número de estudiantes, a diferencia del caso anterior donde se prefija el tiempo para realizar el estudio, en este tipo de censura se espera a que se retiren X estudiantes, este tipo de censura no aplica en nuestro caso debido a que el tiempo de cierre del estudio es fijo y no definimos un número determinado de retiros para finalizar el mismo.

Censura por la izquierda: Es poco común para el análisis de sobrevivencia, ya que se presenta cuando en la primera observación ya ha ocurrido el evento de interés, en muchas ocasiones este tipo de censura se confunde con el truncamiento a la izquierda o la entrada tardía del individuo al estudio, no se presenta en nuestro caso debido a que todos los estudiantes ingresan en el mismo semestre.

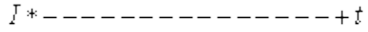
Censura por intervalos: Sucede cuando solo se sabe que al individuo le ocurre el evento de interés entre un instante t_i y un tiempo t_j. No se presenta en nuestro estudio debido a que las observaciones se hacen al iniciar el semestre.

Truncamiento por la izquierda: Ocurre cuando el individuo empieza a observarse después de iniciado el estudio. No se presenta en este caso pues solo se observan individuos que ingresaron en el año 2002.

Truncamiento por la derecha: Ocurre cuando solo se tienen en cuenta los individuos que presentan el evento de interés.

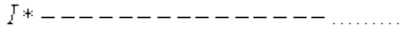
Tipos de observaciones: La combinación de las observaciones previamente indicadas nos llevaría a poder tener en nuestros datos observaciones de diferentes tipos:

a. *No truncada, no censurada*:



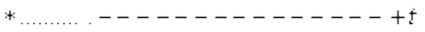
El proceso se inicia en I pero el evento ocurre en t

b. *No truncada, censurada*:



El proceso se inicia en I pero el evento no se presenta durante el seguimiento realizado.

c. *Truncada, no censurada*:



Ya se tenía el proceso antes de entrar en el estudio (el diagnóstico o fecha de inicio está atrasada) y el evento se produce en t.

d. *Truncada, censurada:*

*..... – – – – – – – – – – – – – – – – ..... .

Ya se tenía el proceso antes de entrar en el estudio, como en la situación anterior pero el evento no se presenta durante el seguimiento realizado. (Fernandez, 1995)
En el análisis de la supervivencia asumimos un supuesto básico: los mecanismos del evento y censura son estadísticamente independientes, o el sujeto censurado en C es representativo de los que sobreviven en C. Es decir, los no censurados representan bien a los censurados.

2.2 ELEMENTOS BÁSICOS

En el análisis de supervivencia la variable aleatoria T (tiempo de duración hasta la ocurrencia de un evento), posee una función de densidad de probabilidad f(t), y una función de distribución de probabilidad acumulada $F(t)=P(T \leq t)$, que indica la probabilidad de que el evento de interés ocurra en un tiempo menor o igual al tiempo t, al igual tenemos la función de supervivencia que al contrario de la anterior indica la probabilidad de que el evento de interés ocurra en un tiempo mayor a t, y está dado por

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

Dentro de las funciones básicas y de gran utilidad en el análisis de supervivencia se encuentra la función de riesgo, la cual nos indica la probabilidad de que un individuo que no ha fallado antes del tiempo t lo haga en el siguiente periodo de tiempo de duración Δt y está definida como

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \,|\, T \geq t)}{\Delta t} \qquad (1)$$

Solucionando la probabilidad condicional de la expresión (1), obtenemos:

$$\mathrm{h}\,(\mathrm{t}) = \lim_{\Delta t \to 0} \; \frac{1}{\Delta \mathrm{t}} \frac{\mathrm{P}(\mathrm{t} \leq \mathrm{T} < \mathrm{t} + \Delta \mathrm{t})}{P(T > t)} \qquad (2)$$

Calculando las probabilidades y del hecho que $P(T > t)$ = s (t), obtenemos que:

$$\mathrm{h}\,(\mathrm{t}) = \frac{1}{\mathrm{s(t)}} \lim_{\Delta t \to 0} \; \frac{\mathrm{F(t + \Delta t) - F(t)}}{\Delta \mathrm{t}} \qquad (3)$$

En la expresión (3) aplicando la definición de derivada se obtiene:

$$\mathrm{h(t)} = \frac{f(t)}{s(t)} \qquad (4)$$

Como observamos tenemos una función que relaciona la función de supervivencia y la función de riesgo, ahora podemos hallar la función de riesgo acumulada integrando la expresión (4)

$$\mathrm{H(t)} = \int_0^t h(s)ds \;=\; \int_0^t \frac{f(s)}{S(s)} ds \qquad (5)$$

$$= \int_0^t \frac{f(s)}{1-F(s)} ds$$

Sustituyendo u = 1 - F(s) y du = - f(s) ds se tiene que:

$$\mathrm{H(t)} = \int_0^t \frac{-du}{u}$$

De donde se obtienen los siguientes resultados

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

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

$$h(t) = - d(\ln S(t)) / dt$$

$$f(t)=-dS(t)/dt$$

(Barrera, 2008)

2.3 METODOLOGÍA ESTADÍSTICA

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

a. Los investigadores frecuentemente analizan los datos antes de que todos los estudiantes se hayan graduado, ya que si no habría que esperar muchos años para realizar dichos estudios. Los datos aportados por los estudiantes activos, como se señaló previamente, son observaciones “censuradas” y deben considerarse como tales a la hora de ser analizados.

b. La segunda razón por la que se necesitan métodos especiales de análisis es porque típicamente los estudiantes no inician el estudio o entran al estudio al mismo tiempo.
En la metodología estadística básica se señalaba la existencia de pruebas paramétricas y no paramétricas. En el análisis de sobrevivencia, el análisis de los datos puede ser realizado utilizando técnicas paramétricas y no paramétricas.

- **Paramétricas:**
  - Distribución Exponencial.
  - Distribución de Weibull.

- Distribución Log normal.

- **No paramétricas:**
  - Kaplan-Meier.
  - Nelson-Aalen.

Los modelos paramétricos son importantes en el análisis de datos de sobrevivencia pero esto no quiere decir que sean los únicos que se puedan utilizar. Un modelo muy importante en toda la historia del análisis de sobrevivencia es el de supervivencia exponencial. Históricamente, la distribución exponencial fue el primer modelo utilizado en la distribución de tiempos de vida, ya que esta representaba el tiempo de vida de muchos objetos. Aun la distribución exponencial puede ser útil en varias situaciones. Como es el caso del estudio de los tiempos de vida de artículos fabricados y estudios que involucran la sobrevivencia en enfermedades crónicas.

2.4 MODELOS NO PARAMÉTRICOS

Los métodos estadísticos más utilizados son los no paramétricos. Así, las curvas de sobrevivencia por lo general se producen usando uno de dos métodos: el análisis actuarial o el método del límite de producto de Kaplan-Meier.

El método Kaplan-Meier calcula la sobrevivencia cada vez que un estudiante se retira. El análisis actuarial divide el tiempo en intervalos y calcula la sobrevivencia en cada intervalo. El procedimiento Kaplan-Meier da proporciones exactas de sobrevivencia debido a que utiliza tiempos precisos; el análisis actuarial da aproximaciones, debido a que agrupa los tiempos en intervalos. Antes de que se extendiera el uso de ordenadores, el método actuarial era más fácil de usar para un número muy grande de observaciones.

El método de Kaplan-Meier se utiliza cuando la muestra es menor de 30 y también para muestras mayores de 30 en las que se conocen los tiempos individuales de los censurados y no censurados.

2.4.1 Método de Kaplan - Meier**:** Las técnicas de estimación no paramétricas con datos censurados, se inicia con los aportes de Kaplan y Meier en el año 1958, quienes publicaron resultados obtenidos para observaciones censuradas a la derecha y desarrollaron un estudio de las propiedades básicas de un nuevo estimador que luego se conoció con el nombre de sus creadores, el cual es conocido también como el nombre de estimador limite producto.

El estimador de Kaplan-Meier es uno de los más utilizados en los paquetes estadísticos, es un estimador no paramétrico ya que no tiene ninguna estructura para la función de distribución de probabilidad del tiempo de vida. La característica distintiva del análisis con este método es que la proporción acumulada que sobrevive se calcula para el tiempo de sobrevivencia individual de cada paciente y no se agrupan los tiempos de sobrevivencia en intervalos. Por esta razón es especialmente útil para estudios que utilizan un número pequeño de pacientes. El método de Kaplan-Meier incorpora la idea del tiempo al que ocurren los eventos.

El estimador de la sobrevivencia de Kaplan-Meier se define de la siguiente manera

$$\tilde{S}(t)=\prod_{t_i<t}\frac{n_i-d_i}{n_i} \qquad (6)$$

De donde

n_i = es el número de individuos en riesgo en t_i , es decir, el número de estudiantes activos y no censurados justo antes de t_i

d_i = es el número de estudiantes retirados en el instante t_i

Si se da el caso en que alguna observación censurada cuyo valor coincide con un tiempo de fallo, se hace la hipótesis de que la unidad censurada ocurre inmediatamente después del tiempo de fallo, por lo cual las unidades censuradas en ese instante se contabilizan como unidades de riesgo, esta hipótesis es razonable puesto que un individuo censurado en un tiempo t casi siempre sobrevive después de t. Por lo cual de la ecuación (6) se tiene que

$\tilde{S}(t_0)= 1$ y que $\tilde{S}(t_k)= 0$ si no existen casos censurado.

2.4.1.1 Propiedades**:** El estimador Kaplan – Meier es el estimador no paramétrico máximo verosímil de la función de sobrevivencia, lo cual hace que presente las propiedades de un buen estimador que pueden ser: insesgado, consistente, eficiente y suficiente. El estimador Kaplan – Meier cumple con varias de estas propiedades ya que es consistente (Efron, 1967), y eficiente (Wellner, 1982), además es un estimador de máxima verosimilitud para datos censurados (Peterson, 1997), y es normalmente asintótico (Breslow & Crowley, 1974). (Solano, 2008)

Gracias a estas propiedades se facilita en este estimador el cálculo y la utilización en problemas con datos censurados a la derecha, cuando las estimaciones de S(t) se realizan con datos en ausencia de censura, esta coincide con el estimador no paramétrico de la función de sobrevivencia. Lo anterior hace que el estimador Kaplan – Meier sea un buen estimador para muestras grandes, pero para muestras muy pequeñas estas propiedades ya no son tan robustas.

2.4.1.2 Varianza del estimador Kaplan-Meier: El estimador de Kaplan –Meier da una estimación puntual o un único valor para esta función en cualquier instante t, por lo tanto si se desea tener una precisión de este estimador en diferentes instantes de tiempo sobre diferentes muestras es necesario contar con un buen estimador de la varianza. Para la cual utilizaremos el método delta, el cual se basa en las series de Taylor de primer orden.

En este método se utiliza la aproximación de las series de Taylor con el fin de obtener una función lineal que nos aproxime a una función más complicada como el estimador de la función de sobrevivencia.

Sea f(x) una función de densidad de probabilidad de una variable aleatoria X, la expansión de las series de Taylor de primer orden alrededor de la media es

$$F(x) \cong f(\mu) + (X-\mu)\, f'(\mu) \qquad (7)$$

De la ecuación (7) se obtiene que

$$\text{Var}(f(x)) \cong \{f'(\mu)\}^2 \text{Var}(x-\mu)$$
$$\text{Var}(f(x)) \cong \{f'(\mu)\}^2 \sigma^2$$

Donde σ^2 es la varianza de la variable aleatoria X.

Para ilustrar con un ejemplo se encontrara la varianza de la función logaritmo natural y exponencial

Con la función logaritmo natural

$$\ln(X) \cong \ln(\mu) + (X-\mu)\frac{1}{\mu} \qquad (8)$$

$$\tilde{\text{Var}}(\ln X) \cong \frac{1}{\mu}\tilde{\sigma} \qquad (9)$$

Con la función exponencial

$$\exp(X) \cong \exp(\mu) + (X-\mu)\exp(\mu) \qquad (10)$$

$$\tilde{\mathrm{Var}}(\exp X) \cong \tilde{\sigma}^2\ (\exp(\mu))^2 \qquad (11)$$

(Barrera, 2008)

2.4.1.3 Varianza de función de sobrevivencia: Para calcular el estimador de la varianza por el método delta, primero se calcula el estimador delta del logaritmo natural

$$\tilde{S}(t) = \prod_{ti<t} \frac{ni-di}{ni}$$

$$\ln(\tilde{S}(t)) = \sum_{ti<t} ln \frac{ni-di}{ni} \qquad (12)$$

Sustituyendo $\tilde{p}_i = \frac{ni-di}{ni}$ en la ecuación (12) se tiene

$$\ln(\tilde{S}(t)) = \sum_{ti<t} ln\, \tilde{p}_i \qquad (13)$$

Tomando varianza en ambos lados de la ecuación (13) se tiene que

$$\tilde{\mathrm{Var}}\ (\ln(\tilde{S}(t))) = \sum_{ti<t} Var(ln\, \tilde{p}_i) \qquad (14)$$

Suponiendo la independencia de las variables $\tilde{p}_i$ y además que se distribuyen como variables aleatorias de Bernoulli con probabilidad constante p_i . Por lo cual se tiene como estimador de p_i a $\tilde{p}_i$ y como estimador de la varianza la siguiente expresión $\frac{\tilde{p}_i(1-\tilde{p}_i)}{ni}$, utilizando las ecuaciones (8) y (9) se tiene que

$$\widetilde{Var}(\ln(\tilde{p}_i)) \cong \frac{1}{\tilde{p}_i^2} \frac{\tilde{p}_i(1-\tilde{p}_i)}{ni} \qquad (15)$$

Como $\tilde{p}_i = \frac{ni-di}{ni}$ entonces la ecuación (15) se convierte en

$$\widetilde{Var}(\ln(\tilde{p}_i)) \cong \frac{di}{ni(ni-di)} \qquad (16)$$

Remplazando la ecuación (16) en la ecuación (14) se tiene que

$$\widetilde{Var}\ (\ln(\tilde{S}(t))) \cong \sum_{ti<t} \frac{di}{ni(ni-di)} \qquad (17)$$

La expresión anterior da la estimación de la varianza del logaritmo natural, por lo cual se debe aplicar el método delta sobre la función exponencial como se hizo en (10) y (11) para eliminar el logaritmo natural

$$\widetilde{Var}\ (\exp\ (\ln(\tilde{S}(t)))) \cong (\exp(\ln(\tilde{S}(t))))^2 \sum_{ti<t} \frac{di}{ni(ni-di)} \qquad (18)$$

Simplificando se tiene que

$$\text{Var}\ (\tilde{S}(t)) \cong (\tilde{S}(t))^2 \sum_{t_i<t} \frac{d_i}{n_i(n_i-d_i)} \qquad (19)$$

La desviación estándar se calcula con la raíz cuadrada de la varianza dada en la ecuación (18) (Barrera, 2008)

2.4.1.4 Intervalos de confianza del estimador Kaplan – Meier**:** En este estimador se tiene que en el caso de muestras grandes, a un tiempo fijo t, se distribuye aproximadamente Normal, por lo cual el intervalo de confianza al 100(1- α) % de $S(t)$ para S (t) está dado por:

$$\tilde{S}(t) \pm z_{1-\alpha/2\tilde{\sigma}(t)}$$

Donde $z_{1-\alpha/2}$ es el percentil de una distribución Normal estándar al nivel $1-\alpha/2$, así $\Pr(Z < z_{1-\alpha/2}) = 1-\alpha/2$, con $Z \sim N(0,1)$. En algunos casos al construir intervalos de

esta manera pueden incluir valores fuera del rango [0,1], para evitar estos casos se aplica la distribución normal asintótica a una transformación de S (t) para el cual el rango no está restringido.

2.4.1.5 El estimador de Kaplan Meier ponderado (KMp): El principal problema que se presenta con el estimador Kaplan Meier cuando los datos de estudio contiene una fuerte censura o altísimo porcentaje de observaciones con censura, es que sus estimaciones por lo regular, no solo tienen la tendencia a sobrestimar la fiabilidad o supervivencia de los individuos en estudio con un alto margen de sesgo, si no que van acompañadas de muy poca variabilidad de las estimaciones. Las observaciones de Kaplan-Meier obtenidas son estimaciones sesgadas, ya que el método parte del supuesto de que los individuos con cesura, se conservan activos hasta el siguiente fallo, pues esto es como suponer que el paso del tiempo de un año al siguiente no tiene ningún efecto ni acción sobre las observaciones, por lo cual es de necesidad reducir el sesgo que producen las estimaciones de Kaplan-Meier con datos censurados. En nuestro caso no se presenta este problema debido a que el porcentaje de censura en las observaciones es bajo.

Para esta aparente debilidad del estimador, Bahrawar (2005), propone una modificación al método original que consiste en ponderar o acompañar a las observaciones con censura con un factor o tasa de no censura. Bahrawar et al aplico con esta metodología con datos de trasplantes de corazón en Stanford, donde la variable de respuesta era la supervivencia del paciente después del trasplante. La tasa de censura de la base de datos de Stanford era del 27%. El método de KMp, considera la censura como una parte fundamental del análisis. El factor de ponderación W_j, conocida como tasa de no censura está definida como

$$W_J = \frac{n_j - c_j}{n_j} \quad \text{Con} \quad 0 \le w_j \le 1$$

En el caso que W_j=1, no hay censura en el instante t_j debido a que d_j=0, pero si W_j < 1, en el instante t_j hay por lo menos una censura.
De esta manera el estimador KMp se define de la siguiente manera

$$\tilde{S}(t)=\begin{cases} 1 \text{ si } t=0 \\ \prod_{j:t(j)\le t} W_j\left(\frac{n_j-d_j}{n_j}\right) \\ 0 \text{ si } t \ge t_n \end{cases}$$

Para este caso las estimaciones KMp= $\tilde{S}(t)$ estarían definidas para todo t≥ o, alcanzando el valor 0, aun en el caso que la última observación sea censurada, ya que Bahrawa et al sugieren considerar la última observación censurada como un fallo para garantizar su definición en todo instante. En el caso que no haya censura, $S(t)$ coincide con la función empírica de sobrevivencia de fiabilidad (FES), ya que W_j=1 y $S\hat{}(t)$ coincidiría con KM, que a su vez se convierte en la FES, si todos los datos son completos.

2.4.1.6 La función empírica de supervivencia (FES): Si se tienen n tiempos de fallos ordenados, $t_1 \le t_2 \le t_3 \ldots \le t_n$, donde no se presenta censura, el número de observaciones que sobreviven en el tiempo t_i es n – i. Por lo tanto podemos aceptar que una estimación no paramétrica de la función de supervivencia S(t) en t_i seria la proporción de observaciones que sobrevivan en el instante t_i

$$S\hat{}(t)=\frac{n-i}{n} \qquad i=1,2,3 \ldots \ldots n$$

Con lo anterior observamos que hay una probabilidad cero de sobrevivir más allá del tiempo t_n, como es improbable que ningún valor muestral alcance el tiempo hasta el fallo más alto, esta expresión tiende a subestimar la fiabilidad de la componente. En una

muestra de datos sin censura, el estimador KM coincide con este estimador, usualmente conocido como función empírica de sobrevivencia o fiabilidad (FES).

2.4.2 El estimador de Nelson - Aalen (NA): Este estimador fue propuesto por primera vez por Nelson W. Aalen (1969), y luego por Altschuler en 1970 quién lo descubrió utilizando técnicas de conteo con animales.

Como lo vimos anteriormente la función de riesgo acumulada está dada por $H(t) = \Delta t = -\ln S(t)$, un estimador de esta función se define como

$$\widehat{H}(t) = -\ln \widehat{R}(t)$$

Donde tenemos que $\widehat{R}(t)$ es el estimador Kaplan-Meier de R(t).Nelson-Aalen definió otro posible estimador de esta función conocida como función empírica de riesgo acumulado, definida de la siguiente manera

$$\widehat{H}(t) = \widehat{\Delta}(t) = \sum_{j.t_j \le t} \frac{d_j}{n_j}$$

Donde d_j representa el número de fallos ocurridos en el instante t_j y n_j el número de individuos en riesgo en t_j

De la función empírica de riesgo acumulado tenemos que el cociente d_j / n_j nos da una estimación de la probabilidad condicionada de que una observación que sobrevive hasta justo antes del instante, falle en este mismo instante. A partir de esta cantidad se construyen los estimadores de la función de fiabilidad R(t) y la función de impacto H(t).

$\tilde{H}(t)$ Recolecta para todo $t_j \le t$ las estimaciones puntuales d_j / n_j de la función de riesgo $h(t_j) = h_j$ en cada instante de fallo t_j y acumularlas. El estimador de Nelson-

Aalen de la función de sobrevivencia se obtiene a partir de la relación logarítmica entre Δ(t) y $R(t)$ el cual se define así

$$\tilde{R}(t) = e^{-\tilde{\Lambda}(t)} = e^{-\sum_{j:t_j \le t} \frac{d_j}{n_j}}$$

Para el caso de una variable continua este estimador ha tenido mucha discusión por parte de Nelson (1972), Breslow y Crowley (1974), Efron (1977) y Altschuler (1979). Los cuales llegaron a la conclusión que en este caso $\tilde{H}(t)$ y $\tilde{\Lambda}(t)$ son asintóticamente equivalentes con la excepción de valores altos de t, donde las estimaciones son menos estables, la diferencia entre estos será por lo general pequeña, no existe ninguna razón suficiente para escoger alguna de estas. Las estimaciones $\tilde{H}(t)$ y $\tilde{\Lambda}(t)$ son de gran utilidad en la construcción de gráficas para evaluar la selección de una determinada familia paramétrica de distribuciones. (Barrera, 2008)

2.4.2.1 Estimadores no paramétricos de la función de supervivencia para datos truncados a la izquierda y censurados a la derecha**:** Turnbull fue el primero en referenciar un estimador no paramétrico de la función de sobrevivencia para datos truncados y censurados como puede verse en la revista Journal of the Royal estatistical society (1976)

A partir del estimador $R(t) = P(T > t)$ se pueden obtener estimaciones para otras funciones de interés como son la función de distribución $F(t) = 1 - R(t)$ o la función de riesgo acumulado $H(t) = -\ln(R(t))$.

El estimador de Turnbull al igual que el de Kaplan – Meier es de tipo límite producto, dentro de sus propiedades más importantes es que tiene un comportamiento asintótico, estudiado por Tsai, Jewell, y Wang (1987). Lai y Ying en 1991, estudiaron la

consistencia fuerte uniforme y la convergencia débil del proceso bajo hipótesis más generales

Los estimadores de sobrevivencia de muestras censuradas a la derecha pueden modificarse para obtener muestras truncadas a la izquierda y censuradas a la derecha. En este estudio asociamos al j-ésimo individuo una variable aleatoria X_j la cual indica el instante de entrada al estudio y un tiempo Y_j que indica el fallo o la censura. Se define $t_1 < t_2 < t_3 < ... < t_k$ como los diferentes tiempos de fallo en el caso de datos censurados a la derecha donde d_j es el número de fallos en el instante t_j. Para los estadísticos en el caso de censuras por la derecha es necesario calcular el número de observaciones en riesgo antes del instante t_j, el cual llamaremos r_j que coincide con el número de individuos que están en el estudio en el instante 0 y que tienen un tiempo de participación de al menos t_j, solamente en el caso de cesura a la derecha.

Cuando los datos son truncados a la izquierda, las cantidades r_j se definen como el número de observaciones que entraron al estudio antes del tiempo t_j y que tienen un tiempo de participación de al menos t_j, Con esta definición de r_j se tienen los estimadores de la función de fiabilidad para el caso de datos truncados. Pero debemos tener cuidado con estos estimadores ya que por ejemplo el estimador de límite producto de la función de sobrevivencia en el tiempo t es ahora un estimador por encima de t, condicionado a la supervivencia al menor de los tiempos de entrada X, entonces estimamos $P(T > t / T \geq X)$

2.4.2.2 Estimador de Turnbull**:** El estimador de la función de fiabilidad o de sobrevivencia $S(t) = R(t)$ para el modelo visto anteriormente de truncamiento a la izquierda y censura a la derecha está definido de la siguiente manera

$$\tilde{S}(t)=\prod_{i=1}^{n}\left(1-\frac{1_{\{Y_i\le t,\delta_i=1\}}}{\sum_{j=1}^{n}1_{\{X_j\le Y_i\le Y_J\}}}\right)=\prod_{i=1}^{n}\left(1-\frac{d_i}{r_i}\right)$$

Donde d_i representa el número de fallos en el instante Y_i y r_i es el número de observaciones en riesgo antes de ese mismo instante. El estimador de Turnbull puede presentar problemas en el caso de muestras muy pequeñas o grandes muestras con pocos valores truncados iniciales. Puede presentarse que d_i = r_i en algún y_i, sin importar que éste no sea el estadístico de mayor orden de la serie $y_1, y_{2,} y_3 \ldots y_n$, en ese caso $\tilde{R}(t)=\tilde{S}(t)=0$, para $t\ge y_i$, sin tener en cuenta las observaciones que haya después. Es decir aunque estemos observando supervivencias y fallos después de este punto. Este problema se presenta solamente en el caso de truncamiento.

En el caso que en el estimador de Turnbull haya ausencia de truncamiento coincide con el estimador de Kaplan-Meier y en ausencia de censura coincide con ele estimador de Linden-Bell (1971) y en el caso que no existe censura ni truncamiento el estimador coincide con la función empírica de distribución (FES).

2.4.2.3 Estimador de Nelson Aalen Extendido**:** Nelson (1969) estima la función de riesgo acumulada $H(t)$ en presencia de censura a la derecha, luego Pan y Chappell (1998) extienden este estimador para datos truncados a la izquierda, con el fin de corregir el sesgo producido por la subestimación de la supervivencia mencionada anteriormente.

La estimación de la función de riesgo acumulada está dada mediante la expresión

$$\tilde{H}(t)=\sum_{i=1}^{n}\frac{1_{\{y_i\le t,\delta_i=1\}}}{\sum_{j=1}^{n}1_{\{y_i\le y_j\}}}$$

El cual se puede expresar en el estimador de Nelson para la función de supervivencia de la siguiente manera

$$\tilde{R}(t) = \tilde{S}(t) = e^{-\tilde{H}(t)}$$

Gracias a los trabajos realizados por Nelson-Aalen (1976) y Breslow (1972) este estimador es conocido con el nombre de estimador de Nelson-Aalen o estimador de Breslow mas adelante en el año 1984 Flemyn y Harrington recomiendan este estimador como alternativa al estimador no paramétrico de máxima verosimilitud, dado que el anterior tiene menor error cuadrático medio con datos censurados cuando la probabilidad de sobrevivencia sea al menos 0.2.

Como se mencionó Pan y Chapell extendieron este estimador para datos truncados a la izquierda de la siguiente manera

$$r_i = \sum_{j=1}^{n} 1_{\{x_j \le y_i \le y_j\}}$$

Este estimador es recomendado porque soluciona el problema de subestimación dado en el estimador no paramétrico de máxima verosimilitud cuando hay truncamiento, los autores lo denominan el estimador de Nelson-Aalen extendido. Cabe resaltar que el problema de la subestimación queda casi pero no totalmente solucionado. En síntesis el estimador Nelson Aalen extendido para estimar el riesgo acumulado se presenta de la siguiente manera

$$\tilde{H}_e(t) = \sum_{i=1}^{n} \frac{1_{\{y_i \le t, \delta_i = 1\}}}{\sum_{j=1}^{n} 1_{\{x_j \le y_i \le y_j\}}} = \sum_{y_i \le t} \frac{d_i}{r_i}$$

Para estimar la función de supervivencia a de fiabilidad se utiliza esta expresión a través del estimador

$$\tilde{R}_e(t) = \tilde{S}_e(t) = e^{-\tilde{H}_e(t)}$$

(Barrera, 2008)

2.4.3 Comparaciones de funciones de Supervivencia: Para comparar la sobrevivencia de dos o más grupos de observaciones, debemos hacer todas las comparaciones dos a dos mediante un test estadístico apropiado, con el fin de determinar si todos los grupos presentan la misma supervivencia o cuales grupos son distintos. Las representaciones graficas de las curvas de sobrevivencia dan una alerta sobre posibles diferencias entre éstas pero con las pruebas estadísticas se obtienen unas diferencias más significativas entre las curvas, indicándonos que el factor considerado es importante en el riesgo de que falle.

Para comparar la igualdad de dos o más funciones de fiabilidad con datos censurados se presentan los siguientes contrastes no paramétricos
Test de Log-Rank o también conocido con el nombre de riesgos proporcionales.
Test de Breslow también llamado como el test de Wilcoxon generalizado
Test de Tarone Ware

De estos tres test el de Log-Rank es muy potente para calcular diferencias cuando los logaritmos de las curvas de supervivencias son proporcionales, pero tiene muchos problemas para detectar las diferencias cuando las curvas de supervivencias se cruzan. El test de Breslow detecta las diferencias cuando las curvas de supervivencia se cruzan o se cortan pero solamente al principio por lo cual no es recomendable para un estudio a largo plazo. Y el test de Tarone – Ware es un test intermedio a los otros dos. A continuación se nombran los aspectos más sobresalientes de estos test.

2.4.3.1 Test de log Rank: Es el test más potente cuando el cociente de las funciones de riesgo es aproximadamente constante, es decir que la hipótesis alternativa es de riesgos proporcionales.

Supongamos que se va a comparar la sobrevivencia de dos grupos A_1 y A_2, en donde la función de sobrevivencia es respectivamente S_1 Y S_2, y se tiene una muestra de cada población con su respectivo tamaño n_1 y n_2, en donde n= n_1 + n_2 es el número total de datos en la muestra combinada. Los tiempos de fallo se definen así $t_1 < t_2 < t_3 < ... < t_k$.

Definimos la hipótesis nula de la siguiente manera

$$H_0 = S_1(t) = S_2(T) \qquad \Delta t \leq t \,.$$

Donde t es el tiempo total de observación de la muestra.

En este test de Log-Rank se compara el número de fallos observados dentro de cada grupo A_1 y A_2, además del número de fallos esperados de la hipótesis nula.

Cuando la hipótesis nula es cierta, la probabilidad condicional de fallo en t_j es igual para los dos grupos λ_i, por lo tanto la distribución de probabilidad de (d_{1j}, d_{2j}) esta dada de la siguiente manera

$$\prod_{i=1}^{2}\left[\binom{n_{ij}}{d_{ij}}\lambda_j^{\,d_{ij}}(1-\lambda_i)^{n_j-d_{ij}}\right] = \prod_{i=1}^{2}\left[\binom{n_{ij}}{d_{ij}}\right]\lambda_j^{\,d_{ij}}(1-\lambda_i)^{n_j-d_{ij}}$$

En donde

d_j = número total de fallos ocurridos en el tiempo t_j

n_j = número total del ítems en riesgo antes de t_j

d_{ij}, i=1,2: el número de fallos ocurridos en el tiempo t_j, entre los individuos del grupo i

n_{ij}, i=1,2: el número de ítems en riesgo al principio de t_j, entre los individuos del grupo i.

Con la distribución de probabilidad de (d_{1j}, d_{2j}), se define para cada i el estadístico de Log-Rank así:

$$u_i = \sum_{j=1}^{k} (d_{ij} - e_{ij})$$

En el caso que el valor de k sea suficientemente grande se tiene por el teorema central del límite lo siguiente

$$\frac{\mu_i}{\sqrt{\text{var}(\mu_1)}} = \frac{\sum_{j=}^{1} (d_{ij} - n_{ij})}{\sqrt{\sum_{j=1}^{k} \text{var}(d_{ij})}} \sim N(0,1)$$

Por tanto

$$\frac{\mu_i}{\sqrt{\text{var}(\mu_1)}} \sim x^2{}_1$$

En este test las diferencias observadas en todos los tiempos de fallo tienen igual importancia, sin tener en cuenta el número de ítems en riesgo en cada uno. Sin embargo es más útil tener en cuenta las diferencias observadas al principio de la observación en el estadístico que las observadas al final. Debido que al inicio se observan más casos. De lo escrito anteriormente se define una familia de test para los cuales se tiene el siguiente estadístico

$$u = \frac{\sum_{j=1}^{k} w_j (d_{ij} - n_{ij})}{\sqrt{\sum_{j=1}^{k} w_j{}^2 (d_{ij})}}$$

Donde $w = (w_1, w_2, w_3, ..., w_k)$ es un vector de pesos que pondera las diferencias entre fallos observados y fallos esperados a lo largo del tiempo de observación. Bajo hipótesis nula el estadístico tiene distribución N(0,1). Si se consideran varios vectores pesos se obtienen diferentes test de la siguiente manera:

Si $w = 1$ se obtiene el test de Log-Rank

Si $w_j = n_j$ se obtiene el test de Breslow o wilcoxon generalizado

Si $w_j = \sqrt{n_j}$ se obtiene el test de Tarone-Ware

Si $w_j = \prod_{i=1}^{j} \frac{n_i - d_i + 1}{n_i + 1}$ se obtiene el test de Prentice

Los tres test mencionados anteriormente presentan grandes problemas porque únicamente detectan diferencias cuando se presenta alguna de las siguientes situaciones $S_1(t) < S_2(t)$ ó $S_1(t) > S_2(t)$ para todo t, de lo contrario el valor del estadístico u sería una suma de valores positivos y negativos en el cual el resultado se

aproxima a cero, lo que hace que no sea estadísticamente significativo. Por lo tanto estos test son de utilidad cuando la hipótesis alternativa se corresponde con la hipótesis de riesgo acumulado.

3. APLICACION DEL MODELO KAPLAN - MEIER

Con el fin de hacer un aporte concreto al conocimiento de un problema que afecta a la carrera de matemáticas con énfasis en estadística, aplicamos los conceptos vistos anteriormente al estudio de las cohortes de estudiantes que ingresaron a la universidad del Tolima en los semestres A y B de 2002. Utilizamos para ello los datos facilitados por la coordinación del programa de matemáticas y estadística sobre los 88 estudiantes que ingresaron en los periodos mencionados.
No sobra advertir que el modelo aquí implementado puede ser aplicado para otras cohortes de la carrera y para otros programas de la universidad.

En el trabajo se tiene en cuenta tres variables, la primera es el tiempo que dura cada estudiante durante el estudio, la segunda es el motivo de su retiro, ya sea graduado o no graduado, y la tercera es el género del estudiante si es hombre o mujer. El estudio se hizo en dos partes, en la primera no se tuvo en cuenta el género de los estudiantes por lo cual se tomaron los datos en su totalidad y para la segunda si se tuvo en cuenta el género separando si era hombre o mujer y por último se hacía una comparación de estos dos por medio de test estadísticos.

Para realizar los cálculos se utilizaron los paquetes estadísticos SPSS 20.0 para el estimador Kaplan – Meier, el cual nos arroja resultados importantes como las tablas de sobrevivencia, las medias y medianas de sobrevivencia, comparaciones globales por medio de test y por ultimo una gráfica de las curvas de supervivencia. Con estos resultados se sacan las respectivas conclusiones del estudio, y para el estimador Nelson Aalen se utilizó el paquete estadístico XLSTAT 2013, el cual arroja resultados como la función de riesgo, función de sobrevivencia, limites inferiores y superiores, y sus respectivas graficas con las cuales se pueden sacar las respectivas conclusiones.

A continuación se presentaran los resultados obtenidos aplicando la metodología de Kaplan – Meier, utilizando el paquete estadístico SPSS 20.0. Se tiene en cuenta el

tiempo de sobrevivencia como el tiempo que tarda un estudiante desde que ingresa hasta que sale de la carrera, la mortalidad es la deserción y la censura se toma como la salida de los estudiantes de la carrera por motivos diferentes a la deserción, en este caso la graduación.

Los siguientes son los resultados arrojados por el programa SPSS 20.0, en donde se muestran las diferentes estadísticas y las curvas de sobrevivencia del programa matemáticas con énfasis en estadística de la universidad del Tolima.

Tabla 1. Resumen del procesamiento de los casos

Nº total	Nº de eventos	Censurado	
		Nº	Porcentaje
88	67	21	23,9%

Fuente (autor)

Teniendo en cuenta que:

N^0 total es la cantidad de estudiantes que ingresaron a la carrera durante los semestres A y B en el año 2002.

N^0 de eventos es la cantidad total de estudiantes que desertaron o se retiraron durante el estudio sin haberse graduado.

Los estudiantes censurados son aquellos que se graduaron.

Según los resultados obtenidos en la tabla anterior podemos concluir que

El número total de estudiantes que se matricularon fue 88 durante los semestres A y B de 2002

Durante el estudio desertaron 67 estudiantes, es decir no se graduaron.

De los estudiantes que se matricularon en el año 2002 se graduaron 21, lo que corresponde a un porcentaje del 23,9%.

En el ANEXO A se puede observar la tabla de supervivencia

Tabla 2. Medias y medianas del tiempo de supervivencia

Media				Mediana			
Estimación	Error típico	Intervalo de confianza al 95%		Estimación	Error típico	Intervalo de confianza al 95%	
		Límite inferior	Límite superior			Límite inferior	Límite superior
7,065	,665	5,761	8,369	5,000	1,279	2,493	7,507

Fuente (autor)

Teniendo en cuenta que

La estimación es el tiempo promedio que permanecen los estudiantes durante el tiempo de estudio

El error típico es la desviación estándar del estimador del promedio

Intervalo de confianza son los límites entre los cuales se espera el promedio de sobrevivencia de los estudiantes durante sus estudios.

La confianza es la probabilidad con al que se espera que el promedio de supervivencia este dentro del intervalo (95%)

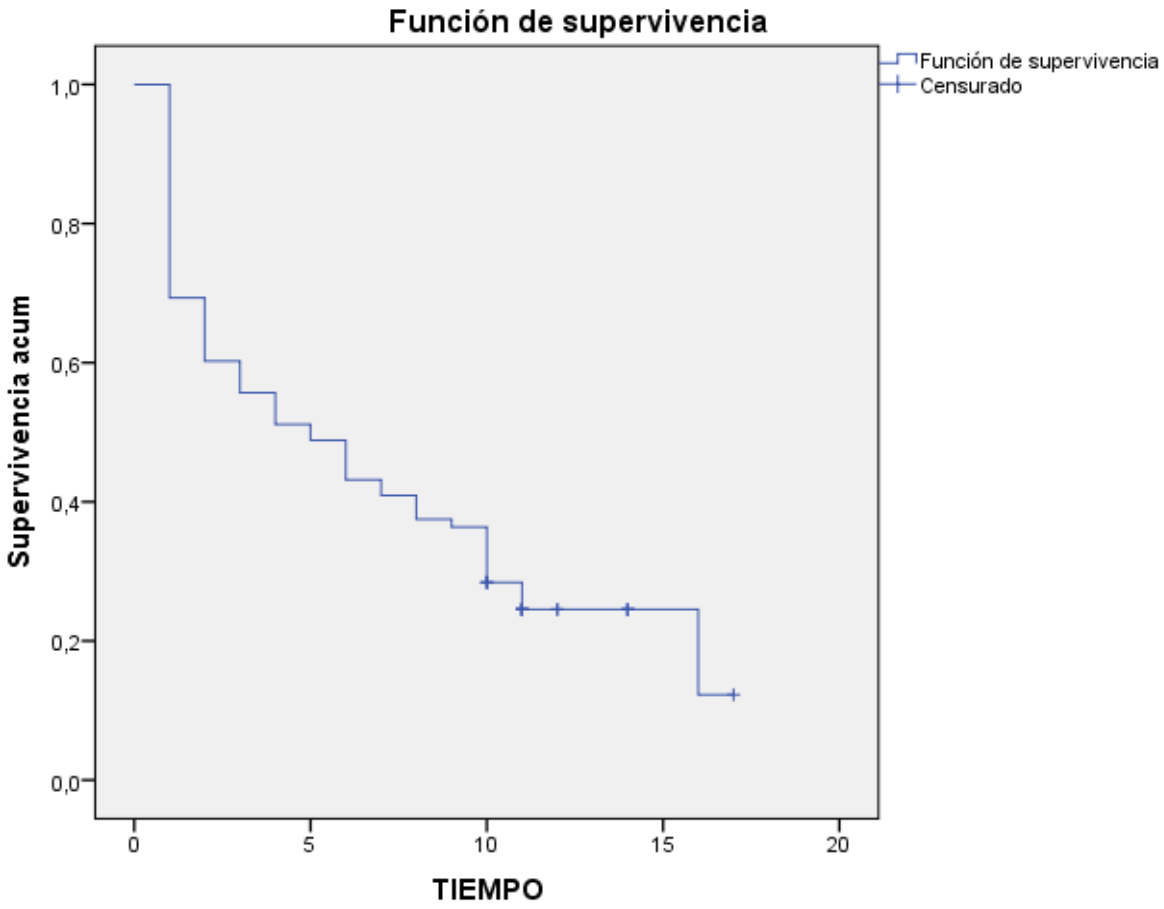
Con los resultados obtenidos en la tabla anterior para las medias del tiempo de sobrevivencia podemos concluir que

Para los estudiantes que ingresan al programa de matemáticas con énfasis en estadística de la universidad del Tolima, el promedio de permanencia es de 7,065 semestres.

El tiempo promedio de permanencia de los estudiantes del programa con una probabilidad del 95% está entre 5.76 y 8.40 semestres aproximadamente.

La siguiente figura 2 nos muestra la curva de sobrevivencia, en donde el eje vertical representa la sobrevivencia acumulada con una probabilidad entre 0 y 1, y el eje horizontal representa el tiempo (semestre) en el cual se retiran los estudiantes del programa.

Figura 2. Funcion de supervivencia



Fuente (autor)

En la figura anterior podemos observar que la probabilidad de retiro de los estudiantes se presenta en mayor cantidad durante los primeros 4 semestres y qué no todos los estudiantes que pasan del 10 semestre se gradúan.

3.1 RESULTADOS TENIENDO EN CUENTA EL GÉNERO

A continuación se mostraran los resultados obtenidos al en el paquete estadístico SPSS 20.0, en el cual se comparan los tiempos promedios de supervivencia de hombres y mujeres

Tabla 3. Resumen del procesamiento de los casos por género

GENERO	Nº total	Nº de eventos	Censurado	
			Nº	Porcentaje
Mujer	**42**	**32**	**10**	**23,8%**
Hombre	**46**	**35**	**11**	**23,9%**
Global	**88**	**67**	**21**	**23,9%**

Fuente (autor)

En la tabla 3 se puede observar que

El número total de estudiantes en estudio es 88 de los cuales 42 son mujeres, se retiraron 32 y se graduaron 10 lo cual corresponde al 23,8% y 46 son hombres, se retiraron 35 y se graduaron 11 lo cual corresponde al 23,9%. Con estos datos observamos que de los estudiantes que ingresan a la carrera matemáticas con énfasis en estadística es más probable que se retiren a que se gradúen a lo largo de su estudio y además que no hay diferencias significativas entre las proporciones de mujeres y hombres que se retiran de la carrera.

En el ANEXO B se puede observar la tabla de supervivencia por género

Tabla 4. Medias y medianas del tiempo de supervivencia por género

GENERO	Media				Mediana			
	Estimación	Error típico	Intervalo de confianza al 95%		Estimación	Error típico	Intervalo de confianza al 95%	
			Límite inferior	Límite superior			Límite inferior	Límite superior

Mujer	7,857	,938	6,018	9,696	7,000	1,389	4,278	9,722
Hombre	6,337	,925	4,523	8,151	3,000	,782	1,468	4,532
Global	7,065	,665	5,761	8,369	5,000	1,279	2,493	7,507

Fuente (autor)

a. La estimación se limita al mayor tiempo de supervivencia si se ha censurado.

En los resultados de la tabla 4 se puede observar que las mujeres tienden a tener un promedio de sobrevivencia mayor que los hombres, ya que el promedio para mujeres es de 7,857 y el intervalo de confianza esta entre 6,018 y 9,696, para los hombres el promedio de supervivencia es de 6,337 y el intervalo de confianza esta entre 4,523 y 8,151. A pesar de estos resultados podemos decir que las diferencias entre la sobrevivencia de las mujeres y los hombres en la carrera matemáticas con énfasis en estadística no es significativa.

3.2 COMPARACIONES GLOBALES

La siguiente tabla se realizan las comparaciones globales, utilizando tres pruebas estadísticas, teniendo en cuenta el género de los estudiantes que ingresan al programa de matemáticas con énfasis en estadística de la universidad del Tolima, en este caso la hipótesis nula se refiere a la igualdad de las funciones de sobrevivencia.

Tabla 5. Prueba de igualdad de distribuciones de supervivencia para diferentes niveles de GENERO**.**

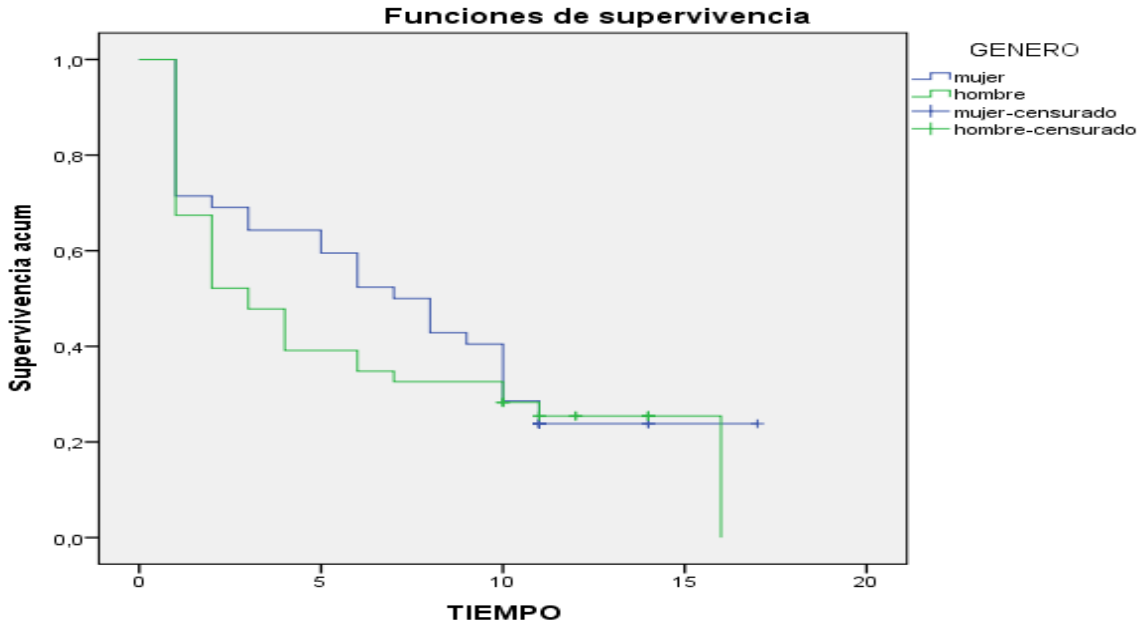
	Chi-cuadrado	Gl	Sig.
Log Rank (Mantel-Cox)	,506	1	,477
Breslow (Generalized Wilcoxon)	,994	1	,319
Tarone-Ware	,763	1	,382

Fuente (autor)

Según los resultados arrojados por el programa en la tabla 5 podemos concluir que a pesar que existe diferencia entre el género de los estudiantes esta no es estadísticamente significativa.

En la siguiente figura 3 se muestra las curvas de sobrevivencia de ambos géneros.

Figura 3. Funciones de supervivencia



En la figura 3 se puede observar que entre el semestre 2 y el semestre 10 hay una separación de las dos curvas de sobrevivencia por lo cual existe una diferencia entre los dos géneros favoreciendo a las mujeres, pero esta diferencia no es estadísticamente significativa, como ya lo habíamos visto en los test anteriores

4. RESULTADOS CON EL MODELO NELSON AALEN

Los siguientes son los resultados y las conclusiones que arroja el programa estadístico XLSTAT en el análisis de supervivencia aplicando el estimador Nelson Aalen, en la primera parte no se tendrá en cuenta el género y en la segunda sí.

Tabla 6. la Estadísticas descriptivas:

Total observado	Eventos	Censurados
88	67	21

Fuente (autor)

Tabla 7. Análisis de Nelson-Aalen:

Tiempo	A riesgo	Eventos	Censuradas	Función de riesgo acumulado	Desviación típica	Límite inferior (95%)	Límite superior (95%)	Función de supervivencia acumulada
1	88	27	0	0,307	0,059	0,191	0,423	0,736
2	61	8	0	0,438	0,075	0,291	0,585	0,645
3	53	4	0	0,513	0,084	0,349	0,678	0,598
4	49	4	0	0,595	0,093	0,412	0,778	0,552
5	45	2	0	0,640	0,099	0,446	0,833	0,528
6	43	5	0	0,756	0,111	0,537	0,974	0,470
7	38	2	0	0,808	0,117	0,578	1,039	0,446
8	36	3	0	0,892	0,127	0,643	1,141	0,410
9	33	1	0	0,922	0,131	0,666	1,178	0,398
10	32	7	3	1,141	0,155	0,838	1,444	0,320

11	22	3	9	1,277	0,173	0,937	1,617	0,279
12	10	0	2	1,277	0,173	0,937	1,617	0,279
14	8	0	6	1,277	0,173	0,937	1,617	0,279
16	2	1	0	1,777	0,529	0,740	2,814	0,169
17	1	0	1	1,777	0,529	0,740	2,814	0,169

Fuente (autor)

En la tabla 7 se tiene en cuenta que

La columna **tiempo** representa la permanencia de los estudiantes durante la carrera

La columna **riesgo** se refiere a los estudiantes que en el semestre indicado se encuentran matriculados y pueden presentar el evento (retiro)

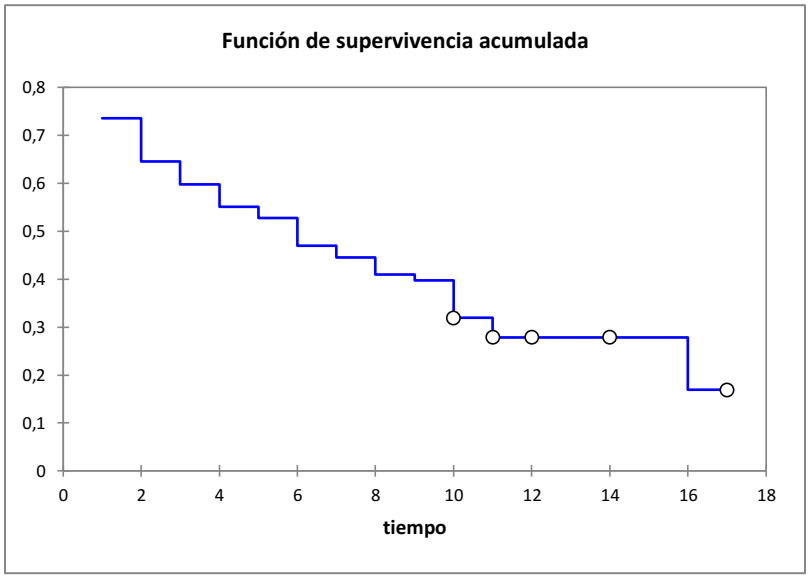
La columna **eventos** representa los estudiantes que en el semestre indicado se han retirado por causas diferentes a la graduación.

La columna **censurada** son los estudiantes que se han graduado en el respectivo semestre.

De la anterior tabla podemos sacar varias conclusiones como por ejemplo la probabilidad de sobrevivencia de los estudiantes antes del quinto semestre es de 53% y antes del décimo semestre es de 32%, dato menor que el anterior. Por tanto cada vez que un estudiante ingresa a un semestre más adelantado, la probabilidad de retiro es mayor.

Además con estos resultados se puede estimar la rapidez con que se presenta el evento de interés por ejemplo en el quinto semestre su estimación es de 64%, es decir que la probabilidad de que un estudiante que no se ha retirado en el quinto semestre lo haga en el sexto es del 64%.

Figura 4. Función de supervivencia acumulada



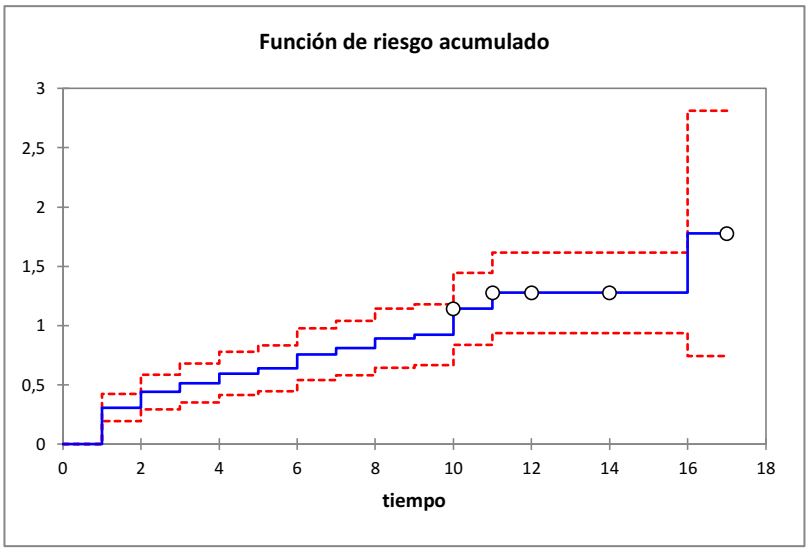
Fuente (autor)

En la figura 4 podemos observar que

La sobrevivencia de los estudiantes está en un rango entre el primero y séptimo semestre.

Entre el semestre 11 y 16 la probabilidad de sobrevivencia de los estudiantes es casi igual, a diferencia de los semestres anteriores que la probabilidad de sobrevivencia es más grande.

Figura 5. Función de riesgo acumulado

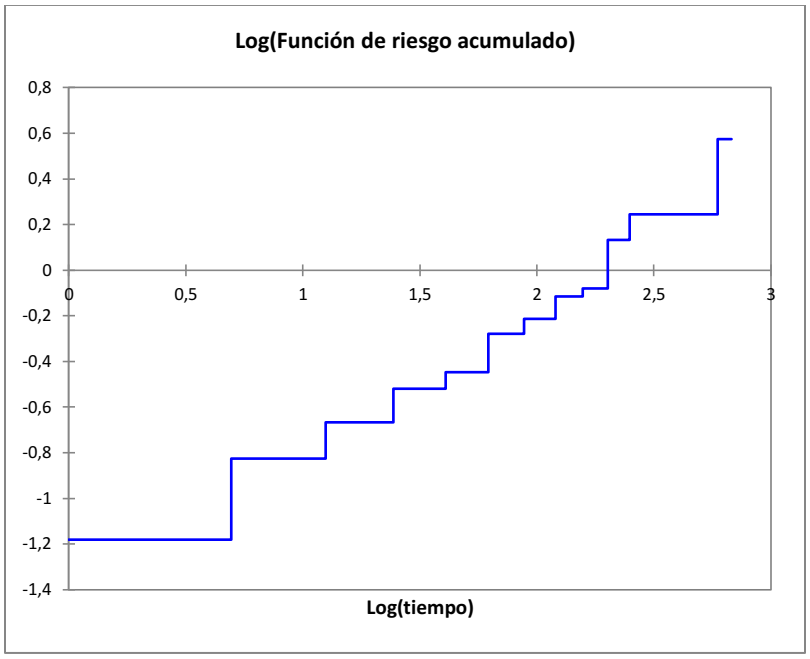


Fuente (autor)

De la figura 5 se tiene que

La curva del centro representa la sobrevivencia de los estudiantes durante el estudio, y las curvas exteriores representan los límites inferiores y superiores, los cuales indican la probabilidad máxima y mínima de que un estudiante se retire en cada uno de los semestres matriculados. Con esta figura se puede concluir que desde el primer semestre los estudiantes empiezan a retirarse de la carrera y a medida que avanzan los semestres la probabilidad de que un estudiante se retire es creciente, en el único intervalo que no crece significativamente es entre los semestres 11 y 16.

Figura 6. Log (Función de riesgo acumulado)



Fuente (autor)

4.1 RESULTADOS POR GENEROS

Después de hacer el análisis general, se hará el análisis separado por el género.

Primero se realiza el análisis de los datos para mujeres

Tabla 8. Estadísticas descriptivas (mujer):

Total observado	Eventos	Censurados
42	32	10

Fuente (autor)

En total durante el estudio se analizaron 42 mujeres de las cuales 32 presentaron el evento de interés (retiro) y 10 fueron censurados (graduación)

Tabla 9. Estadísticas descriptivas (hombre):

Total observado	Eventos	Censurados
46	35	11

Fuente (autor)

En total durante el estudio se analizaron 46 hombres de los cuales 35 presentaron el evento de interés (retiro) y 11 fueron censurados (graduación)

Tabla 10. Análisis de Nelson Aalen (mujer):

tiempo	A riesgo	Eventos	Censuradas	Función de riesgo acumulado	Desviación típica	Límite inferior (95%)	Límite superior (95%)	Función de supervivencia acumulada
1	42	12	0	0,286	0,082	0,124	0,447	0,751
2	30	1	0	0,319	0,089	0,145	0,493	0,727
3	29	2	0	0,388	0,101	0,189	0,587	0,678
5	27	2	0	0,462	0,114	0,238	0,686	0,630

6	25	3	0	0,582	0,134	0,320	0,844	0,559
7	22	1	0	0,628	0,141	0,351	0,904	0,534
8	21	3	0	0,770	0,163	0,450	1,091	0,463
9	18	1	0	0,826	0,173	0,488	1,164	0,438
10	17	5	0	1,120	0,217	0,695	1,545	0,326
11	12	2	7	1,287	0,247	0,803	1,771	0,276
14	3	0	2	1,287	0,247	0,803	1,771	0,276
17	1	0	1	1,287	0,247	0,803	1,771	0,276

Fuente (autor)

Tabla 11. Análisis de Nelson-Aalen (hombre):

tiempo	A riesgo	Eventos	Censuradas	Función de riesgo acumulado	Desviación típica	Límite inferior (95%)	Límite superior (95%)	Función de supervivencia acumulada
1	46	15	0	0,326	0,084	0,161	0,491	0,722
2	31	7	0	0,552	0,120	0,317	0,787	0,576
3	24	2	0	0,635	0,134	0,373	0,897	0,530
4	22	4	0	0,817	0,162	0,500	1,134	0,442
6	18	2	0	0,928	0,180	0,576	1,280	0,395
7	16	1	0	0,991	0,190	0,618	1,364	0,371
10	15	2	3	1,124	0,212	0,708	1,540	0,325
11	10	1	2	1,224	0,235	0,764	1,684	0,294
12	7	0	2	1,224	0,235	0,764	1,684	0,294
14	5	0	4	1,224	0,235	0,764	1,684	0,294
16	1	1	0	2,224	1,027	0,211	4,237	0,108

Fuente (autor)

En las tablas 10 y 11 se puede observar que entre los semestres 2 y 7 la probabilidad de supervivencia es mayor para las mujeres, por ejemplo en el semestre seis la

probabilidad de supervivencia de las mujeres es de 56% mientras que de los hombres es de 40%, después del semestre 10 las probabilidades son casi iguales por ejemplo para las mujeres en el semestre 14 es de 28% y para los hombres es de 29%. Además en la función de riesgo se observa que tienen más riesgo de retirarse del estudio los hombres que las mujeres durante todos los semestres, especialmente antes del semestre 10.

Figura 7. Función de supervivencia Acumulada (mujer)

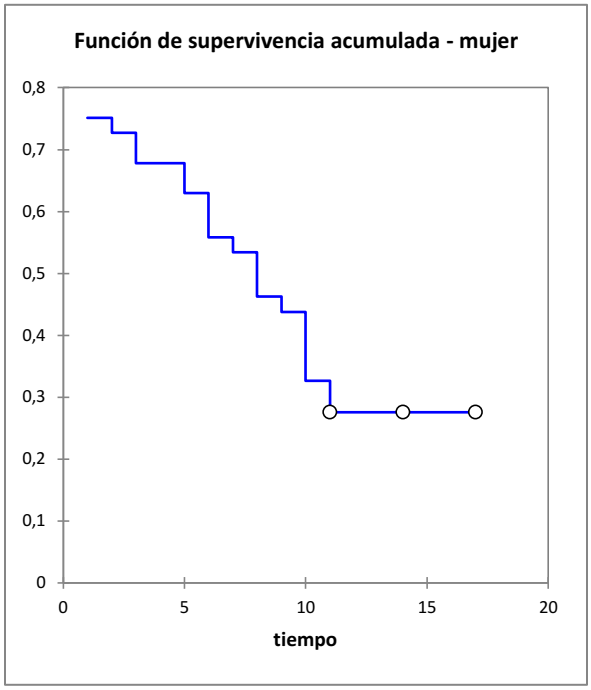
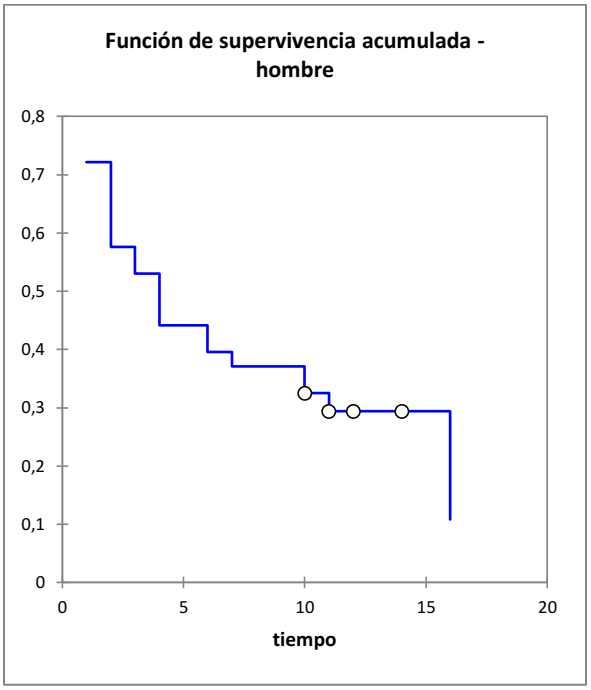


Figura 8. Función de supervivencia acumulada (hombre)



Fuente (autor)

En las figuras 7 y 8 se observa la función de sobrevivencia para los dos géneros. De la cual podemos decir que

Entre los semestres 2 y 10 las mujeres presentan una sobrevivencia más grande que los hombres pero del semestre 10 en adelante la probabilidad de sobrevivencia es similar como lo habíamos notado en las tablas 12 y 13.

Figura 9. Función de riesgo Acumulado (mujer)

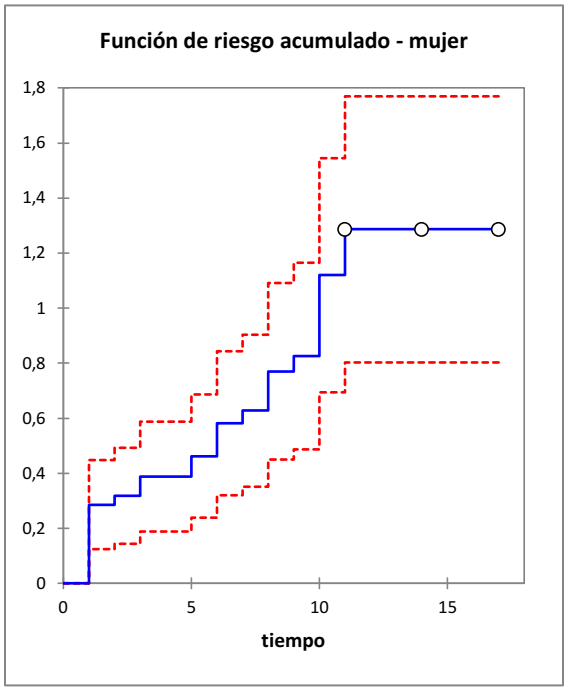
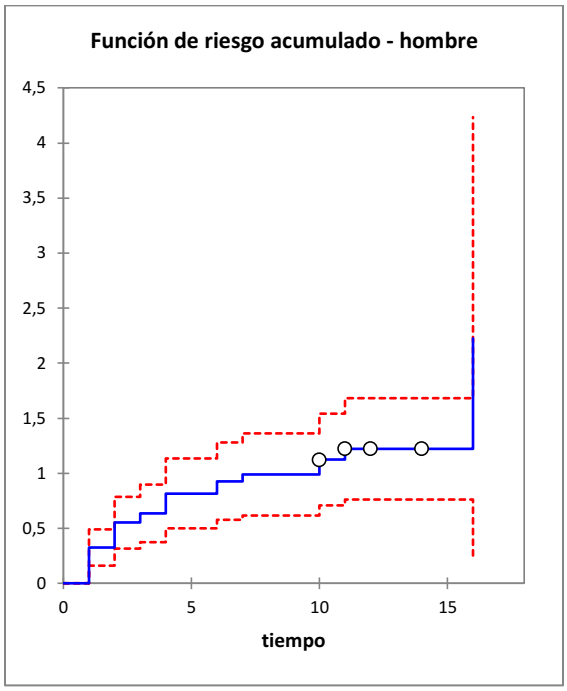


Figura 10. Función de riesgo acumulado (hombre)



Fuente (autor)

En las figuras 9 y 10 al igual que en las tablas 10 y 11 se observa que antes del semestre 10 la probabilidad de retiro es más grande en los hombres y que a medida que avanza el análisis esta probabilidad se va igualando para los dos géneros.

Figura 11. Log (Función de riesgo acumulado) - mujer

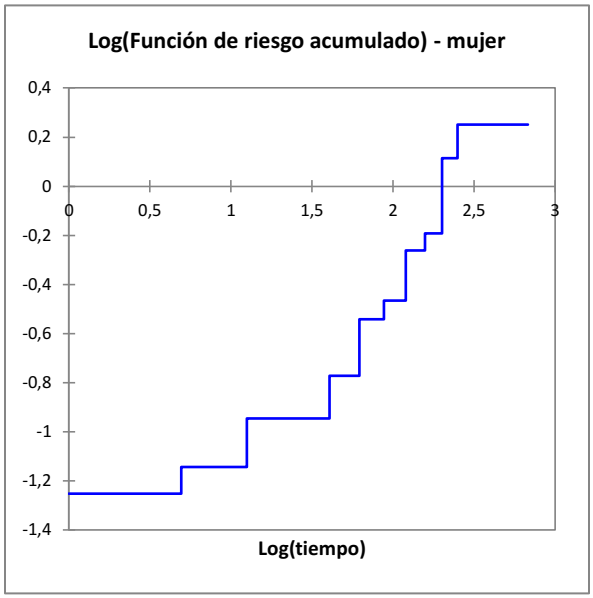
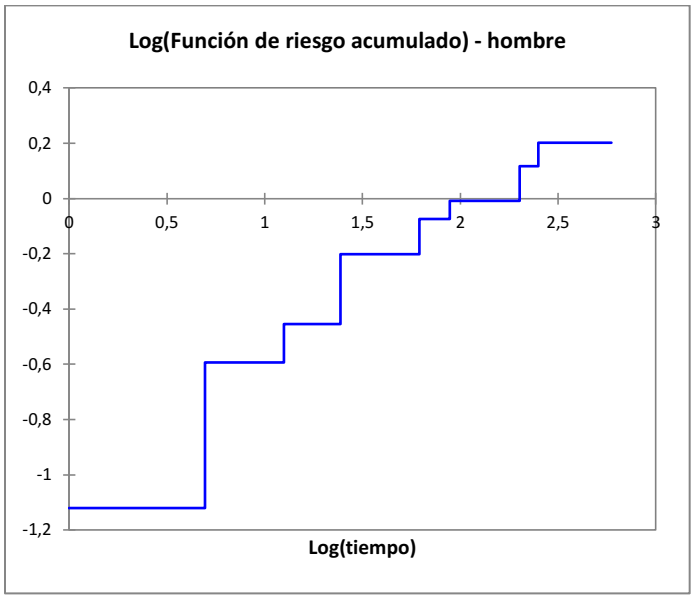


Figura 12. Log (Función de riesgo) acumulado)- hombre



Fuente (autor)

4.2 COMPARACIONES GLOBALES

A continuación se muestran los resultados obtenidos al comparar los géneros de los estudiantes por medio de los estadísticos Long – Rank, Wilcoxon y Tarone – Ware, utilizando el estimador de supervivencia de Nelson Aalen.

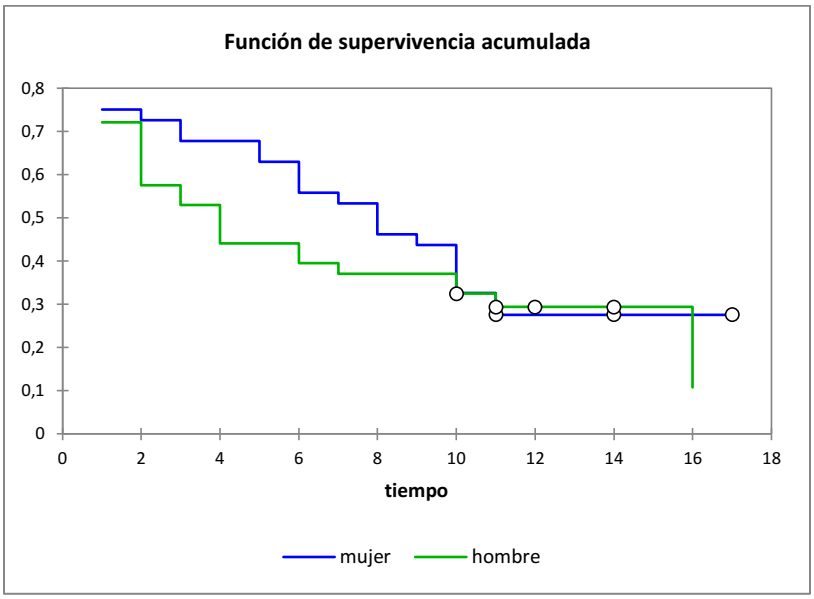
Tabla 12. Prueba de igualdad de las funciones de supervivencia acumuladas (GDL = 1):

Estadística	Valor observado	Valor crítico	p-valor	alfa
Log-Rank	0,506	3,841	0,477	0,050
Wilcoxon	0,994	3,841	0,319	0,050
Tarone-Ware	0,763	3,841	0,382	0,050

Fuente (autor)

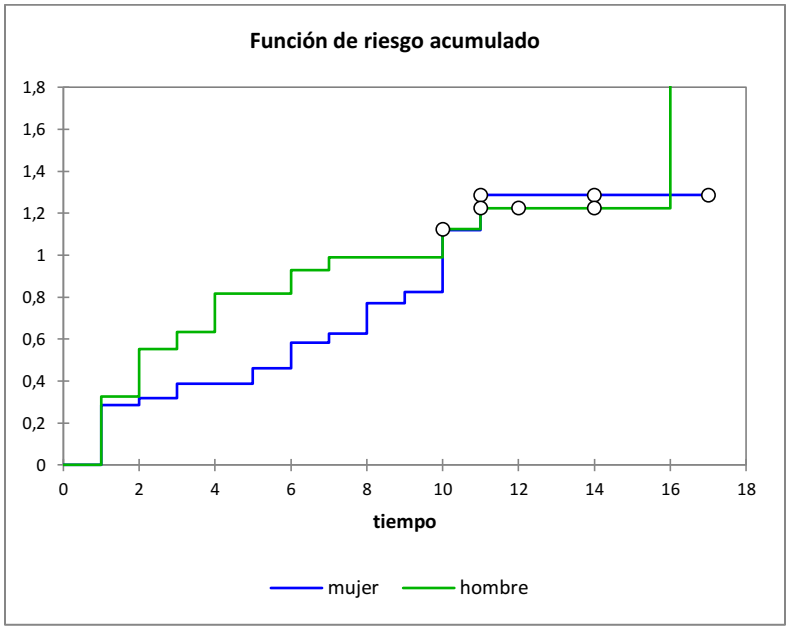
Como se puede observar en la tabla 12 los resultados con el estimador Nelson Aalen son iguales a los obtenidos con el estimador de Kaplan-Meier.

Figura 13. Función de supervivencia acumulada por género



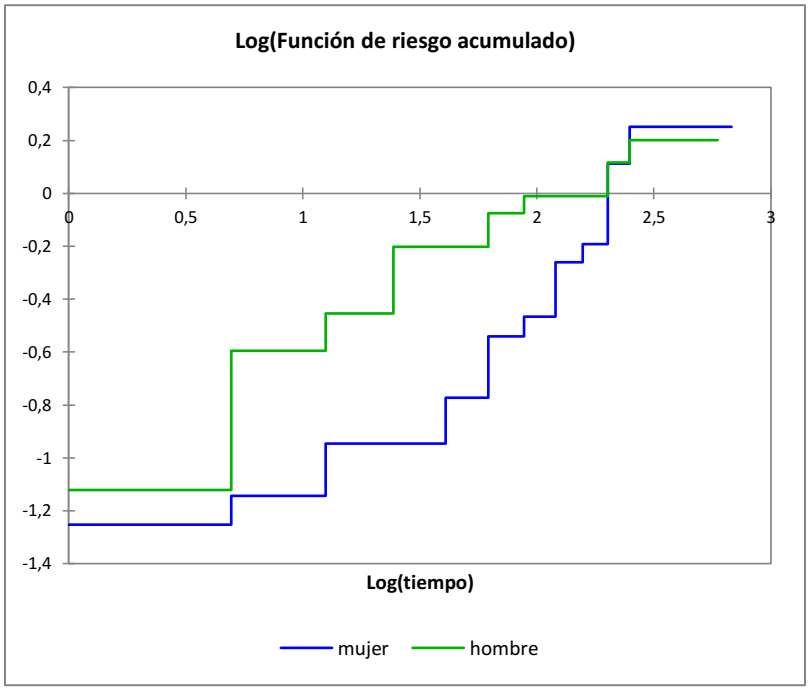
Fuente (autor)

Figura 14. Función de riesgo acumulado por género



Fuente (autor)

Figura 15. Log (Función de riesgo acumulado) por género



Fuente (autor)

Con las figuras 13, 14 y 15 podemos observar que al igual que en el estimador Kaplan-Meier, se concluye que entre el semestre 2 y el semestre 10 hay una separación de las dos curvas de sobrevivencia por lo cual existe una diferencia entre los dos géneros a favor de las mujeres, pero esta diferencia no es estadísticamente significativa, además que en estos semestres es más grande la probabilidad de retiro para los hombres que para las mujeres.

5. CONCLUSIONES

- La permanencia promedio de un estudiante que ingresa al programa de matemáticas con énfasis en estadística es de 7.06 semestres y este valor esta entre 5.76 y 8.40 con un 95% de confianza.

- En el programa de matemáticas con énfasis en estadística las mujeres tienen mejor promedio puntual que los hombres, pero las diferencias entre ellos no son significativas estadísticamente.

- La probabilidad de que un estudiante del programa deserte es mayor entre el segundo y el séptimo semestre.

- De los estudiantes matriculados en el programa se gradúan aproximadamente el 29%.

REFERENCIAS

Aguayo, C. (2007). *Fabis.* Recuperado el 10 de Mayo de 2013, de http://www.fabis.org/html/archivos/docuweb/SuperviKM_1r.pdf

Barrera, M. (2008). *Analisis de supervivencia aplicado a la desercion estudiantil en la uiversidad tecnologica de Pereira*.

Colosimo, A. (2000). *Analisis de sobrevivencia aplicada.* Projeto Fisher.

Fernandez, P. (1995). *fisterra.com*. Recuperado el 10 de marzo de 2013, de http://www.fisterra.com/mbe/investiga/supervivencia/supervivencia.asp

Solano, H. (2008). *Análisis de supervivencia en fiabilidad. Predicción en condiciones de alta censura y truncamiento: el caso de las redes de suministro de agua potable*.. Carrión García, A. Dir. ; Debón Aucejo, AM.

ANEXOS

ANEXO A.

Tabla . Tabla de supervivencia

	Tiempo	Estado	Proporción acumulada que sobrevive hasta el momento		Nº de eventos acumulados	Nº de casos que permanecen
			Estimación	Error típico		
1	1,000	retirado	.	.	1	87
2	1,000	retirado	.	.	2	86
3	1,000	retirado	.	.	3	85
4	1,000	retirado	.	.	4	84
5	1,000	retirado	.	.	5	83
6	1,000	retirado	.	.	6	82
7	1,000	retirado	.	.	7	81
8	1,000	retirado	.	.	8	80
9	1,000	retirado	.	.	9	79
10	1,000	retirado	.	.	10	78
11	1,000	retirado	.	.	11	77
12	1,000	retirado	.	.	12	76
13	1,000	retirado	.	.	13	75
14	1,000	retirado	.	.	14	74
15	1,000	retirado	.	.	15	73
16	1,000	retirado	.	.	16	72
17	1,000	retirado	.	.	17	71
18	1,000	retirado	.	.	18	70
19	1,000	retirado	.	.	19	69
20	1,000	retirado	.	.	20	68

21	1,000	retirado	.	.	21	67
22	1,000	retirado	.	.	22	66
23	1,000	retirado	.	.	23	65
24	1,000	retirado	.	.	24	64
25	1,000	retirado	.	.	25	63
26	1,000	retirado	.	.	26	62
27	1,000	retirado	,693	,049	27	61
28	2,000	retirado	.	.	28	60
29	2,000	retirado	.	.	29	59
30	2,000	retirado	.	.	30	58
31	2,000	retirado	.	.	31	57
32	2,000	retirado	.	.	32	56
33	2,000	retirado	.	.	33	55
34	2,000	retirado	.	.	34	54
35	2,000	retirado	,602	,052	35	53
36	3,000	retirado	.	.	36	52
37	3,000	retirado	.	.	37	51
38	3,000	retirado	.	.	38	50
39	3,000	retirado	,557	,053	39	49
40	4,000	retirado	.	.	40	48
41	4,000	retirado	.	.	41	47
42	4,000	retirado	.	.	42	46
43	4,000	retirado	,511	,053	43	45
44	5,000	retirado	.	.	44	44
45	5,000	retirado	,489	,053	45	43
46	6,000	retirado	.	.	46	42
47	6,000	retirado	.	.	47	41
48	6,000	retirado	.	.	48	40
49	6,000	retirado	.	.	49	39
50	6,000	retirado	,432	,053	50	38

51	7,000	retirado	.	.	51	37
52	7,000	retirado	,409	,052	52	36
53	8,000	retirado	.	.	53	35
54	8,000	retirado	.	.	54	34
55	8,000	retirado	,375	,052	55	33
56	9,000	retirado	,364	,051	56	32
57	10,000	retirado	.	.	57	31
58	10,000	retirado	.	.	58	30
59	10,000	retirado	.	.	59	29
60	10,000	retirado	.	.	60	28
61	10,000	retirado	.	.	61	27
62	10,000	retirado	.	.	62	26
63	10,000	retirado	,284	,048	63	25
64	10,000	graduado	.	.	63	24
65	10,000	graduado	.	.	63	23
66	10,000	graduado	.	.	63	22
67	11,000	retirado	.	.	64	21
68	11,000	retirado	.	.	65	20
69	11,000	retirado	,245	,046	66	19
70	11,000	graduado	.	.	66	18
71	11,000	graduado	.	.	66	17
72	11,000	graduado	.	.	66	16
73	11,000	graduado	.	.	66	15
74	11,000	graduado	.	.	66	14
75	11,000	graduado	.	.	66	13
76	11,000	graduado	.	.	66	12
77	11,000	graduado	.	.	66	11
78	11,000	graduado	.	.	66	10
79	12,000	graduado	.	.	66	9
80	12,000	graduado	.	.	66	8

81	14,000	graduado	.	.	66	7
82	14,000	graduado	.	.	66	6
83	14,000	graduado	.	.	66	5
84	14,000	graduado	.	.	66	4
85	14,000	graduado	.	.	66	3
86	14,000	graduado	.	.	66	2
87	16,000	retirado	,123	,090	67	1
88	17,000	graduado	.	.	67	0

Fuente (autor)

ANEXO B

Tabla 5. Tabla de sobrevivencia

GENERO	Tiempo	Estado	Proporción acumulada que sobrevive hasta el momento		Nº de eventos acumulados	Nº de casos que permanecen
			Estimación	Error típico		
	1	1,000	Retirado	.	.	1	41
	2	1,000	Retirado	.	.	2	40
	3	1,000	Retirado	.	.	3	39
	4	1,000	Retirado	.	.	4	38
	5	1,000	Retirado	.	.	5	37
Mujer	6	1,000	Retirado	.	.	6	36
	7	1,000	Retirado	.	.	7	35
	8	1,000	Retirado	.	.	8	34
	9	1,000	Retirado	.	.	9	33
	10	1,000	Retirado	.	.	10	32
	11	1,000	Retirado	.	.	11	31

	12	1,000	Retirado	,714	,070	12	30
	13	2,000	Retirado	,690	,071	13	29
	14	3,000	Retirado	.	.	14	28
	15	3,000	Retirado	,643	,074	15	27
	16	5,000	Retirado	.	.	16	26
	17	5,000	Retirado	,595	,076	17	25
	18	6,000	Retirado	.	.	18	24
	19	6,000	Retirado	.	.	19	23
	20	6,000	Retirado	,524	,077	20	22
	21	7,000	Retirado	,500	,077	21	21
	22	8,000	Retirado	.	.	22	20
	23	8,000	Retirado	.	.	23	19
	24	8,000	Retirado	,429	,076	24	18
	25	9,000	Retirado	,405	,076	25	17
	26	10,000	Retirado	.	.	26	16
	27	10,000	Retirado	.	.	27	15
	28	10,000	Retirado	.	.	28	14
	29	10,000	Retirado	.	.	29	13
	30	10,000	Retirado	,286	,070	30	12
	31	11,000	Retirado	.	.	31	11
	32	11,000	Retirado	,238	,066	32	10
	33	11,000	Graduado	.	.	32	9
	34	11,000	Graduado	.	.	32	8
	35	11,000	Graduado	.	.	32	7
	36	11,000	Graduado	.	.	32	6
	37	11,000	Graduado	.	.	32	5
	38	11,000	Graduado	.	.	32	4
	39	11,000	Graduado	.	.	32	3

Hombre	40	14,000	Graduado	.	.	32	2
	41	14,000	Graduado	.	.	32	1
	42	17,000	Graduado	.	.	32	0
	1	1,000	Retirado	.	.	1	45
	2	1,000	Retirado	.	.	2	44
	3	1,000	Retirado	.	.	3	43
	4	1,000	Retirado	.	.	4	42
	5	1,000	Retirado	.	.	5	41
	6	1,000	Retirado	.	.	6	40
	7	1,000	Retirado	.	.	7	39
	8	1,000	Retirado	.	.	8	38
	9	1,000	Retirado	.	.	9	37
	10	1,000	Retirado	.	.	10	36
	11	1,000	Retirado	.	.	11	35
	12	1,000	Retirado	.	.	12	34
	13	1,000	Retirado	.	.	13	33
	14	1,000	Retirado	.	.	14	32
	15	1,000	Retirado	,674	,069	15	31
	16	2,000	Retirado	.	.	16	30
	17	2,000	Retirado	.	.	17	29
	18	2,000	Retirado	.	.	18	28
	19	2,000	Retirado	.	.	19	27
	20	2,000	Retirado	.	.	20	26
	21	2,000	Retirado	.	.	21	25
	22	2,000	Retirado	,522	,074	22	24
	23	3,000	Retirado	.	.	23	23
	24	3,000	Retirado	,478	,074	24	22
	25	4,000	Retirado	.	.	25	21

26	4,000	Retirado	.	.	26	20
27	4,000	Retirado	.	.	27	19
28	4,000	Retirado	,391	,072	28	18
29	6,000	Retirado	.	.	29	17
30	6,000	Retirado	,348	,070	30	16
31	7,000	Retirado	,326	,069	31	15
32	10,000	Retirado	.	.	32	14
33	10,000	Retirado	,283	,066	33	13
34	10,000	Graduado	.	.	33	12
35	10,000	Graduado	.	.	33	11
36	10,000	Graduado	.	.	33	10
37	11,000	Retirado	,254	,065	34	9
38	11,000	Graduado	.	.	34	8
39	11,000	Graduado	.	.	34	7
40	12,000	Graduado	.	.	34	6
41	12,000	Graduado	.	.	34	5
42	14,000	Graduado	.	.	34	4
43	14,000	Graduado	.	.	34	3
44	14,000	Graduado	.	.	34	2
45	14,000	Graduado	.	.	34	1
46	16,000	Retirado	,000	,000	35	0

Fuente (autor)

En las tablas anteriores

La columna definida como **tiempo** indica el tiempo en el cual los diferentes estudiantes de la carrera están en el seguimiento, de forma ascendente.

La columna **estado** nos indica si el estudiante al final del estudio se había retirado o graduado (censura)

La columna **proporción acumulada** se define como la proporción de casos para los que no han tenido lugar el evento en cada tiempo.

La columna **número de eventos acumulados** se refiere al número de estudiantes que se han retirado hasta el momento.

La columna **número de casos que permanecen** son los estudiantes que no han presentado el evento de interés, en este caso que se hayan retirado de la carrera sin haberse graduado.